

Développement de méthodes de fouille de données basées sur les modèles de Markov cachés du second ordre pour l'identification d'hétérogénéités dans les génomes bactériens

Catherine Eng

▶ **To cite this version:**

Catherine Eng. Développement de méthodes de fouille de données basées sur les modèles de Markov cachés du second ordre pour l'identification d'hétérogénéités dans les génomes bactériens. Médecine humaine et pathologie. Université Henri Poincaré - Nancy 1, 2010. Français. NNT : 2010NAN10041 . tel-01746308

HAL Id: tel-01746308
https://hal.univ-lorraine.fr/tel-01746308

Submitted on 29 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Nancy-Université
Université Henri Poincaré

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

➢ Contact SCD Nancy 1 : theses.sciences@scd.uhp-nancy.fr

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

Nancy-Université
Université Henri Poincaré

paul verlaine université - metz

Ecole Doctorale BioSE (Biologie, Santé, Environnement)

Thèse

Présentée et soutenue publiquement pour l'obtention du titre de

DOCTEUR DE l'UNIVERSITE HENRI POINCARE

Mention : « Sciences de la Vie et de la Santé »

par Catherine ENG

Développement de méthodes de fouille de données fondées sur les modèles de Markov cachés du second ordre pour l'identification d'hétérogénéités dans les génomes bactériens

Soutenance le 15 juin 2010

Membres du jury :

Rapporteurs :

M. GASCUEL O.	Directeur de Recherche, LIRMM UMR CNRS 5506, Université Montpellier
M. VARRÉ J.S.	Maître de conférences HDR, LIFL UMR USTL et CNRS 8022, Université Lille

Examinateurs :

M. AYMERICH S.	Directeur de Recherche, Institut Micalis UMR INRA 1319 et AgroParisTech, Jouy-en-Josas
M. BRÉHÉLIN L.	Chargé de Recherche, LIRMM UMR CNRS 5506, Université Montpellier
M. LEBLOND P.	Professeur, UMR INRA 1128 Génétique et Microbiologie, Nancy-Université, directeur de thèse
M. MARI J.F.	Professeur, LORIA UMR CNRS 7503, Nancy-Université, co-directeur de thèse
M^{elle} THIBESSARD A.	Maître de conférences, UMR INRA 1128 Génétique et Microbiologie, Nancy-Université, co-encadrant

Laboratoire de Génétique et Microbiologie – UMR UHP/INRA 1128 – IFR110
Laboratoire Lorrain de Recherche en Informatique et ses Applications – UMR CNRS 7503 et INRIA Lorraine

INRA Institut National de la Recherche Agronomique

Loria

Remerciements

Je tiens à remercier toutes les personnes qui m'ont soutenue tout au long de cette thèse très enrichissante.

Tout d'abord, je remercie le Professeur *Bernard Decaris* et le Professeur *Pierre Leblond* pour m'avoir permis de réaliser ces travaux au sein du laboratoire de Génétique et Microbiologie - UMR INRA/UHP 1128 de l'Université Nancy 1. Notamment à *Pierre* qui a fait confiance à mon travail de recherche. Je le remercie pour ses questions et ses idées pertinentes sans compter son investissement et son soutien pendant les moments difficiles.

Un immense merci au Professeur *Jean-François MARI* qui m'a donné des outils intéressants et un environnement de travail au sein du LORIA – Laboratoire Lorrain de Recherche en Informatique et ses Applications. Je le remercie pour les discussions enrichissantes sur les HMM, de m'avoir permis de comprendre leur intérêt et surtout de la confiance accordée à mon travail. Il a toujours prêté une oreille attentive à mes interrogations. Son efficacité et sa quiétude m'ont fait apprécier le travail réalisé à ses côtés.

Un grand merci à *Annabelle Thibessard*, Maître de conférences, une encadrante qui a donné un surnom affectueux aux HMM, les « Hmeumeux ». Ces quatre années m'ont permis de la voir s'épanouir dans son travail comme dans son rôle de mère. Je la remercie de m'avoir transmis sa passion pour l'enseignement et la recherche ainsi sa rigueur de travail la caractérisant bien.

Un merci à tous les trois pour ces quatre années de travail et parfois de doute. Je n'oublie pas *Bertrand Aigle*, Maître de conférences HDR, et *Charu Astana*, une indienne ingénieuse avec lesquels j'ai apprécié de travailler sur l'analyse des promoteurs. Hare Krishna …

Un merci particulier à *Olivier Gascuel*, Directeur de Recherche (LIRMM-CNRS), à *Laurent Bréhélin*, Chargé de recherche (LIRMM-CNRS), à *Jean-Stéphane Varré*, Maître de conférences (HDR, LIFL) et à *Stéphane Aymerich*, Directeur de Recherche (INRA, Thiverval-Grignon) qui ont accepté d'être rapporteurs et examinateurs de ce travail et surtout pour leurs remarques ainsi que le temps accordé à ce manuscrit.

Qu’aurait été l’environnement au laboratoire sans ces personnages ? Je remercie *Laurence* de son soutien, j’ai eu la chance d’intégrer au même moment qu’elle le laboratoire. Je regretterai notamment ses réserves de sucreries dont certains gourmands se souviennent encore de leurs effets secondaires. Un grand merci à *Luisa* qui a toujours soutenue les personnes en leur apportant son sourire et sa joie de vivre ; à *Emilie* et *Isolde*, de nouvelles venues qui ont partagé notre bureau et qui ont su l’égayer. Je n’oublie pas *Ludovic*, le seul mâle de ce bureau qui a su explorer son côté féminin. De même, il me faudrait plus que ces quelques lignes pour remercier toutes les personnes du laboratoire et du service informatique (*Christian* et *Nicolas*) que j’apprécie et qui se reconnaîtront car ils ont su établir une atmosphère de travail. Je n’oublie pas les personnes qui ont transité plus ou moins longtemps au laboratoire : *Séverine*, *Céline*, *Carole*.

Qu’aurait été mon travail sans une saine vie sportive ? Ne dit on pas « *Mens sana in copore sano* » ? Je remercie tous les badistes et amis sportifs qui ont été mes partenaires privilégiés en me permettant de souffler et d’évacuer mon stress. Je n’oublie pas *Chantal*, *Laëtitia* et *Michel*, le trio d’enfer qui ne cessent de se demander quand je finirai cette thèse. Un merci particulier à *Chantal* pour sa lecture attentive.

Qu’aurait été ma thèse sans le soutien de mon compagnon de vie et d’aventures ? Toujours présent dans les moments de joie, de peine et d’anxiété, il s’est tenu en permanence à mes côtés jusqu’à des heures reculées de la nuit, réclamant des câlins … merci mon fidèle chat Astuce … ou plutôt à son maître *Alexandre*.

Qu’aurais je été sans ma famille, et sans son soutien sans faille ? Fort d’une histoire commune, nous sommes tous parvenus à nous épanouir personnellement. Je remercie ma chère maman et *Kham-Souk*, vous avez été mes modèles cachés ainsi que *Sylvie*, *Fa yuan et Stéphanie* pour leur soutien immuable. Je leur dédie cette thèse.

Table des figures

Figure 1 : Expression de l'information génétique chez les bactéries. 3
Figure 2 : Exemple d'un modèle de Markov caché expliquant une suite de symboles observés. 10
Figure 3 : Représentation sous forme de réseaux bayésiens dynamiques de différents M*m*M*d* 11
Figure 4 : Deux structures différentes de HMM. 15
Figure 5 : Visualisation de la succession des probabilités *a posteriori* des états cachés déroulées dans le temps. 16
Figure 6 : Processus d'analyse de *CarottAge*. 18
Figure 7 : Exemple de segmentation d'une séquence génomique avec différents modèles HMM 19
Figure 8 : Evolution des probabilités *a posteriori*, P_i et P_{ii} pour plusieurs HMM 20
Figure 9 : Distribution des chromosomes linéaires (traits) et circulaires (cercles) chez les procaryotes (figure dérivée de Volff et Altenbuchner 2000). 23
Figure 10 : Cycle de différenciation morphologique des Streptomyces (Angert, 2005). 25
Figure 11 : Etapes de transcription chez les bactéries. 26
Figure 12 : Structure des quatre régions fonctionnelles conservées du facteur σ^{70} d'*E. coli* (Lonetto *al.*, 1992). 27
Figure 13 : Schéma d'interaction d'un ARN polymérase avec une région régulatrice permettant l'initiation de la transcription (d'après de Busby et Ebright, 1999). 28
Figure 14 : Alignement des régions promotrices du régulon σ^R (Paget *et al.*, 2001). 30
Figure 15 : Mécanismes de parasexualité chez les bactéries. 55
Figure 16 : Arbre phylogénétique basé sur l'analyse du polymorphisme du gène *sodA* chez les espèces du genre *Streptococcus* (Poyart *et al.*, 1998). 60
Figure 17 : Évolution des Lactobacillales par acquisition et perte de gènes 62
Figure 18 : Distribution phylogénétique de gènes acquis par HGT. 66
Figure 19 : Réseau phylogénétique (d'après Huson et Bryant, 2005). 67
Figure 20 : Processus de fouille de données pour la recherche de régions « atypiques ». 72
Figure 21 : Schématisation du processus d'extraction des régions atypiques. 74
Figure 22 : Visualisation par *ARTEMIS* d'une région atypique. 74
Figure 23 : Schéma relationnel de la base de données des firmicutes et du traitement des régions « atypiques ». 76

Figure 24 : Arbre phylogénétique des 70 espèces de la base de données des firmicutes basé sur les gènes d'ARNr 16S généré par le logiciel *PhyloDraw* (Choi *et al.,* 2000).....79
Figure 25: Localisation des régions identifiées par le HMM (M2M0) dans les éléments mobiles de *S. thermophilus.* 82
Figure 26 : Localisation des 362 gènes « atypiques » sur le génome de *S. thermophilus* CNRZ1066. 84
Figure 27 : Analyse compositionnelle des gènes « atypiques ». 86
Figure 28 : Distribution des gènes dans les différentes catégories fonctionnelles chez *S. thermophilus* CNRZ1066. 89
Figure 29 : Organisation du cluster CRISPR2 présent chez *S. thermophilus* CNRZ1066 (Bolotin *et al.*, 2005).....91
Figure 30 : Région exogène de 17 kb identifiée dans le génome de *S. thermophilus* (Bolotin *et al.*, 2004).....91
Figure 31 : Distribution des gènes « atypiques » chez les firmicutes. 93
Figure 32 : Nombre cumulatif de gènes « atypiques » en fonction de leur occurrence chez les firmicutes.....93
Figure 33 : Comparaison de la distribution des gènes variables et « atypiques » de *S. thermophilus* CNRZ1066 au sein des 47 souches utilisées dans les expériences CGH. 94
Figure 34 : Dispersion intra et inter spécifique des gènes « atypiques ».....96
Figure 35 : Visualisation sous ARTEMIS de quatre régions « atypiques » bornées par une séquence répétée de 139 nt. 102
Figure 36 : Visualisation sous ARTEMIS de deux régions « atypiques ».....103
Figure 37 : Modèle de tolérance des gènes exogènes facilitée par les protéines H-NS (Navarre *et al.*, 2007).....106
Figure 38 : Diagramme de Venn résumant l'analyse des gènes extraits par le M2M2 sur la base des critères de l'annotation (*annotation HGT*), de la variabilité au sein de l'espèce (*CGH-set*) et de la distribution au sein des espèces de firmicutes (*Firmicutes_s7*). 107
Figure 39 : Schéma d'extension des *iPeak-motifs*. 113
Figure 40 : Étape d'analyse de la distribution des gènes au sein des firmicutes. 121

Liste des tableaux

Tableau 1 : Exemple de distribution d'un état caché après apprentissage 18
Tableau 2 : Exemple de fichier *dump* des probabilités *a posteriori*, P_i, utilisé par ARTEMIS 20
Tableau 3 : Recherche de SFBS putatifs sur *B. subtilis* 48
Tableau 4 : Les SFBS des facteurs sigma de *B. subtilis* identifiés dans la DBTBS 48
Tableau 5 : SFBS identifiés dans les régions promotrices 49
Tableau 6 : Facteur d'enrichissement des SFBS potentiels 50
Tableau 7 : Identification de gènes homologues à des gènes de virulence de *S. pyogenes* SF370 (M1) et *S. pneumoniae* TIGR4 chez *S. thermophilus* CNZ1066 (Bolotin *et al*., 2004) 63
Tableau 8 : Taille et composition des éléments ICE et CIME utilisés par rapport au génome de *S. thermophilus* CNRZ1066 81
Tableau 9 : Analyse des chimères *in silico* 82
Tableau 10 : Longueur (en pb) des 362 gènes « atypiques » et 1553 CDS du reste du génome 87
Tableau 11 : Prédiction de la localisation cellulaire des protéines 88
Tableau 12 : Annotation fonctionnelle de gènes ou séquences suggérant une acquisition par HGT 90
Tableau 13 : Gènes « atypiques » et spécifiques de *S. thermophilus* et présents dans les 47 souches 99
Tableau 14 : Gènes « atypiques » rencontrant les 3 critères de l'analyse 108
Tableau 15 : Temps de calcul d'une recherche d'un motif composite à partir d'un consensus de classe 112
Tableau 16 : Consensus de classes correspondant à des motifs régulateurs 114
Tableau 17 : Regroupement proposé pour quelques motifs SFBS similaires ou peu divergents dans l'ensemble des 812 SFBS potentiels 116
Tableau 18 : Sensibilité et spécificité des différents modèles de Markov sur les 73 gènes variables (*CGH-set*) 118
Tableau 19 : Sensibilité et spécificité des méthodes pour les 73 gènes variables (*CGH-set*) 118

Sommaire

INTRODUCTION 1

MODÈLES DE MARKOV CACHÉS 7

1.1 DÉFINITIONS 8
1.1.1 Processus et suites stochastiques 8
1.1.2 Modèles de Markov cachés ou HMM 8
1.1.3 Dépendances à l'intérieur d'un modèle 10
1.1.4 Particularités propres au domaine biologique : les kmers 12
1.2 LES HMM POUR L'ANALYSE DE DONNÉES BIOLOGIQUES 13
1.3 NOTRE APPROCHE DE LA MODÉLISATION 14
1.3.1 Les spécificités des modèles HMM d'ordre 2 utilisés 14
1.3.2 CarottAge 15
1.3.3 Les étapes du processus de fouille de données à l'aide de CarottAge pour une application biologique 17

APPROCHE HYBRIDE STOCHASTIQUE ET COMBINATOIRE POUR LA RECHERCHE ET LA MODÉLISATION DE MOTIFS DE RÉGULATION GÉNIQUE 21

INTRODUCTION 22
2.1 MODÈLE D'ÉTUDE : *STREPTOMYCES* 22
2.1.1 Originalités génomiques 23
2.1.2 Cycle cellulaire complexe 24
2.2 INITIATION DE LA TRANSCRIPTION CHEZ LES BACTÉRIES 25
2.2.1 Les facteurs sigma chez S. coelicolor A3(2) 28
2.2.2 Caractéristiques des sites de reconnaissance des facteurs sigma (SFBS) 29

2.3 FOUILLE DE DONNÉES 31

ANALYSE DU GÉNOME DE *BACILLUS SUBTILIS* 47
2.4 COMPLÉMENTS DE RÉSULTATS POUR L'IDENTIFICATION DE SFBS PUTATIFS SUR *BACILLUS SUBTILIS* 47

UTILISATION DES HMM D'ORDRE 2 (M2M2) POUR LA DÉTECTION DE GÈNES ACQUIS PAR TRANSFERT HORIZONTAL, HGT 53

INTRODUCTION 54
3.1 MÉCANISMES DE TRANSFERT HORIZONTAL DE GÈNES (HGT) 54
3.2 MAINTIEN DE L'INFORMATION GÉNÉTIQUE 57
3.3 AMÉLIORATION DES SÉQUENCES ACQUISES PAR TRANSFERT HORIZONTAL 58
3.4 MODÈLE D'ÉTUDE : *STREPTOCOCCUS THERMOPHILUS* 59
3.4.1 Le genre Streptococcus 59
3.4.2 Le « core » génome et le génome accessoire de Streptococcus thermophilus 61
3.4.3 Flux de gènes chez S. thermophilus 62
3.5 MÉTHODES DE DÉTECTION DES GÈNES EXOGÈNES 64
3.5.1 Méthodes phylogénétiques 64
3.5.2 Méthodes paramétriques 67

MATÉRIEL ET MÉTHODES 71
3.6 EXTRACTION DES POSITIONS ET DES RÉGIONS « ATYPIQUES » 71
3.6.1 Traitement des probabilités a posteriori *72*
3.6.2 Détection et extraction des régions « atypiques » *73*
3.7 CRÉATION D'UNE BASE DE DONNÉES 75
3.8 DÉFINITION D'UN NOUVEAU PARAMÈTRE : LA VALEUR DE DISPERSION 76
3.8.1 Réalisation d'un arbre phylogénétique des firmicutes *77*
3.8.2 Calcul de la valeur de dispersion pour un gène *79*

RÉSULTATS : ANALYSE DU GÉNOME DE *S. THERMOPHILUS* 81
3.9 ANALYSES PRÉLIMINAIRES DE SÉQUENCES CHIMÉRIQUES 81
3.10 ANALYSE GLOBALE DU GÉNOME DE *S. THERMOPHILUS* CNRZ1066 83
3.10.1 Extraction et analyse de gènes « atypiques » *83*
3.10.2 Annotation des gènes « atypiques » *89*
3.10.3 Distribution phylogénétique des gènes « atypiques » *91*
3.10.4 Gènes d'adaptation chez S. thermophilus *99*
3.10.5 Ilots génomiques S. thermophilus *100*

CONCLUSION 104
3.11 CDS « ATYPIQUES » DE PETITE TAILLE ET PSEUDOGÈNES 104
3.12 ENRICHISSEMENT EN BASES A ET T DES SÉQUENCES « ATYPIQUES » 105
3.13 VALIDATION DU M2M2 COMME MÉTHODE D'IDENTIFICATION DE GÈNES ISSUS DU TRANSFERT HORIZONTAL 106
3.14 SCÉNARIO ÉVOLUTIF DU GÉNOME DE *S. THERMOPHILUS* 109

DISCUSSION ET PERSPECTIVES 111

4.1 FOUILLE DE DONNÉES POUR LA RECHERCHE DES SITES DE RÉGULATION 111
4.1.1 Avantages de la méthode : grande flexibilité de recherche *112*
4.1.2 Amélioration de la stratégie *115*
4.2 FOUILLE DE DONNÉES DÉVELOPPÉE POUR L'ANALYSE DES GÈNES « ATYPIQUES » 117
4.2.1 Méthodes HMM vs. autres méthodologies de détection du HGT *117*
4.2.2 Distribution des gènes dans un arbre phylogénétique : valeur de dispersion *119*
4.3 LES HMM DU SECOND ORDRE POUR LA FOUILLE DE DONNÉES GÉNOMIQUES 121
4.4 DÉFINITION AUTOMATIQUE D'UNE TOPOLOGIE DE HMM D'ORDRE 2 123

RÉFÉRENCES BIBLIOGRAPHIQUES 125

ANNEXE WEB : LISTE DES FICHIERS ACCESSIBLES PAR URL 151

ANNEXES 153

Introduction

Le progrès technologique a stimulé les recherches dans le domaine de la biologie en rendant possible l'accès aux séquences biologiques (ADN, protéines). L'obtention de séquences génomiques est devenue rapide et relativement peu coûteuse. Les capacités de séquençage des acides nucléiques ont considérablement augmenté depuis l'avènement de la méthode enzymatique de Sanger (Sanger *et al.*, 1974 ; Sanger *et al.*, 1977). Ainsi, l'utilisation de séquenceur automatique puis le développement de nouvelles technologies, comme le pyroséquençage, ont rendu possible l'accès à la séquence d'un génome bactérien entier en quelques jours seulement (hors étapes d'assemblage et d'analyse), alors qu'un projet de séquençage s'étalait sur plusieurs années, il y a encore une décennie. Ces données biologiques s'accumulent de manière exponentielle dans les banques de données. A la dernière mise à jour GOLD (24 janvier 2010) de la base de données de génomes (Liolios *et al.*, 2008), 1198 génomes d'organismes vivants ont entièrement séquencés et publiés - dont essentiellement 997 génomes bactériens - et 3817 projets de séquençage bactériens sont en cours. Ces données représentent une masse impressionnante d'information qu'il est essentiel de stocker, de caractériser et d'analyser. Cela nécessite l'aide notamment d'outils informatiques et mathématiques (probabilité, statistiques et méthode de classification). La tâche de décryptage par des méthodes expérimentales (*in vitro* et/ou *in vivo*) reste très longue et la bioinformatique permet de traiter *in silico* des données afin de rechercher un ensemble d'informations qu'il faut ensuite valider *in vivo*, ce qui aboutit à un gain de temps. L'analyse *in silico* est prédictive et peut être assimilée à de la fouille de données puisqu'elle consiste à extraire de l'information à partir d'une masse de données sous forme de règles, d'associations, de tendances cachées ou de structures particulières qui ne sont pas forcément celles que nous imaginions *a priori*. La fouille de données est un domaine de recherche où se rencontrent des méthodes issues de l'intelligence artificielle, de la théorie de l'information, de l'analyse de données traditionnelles (analyse factorielle, classification automatique, analyse discriminante) et surtout - en ce qui nous concerne - des méthodes de modélisation probabilistes et statistiques tels que les modèles de Markov cachés (HMM pour Hidden Markov Model). Les méthodes de fouille de données sont aussi dites descriptives ou exploratoires. Les méthodes de fouille de données descriptives visent à mettre en évidence des informations cachées par le bruit induit par la variabilité et le volume des données. Quant aux méthodes de fouille de données dites prédictives ou explicatives, elles visent à extrapoler de nouvelles informations à partir des données existantes. Dans le cas des données génomiques, l'information génomique est cryptée à l'intérieur de la séquence d'ADN sous forme de milliers à des centaines de milliards de « lettres ». Nous abordons la fouille de données descriptive sous l'angle de la modélisation stochastique des données de séquences. Il s'agit de cerner l'incertain en modélisant la variabilité d'un phénomène qui n'est pas encore totalement explicable, afin de faire apparaître des curiosités statistiques susceptibles de contenir des informations.

L'avantage d'une telle méthode est de fournir une nouvelle description des données qui peut différer de celle proposée *a priori*. Par conséquent, de nouvelles caractéristiques peuvent ainsi être révélées.

Cette thèse est le résultat d'une collaboration entre le Laboratoire de Génétique et Microbiologie (LGM) et le Laboratoire LOrrain de Recherche en Informatique et ses Applications (LORIA). Cette collaboration a débuté sous forme d'un projet commun financé par la région lorraine, l'État et différents organismes de recherche. Le LGM s'intéresse à deux organismes : les bactéries *Streptomyces coelicolor* et *Streptococcus thermophilus*. L'équipe Orpailleur du LORIA s'intéresse à la fouille de données. Cette équipe avait développé une boîte à outils – *CarottAge* – autours de modèles de Markov cachés du second ordre (HMM2) pour modéliser les successions de culture dans des territoires agricoles. La première idée de la collaboration a été d'utiliser *CarottAge* pour fouiller le génome des bactéries modèles du LGM. Ce travail n'a pas été immédiat comme le montrent les chapitres suivants qui sont consacrés à l'adaptation de *CarottAge* au contexte des données biologiques. Deux idées fortes ont été (i) d'associer des méthodes stochastiques avec des méthodes combinatoires et (ii) de se servir du contexte des observations dans une séquence génomique, en considérant des k*mers* plutôt que des nucléotides. Un k*mer* correspond à la concaténation de plusieurs nucléotides. La prise en compte de l'information contextuelle pour lever les ambiguïtés de l'analyse structurale avait été montrée dès le début des analyses sur la reconnaissance des formes en 1980.

Compte tenu des nombreux travaux en modélisation stochastique à l'aide de HMM du premier ordre dans le domaine biologique, cette thèse s'est orientée vers l'exploration – et la comparaison – de plusieurs types de HMM ayant des dépendances variables au niveau de la chaîne des états cachés et de la séquence d'observations. Nous avons comparé des chaînes de Markov du premier et du second ordre (ou d'ordre 2) entre elles et utilisé des chaînes de nucléotides ou de k*mers* dont les probabilités d'observation dépendent – ou non – des observations précédentes. Suivant le modèle, nous montrons que les HMM d'ordre 2 identifient des hétérogénéités qui se présentent sous la forme de régions courtes (quelques nucléotides) ou de régions plus longues (milliers de nucléotides). Les régions courtes présentent un enrichissement dans les régions intergéniques, et les régions longues (taille de l'ordre du gène) comportent des gènes codant des fonctions connues pour être fréquemment transférées (transporteurs spécifiques impliqués dans la résistance aux antibiotiques, système de restriction-modification, éléments génétiques mobiles). Ces premières observations nous ont amené à modéliser des motifs de régulation (régions courtes) chez *Streptomyces coelicolor* et à détecter des gènes « atypiques » (régions longues) dont une partie pourrait être potentiellement des gènes exogènes chez *Streptococcus thermophilus*.

Les signaux de régulation et de traduction

La figure 1 représente le dogme central de la biologie moléculaire. La séquence d'ADN est transcrite en molécule d'ARN, elle-même traduite en protéine. Celle-ci se replie en structure spécifique et assure une fonction particulière au sein de la cellule bactérienne.

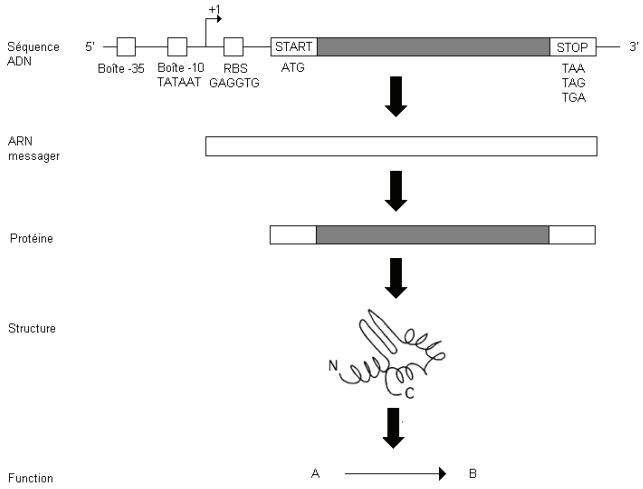


Figure 1 : Expression de l'information génétique chez les bactéries.
L'expression génique débute par la transcription de l'ADN en une molécule d'ARN depuis la région promotrice. Ensuite, la séquence d'ARN est traduite en séquence protéique.

Les signaux de transcription (par exemple, les boîtes -35 et -10 composant le promoteur) et de traduction (RBS pour Ribosomal Binding Site) confortent l'existence d'une séquence codante prédite *in silico*. Ces signaux sont des motifs de régulation de l'expression génique du gène aboutissant à la synthèse d'une protéine. La transcription constitue la première étape de l'expression génique ; elle est composée de trois phases (initiation, élongation et terminaison) qui sont soumises à une régulation. Ainsi, l'expression du gène débute par la liaison plus ou moins forte de facteurs protéiques à leurs sites de reconnaissance, TFBS (Transcription Factor Binding Sites). Ces facteurs sont nécessaires à l'initiation et/ou à l'activation de la transcription. Les facteurs sigma sont des facteurs de transcription et plus précisément, ce sont des facteurs d'initiation à la transcription. Ils reconnaissent des sites cibles, des SFBS (Sigma Factor Binding Sites) qui sont en général, des motifs composites de la forme boîte$_{-10}$-séparateur-boîte$_{-35}$. Les consensus des boîtes -10 et -35 des SFBS sont plus ou moins flexibles ainsi que la taille de leur séparateur.

Chez *Bacillus subtilis*, organisme modèle pour l'étude des bactéries Gram positives, la recherche et la prédiction de SFBS pour le facteur σ^A (TTGACA-N$_{14}$-TATAAT), par une méthodologie fondée sur les HMM, ont été analysées par Jarmer et ses collègues (2001). La topologie du HMM développé a l'avantage de prendre en considération simultanément la présence des deux boîtes et de permettre une variabilité au niveau de la taille du séparateur. Ce HMM nécessite au préalable un apprentissage sur un ensemble de séquences correspondant à des SFBS caractérisés expérimentalement. Dans la plupart des cas, ces données expérimentales sont manquantes ou insuffisantes. Cette thèse, propose une approche qui effectue l'estimation des paramètres du HMM d'ordre 2 sur la totalité du génome et ne nécessite pas au préalable un ensemble de données caractérisées. De plus, nous avons développé une stratégie originale en combinant :

- deux méthodes stochastiques (HMM d'ordre 2 et *R'MES*) afin d'identifier des motifs pertinents (par le HMM2) et de valider leur caractère exceptionnel (par *R'MES*),
- une méthode de classification pour regrouper les motifs identifiés,

- une méthode combinatoire pour la reconstitution de motifs composites robustes.

Cette stratégie a été appliquée à l'analyse de l'organisme modèle *Streptomyces coelicolor* qui possède un réseau de régulation génique important (pas moins de 12 % des gènes coderaient pour des régulateurs) dont au moins 65 facteurs sigma (Bentley *et al.*, 2002). Seule une dizaine de ces facteurs sigma ainsi que leurs sites de reconnaissance (ou SFBS) ont été caractérisés expérimentalement (Ainsa *et al.*, 1999 ; Bibb *et al.*, 2000 ; Kelemen *et al.*, 1998 ; Kormanec et Sevcikova, 2002 ; Lee *et al.*, 2004 ; Paget *et al.*, 2003 ; Sevcikova *et al.*, 2005 ; Takano *et al.*, 2005).

Les mécanismes de transfert horizontaux entre espèces

Par ailleurs, l'exploration des données génomiques au travers de l'investigation et la comparaison des génomes (gènes conservés, gènes essentiels et accessoires) apportent une meilleure compréhension de leur évolution ; les flux de gènes reflètent une adaptation à un environnement donné. Les gènes exogènes correspondent à une acquisition horizontale, HGT (Horizontal Gene Transfer), c'est-à-dire qu'ils ne proviennent pas d'une hérédité verticale (ascendant vers descendant). Les caractéristiques de ces séquences sont variables et dépendent du caractère plus ou moins récent de l'acquisition à l'échelle évolutive et des mécanismes d'acquisition. Notre organisme modèle, *Streptococcus thermophilus* est une bactérie dont l'unique niche écologique connue est le milieu lait. Cette bactérie a été décrite comme une espèce à évolution clonale, c'est-à-dire présentant peu d'hétérogénéité génétique. Cette faible diversité serait due à son émergence récente entre 3 000 et 30 000 ans (Bolotin *et al.*, 2004) à partir de l'ancêtre de *Streptococcus salivarius* (bactérie commensale de la cavité buccale de l'homme). Les travaux de cette thèse suggèrent que le transfert horizontal est néanmoins très fréquent chez cette espèce, diversifiant fortement le génome accessoire afin de s'adapter à sa nouvelle niche écologique. Ces gènes exogènes sont contenus dans un ensemble de gènes « atypiques » dont l'extraction a été réalisée par l'analyse des états cachés d'un HMM d'ordre 2. En effet, nous avons constaté qu'un état caché capturait principalement ce signal. La procédure d'estimation des paramètres du HMM d'ordre 2 est similaire à celle proposée par Nicolas et ses collègues (2002). L'hypothèse de leur origine exogène a été confortée par la nature des fonctions codées (transporteurs spécifiques impliqués dans la résistance aux antibiotiques, systèmes de restriction-modification, synthèse d'exopolysaccharides, éléments génétiques mobiles) et d'autre part par l'analyse de leur distribution au sein de 47 souches de *S. thermophilus* (données d'hybridation de génomique comparative) et au sein d'un arbre phylogénétique des firmicutes développé à cet effet. En effet, la distribution sporadique d'une séquence dans un arbre phylogénétique d'espèce robuste souligne son caractère accessoire et peut donc suggérer l'échange (perte ou gain) fréquent de cette séquence au cours de l'évolution. Ce dernier traitement a nécessité la génération et l'analyse d'une base de données ainsi que la mise en place d'un indice de dispersion des gènes atypiques dans l'arbre des firmicutes. Cette partie peut être vue comme une généralisation des travaux de Nicolas et ses collègues (2002) qui se sont appuyés sur des modèles de Markov cachés d'ordre 1 où les symboles correspondent à des nucléotides afin de détecter des hétérogénéités dans le génome

de *Bacillus subtilis*. Les travaux de cette thèse s'intéressent aux modèles de Markov cachés d'ordre 2 où les symboles utilisés correspondent à des nucléotides ou à des k*mers*.

Un avantage majeur de l'utilisation des HMM dans les deux applications présentées est leur capacité à segmenter la séquence génomique en régions aux propriétés statistiques homogènes. Cette thèse explore l'intérêt des différents ordres des modèles de Markov cachés, au niveau de la chaîne des états cachés et/ou au niveau de la chaîne d'observations, dans l'identification d'informations biologiques (signaux SFBS ou gènes « atypiques ») dans les deux génomes bactériens modèles. Le manuscrit est organisé en quatre chapitres.
Le chapitre 1 présente un outil de fouille de données, les HMM. Il débute par une définition des principaux aspects mathématiques avec une description des différentes dépendances au niveau de la suite des états cachés et de la chaîne d'observations. Il se termine par une description de notre approche de la modélisation et une description détaillée du processus de construction des HMM d'ordre 2 par le logiciel *CarottAge* ainsi que les modifications apportées dans le cadre de l'analyse de séquences génomiques.
Les chapitres 2 et 3 correspondent à deux champs d'application des HMM d'ordre 2 pour l'analyse d'hétérogénéités des génomes bactériens. Le chapitre 2 débute par une introduction à la problématique biologique de l'identification de SFBS dans les génomes bactériens et particulièrement chez *S. coelicolor*. Ensuite, la seconde partie comporte un état de l'art des méthodes de détection *in silico* des TFBS, la stratégie développée et les résultats obtenus pour l'analyse de deux génomes bactériens ayant des caractéristiques génomiques différentes, *S. coelicolor* et *B. subtilis*. Cette partie correspond à un article publié dans une revue internationale. Enfin, la dernière partie de ce chapitre 2, quant à elle, fournit une description plus approfondie des résultats et de la fouille de données réalisée chez *B. subtilis*. Le chapitre 3 correspond à l'application des HMM d'ordre 2 pour l'identification de gènes exogènes. Il débute par une introduction et un état de l'art des méthodes de détection des gènes issus du transfert horizontal. Puis, il présente la stratégie développée ainsi que nos résultats d'analyse du génome de *S. thermophilus*.
Le chapitre 4 présente une discussion des stratégies développées fondées sur les HMM d'ordre 2 ainsi qu'une comparaison des performances entre différents ordres des HMM et/ou avec d'autres programmes. En outre, les perspectives d'amélioration des stratégies développées y sont proposées.

Publications de ce travail

Publication internationale dans des revues avec comité de lecture :
Eng C., Asthana C., Aigle B., Hergalant S., Mari JF. and *Leblond P*. A new data mining approach for the detection of bacterial promoters combining stochastic and combinatorial methods.
Journal of Computational Biology, 2009. Nov. 2009. Vol. 16.

Publication n°2 (en préparation) :
Eng C., Thibessard A., Danielsen M., Rasmussen T., Mari JF. and Leblond P. Detection of horizontal gene transfer in *Streptococcus thermophilus* using data mining based on stochastic method.

Communications nationales et internationales (orales et posters)

Eng C., Thibessard A., Mari JF. and *Leblond P*. HMM and horizontal transfert in Streptococcus genomes. *Journée Bioinformatique du LORIA,* 2007 (France). Communication orale.

Eng C., Thibessard A., Mari JF. and *Leblond P*. Data Mining using Hidden Markov Models (HMM2) to detect DNA sequences acquired by horizontal transfer into *Streptococcus* genomes. *15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB),* 2007, Vienne (Autriche). Poster.

Eng C., Thibessard A., Mari JF. and *Leblond P*. Méthodes de fouille de données appliquées à l'analyse de génomes de bactéries du genre *Streptococcus*. *Quatrième rencontre des Microbiologistes de l'INRA,* 2006, Dourdan (France). Poster.

Eng C., Thibessard A., Mari JF. and *Leblond P*. Modèle HMM2 pour la détection du transfert horizontal dans les génomes de bactéries lactiques. *Journée Bioinformatique du LORIA*, 2005 (France). Communication orale.

Eng C., Thibessard A., Mari JF. and *Leblond P*. Data Mining using Hidden Markov Models (HMM2) to detect heterogeneities into bacterial genomes. *Journées Ouvertes Biologie, Informatique, Mathématiques (JOBIM),* 2005, Lyon (France).

Chapitre 1

Modèles de Markov cachés

1.1 DÉFINITIONS 8
1.1.1 Processus et suites stochastiques 8
1.1.2 Modèles de Markov cachés ou HMM 8
1.1.3 Dépendances à l'intérieur d'un modèle 10
1.1.4 Particularités propres au domaine biologique : les kmers 12
1.2 LES HMM POUR L'ANALYSE DE DONNÉES BIOLOGIQUES 13
1.3 NOTRE APPROCHE DE LA MODÉLISATION 14
1.3.1 Les spécificités des modèles HMM d'ordre 2 utilisés 14
1.3.2 CarottAge 15
1.3.3 Les étapes du processus de fouille de données à l'aide de CarottAge pour une application biologique 17

Introduction

La théorie des modèles de Markov cachés (HMM pour Hidden Markov Model) a été introduite au début des années 1970 (Baum *et al*., 1966). Les premières applications furent dans le domaine de la reconnaissance vocale (Baker *et al.*, 1975 ; Rabiner, 1989) où ils se sont rapidement imposés et développés comme des modèles standards. Une classification de leurs applications dans différents domaines tels que la climatologie, l'économétrie, la reconnaissance d'image et la biologie a été proposée par Cappé (2001). Les HMM sont une généralisation des modèles de Markov. Ce sont des modèles stochastiques qui permettent d'analyser des processus aléatoires dépendant du temps ou de l'espace. La modélisation de la variabilité d'un phénomène par des méthodes stochastiques ne représente pas parfaitement la réalité mais vise à la décrire au mieux.

L'utilisation de méthodes stochastiques et plus particulièrement, de méthodes Markoviennes en bioinformatique peut prendre des aspects multiples : profil HMM (Krogh *et al*., 1994 ; Eddy, 1998), modèles semi-Markoviens (Kulp *et al.*, 1996 ; Burge et Karlin, 1997), … nous nous sommes volontairement limités à la description des HMM pour la fouille de données comme outils d'extraction de connaissances quand un minimum d'informations sur la structure des objets recherchés est donnée *a priori*.

Ce premier chapitre se décompose en deux parties. La première partie présente une définition mathématique des différents types de HMM utilisés. Puis, la seconde partie décrit le processus de génération des HMM obtenu par l'outil *CarottAge*.

1.1 Définitions

1.1.1 Processus et suites stochastiques

Les variables aléatoires sont notées en majuscules alors que leurs réalisations sont notées en minuscules ; $X_t = q_t$ signifiant que q_t est la valeur au temps t de la variable X_t. Pour définir qu'une suite de n variables (X_t) prend la suite de valeurs (q_t), on écrira : $X_1^n = q_1^n$

Définition d'un processus stochastique
On appelle processus stochastique une suite de variables aléatoires $(X_t)_{t\in T}$. Si T est dénombrable, le processus est alors dit discret.

Définition d'une suite stochastique
Soit S un ensemble fini appelé ensemble des états du système. Entre les instants t et $t+1$, le système passe aléatoirement d'un état à un autre, soit u ce nouvel état. On définit la suite stochastique $(X_1, X_2, \ldots, X_t, \ldots.)$ à ensemble discret d'états par la relation : $(X_t = u)$, si et seulement si le système est dans l'état $u \in S$ au temps t.

Définition d'une chaîne de Markov
Une chaîne de Markov est une suite stochastique à un ensemble discret d'états, telle que :

- une suite du premier ordre :

$$P\left(X_t = q_t \middle| X_1^{t-1} = q_1^{t-1}\right) = P\left(X_t = q_t \middle| X_{t-1} = q_{t-1}\right) \quad (1)$$

- une suite du second ordre :

$$P\left(X_t = q_t \middle| X_1^{t-1} = q_1^{t-1}\right) = P\left(X_t = q_t \middle| X_{t-1} = q_{t-1}, X_{t-2} = q_{t-2}\right) \quad (2)$$

Quand ces probabilités ne dépendent pas du temps, la chaîne est dite homogène.

1.1.2 Modèles de Markov cachés ou HMM

Les modèles de Markov cachés comportent deux processus stochastiques. L'un est un processus observable, noté $(O_t)_{t\in\{1,2,\ldots\},}$ qui représente la suite des symboles observés $(o_t)_{t\in\{1,2,\ldots\}}$. Le second processus est non observable, noté $(X_t)_{t\in\{1,2,\ldots\}}$. Il correspond à une chaîne de Markov constituée des états de S, $(q_t)_{t\in\{1,2,\ldots\}}$.

HMM du premier ordre (HMM1)

Un modèle de Markov caché du premier ordre est défini par :

a) Un ensemble ***S*** de *N* états ;

b) Un alphabet ***C*** de *M* symboles observables ;

c) Une matrice de transitions ***A*** entre les états définie par :

$\boldsymbol{A} = (a_{ij}) = P(X_t = j \mid X_{t-1} = i),\ i, j \in S^2$ et $\sum_{j \in S} a_{ij} = 1$

d) Une matrice de probabilité d'émission ***B*** qui représente les *N* densités b_i sur les *M* observations faites sur les états cachés. On note :

$\boldsymbol{B}(i,o) = b_i(o) = P(O_t = o \mid X_t = i),\ i \in S,\ o \in C$ et $\sum_{o \in C} b_i(o) = 1$

e) La distribution des états initiaux :

$$\pi_i = P(X_1 = i), i \in S$$

Si on définit un unique état de début, la distribution de l'état initial peut être représentée par les transitions de l'état initial vers les différents états.

Le processus caché

La probabilité de la suite d'états $(q_t)_{t \in \{1,2,\ldots\}}$ dans une chaîne d'ordre 1 est définie par :

$$\begin{aligned} P(X_1^T = q_1^T) &= P(X_1 = q_1) P\left(X_2 = q_2 \middle| X_1 = q_1\right) P\left(X_3 = q_3 \middle| X_2 = q_2\right) \times \ldots \times \\ &\quad P\left(X_T = q_T \middle| X_{T-1} = q_{T-1}\right) \\ &= \pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1}, q_t} \end{aligned} \tag{3}$$

Dans ce cas, le processus modélisé a une mémoire très limitée car l'état q_t atteint au temps *t* ne dépend que de l'état q_{t-1} atteint au temps *t-1* et pas des états atteints aux temps précédents.

Dans le cas d'un HMM d'ordre 2, la probabilité de l'état caché *k* au temps *t* sachant que les états cachés étaient *j* et *i* respectivement au temps *t-1* et *t-2* est définie par :

$a_{ijk} = P(X_t = k \mid X_{t-1} = j, X_{t-2} = i),\ i, j, k \in S^3$

La probabilité de la suite d'états $(q_t)_{t \in \{1,2,\ldots\}}$ dans une chaîne d'ordre 2 est définie par :

$$\begin{aligned} P(X_1^T = q_1^T) &= P(X_1 = q_1) P\left(X_2 = q_2 \middle| X_1 = q_1\right) P\left(X_3 = q_3 \middle| X_2 = q_2, X_1 = q_1\right) \\ &\quad P\left(X_4 = q_4 \middle| X_3 = q_3, X_2 = q_2\right) \times \ldots \times P\left(X_T = q_T \middle| X_{T-1} = q_{T-1}, X_{T-2} = q_{T-2}\right) \\ &= \pi_{q_1} a_{q_1 q_2} \prod_{t=3}^{T} a_{q_{t-2} q_{t-1} q_t} \end{aligned} \tag{4}$$

Sans perdre de généralité, on introduit deux états supplémentaires notés *début* et *fin* atteints respectivement aux temps *0* et *T+1* afin d'éviter l'utilisation des probabilités initiales.

Le processus observé

Le processus observé est gouverné par la suite des états cachés. Dans le cas général, les symboles prennent des valeurs dans un alphabet *C* de taille 4 dans le cas d'une séquence nucléotidique (les quatre acides aminés), 64 dans le cas de 3*mers* ou 20 dans le cas d'une séquence protéique. Dans le cas le plus simple, la probabilité d'émission des symboles est uniquement dépendante de l'état caché sous-jacent. On verra dans le prochain paragraphe, qu'il est aussi possible de prendre en considération les quelques observations précédentes.

Un des problèmes biologiques est de pouvoir rendre compte de l'hétérogénéité observée tout le long d'une séquence d'ADN. Les HMM permettent de rendre compte de l'existence de sous séquences de nature différente dans une séquence génomique. Ils supposent que la séquence peut être découpée en fragments homogènes (figure 2). La séquence est alors homogène par morceaux. La taille, la position et la composition de ces fragments sont *a priori* inconnues. Le modèle dispose d'un nombre fini d'états qui s'adaptent à chacun de ces fragments. La succession des fragments est une chaîne de Markov.

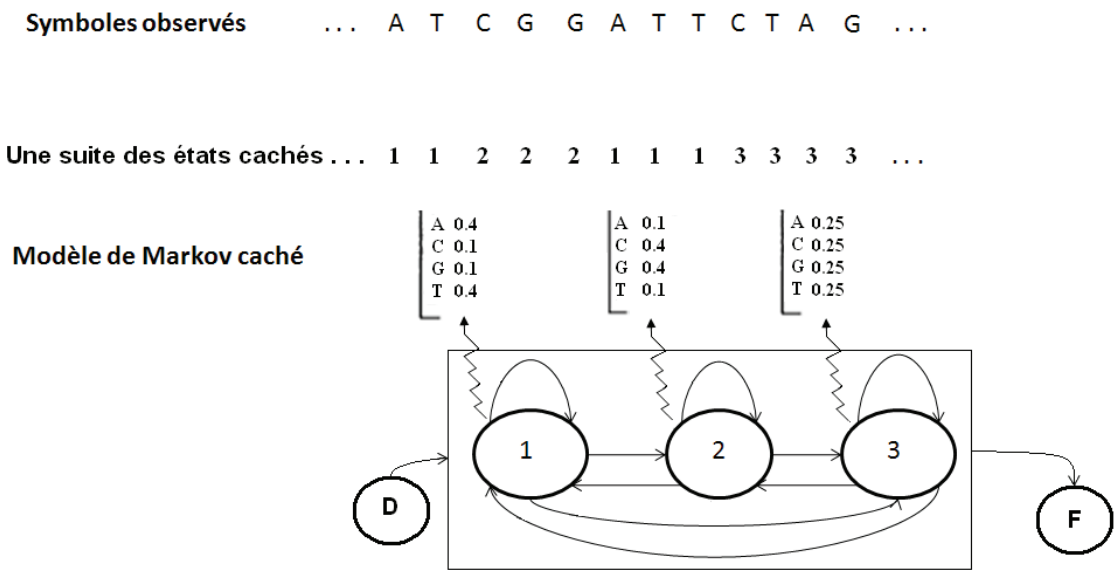


Figure 2 : Exemple d'un modèle de Markov caché expliquant une suite de symboles observés. Une suite des états cachés est associée à la séquence observée. Le HMM possède 3 états connectés. Les flèches représentent les transitions entre les états. Le cercle contenant la lettre « D » ou la lettre « F » représente respectivement, l'état de début et de fin. A chaque état caché est associée une distribution de probabilités de chaque symbole indiquée par les flèches en dents de scie. L'algorithme de Viterbi (Forney, 1973) peut être utilisé pour construire la suite d'états cachés la plus probable *a posteriori*.

1.1.3 Dépendances à l'intérieur d'un modèle

Un HMM peut être représenté par un réseau bayésien dynamique qui contient des dépendances entre les variables qui le composent (figure 3). Par la suite, on notera M*m*M*d*, un HMM dans lequel *m* est associé à la dépendance – l'ordre – de la suite des états cachés et *d* à la dépendance de la suite des symboles observés.

On appellera HMM2 un modèle de type M2M0 et HMM1 un modèle de type M1M0.

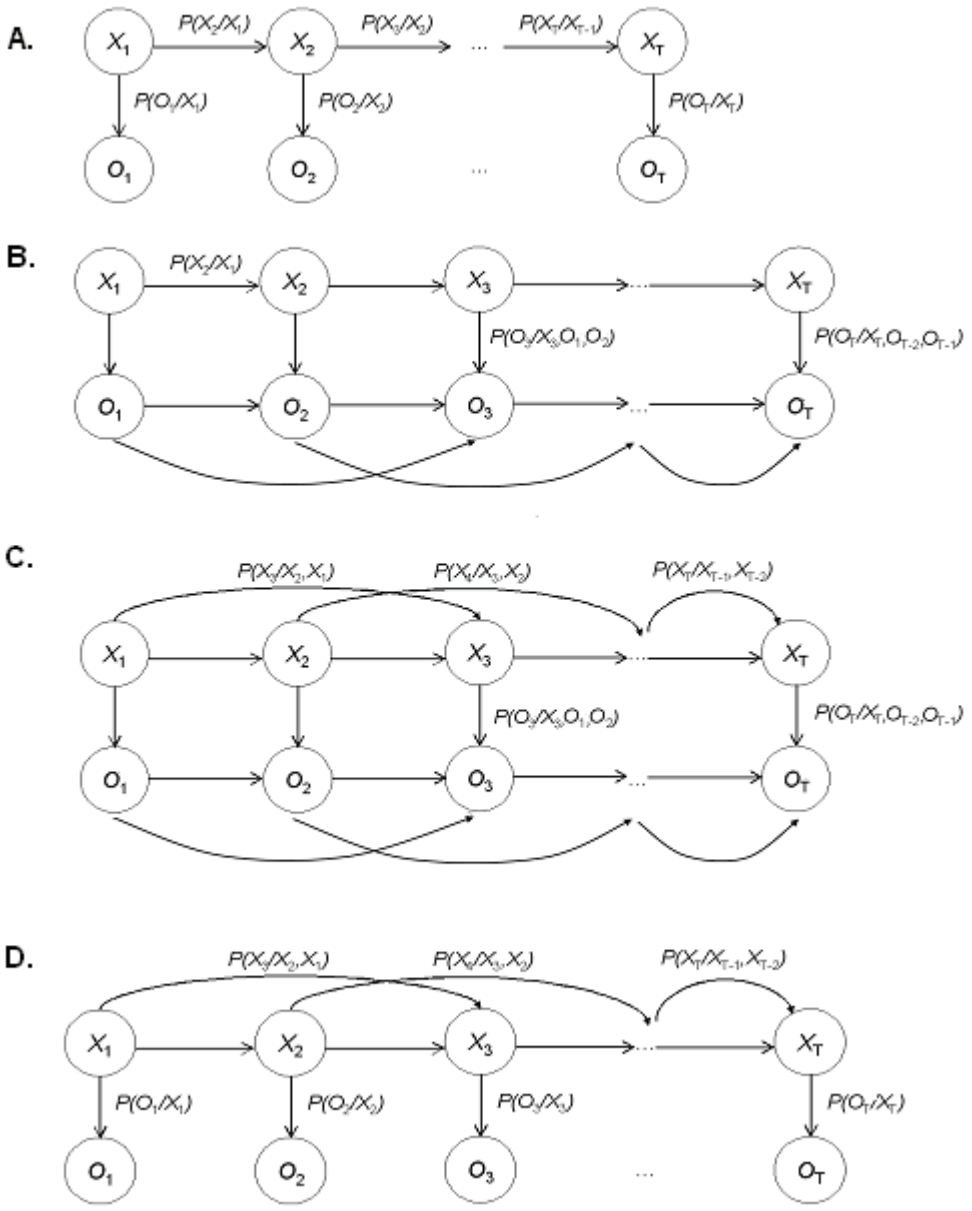


Figure 3 : Représentation sous forme de réseaux bayésiens dynamiques de différents M*m*M*d*.
A. Un M1M0 : le processus caché est une chaîne de Markov d'ordre 1 et la probabilité d'émission du symbole, O_t, est dépendante de X_t.
B. Un M1M2 : le processus caché est une chaîne de Markov d'ordre 1 (comme A) et la probabilité d'observation de O_t dépend de O_{t-1}, O_{t-2}, et de X_t.
C. Un M2M2 : le processus caché est une chaîne de Markov d'ordre 2 et la probabilité de O_t dépend de O_{t-1} et O_{t-2} et de X_t.
D. Un M2M0 : le processus caché est une chaîne de Markov d'ordre 2 (comme C), la probabilité d'observation de O_t dépend uniquement de X_t.

Dépendances dans la suite des états cachés

- Pour un modèle de Markov caché d'ordre 1 (M1M*d*) (figure 3.A et 3.B), la probabilité de l'état X_t est dépendante de l'état précédent l'état X_{t-1}. La probabilité de transition entre les états cachés est donnée par la matrice de transition à deux dimensions $\boldsymbol{A}$.

- Pour un modèle de Markov caché d'ordre m (MmMd) avec $m \geq 2$, la probabilité de l'état X_t est dépendante des m états précédents. La matrice $\boldsymbol{A}$ possède $m+1$ dimensions. Par exemple, dans un modèle de Markov caché d'ordre 2 (M2M*d*) (figure 3.C et 3.D), la matrice de transition $\boldsymbol{A}$ est définie par :
  $\boldsymbol{A} = (a_{ijk}) = P(X_t = k \mid X_{t-1} = j, X_{t-2} = i),\ i, j, k \in S^3$

Dépendances dans la suite des observations

- Pour un modèle MmM0, la distribution des symboles observés o_t est uniquement dépendante de l'état caché i (figure 3.A et 3.D). Les symboles sont générés selon une loi b_i :

$$b_i(o) = P(O_t = o \mid X_t = i),$$

avec $o \in C$ et $i \in S$

- Pour un modèle MmMd où $d \geq 1$ (figure 3.B et 3.C), la probabilité d'émettre un symbole o dépend des d symboles précédents, noté a, et de l'état caché i et est définie par :

$$b_i(o,a) = P\left(O_t = o \middle| X_t = i, O_{t-d}^{t-1} = a\right), i \in S, o \in C, a \in C^d$$

Avec la condition de normalisation :

$$\sum_{o \in C} b_i(o,a) = 1, \forall i \in S, \forall a \in C^d \qquad (5)$$

1.1.4 Particularités propres au domaine biologique : les k*mers*

Dans les HMM, chaque état caché comporte une densité de probabilité d'émission de symboles. Nous nous sommes intéressés à l'émission de symboles correspondant à des k*mers*. Un k*mer* peut être vu comme l'observation d'un nucléotide y_t à l'instant t avec un contexte spécifique $y_{t-k+1}, \ldots, y_{t-2}, y_{t-1}$ défini par les $k-1$ nucléotides qui ont été observés aux instants précédents $t-k+1, \ldots, t-1$. De manière similaire, une séquence d'ADN peut être vue comme une séquence de k*mers* se chevauchant avec un déplacement consécutif de longueur l. Par exemple, la séquence *##TAGGCTA* peut être vue comme une séquence de 3*mers* et un déplacement de 1 ($k = 3$ et $l = 1$) : ##T-#*TA-TAG-AGG-GGC-GCT-CTA* ; le symbole « # » représente le contexte vide.

Les formules d'estimation du HMM2 implémentées par l'algorithme EM (Mari *et al.,* 1997) généralisent les formules classiques de l'estimation d'un HMM1. Elles sont données dans l'article *JCB* publié dans la prochaine partie.

1.2 Les HMM pour l'analyse de données biologiques

Les HMM peuvent être employés pour la reconnaissance de formes, pour la fouille de données et pour la modélisation du bruit de fond dans un génome selon trois principes différents.

En reconnaissance de formes, un HMM peut être utilisé pour modéliser un élément bien caractérisé expérimentalement (motif, domaine) (Krogh *et al.*, 1994 ; Kulp *et al.*, 1996 ; Yada et Hirosawa, 1996 ; Burge et Karlin, 1997 ; Eddy, 1998 ; Sonnhammer *et al.*, 1998). On utilise autant de HMM que de formes à reconnaître. Dans ce cas, les paramètres des HMM sont estimés à partir d'un jeu de données représentatif de chaque forme. On cherchera alors à déterminer à l'aide d'une séquence inconnue d'observations, soit la probabilité d'observer cette séquence sachant le modèle (discrimination), soit à associer les différents fragments de la séquence aux états cachés (figure 2) grâce à l'alignement calculé par l'algorithme de Viterbi (Forney, 1973) entre la suite d'états la plus probable et la suite d'observations. N'ayant pas de discrimination à faire entre plusieurs modèles, nous n'avons pas utilisé cet algorithme.

En fouille de données, un unique HMM permet d'extraire de l'information contenue dans une séquence biologique sans *a priori* sur son contenu génétique, c'est-à-dire sans *a priori* sur le phénomène biologique recherché. On cherche alors à extraire les probabilités *a posteriori* des états cachés tout le long de la séquence d'observations afin de localiser des zones atypiques et de les relier à des phénomènes biologiques.

Enfin, les HMM peuvent être utilisés pour représenter un bruit de fond moyen associé à un génome. Dans ce cas, ils sont utilisés pour calculer des statistiques sur le nombre moyen d'occurrences d'un motif nucléotidique et les comparer aux statistiques observées (Pesole *et al.*, 1992 ; Krogh *et al.*, 1994 ; Yada *et al.*, 1998 ; Azad et Lawrence, 2005). Ceci permet de déceler les motifs sur ou sous-représentés dont l'apparition ne tient pas compte du hasard. Nous n'avons pas utilisé les HMM d'ordre 2 dans cette optique.

L'information contextuelle associée à un nucléotide est représentée par un k*mer*. En augmentant la longueur des k*mers*, le contexte pris en compte augmente aussi. Nous avons exploré l'avantage de cette caractéristique et comparé cette modélisation à la modélisation à l'aide de $M1M_{k\text{-}1}$ et $M2M_{k\text{-}1}$ pour l'exploitation des données biologiques. Nous avons remarqué expérimentalement que les probabilités *a posteriori* des états cachés étaient lissées dans les M2M*d*, vraisemblablement à cause des conditions de normalisation supplémentaires introduites dans l'estimation des probabilités conditionnelles (Eq. 5). L'avantage d'utiliser un k*mer* est d'éviter ce lissage et de permettre la détection de motifs courts.

Nous avons aussi analysé un seul ou les deux brins de l'ADN. En effet, les motifs de régulation sont généralement observés et localisés en amont des séquences codantes qui peuvent se trouver, soit dans le brin sens, soit dans le brin complémentaire. Le traitement des deux brins est dans ce cas judicieux.

1.3 Notre approche de la modélisation

1.3.1 Les spécificités des modèles HMM d'ordre 2 utilisés

HMM d'ordre 2 pour la classification non supervisée

Ce travail présente une utilisation des HMM d'ordre 2 (HMM2 et M2M2) pour segmenter la séquence génomique en fonction d'une composition locale afin d'identifier des hétérogénéités courtes (motifs de régulation) ou longues (HGT).

Les processus d'analyse comportent la construction de HMM d'ordre 2, leur estimation et le *decoding* (calcul des probabilités *a posteriori* des états cachés). Les informations biologiques sur les gènes exogènes sont insuffisantes pour générer une topologie précise d'un HMM pour leur identification. Dans l'apprentissage des HMM d'ordre 2, l'intégralité de la séquence génomique est utilisée et l'estimation se fait sans segmentation *a priori* de la séquence génomique. Ce point diffère d'une utilisation des HMM dans une optique de reconnaissance de formes qui nécessite des séquences représentatives (séquence de famille de protéines, CDS, biais compositionnel). L'apprentissage consiste à déterminer les paramètres du modèle qui maximisent sa vraisemblance pour la séquence d'observations. L'algorithme Baum-Welch (Baum *et al.*, 1970) est utilisé à cette fin. Après l'obtention du modèle paramétré, le *decoding* est réalisé dans la dernière étape d'apprentissage pour obtenir à chaque position de la séquence, la probabilité de l'état caché s_i ou d'une transition au départ de l'état s_i du type (s_i, s_i) ou (s_i, s_i, s_i). Ce processus est similaire à celui employé pour segmenter le génome de *Bacillus subtilis* par des HMM d'ordre 1 afin d'identifier des séquences exogènes (Nicolas *et al.*, 2002). Ces HMM d'ordre 1 utilisent un alphabet composé de quatre symboles observables (les nucléotides) avec une dépendance de deux pour la chaîne d'observations et le *decoding* est similaire au notre. Néanmoins, nos modèles diffèrent des précédents par l'ordre des modèles et dans notre cas, la possibilité d'utiliser des k*mers* (cf. section 1.1.4).

Biais compositionnel

Le génome de *S. thermophilus* CNRZ1066 comporte une inversion du GC skew cumulatif au niveau du terminus (*Term*) et de l'origine (*oriC*) de réplication (cf. section 3.10.1, figure 26). Le calcul du GC skew cumulatif le long de la séquence génomique est exprimé par le ratio (nombre de G – nombre de C) / (Nombre de G + Nombre de C). Nous avons observé que ce GC skew cumulatif inversé influence fortement le modèle. En effet, des analyses préliminaires sur un brin du génome complet de *S. thermophilus* CNRZ1066 ont révélé que nos modèles M2M2 segmentaient toujours le génome en deux régions homogènes. L'observation de l'état de *decoding* montrait que les probabilités *a posteriori* étaient faibles sur la première partie du génome puis devenaient plus élevées le long de la seconde moitié du génome. La position de transition observée entre les probabilités *a posteriori* basses et élevées correspondait à la position de l'inversion obtenue par le calcul du GC skew. Ce biais compositionnel indique la richesse d'un brin en G par rapport au C et résulterait d'un biais de réplication. Le GC skew permet d'identifier *in silico* l'origine et le terminus de réplication.

Afin de s'affranchir de ce biais, le point d'inflexion du GC skew a été déterminé en observant le GC skew cumulatif le long de la séquence génomique. Ensuite, une séquence génomique a été générée *in silico* en concaténant les brins sens des deux bras allant de la position 1 jusqu'à la position du point d'inflexion et de la séquence complémentaire à partir de la fin de la séquence génomique jusqu'au point d'inflexion. Par le même procédé, deux séquences chimériques ont été générées pour l'analyse de *B. subtilis* afin d'identifier des SFBS (cf. section 2.4). Ces deux séquences ont permis l'estimation des paramètres de deux modèles HMM2 appelés par la suite : HMM2+ et HMM2-.

Topologie du modèle

La génération de la topologie d'un HMM est une étape importante et délicate car elle doit être conforme à la représentation biologique que l'on veut modéliser. Une représentation de plus en plus fine du modèle biologique entraine une capture de l'information de plus en plus importante (Eddy, 1998 ; Krogh *et al.*, 1998). Dans le cadre de notre analyse, nous avons utilisé un modèle linéaire (gauche-droite) et un modèle ergodique (figure 4).

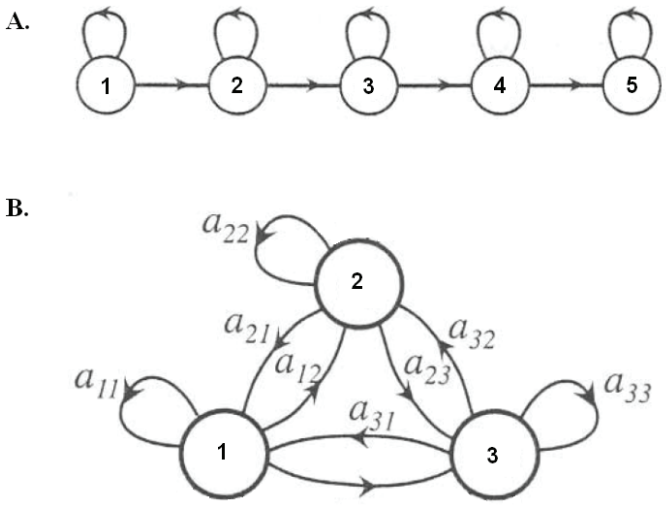


Figure 4 : Deux structures différentes de HMM.

A. HMM linéaire ou « gauche-droite » minimal, les transitions possibles sont d'un état caché vers l'état caché suivant ou sur lui-même. A partir de l'état caché suivant, il n'est pas possible de revenir dans l'état caché précédent. Nous utilisons ce modèle pour établir une estimation initiale de densités utilisées dans un modèle ergodique

B. HMM ergodique : toutes les transitions entre états sont possibles. Nous utilisons ce HMM pour segmenter le génome.

1.3.2 CarottAge

CarottAge est un logiciel libre sous la licence GNU Public License qui génère des HMM pour l'analyse de successions d'observations continues ou discrètes. *CarottAge* est un acronyme composé du mot *carotte* qui est la traduction de Markov en russe et du mot *Age*. C'est un ensemble de programmes développés en C++. Les agronomes l'utilisent pour représenter la succession des cultures ainsi que leur évolution temporelle et spatiale. La figure 5 représente la succession des états cachés déroulée dans le temps mettant en jeu des données agricoles.

Cette analyse couplée à une visualisation permet d'observer les alternances de cultures. Une culture peut être représentée par un état caché (maïs, colza, orge, …) ou une distribution (état de réserve noté « ? »).

Nous avons utilisé le logiciel *CarottAge* pour la génération et l'estimation des paramètres des HMM pour l'analyse des données biologiques. Ce logiciel a été choisi car il permettait de générer des HMM d'ordre 2 avec peu de modifications. En effet, la principale modification s'est effectuée dans la définition de l'alphabet où les symboles utilisés sont des nucléotides ou des k*mers* à la place des différents types de culture. En outre, les logiciels permettant de générer des HMM pour l'analyse des séquences biologiques fournissent seulement la possibilité de créer des HMM d'ordre 1 (HMMER, SHOW) (Eddy, 1998 ; Nicolas *et al.*, 2002). Le second atout de *CarottAge* est qu'il offre la possibilité de travailler avec un alphabet variable, soit de nucléotides, soit de k*mers* contrairement aux logiciels précédents qui n'utilisent qu'un alphabet de 4 symboles (les nucléotides) ou de 20 symboles (les acides aminés pour HMMER). Enfin, *CarottAge* permet de travailler sur une ou plusieurs séquences génomiques complètes.

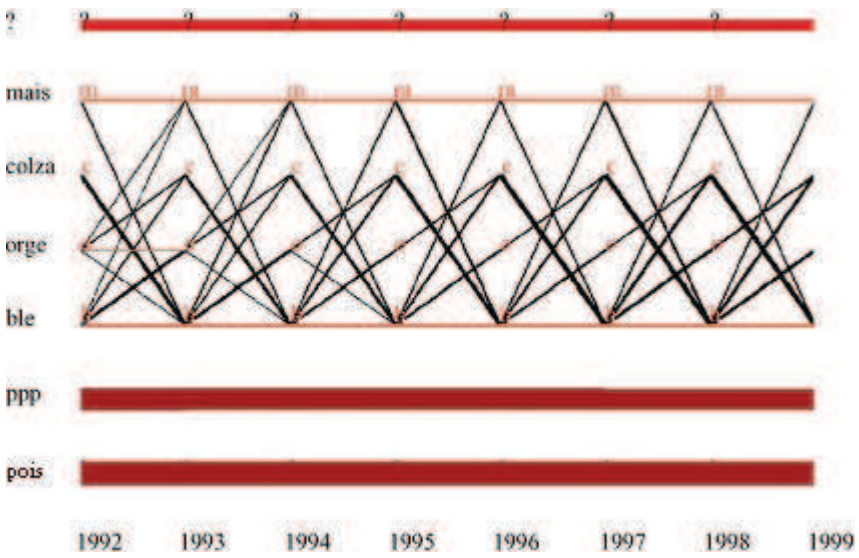


Figure 5 : Visualisation de la succession des probabilités *a posteriori* des états cachés déroulées dans le temps.

Decoding des états d'un HMM2 appris sur les occupations des terres agricoles lorraines de 1992-1999 (Le Ber *et al.*, 2006). L'épaisseur et la couleur des traits représentent les probabilités de transition entre les états cachés. Les états cachés sont les différentes cultures (maïs, colza, orge, blé, pois) ou l'état de réserve - « ? » - qui représente une distribution de toutes les cultures. L'état « ppp » correspond aux prairies permanentes. Contrairement aux données biologiques, les données agronomiques sont spatialisées et ont plus de modalités. En revanche, les séquences sont plus courtes.

1.3.3 Les étapes du processus de fouille de données à l'aide de *CarottAge* pour une application biologique

La fouille de données à l'aide d'un HMM d'ordre 2 implémenté dans le logiciel *CarottAge* se décompose en quatre principales étapes d'analyse (figure 6).

1. Edition de la topologie du modèle et de ses paramètres.

La définition initiale de la topologie du HMM d'ordre 2 est linéaire avec les transitions équiprobables entre chaque état et une distribution uniforme des probabilités d'émission des symboles dans chaque état. L'utilisation d'un modèle linéaire va permettre une première estimation des densités associées à chaque état caché qui sera ensuite utilisée comme graine pour l'estimation des densités du modèle ergodique.
Le tableau 1 liste les structures du modèle généré. La distance entre les états cachés est calculée selon la formule de la divergence de Kullback-Leibler (Kullback et Leibler, 1951) qui mesure la dissimilarité entre deux distributions de probabilités et qui permet de retenir les topologies ayant un état caché distant des autres.

2. Estimation itérative des paramètres du modèle.

L'estimation des paramètres du modèle généré est réalisée par l'algorithme Baum-Welch (EM) pour un M2M*d* à l'aide des données observées. Une première estimation des paramètres est effectuée sur un modèle linéaire afin d'obtenir une segmentation de la séquence en régions homogènes (figure 7.A). Le modèle linéaire est transformé en modèle ergodique (figure 7.B). Les paramètres du modèle ergodique sont estimés par l'algorithme Baum-Welch sur les mêmes données.

3. Etape de *decoding*.

L'étape de *decoding* profite de la dernière itération de EM pour calculer les probabilités *a posteriori* des états cachés. Cela permet une réduction du temps de calcul. Nous avons comparé les variations de la probabilité *a posteriori* des états cachés suivant que nous utilisions des 3*mers* dans un modèle M2M0 ou des nucléotides dans un modèle M2M2. Il est alors possible d'extraire trois types de probabilités *a posteriori* du modèle sachant la séquence d'observations (o_t). Dans toutes les formules, s_i est l'état caché atteint aux temps t :

- Type 1 : la probabilité *a posteriori* de l'état caché s_i :
  $P_i = P\left(X_t = s_i \mid O_1^T = o_1^T\right)$.
- Type 2 : la probabilité *a posteriori* de la transition de l'état caché s_i vers s_i :
  $P_{ii} = P\left(X_t = s_i \mid X_{t-1} = s_i, O_1^T = o_1^T\right)$.
- Type 3 : la probabilité *a posteriori* de la transition $s_i s_i$ vers s_i :
  $P_{iii} = P\left(X_t = s_i \mid X_{t-1} = s_i, X_{t-2} = s_i, O_1^T = o_1^T\right)$.

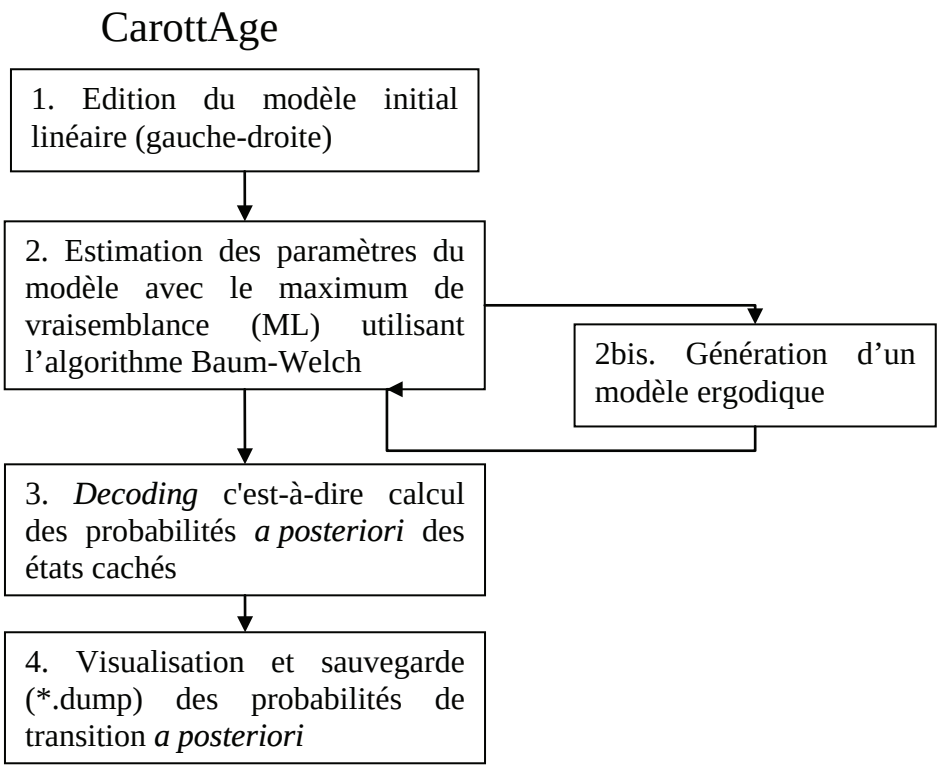


Figure 6 : Processus d'analyse de *CarottAge*.

Tableau 1 : Exemple de distribution d'un état caché après apprentissage

P.d.f. at state 3
Nb. of symbols = 64 MAX = 300
0 : 0.43575 : G C T
1 : 0.42079 : A A A
2 : 0.41986 : C A A
3 : 0.41275 : T T T
4 : 0.40176 : C G A
5 : 0.39718 : G G A
6 : 0.39586 : T G A
7 : 0.39528 : C C T
8 : 0.38595 : G T T
.....

Kullback-Leibler distance

3	0.34373			
4	0.013491	0.24541		
5	0.0012188	0.35959	0.015597	
6	0.0067246	0.26648	0.0033063	0.0083118
	2	3	4	5

La densité des probabilités d'émission des 64 symboles dans l'état 3 est donnée ainsi que la matrice – symétrique dont on ne présente que la moitié – des distances Kullback-Leibler entre les 5 états cachés d'un HMM2 ergodique. Le numéro des états est fourni dans la première colonne et la dernière ligne. Le rectangle identifie l'état caché le plus éloigné des 4 autres. Par la suite, cet état sera qualifié de « divergent ».

La figure 8 compare pour un état caché, la probabilité *a posteriori* P_{ii}, (de type 2), pour un modèle M2M0 de k*mers* avec la probabilité *a posteriori* $P_{i,}$ (de type 1), issu un modèle M2M2 de nucléotides. On remarque tout de suite la différence entre ces deux courbes où l'une est plus lisse et va nous permettre d'extraire des hétérogénéités longues (plusieurs milliers de paires de bases). L'autre sera employée à extraire des motifs courts (quelques nucléotides).
L'extraction des probabilités *a posteriori* s'avère utile pour l'analyse des régions d'intérêt et/ou pour avoir une visualisation simultanée avec des annotations du génome d'intérêt. Le logiciel ARTEMIS permet une représentation conviviale des génomes avec leur annotation. Il permet aussi simplement la visualisation des probabilités *a posteriori* de l'état divergent étudié. L'exploration des sorties de *decoding* va nécessiter le développement de procédés de post-traitement distincts et efficaces des probabilités *a posteriori* qui seront détaillés dans les chapitres relatant les applications (cf. section 2.3 et 3.6).

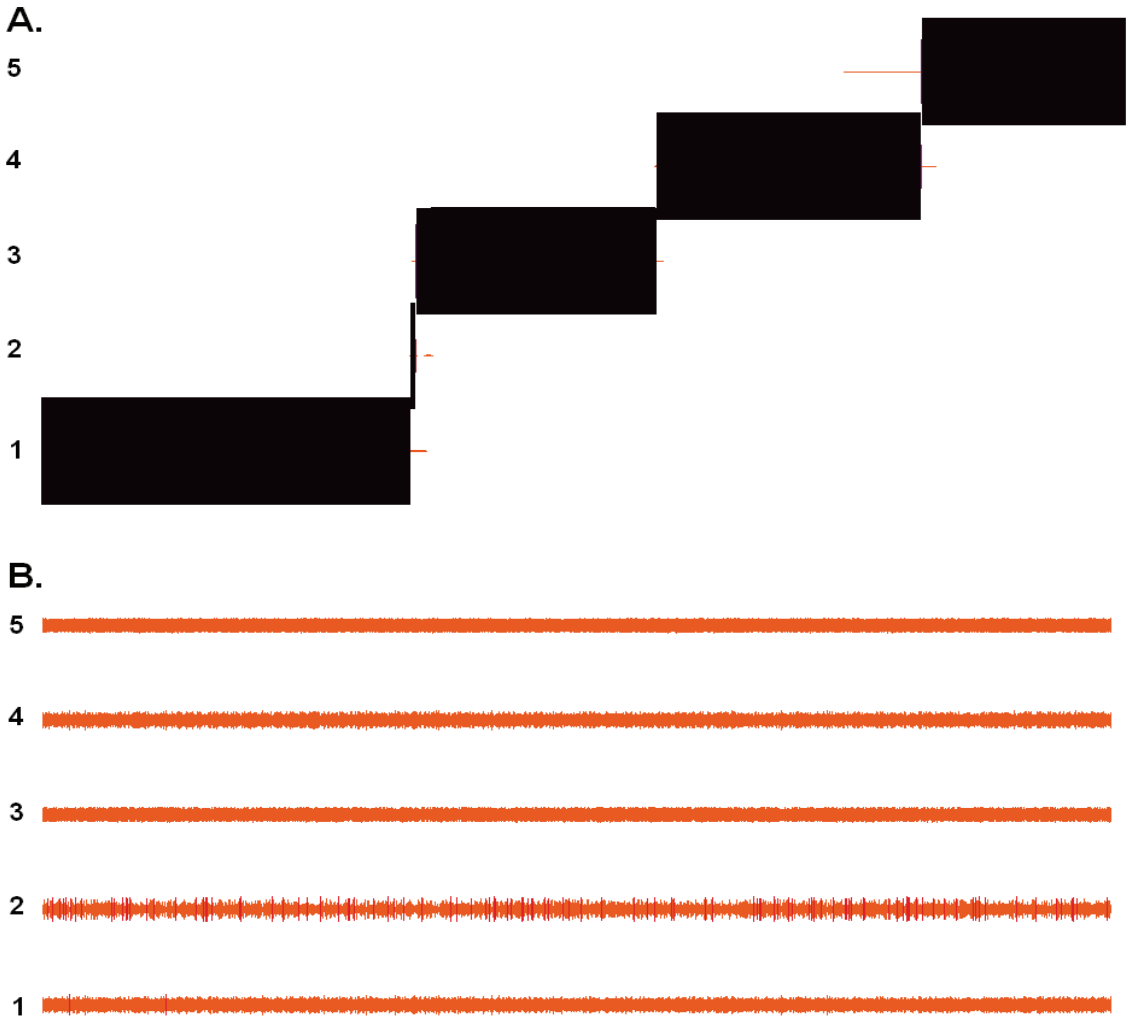


Figure 7 : Exemple de segmentation d'une séquence génomique avec différents modèles HMM
Les traits verticaux représentent les probabilités *a posteriori* pour chaque état à chaque position.
A. Segmentation avec un modèle linéaire à 5 états et 3*mers*.
B. Segmentation avec un modèle ergodique à 5 états et 3*mers*.

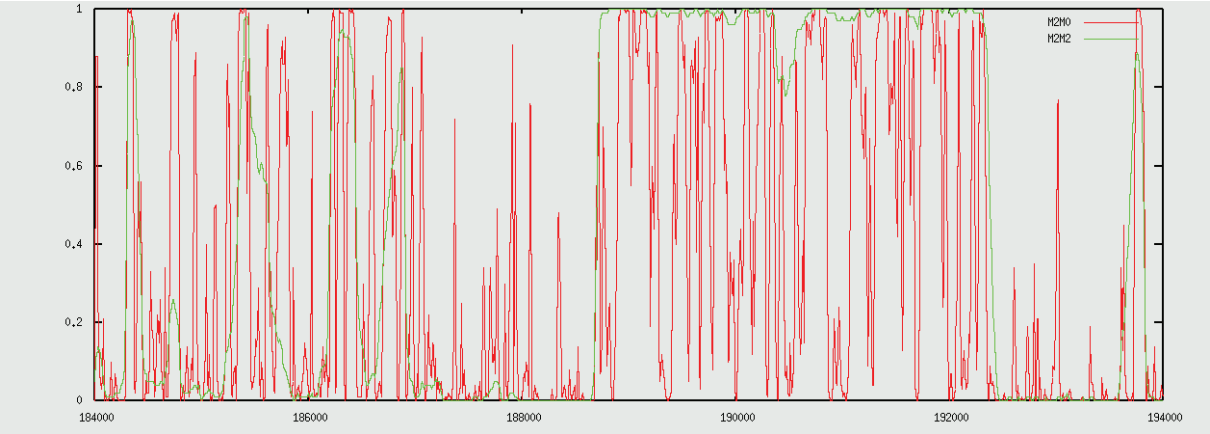


Figure 8 : Evolution des probabilités *a posteriori*, P_i et P_{ii} pour plusieurs HMM
La courbe rouge représente la probabilité *a posteriori* P_{ii}, dans un M2M0 de 3*mers* et la courbe verte représente la probabilité *a posteriori* P_i dans un M2M2 de nucléotides.

Tableau 2 : Exemple de fichier *dump* des probabilités *a posteriori*, P_i, utilisé par ARTEMIS

#1 -> 549 | 5 states
1 0.00 0.00 0.00 0.00 0.00
2 0.01 0.00 0.07 0.00 0.91
3 0.30 0.02 0.01 0.01 0.67
4 0.43 0.00 0.04 0.01 0.52
5 0.48 0.01 0.04 0.01 0.46
6 0.50 0.01 0.06 0.02 0.41
7 0.48 0.00 0.11 0.21 0.20
8 0.18 0.04 0.00 0.43 0.35
9 0.39 0.00 0.24 0.37 0.00
10 0.26 0.26 0.10 0.29 0.09
11 0.25 0.52 0.13 0.00 0.09
……..

Chapitre 2

Approche hybride stochastique et combinatoire pour la recherche et la modélisation de motifs de régulation génique

2.1 MODÈLE D'ÉTUDE : *STREPTOMYCES* 22
2.1.1 Originalités génomiques *23*
2.1.2 Cycle cellulaire complexe *24*
2.2 INITIATION DE LA TRANSCRIPTION CHEZ LES BACTÉRIES 25
2.2.1 Les facteurs sigma chez S. coelicolor A3(2) *28*
2.2.2 Caractéristiques des sites de reconnaissance des facteurs sigma (SFBS) *29*

2.3 FOUILLE DE DONNÉES 31

ANALYSE DU GÉNOME DE *BACILLUS SUBTILIS* 47
2.4 COMPLÉMENTS DE RÉSULTATS POUR L'IDENTIFICATION DE SFBS PUTATIFS SUR *BACILLUS SUBTILIS* ... 47

Cette partie concerne la stratégie développée pour la modélisation des sites de reconnaissance des facteurs de transcription (TFBS pour transcription facteur binding site). Ce chapitre s'article autour d'une publication parue dans la revue *Jounal of Computational Biology* (*JCB*) et est organisé de la manière suivante.

La partie I donne une description biologique de la structure des TFBS dans les génomes bactériens. Cette partie est plus détaillée que dans l'article publié dans *JCB.*

En revanche, la partie II reprend l'article publié dans *JCB* et commence par présenter un état de l'art des méthodes de détection des TFBS. Ensuite, elle décrit notre système de fouille de données pour l'identification de TFBS sans connaissance *a priori* du contenu génétique. La fouille de données associe une méthode stochastique fondée sur les HMM d'ordre 2 pour l'identification de motifs élémentaires et une méthode de classification de ces motifs

permettant de définir des graines pour la recherche de motifs composites sur-représentés dans le génome. Notre approche permet l'identification de motifs composites dans une séquence génomique en tenant compte de la variabilité des motifs et du séparateur. Nous présentons les résultats obtenus sur la bactérie *S. coelicolor*.
La partie III développe les résultats obtenus sur *Bacillus subtilis*. L'existence de cette dernière partie est motivée par un effort de précisions par rapport aux résultats décrits dans l'article pour l'analyse du génome de *B. subtilis*.

Introduction

La flexibilité de l'expression génétique est essentielle dans l'adaptation des bactéries à leur environnement. L'expression génétique comprend la transcription et la traduction de l'information de chaque gène et aboutit à la synthèse d'une protéine. Elle assure aussi la stabilité des transcrits et des produits. Dans la transcription, les différentes étapes que sont l'initiation, l'élongation et la terminaison, sont soumises à une régulation. L'expression du gène débute par la liaison de multiples facteurs protéiques, appelés facteurs de transcription (TF pour Transcription Factor) à leur site de reconnaissance, les TFBS (Transcription Factor Binding Sites). Les facteurs de transcription sont des protéines nécessaires à l'initiation de la transcription et/ou à l'activation/répression. L'organisme modèle utilisé en laboratoire est la bactérie *Streptomyces coelicolor* A3(2) qui possède un important réseau de régulation transcriptionnelle incluant au moins 60 facteurs sigma (SF pour Sigma Factor) impliqués dans la différenciation morphologique et biochimique. Cependant, seule une dizaine de ces sites de reconnaissance ont été caractérisés expérimentalement (Potuckova et *al.*, 1995 ; Paget *et al.*, 1998, Paget *et al.*, 1999 ; Cho *et al.*, 2001 ; Bibb *et al.*, 2000, Kim *et al.*, 2008). La détection des TFBS *in silico* constitue un défi important et complexe, en particulier à cause du caractère variable des motifs de reconnaissance. Les motifs constituant le TFBS présentent des variations en termes de séquence, d'organisation et d'occurrence. Ces variabilités à différents niveaux rendent difficile leur prédiction par des outils de bioinformatique.

2.1 Modèle d'étude : *Streptomyces*

Les *Streptomyces* sont des bactéries Gram positives à haut contenu en bases G et C et font partie des *Actinomycètes*. Les *Streptomyces* sont des bactéries filamenteuses dont l'habitat naturel est le sol. Ces bactéries possèdent l'un des plus grands génomes bactériens (~10 Mpb) (Bentley *et al.*, 2002), leur chromosome est linéaire (figure 9) et présente une composition en bases G et C élevée (70-74 %). En plus de ces caractéristiques structurales et compositionnelles, les *Streptomyces* ont un métabolisme secondaire complexe et original. Ils synthétisent une diversité de métabolites secondaires ayant des structures chimiques variées. L'application la plus connue des métabolites secondaires est la lutte anti-infectieuse. Les *Streptomyces* et d'autres genres relativement proches produisent à eux seuls plus de 50 % des antibiotiques commerciaux (Demain, 1999 ; Rigali *et al.*, 2008). Il faut ajouter à cela que, les *Streptomyces* sécrètent beaucoup d'autres métabolites ayant une activité utile pour l'homme

comme antifongique, anti-parasitique, antivirale, anti-tumorale, immunosuppressive, insecticide ou herbicide.

2.1.1 Originalités génomiques

La circularité des chromosomes a longtemps été considérée comme le paradigme chez les bactéries. Comme l'indique cette étude qui porte sur 44 espèces (Volff *et al.*, 2000) seules quatre espèces (incluant *Streptomyces*) présentent un chromosome linéaire (Ferdows et Barbour, 1989) (figure 9). Ces quatre espèces n'étant pas phylogénétiquement liées, cela suggère une apparition sporadique de la linéarité.

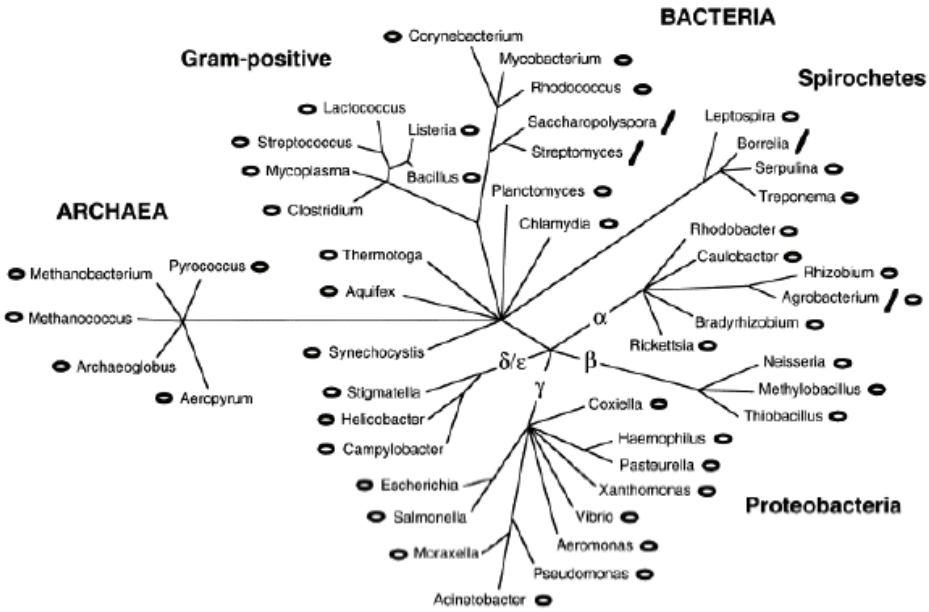


Figure 9 : Distribution des chromosomes linéaires (traits) et circulaires (cercles) chez les procaryotes (figure dérivée de Volff et Altenbuchner 2000).

Une autre originalité des *Streptomyces* provient de leur taille. Les génomes des eucaryotes sont de grandes tailles et présentent une grande dispersion entre 10^7 et 10^{11} paires de bases (Gregory, 2005). Quant aux bactéries, elles présentent des génomes de petites tailles et variables de $0{,}6x10^6$ à 10^7 paires de bases. L'un des plus petits génomes bactériens séquencés est celui de *Mycoplasma genitalium* qui a une taille de 580 Kpb et qui comporte 517 gènes dont 37 gènes codant pour des ARN de structure (ARN ribosomiques et de transfert) (Fraser *et al.*, 1995). Ce dernier se développe dans les systèmes génitaux et respiratoires humains. A l'opposé, le plus grand génome bactérien séquencé est celui de *Sorangium cellulosum* 'So ce 56' qui a une taille de 13 Mpb comprenant pas moins de 9 703 gènes (NC_010162[1]) (Schneiker *et al.*, 2007). Notre organisme modèle, *Streptomyces coelicolor* A3(2), fait partie des plus grands génomes bactériens séquencés, et atteint une taille d'environ 8,7 Mpb contenant pas moins de 7 912 gènes (Bentley *et al.*, 2002).

A la différence des eucaryotes, le génome des bactéries montre une corrélation entre le nombre de gènes et la taille du génome (Mira *et al.*, 2001). Cette corrélation est aussi

[1] Numéro d'identifiant sur le NCBI (http://www.ncbi.nlm.nih.gov)

observée dans 87 génomes séquencés de procaryotes (Konstantinidis et Tiedje, 2004). Le nombre de gènes est corrélé, vraisemblablement, au style de vie (bactérie intracellulaire ou vie autonome) et à la complexité de l'organisme. Cela semble être le cas chez *S. coelicolor* qui présente un cycle cellulaire complexe au cours duquel une différenciation morphologique accompagnée d'une différentiation métabolique sont observées (Chater, 1993). De même, chez l'organisme modèle *Bacillus subtilis* qui possède 4 423 gènes pour une taille de génome de 4,21 Mpb. L'habitat de *B. subtilis* est le sol (Kunst *et al*., 1997) et cette bactérie présente une différenciation cellulaire complexe par la formation d'endospores, au cours duquel de nombreux métaboliques ou enzymes sont sécrétés (hydrolases et antibiotiques). De plus, *B. subtilis* est capable de se différencier en un état physiologique particulier, l'état de compétence naturelle qui permet d'internaliser des fragments d'ADN présents dans l'environnement. L'état de compétence chez *B. subtilis* ne requiert pas moins de 40 gènes (Solomon et Grosmann, 1996). En revanche, *Mycoplasma genitalium* possède peu de gènes et elle évolue dans un environnement stable (mode de vie parasitaire).

Dans la banque de données GenBank (Dennis *et al.*, 2008), seuls trois génomes de *Streptomyces* ont été entièrement séquencés, annotés et sont disponibles directement avec les identifiants entre parenthèses ; *Streptomyces avermitilis* MA-4680 (NC_003155), *Streptomyces coelicolor* A3(2) (NC_003888) et *Streptomyces griseus* subsp. *griseus* NBRC 13350 (NC_010572). Un quatrième génome de *Streptomyces*, *Streptomyces scabies*[2] est accessible mais ne comporte pas d'annotation. Le nombre réduit de séquences génomiques disponibles pour le genre *Streptomyces* reflète, vraisemblablement, les difficultés de séquençage et d'assemblage liées à la composition nucléotidique particulière de ce grand génome bactérien (haut contenu en bases G et C). La présence d'une grande quantité de nucléotides G et C se traduit notamment par des régions à composition monotone et structurées facilitant la formation de structures secondaires difficiles à séquencer. En plus de cette caractéristique compositionnelle, ces génomes montrent un fort niveau de redondance (gènes dupliqués, paralogues issus du transfert horizontal) engendrant des difficultés supplémentaires à l'assemblage des séquences génomiques.

2.1.2 Cycle cellulaire complexe

Les *Streptomyces* présentent un cycle cellulaire complexe accompagné d'une différenciation morphologique poussée et d'une réorientation du métabolisme vers la production de nombreux composés d'intérêt (Chater 1993). Dans des conditions nutritives favorables, la spore germe (figure 10.A) et forme des hyphes (figure 10.B). Ces hyphes forment un mycélium végétatif (figure 10.C) qui se développe jusqu'à l'épuisement des nutriments dans le milieu. Cette carence nutritive va entraîner le développement d'hyphes aériens (figure 10.D) à partir de la surface de la colonie. Enfin, ces hyphes aériens se différencient en longues chaînes de spores par septation régulière (figure 10.E, figure 10.F). En concomitance avec ce cycle de différenciation morphologique, de nombreux métabolites secondaires sont

[2] *Streptomyces scabies* : http://www.sanger.ac.uk/Projects/S_scabies/

synthétisés procurant à la bactérie des éléments de lutte contre d'autres organismes. L'analyse du génome de *S. coelicolor* A3(2) indique que ce dernier contient au moins dix huit clusters codant pour des complexes enzymatiques caractéristiques du métabolisme secondaire (biosynthèse des antibiotiques, sidérophores, pigmentations) (Bentley *et al.*, 2002). Le séquençage du génome de *S. avertimilis* a permis d'identifier trente voies de biosynthèse de métabolites secondaires (Omura *et al.*, 2001 ; Ikeda *et al.*, 2003). Chez *S. ambofaciens*, bactérie modèle du laboratoire, la séquence partielle du génome a révélé au moins douze voies de biosynthèse dont la plupart sont spécifiques de cette espèce indiquant une diversité importante au sein du genre *Streptomyces*. La régulation du cycle cellulaire complexe nécessite un important réseau de régulation. En effet, chez *S. coelicolor* A3(2), 12,3 % des gènes codent des fonctions de régulation (965 sur 7825 gènes prédits). Parmi eux, 65 facteurs sigma ont été prédits et sont recrutés par l'ARN polymérase afin d'initier la transcription de manière spécifique (Bentley *et al.*, 2002).

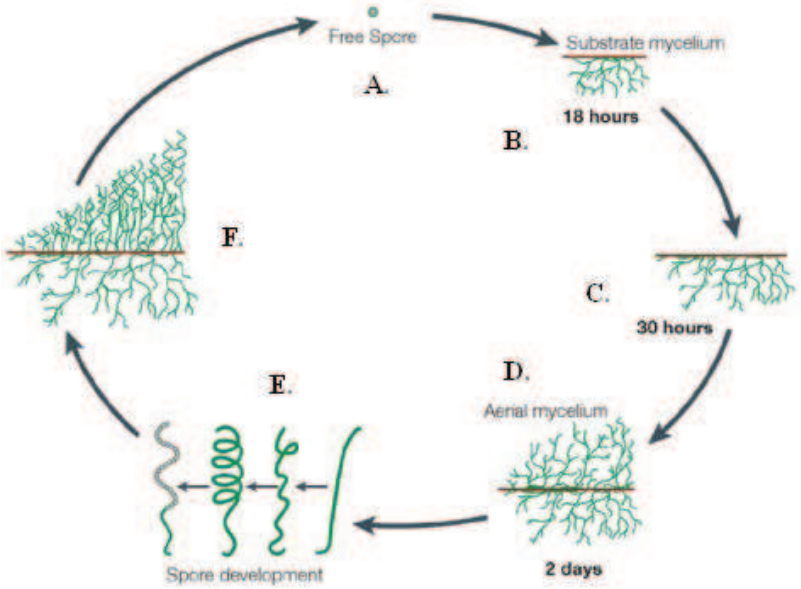


Figure 10 : Cycle de différenciation morphologique des Streptomyces (Angert, 2005).

2.2 Initiation de la transcription chez les bactéries

L'étape clé de la transcription des gènes chez les bactéries est l'initiation qui résulte de l'interaction de l'ARN polymérase avec la région promotrice d'un gène et l'activation de l'ARN polymérase permettant l'ouverture de l'ADN double hélice. L'ARN polymérase des bactéries est formée de quatre sous unités ($2\alpha\beta\beta'$) et est appelée l'ARN polymérase « core » (figure 11). Les sous-unités α forment un dimère qui se lie aux sous-unités β et β'et à des protéines activatrices de la transcription (Zhang *et al.*, 1998). La sous-unité α comporte deux domaines fonctionnels (domaine C-terminal, αCTD et N-terminal, αNTD) séparés par un « linker » flexible d'environ 14 nt (Jeon *et al.*, 1997). La spécificité de la liaison de l'ARN polymérase avec le promoteur résulte du recrutement de l'ARN polymérase « core » au site promoteur par la sous unité sigma (σ). L'ARN polymérase « core » avec la sous unité sigma

est appelée holoenzyme (2αββ'σ). C'est la spécificité d'interaction qui détermine la force promotrice c'est-à-dire l'efficacité d'initiation de la transcription (Haldenwang, 1995). Le facteur σ reconnait de manière spécifique un site cible composé généralement des boîtes dites « -10 » et « -35 » (ou SFBS pour Sigma Factor Binding Sites). Ces annotations correspondent à la position des boîtes par rapport au site de début de la synthèse de l'ARN messager ou ARNm (+1 de transcription) dans les régions promotrices des gènes. Cette reconnaissance spécifique permet une forte liaison entre l'holoenzyme à la séquence d'ADN. L'ARN polymérase recouvre une région de l'ADN qui s'étend de -50 à +20 nucléotides et l'ouverture de la double hélice d'ADN est réalisée au quatrième nucléotide composant le motif de la boîte -10. Cette boîte est formée d'appariements de faible stabilité (A et T) qui facilitent la séparation des deux brins d'ADN. Le facteur σ est relargué immédiatement après l'activation de la transcription car il n'est pas nécessaire à la poursuite de la transcription et peut être ainsi recruté par une autre ARN polymérase « core ».

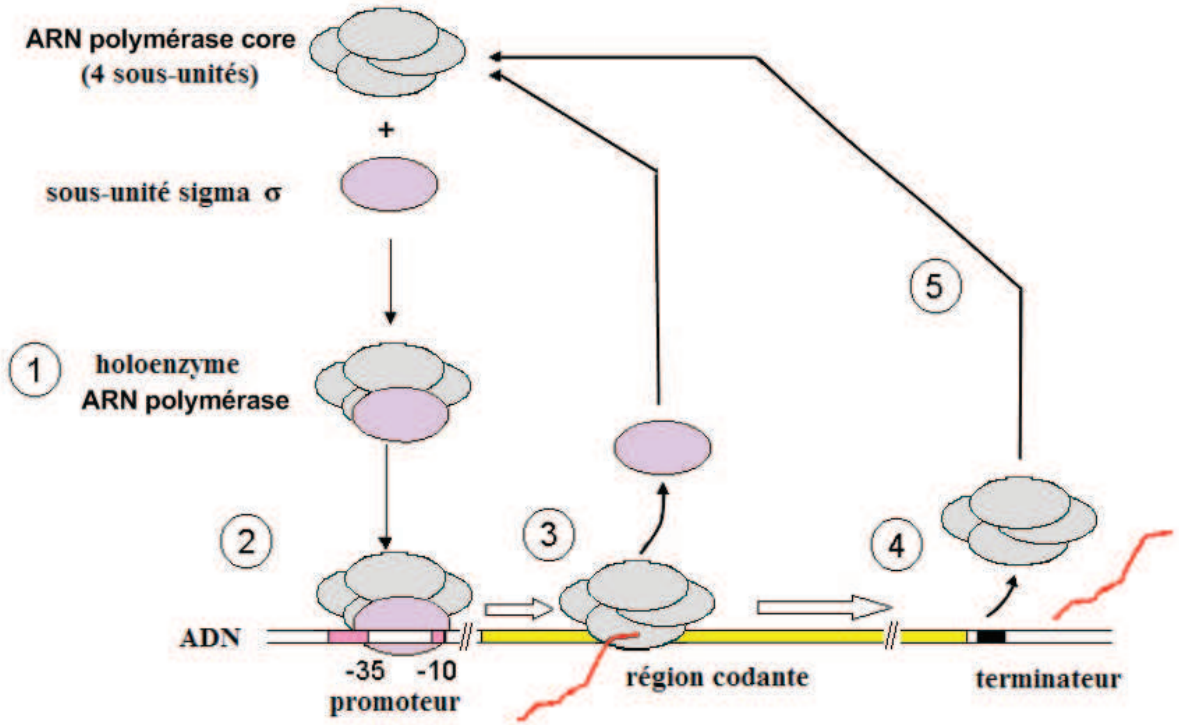


Figure 11 : Etapes de transcription chez les bactéries.

1) L'association entre l'ARN polymérase « core » (2αββ') et le facteur sigma (σ) forme l'holoenzyme.
2) Le facteur σ reconnaît les sites localisés dans le promoteur (boîte -10 et -35) et permet la fixation de l'holoenzyme au promoteur.
3) La libération du facteur σ.
4) La transcription est achevée lorsque l'ARN polymérase « core » rencontre un terminateur de transcription.
5) Le cycle recommence et l'ARN polymérase « core » peut recruter une autre sous unité σ (identique ou différente).

L'arrêt de la transcription découle de deux procédures distinctes. La première utilise une séquence spécifique appelée terminateur. Ce terminateur est constitué d'une séquence riche en nucléotides G et C qui conduit à la formation de structure de type tige boucle. Cette structure est suivie de 6 résidus d'uracile qui provoquent la libération de l'ARN polymérase. Le second

processus nécessite une protéine particulière appelée facteur rho (ρ). Ce dernier s'attache à l'ARNm à des sites spécifiques localisés en amont des terminateurs. Ils sont d'une longueur comprise entre 50 à 90 bases et sont riches en résidus C. Le facteur rho se déplace le long du brin d'ARN nouvellement synthétisé et décroche l'ARN polymérase à l'arrêt ou en pause sur une structure de type tige boucle (Stewart *et al.*, 1986).

Le facteur σ^{70} chez *E. coli* comprend quatre régions formées de sous régions (figure 12) (Lonetto *et al.*, 1992). La région 1.1 inhibe la liaison du facteur σ en absence de contact avec l'ARN polymérase core (Dombroski *et al.*, 1992). La région 2.1 chevauche une région nécessaire pour la liaison à l'ARN polymérase core (Lesley *et al.*, 1989). Quant à la région 2.4, elle interagit avec le consensus de la boîte -10 et la région 4.2 interagit avec le consensus de la boîte -35. Les séquences protéiques des régions 2 et 4 sont très conservées au sein de 31 facteurs sigma de la famille du facteur σ^{70}, ce qui supportent leur rôle dans la reconnaissance des régions promotrices (Lonetto *et al.*, 1992).

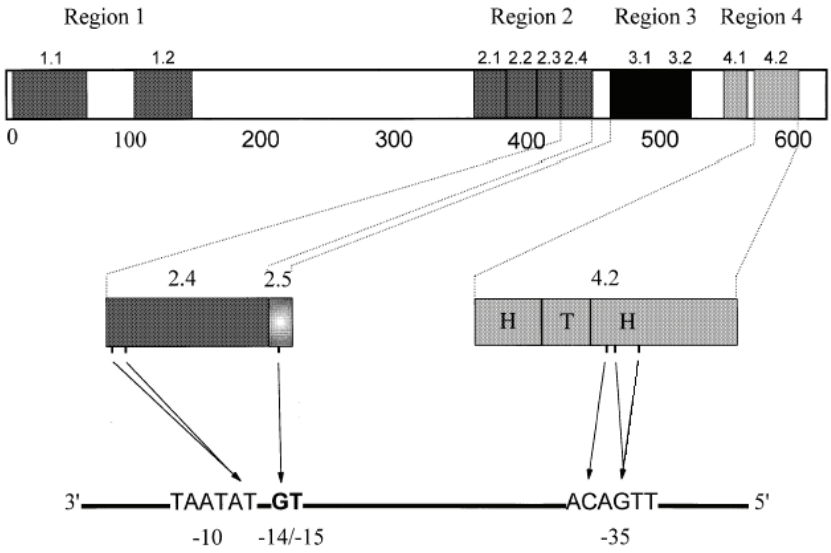


Figure 12 : Structure des quatre régions fonctionnelles conservées du facteur σ^{70} d'*E. coli* (Lonetto *al.*, 1992).

Les séquences consensus du promoteur, boîtes -10 et -35, sont reconnues par les régions 2.4 et 4.2 respectivement du facteur sigma. La région 2.5 représente arbitrairement une région localisée entre la région 2 et 3 et elle pourrait interagir avec le motif 5'-GT-3' en amont de la boîte -10.

Les facteurs sigma appartiennent au groupe des activateurs de classes II car ils interagissent directement avec l'ARN polymérase et activent la transcription (Barnard *et al.*, 2004). Il existe des activateurs de classe I qui se distinguent des classe II par la reconnaissance de sites cibles localisés dans les régions -60 à -90 et activent la transcription par interaction avec le domaine αCTD de l'ARN polymérase (ex, CRP ou CAP chez *E. coli*) (Schultz *et al.*, 1991). La protéine CAP (Catabolism Activation Protein) intervient dans la régulation des gènes du catabolisme des sucres. La concentration d'AMP cyclique augmente quand la quantité de glucose diminue, et l'AMPc se fixe sur la protéine CAP. Ce complexe se lie à l'ADN (sites localisés en position -60) et recrute l'unité α de l'ARN polymérase (Schultz *et al.*, 1991).

Cette association stimule l'initiation à la transcription de l'ARN polymérase et donc l'expression des gènes du catabolisme.

Les activateurs de classe I et II sont définis par leurs interactions directes avec l'ARN polymérase. D'autres motifs d'ADN peuvent faciliter la reconnaissance du promoteur par l'ARN polymérase tels que les UP éléments (Upstream promoter element) (Ross *et al.*, 1993) (figure 13). Chez *B. subtilis*, un enrichissement en dinucléotides AA et TT entre la région -36 et -70 des promoteurs reconnus par le facteur σ^A influencerait l'affinité de la fixation avec le domaine α de l'ARN polymérase (Helmann *et al.*, 1995 ; Frederick *et al.*, 1995). La contribution à l'activité de transcription de ces éléments UP ou UAS (Upstream Activation Sequences) chez *B. subtilis* a été confirmée par l'analyse du facteur σ^D (σ^{38}). Le facteur σ^D reconnait les boîtes -10 (GCCGATAT) et -35 (TAAA), et contrôle l'expression des gènes de synthèse des flagelles. Un élément UP est présent dans la région -42 et -74 au voisinage des boîtes SFBS du facteur σ^D. Ces éléments UP sont riches en nucléotide A et T. A travers des expériences de footprinting, qui permet d'identifier des régions non protégées par une interaction protéine-ADN en dégradant l'ADN, les auteurs suggèrent que les boîtes -10 et -35 confèrent une haute spécificité du promoteur quant aux éléments UP, ils confèreraient une haute activité du promoteur (Frederick *et al.*, 1995). L'analyse de l'élément UP du promoteur du gène *rrnB* chez *E. coli* montre que l'élément UP est composé de deux sous sites consensus, un distal (positions -57 à -47) et l'autre proximal (positions -46 à -38) (Estrem *et al.*, 1999). Les auteurs suggèrent que la stimulation par le site proximal requiert une fixation d'un domaine αCTD alors que la stimulation par le site distal nécessite l'interaction des deux domaines αCTD à chacun des deux sites.

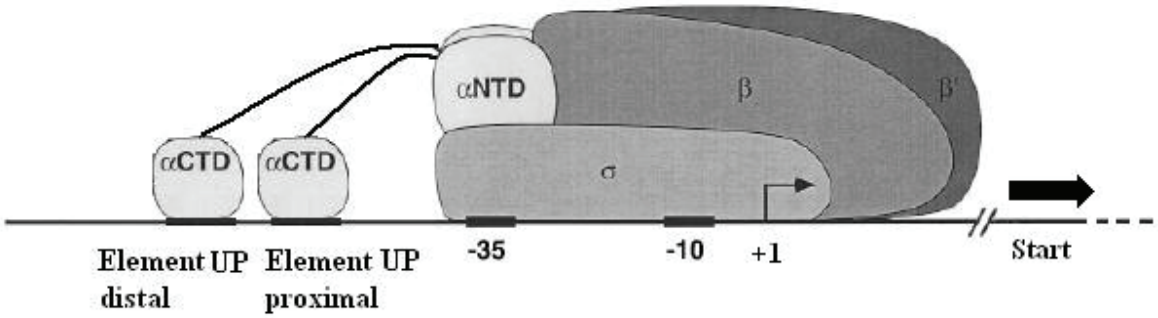


Figure 13 : Schéma d'interaction d'un ARN polymérase avec une région régulatrice permettant l'initiation de la transcription (d'après de Busby et Ebright, 1999).

2.2.1 Les facteurs sigma chez *S. coelicolor* A3(2)

Le facteur σ principal est impliqué dans l'expression des gènes de « ménage » tandis que d'autres facteurs σ alternatifs contrôlent des ensembles de gènes impliqués dans une réponse à un stimulus spécifique. *S. coelicolor* dispose de plus de 65 SF prédits (Bentley *et al.*, 2002). Quatre CDS codent des isoformes ($\sigma^{HrdA,B,C,D}$) du facteur sigma principal (σ^{HrdB}, Buttner *et al.*, 1990 ; Brown, 1992). Le SF, σ^{WhiG}, est un homologue d'une protéine de la famille des

facteurs sigma régulant la mobilité et la chemotaxie. Il s'est avéré essentiel au processus de septation du mycélium aérien pour la formation des spores (Chater, 2000). Neuf homologues du facteur général de réponse au stress, le facteur σ^B de *B. subtilis* ont été distingués ainsi que 51 facteurs sigma ayant une fonction extracytoplasmidique (ECF). Les facteurs sigma ECF sont spécifiques dans la réponse à un stimulus environnemental et activent des gènes impliqués dans des processus tels que le stress oxydatif, l'homéostasie de la paroi cellulaire et le développement du mycélium aérien (Paget *et al*., 2002). Actuellement, seuls cinq facteurs sigma ECF ont été caractérisés : les facteurs σ^U et σ^{BldN} qui ont un rôle dans la différenciation morphologique (formation des hyphes aériens), le facteur σ^E qui serait impliqué dans la structuration de la paroi cellulaire (Paget *et al.*, 1999a), le facteur σ^R qui régule des gènes en réponse au stress oxydant (Paget *et al.*, 2000), et le facteur σ^{LitS} qui serait impliqué dans la synthèse de carotenoïdes et de gènes de protection des dommages photo-oxydants (Takano *et al.*, 2005). Sur les neuf homologues du facteur σ^B de *B. subtilis*, le facteur σ^F contrôle des gènes impliqués dans la phase tardive de la sporulation (Potuckova *et al.*, 1995 ; Kelemen, 1996) et le facteur σ^H joue un rôle dans la différenciation morphologique (Cho *et al*., 2001) et dans la réponse au stress osmotique (Lee *et al*., 2005).
Le régulon du facteur σ^R (ensemble de gènes fonctionnels n'appartenant pas à un même opéron dont l'expression est contrôlée par le facteur σ^R) comporte 30 gènes caractérisés expérimentalement avec l'identification des sites de liaison, SFBS (Paget *et al.*, 2000). Le régulon σ^R représente le régulon de *S. coelicolor* le mieux défini à ce jour car le consensus SFBS est conservé parmi un nombre important de gènes.

2.2.2 Caractéristiques des sites de reconnaissance des facteurs sigma (SFBS)

Comme dans tout le règne bactérien, les SFBS montrent une grande flexibilité de leur motif de reconnaissance. C'est cette flexibilité associée à la finesse de la régulation transcriptionnelle qui rend leur identification *in silico* délicate.

Cinq niveaux de variabilité peuvent être distingués :

i) La longueur des boîtes -35 et -10. Celles-ci varient par exemple dans le cas des SFBS de *S. coelicolor* de 3 nucléotides (GTT) pour la boîte -10 du facteur σ^R à 8 nucléotides (SCCAGNNW[3]) pour la boîte -10 du facteur σ^{WhiG}.

ii) Le degré de conservation des motifs SFBS. Cette conservation peut être forte comme dans le cas du SFBS de σ^R chez *S. coelicolor* (GGAAT-N_{18}-GTT) (figure 14). En revanche, le SFBS de σ^B (GNTT-$N_{12\text{-}15}$-GGGTAT) est moins strict au niveau du motif de la boîte -35 (GNTT). Cette variation s'exprime au travers du « N » qui indique une substitution quelconque. De même, le consensus SFBS de σ^{WhiG} est très peu conservé car les deux boîtes présentent des motifs très flexibles (TRVR-$N_{14\text{-}16}$-SCCAGNNW).

iii) La longueur du séparateur entre les deux motifs des boîtes -35 et -10. Le séparateur correspond au nombre optimum de nucléotides entre les deux boîtes pour permettre une

[3] N = A/C/G/T; S = G/C; R = A/G; V = A/C/G; W = A/T.

reconnaissance efficace par le SF. La taille de ce séparateur peut être stricte comme le cas du consensus SFBS de σ^R (18 nt) ou plus variable comme pour σ^B (entre 12 et 15).

iv) La variation des motifs SFBS d'un génome à un autre. Les facteurs sigma bien caractérisés expérimentalement dans une espèce pourrait présenter un orthologue dans d'autres espèces ce qui faciliterait leur identification. Cependant, la conservation des boîtes n'est pas toujours évidente. Ainsi, les facteurs homologues σ^{70} de *E. coli* et *B. subtilis* présentent un consensus SFBS identique (TTGACA-N_{17}-TATAAT) alors que celui de *S. coelicolor* (TTGWCA-$N_{17\text{-}18}$-TAGART) est différent (Brown, 1992 ; Buttner *et al.*, 1990).

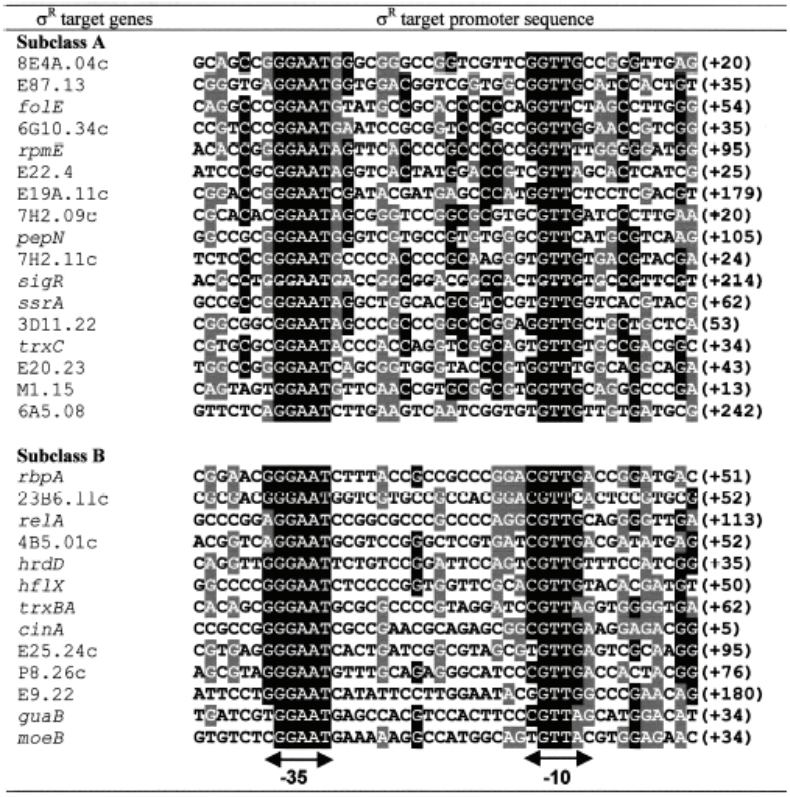


Figure 14 : Alignement des régions promotrices du régulon σ^R (Paget *et al.*, 2001).
Cet alignement utilisant clustalW met en évidence les deux motifs conservés représentant les boîtes -35 (GGAAT) et -10 (GTT) avec une longueur du séparateur de 18 nt. Les valeurs entre les parenthèses indiquent la localisation du codon start de la traduction par rapport au SFBS.

En addition du degré de variabilité des SFBS, l'identification d'un SFBS peut être rendue plus complexe par la présence de plusieurs sites de régulation en amont d'un gène. Par exemple, le gène codant le facteur σ^N présente deux promoteurs. L'un des deux sites est reconnu par le facteur σ^N (SigNP2), c'est une autorégulation, alors que le second, SigNP1, nécessite probablement un autre facteur de transcription pour activer sa transcription (Kate *et al.*, 2007). L'intérêt des sites multiples est la possibilité d'induire une régulation fine des gènes.

2.3 Fouille de données

Publication n°1 :

A new data mining approach for the detection of bacterial promoters combining stochastic and combinatorial methods.

Eng C., Asthana C., Aigle B., Hergalant S., Mari JF. and *Leblond P.*

***Journal of Computational Biology* 16**, 9 (2009) 1211-1225

1. INTRODUCTION

The versatility of gene expression is essential for the adaptation of any living organism to its environment. A key level of control of gene expression is the first transcription step (i.e., initiation), promoted by interaction of RNA polymerase (RNAP) with gene promoter. RNAP holoenzyme is recruited at a given promoter through the recognition of a promoter by a transcriptional factor, called ‘‘sigma (s) factor,” which is a variable subunit of RNAP holoenzyme. Additional regulators can modulate the efficiency of this interaction and alter the transcriptional level.

The number of σ factors is variable in bacteria, and like other regulatory genes, it is related to the ecological niche occupied by the bacterium, and reflects its ability to cope with environmental changes including biotic competition. Beside the principal σ factor involved in the expression of the housekeeping genes, the other alternative σ factors are assumed to control sets of genes (called ‘‘regulons’’) involved in response to specific environmental stimuli. Most of the σ factors in bacteria are related to their homologue σ^{70} from *Escherichia coli*. In the free living soil bacterium Streptomyces model used in this study, not less than 60 σ factors are encoded and are involved in various responses leading to morphological as well as biochemical differentiation (Bentley et al., 2002; Ikeda et al., 2003). Furthermore, fewer than 10 σ factors binding sites have been experimentally characterized, including by determining their DNA binding sites (Bibb et al., 2000; Cho et al., 2001; Paget et al., 1998, 1999; Potuckova et al., 1995). The best characterized gene network, called the ‘‘SigR regulon,’’ is involved in oxidative stress response and includes about 30 genes (Paget et al., 2001).

The σ factors usually recognize two DNA boxes (about 6 bp long), constituting what is usually called the ‘‘promoter.’’ The bacterial promoters are located approximately at 10 and 35 bp upstream of the transcription start site. The spacer between these two boxes has a variable size from 16 to 20 bp for the major σ^{70} family. The sequence of a given promoter is typified by several levels of flexibility in addition to that of the spacer length. Mismatches can be tolerated and even allow for the modulation of promoter strength at some specific genes of the regulon. In some cases, an extended -10 promoter box may be observed and may substitute for the absence of a clear -35 element.

Experimental procedures are efficient to identify individual promoters but not conceivable for sets of genes at the whole genome scale. This motivated the search for

computational methods based on the knowledge gained about the properties of known promoters or based on an efficient representation of DNA motifs by means of combinatorial or stochastic methods. There are three problems to solve when mining sigma factor binding sites (SFBSs) and generally transcriptional factor binding sites (TFBSs): (i) what kind of knowledge is available, (ii) how to locate a TFBS, and (iii) how to represent the variability of the extracted sequences. The use of prior knowledge about promoter sequences gives relevant data but can be extrapolated only to the same genome or to the more closely related strains or species. DNA stability can also be used as a characteristic of promoter regions (Kanhere and Bansal, 2005; Rangannan and Bansal, 2007).

The widely used approach for mining TFBSs is based on position weight matrices (PWM) trained on sets of nucleotide sequences known to be recognized by transcription factors (Hiard et al., 2007; Munch et al., 2005; Robison et al., 1998; Stormo, 2000). The limit of the method is the flexibility of the consensuses depending directly on the quantity of input biological data. In order to optimize the method, extra biological knowledge can be introduced into the parameters such as the UP-elements (A/T-rich regions upstream of the -35 box) (Typas and Hengge, 2005) considered in the Beadle method (Maetschke et al., 2006). Jacques et al. (2006) recently reported the use of matrices representing the genomic distribution of hexanucleotides pairs (*box1–N_x–box2*, where *box1* and *box2* ≥ 6) by estimating the ratio of hits in intergenic regions to the whole genome. Notably, the authors concluded that promoters are over-represented in intergenic regions. Other methods are trained with established descriptions of promoters such as Hidden Markov Model (HMM) (Jarmer et al., 2001; Petersen et al., 2003), support vector machine (SVM) (Gordon et al., 2006), or neural networks (Burden et al., 2005).

Statistical approaches are mostly based on the search for exceptional DNA motifs in subset of DNA sequences enriched or expected to be enriched in promoter sequences such as the upstream regions of coregulated coding sequences (Bailey and Elkan, 1994; Buhler and Tompa, 2001; Lawrence et al., 1993; Van Helden et al., 2000; Segal and Sharan, 2005) or motifs significantly conserved in upstream regions of orthologous genes (Blanchette and Tompa, 2002; Loots et al., 2002; McCue et al., 2001; McGuire et al., 2000; Rajewsky et al., 2002; Siddharthan et al., 2005; Wang and Stormo, 2003). As an example, Sigffrid (Touzain et al., 2008) can find motifs significantly conserved in upstream regions of orthologous genes in different species. But the statistical methods cannot be used for organisms with insufficient biological knowledge or with raw genome sequences.

Other methods are based on an exact description of the promoters defined as structured motifs (*box1–spacer–box2*). Vanet et al. (2000) specified an algorithm based on a suffix-tree to represent the input data and take them into account for the flexibility of the spacer by allowing jumps in the tree. This algorithm was improved by Carvalho et al. (2004) in terms of time and space but not sufficiently in terms of computation time. A similar approach is proposed by Eskin and Pevzner (2002), who first consider a composite pattern (*box1–Nx–box2*) as one larger pattern (*box1* + *box2*) and then use mismatch trees to split pattern spaces in order to rule out weak subspaces. This method shows a good efficiency to identify long patterns. However, the composite patterns proposed by both approaches are

rather long compared to the number of conserved motifs in the known promoters. Similar approaches were used by Studholme et al. (2004) and Mwangi and Siggia (2003), which compared the probabilities of all the patterns *box1–N_x–box2* with those of the motifs (*box1* and/or *box2*) upstream of coding sequences. Then, the composite patterns were clustered using sequence similarities to generate weight matrices (Li et al., 2002; Mwangi and Siggia, 2003; Studholme et al., 2004).

HMM has been used (Besemer et al., 2001; Churchill, 1989; Jarmer et al., 2001; Krogh et al., 1994; Liu et al., 2001; Nicolas et al., 2002; Thijs et al., 2001; Yada et al., 1998) for motif detection in two ways. The first approach is to assume that some information is encoded in the DNA sequence and is hidden by a background sequence that must be adequately modeled by an HMM (Liu et al., 2001; Thijs et al., 2001; Yada et al., 1998). These background models are capable of generating non-motif sequences from the genome under investigation. The second approach is to use HMMs for pattern matching or motif discovery. Here, a first-order HMM is developed as model codons and/or different patterns found in the genome (Besemer et al., 2001; Jarmer et al., 2001; Nicolas et al., 2002). The training procedure can be supervised (Jarmer et al., 2001) or unsupervised (Besemer et al., 2001), and the model can be targeted towards general motif search or carefully built for particular motifs like the SigA recognition site model (Jarmer et al., 2001).

Some review articles have carried out critical evaluations of motif discovery techniques and highlighted the limitations and potentials of the different methodologies (Osada et al., 2004; Tompa et al., 2005). In their review article, Hu et al. (2005) suggest that an ensemble algorithm approach, using analysis from different programs (methods), can complement one another's results and improve the prediction potential of DNA motifs.

We propose a new data mining method based on second-order HMM (HMM2) and combinational methods for SFBS prediction that voluntarily implements a minimum amount of knowledge. The original features of the presented methodology include (i) the use of k*mers* as observations, (ii) the Expectation Maximization estimation on the entire genome without *a priori* knowledge of their genetic content, (iii) an automatic peak extraction algorithm that captures the short DNA motifs underlying the peaks of the state *a posteriori* probability, and (iv) a suite of algorithms for finding SFBS and potential TFBS motifs.

On some points, our data mining method is similar to the work of Nicolas et al. (2002). Both methods use one HMM to model the entire genome. The parameter estimation is done in both cases by the EM algorithm. Both methods look for attributing biological characteristics to the states by analyzing the state output *a posteriori* probability. But our method differs on the following points: we use (i) an HMM2 that has proven interesting capabilities in modelling short sequences, and (ii) an original peak extraction algorithm to locate some short nucleotides sequences that could be a part of TFBS motifs (*box1* or *box2*). These sequences are further reassembled in TFBS composite motifs of the form *box1–spacer–box2* by means of combinatorial methods.

2. METHODS AND ALGORITHMS

2.1. HMM2 specifications

HMM2 was introduced by He (1988). As shown in speech recognition, modeling of the hidden process by a second-order Markov chain exhibited a good capability in representing short segments (Du Preez, 1998; Mari et al., 1997). In Data Mining (Mari and Le Ber, 2006) and Ecology (Le Ber et al., 2006), better performances were achieved by considering the contextual information defined by the neighboring observations. In our case (genomics), the nucleotides at index *t-1, … , t-k+1* define the contextual information that leads to the definition of k*mers*. The higher is the *k* value, the more important is the contextual information. It should be noted that *k* is not the order of the hidden Markov chain.

2.1.1. HMM2 mathematical definitions. An HMM2 is defined by:

- a set S of N states, $\boldsymbol{S} = \{s_1, s_2, \ldots, s_N\}$;
- a transition matrix $\boldsymbol{A} = (a_{i_1 i_2 i_3})$ over S x S x S where $a_{i_1 i_2 i_3}$ is the *a priori* transition probability $P(X_t = s_{i_3} \mid X_{t-2} = s_{i_1}, X_{t-1} = s_{i_2})$ for the hidden process to be in state s_{i_3} at index *t* assuming it was in the state s_{i_2} at index *t*-1 and s_{i_1} at index *t*-2;

a matrix ***B*** that represents the *N* discrete probability functions (*pdf*) over the set of *M* output symbols (the k*mer*). $\boldsymbol{B}(i,o) = P(O_t = o \mid X_t = s_i)$. $\boldsymbol{B}(i,o)$ is the conditional probability of symbol (k*mer*) *o* assuming the state s_i.

An HMM2 model is defined by a second-order Markov chain *X* that governs a set of *pdf* of output symbols. We have investigated the use of k*mer* as output symbols instead of nucleotides. A k*mer* may be viewed as a single nucleotide y_t observed at index *t* with a specific context $y_{t-k+1}, \ldots, y_{t-2}, y_{t-1}$ made of k-1 nucleotides that have been observed at index *t-k*+1, . . . , *t*-1. Similarly, a DNA sequence can be viewed as a sequence of overlapping k*mer* that an HMM analyzes with a consecutive shift of ℓ. For example, a sequence ##TAGGCTA can be viewed as a sequence having the same length of 3-*mer* (*k*=3 and ℓ=1): ##T-#TA-TAG-AGG-GGC-GCT-CTA, where # represents an empty context.

The HMM2 training is performed by the EM algorithm. The EM algorithm is an efficient iterative procedure (Dempster *et al.* 1977) that locally maximizes the Maximum Likelihood (ML) estimate of P(O = *o* | *HMM2*). Unfortunately, the maximum value depends on the initial conditions. The HMM2 estimation formulas implemented by the EM algorithm (Baum et al., 1970) generalize the classical formulas that are used to estimate a first-order HMM. Basically, given a sequence of symbols $o_1, o_2, \ldots, o_T$ the second-order EM algorithm performs the expected count of the transition $s_{i_1}, s_{i_2}, s_{i_3}$ at index *t-2, t-1, t*:

$$\eta_t(i_1, i_2, i_3) = \mathrm{P}\left(\mathrm{X}_{t-2} = s_{i_1}, \mathrm{X}_{t-1} = s_{i_2}, \mathrm{X}_t = s_{i_3} \middle| O_1^T = o_1^T\right) \quad (1)$$

The re-estimated second-order *a priori* transition probabilities are normalised as follows:

$$\overline{a_{i_1 i_2 i_3}} = \frac{\sum_t \eta_t(i_1, i_2, i_3)}{\sum_{i,t} \eta_t(i_1, i_2, i)}. \quad (2)$$

Whereas the re-estimated first-order *a priori* transition probabilities are normalised as follows:

$$\overline{a_{i_2 i_3}} = \frac{\sum_{i_1,t} \eta_t(i_1, i_2, i_3)}{\sum_{i_1,i,t} \eta_t(i_1, i_2, i)}. \quad (3)$$

The re-estimated output probability $\boldsymbol{B}(i,o) = P(O_t = o | X_t = s_i)$ is computed like in a HMM1:

$$\overline{\mathrm{B}(i,o)} = \frac{\sum_{i_1,i_2,t} \eta_t(i_1, i_2, i)\ 1_{o=o_t}}{\sum_{i_1,i_2,t} \eta_t(i_1, i_2, i)}. \quad (4)$$

where $1_{o=o_t}$ means 1 if the condition $o = o_t$ is true, zero otherwise.

When o is a k*mer*, $o = w_1\ w_2 \ldots w_k$ of k nucleotides $w_1, \ldots, w_k$, additional normalisation constraints are usually (Bize et al., 1999) introduced to get the output conditional probability $P(w_k \mid w_1, \ldots, w_{k-1}, X_t = s_i)$ that expresses the dependencies between the k nucleotides on state s_i. This probability is re-estimated as follows:

$$\overline{\mathrm{B}(i, w_k, \ldots, w_1)} = \frac{\sum_{i_1,i_2,t} \eta_t(i_1, i_2, i)\ 1_{y_t=w_k, y_{t-1}=w_{k-1}, \ldots, y_{t-k+1}=w_1}}{\sum_{i_1,i_2,t} \eta_t(i_1, i_2, i)\ 1_{y_{t-1}=w_{k-1}, \ldots, y_{t-k+1}=w_1}}. \quad (5)$$

We have observed (data not shown) that these constraints dramatically smooth the state *a posteriori* probability $P(X_t = s_i \mid Y_1^T = y_1^T)$. On the other hand, when these constraints are not implemented, the model cannot generate coherent sequences except if the k*mer* are not overlapping. Because, we do not use HMM to generate sequences of nucleotides, we do not implement these constraints.

Following Nicolas et al. (2002), we used a single ergodic HMM2—all the states are connected together—to model the entire genome, and estimated its parameters using Eqs. (2) and (4) of the second order EM algorithm. During the last iteration of the second-order EM algorithm, we performed the *a posteriori* decoding. At each index t (at each k*mer* position in the DNA sequence), we computed the *a posteriori* probability $P(X_t=s_i,\ X_{t-1}=s_{ij} \mid$ ''the sequence'') for a specific state s_i, and look for the fluctuations of this probability along the genome (cf. section 1.3). The sudden increase in this *a posteriori* probability locates a short DNA sequence that is typical of state s_i (called ''heterogeneity''). Obviously, to extract such heterogeneity, the *pdf* associated with the hidden states must be different. We used the Kullback-Leibler (1951) information number—called ''divergence''—between two

distributions as a distance between two states. We have designed a training strategy by successively training an ergodic HMM2 with an increasing number of states until a state appears to be distant enough from the other states. This state—called ‘‘best decoding state’’—will be used further to detect local heterogeneities. The other *a posteriori* probabilities associated to the non decoding states are disregarded.

2.1.2. HMM2 biological input preprocessing. In contrast to the pattern matching paradigm, in which the HMMs are used to discriminate between various and well-identified motifs, only two models were specified in our experiments. These models do not incorporate any *a priori* knowledge of the genetic structure of the genome of interest. The complete bacterial genome is used to train (maximum likelihood estimation using the second-order EM algorithm) two specific HMM2: HMM2+ and HMM2-. Ideally, when a marked GC skew is observed, as in the case of *B. subtilis* (Kunst et al., 1997), the HMM2± models were constructed to incorporate the biased base composition of DNA strands relative to the position of the replication origin. In contrast, when the genome does not show a definite GC skew, as in the case of *S. coelicolor* (Bentley et al., 2002), the HMM2+ and HMM2- models were constructed corresponding to the 50 to 30 sequence of the linear chromosome and its reverse complement, respectively. The best decoding state is identified for both HMM2+ and HMM2- models.

2.1.3. Automatic peaks detection and underlying sequence extraction. A peak is detected as local maximum of the *a posteriori* probability values generated over the genome, characterized by its width and its height (Fig. 1). The search of *a posteriori* probability peaks for a given state is performed using a sliding window 200-nucleotide-long window with an overlap of 100 nucleotides. The height of a peak is defined by its maximum value, which must be higher than the mean plus a given percentage of the dynamics (maximum–minimum values). This percentage is called ‘‘*peak-variation*.’’ Once a peak is located, the short DNA sequence underlying the peak (called ‘‘*Peak-motifs*’’) is automatically extracted. Both HMM2+ and HMM2- models are processed to get a set of total *Peak-motifs*. Peaks in the intergenic regions are called ‘‘*iPeaks*,’’ and the underlying short DNA sequence is called a ‘‘*iPeak-motif*.’’

2.1.4. Clustering the iPeak-motifs. In order to detect similar *iPeak-motifs*, we used an unsupervised clustering algorithm strategy (the ClusterMean algorithm; Algorithm 1), with the mean distance hierarchical grouping algorithm described by Ward (1963). This procedure builds a hierarchy where the closest sequences are in the same cluster. The distance between two sequences is defined as the inverse of their similarity, and the distance between two clusters C_1 and C_2 is defined as the mean distance between the pairs ($sequence_{i1}$, $sequence_{i2}$), where $sequence_{i1}$ belongs to C_1 and $sequence_{i2}$ belongs to C_2. The algorithm builds a hierarchy of distinct (non-overlapping) clusters that are represented by a consensus motif. Each cluster consensus motif was generated by Multalin (Corpet, 1988), and its significance was statistically validated by R’MES, a tool for finding exceptional motifs in sequence

(Hoebeke and Schbath, 2006). R'MES score is a statistical score of exceptionality, where frequent motifs have higher scores.

Algorithm 1 The ClusterMean algorithm

1. Make a N square distance matrix (N = all *iPeak-motifs*). The distance, d¼1/s, where s is the Smith and Waterman (1981) score for the local alignment between a given pair of *iPeaks*.
2. Cluster *iPeaks* into clusters using the mean distance hierarchical grouping, based on their distance values between clusters.
3. Find the consensus sequence representing each cluster of motifs using the programme Multalin.
4. Check for redundancy (group clusters with same consensus). Select statistically significant cluster consensus using the programme R'MES to verify whether a consensus is a good representation of its cluster.

2.2. SFBS consensus search algorithm

A SFBS is structured as *box1-spacer-box2* where the spacer ranges from 3 to 25 (Algorithm 2). In order to retrieve SFBSs, the basic idea of the mining strategy, is to select a cluster having a high R'MES score consensus, extend all the sequences belonging to this cluster and look for over-represented motifs using R'MES. The consensus of the cluster acts for *box1*, the shorter motifs spaced with appropriate spacer value(s) act for *box2*.

Algorithm 2 Modeling SFBS motif

```
foreach box1 (consensus motif of a cluster) do // Step1: Fixing box1
        extend all the sequences in the cluster to length of 50;
        use R'MES to find statistically significant over-represented boxes of length,
        l_box2 (these boxes act as potential box2);
                foreach1 box2 given by R'MES do // Step2: Finding box2
                        foreach2 spacer value in a given range do
                                scan_for_match (Dsouza et al., 1997) ''box1-spacer-box2'' in the whole
                                genome
                                compute the number of occurrences
                                if the number of occurrences _threshold then
                                        log on a file ''box1-spacer-box2'', the number of occurrences found,
                                        the value of spacer
                                        and the R'MES score.
                                        (e.g. GAAT-spacer-GGT, number of occurrences = 2, spacer = 6,8).
                                endif
                        end foreach2
                end foreach1
        Rank the TFBSs based on their number of occurrences and their R'MES score in descending order
        Select the N first
end foreach
Compile together the files, this gives a list of putative SFBSs and other TFBSs (than SFBS) that can be probed
for further investigation
Define a class with the genes that are found downstream each TFBS by scanning the whole genome (or
intergenic region) with the pattern matching program: scan_for_match.
```

2.3. General parameters selection using the set-sco

In order to tune (i) the HMM2 parameters (order, topology and number of iterations of the EM algorithm), (ii) the peak extraction algorithm parameters, (iii) the number of classes

during the clustering process, and (iv) the thresholds used by R'MES to detect over-represented motifs, we specified a manageable validation dataset. For this purpose, the *S. coelicolor* A3(2) genome (8.7 Mb) was fragmented into 174 non-overlapping segments of about 50 kilobases. We defined our validation data (called ''*set-sco*'') as the minimal set of fragments that includes all the genes regulated by the sigma factor SigR involved in oxidative stress response (Paget et al., 2001) scattered along the genome. SigR regulon includes at least 30 genes and is the largest regulon experimentally validated in Streptomyces. The *set-sco* covered 1.15 MB and contained 1135 genes, including the 30 SigR-regulated genes. Out of these 30 SigR genes, 25 have their -35 box starting in an intergenic DNA sequence (Table 1), while the remaining five genes have -35 box in the upstream coding sequences (Paget et al., 2001).

2.3.1. HMM order. Table 2 shows the specificity and the sensitivity at binding sites level by specifying the following values: TP (true positives), the number of binding sites predicted as binding sites; TN (true negatives), the number of nonbinding site predicted as nonbinding sites; FP (false positives), the number of nonbinding sites predicted as binding sites; and FN (false negatives), the number of binding sites predicted as nonbinding sites (Annexe A).
The specificity and the sensitivity of the SigR binding sites (-35 box) recognition using the *set-sco* are 0.24 and 0.8, respectively, with the HMM2 analysis. Using a HMM1 model, only 17 over 25 entire known sigma factor SigR binding sites (-35 box) localized in intergenic regions are detected, in contrast to the HMM2 model where all -35 boxes have been found (± 2 nt; Table 1). The specificity of HMM1 (0.23) and HMM2 (0.24) are comparable. However, the HMM2 model makes less false negatives errors and shows a better sensitivity (0.8) compared to the HMM1 (0.68).

2.3.2. HMM topology. We found in *set-sco* analysis that four states and 3-*mers* HMM2 model was the best topology. The divergence matrix identified two different decoding states for HMM2+ and HMM2- models. The number of training iterations was fixed to five and gave the best enrichment in intergenic regions relative to the coding sequences (3.8-fold; for additional data, see *set-sco* analysis; see online Supplementary Material at www.liebertonline.com).

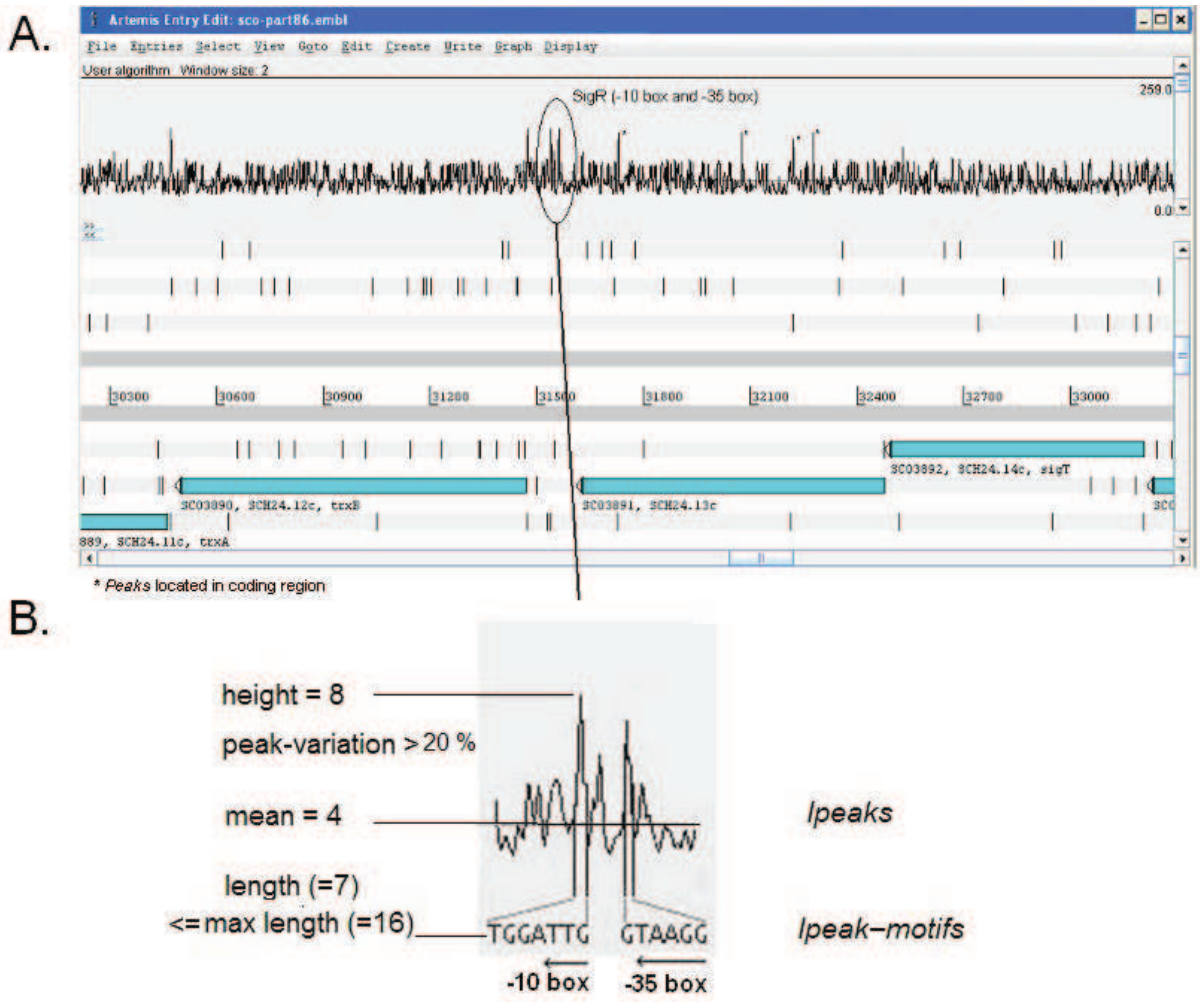


FIG. 1. Visualization of the HMM2 *a posteriori* output of *S. coelicolor*. (**A**) Figure visualized with Artemis (Rutherford et al., 2000). The top graph shows HMM2 output; the bottom shows the annotated physical sequence (using the EMBL file). (**B**) Zoom of SCO3890 *iPeaks*. The *iPeak-motifs* show the sigma factor SigR -35 box (GGAAT) and -10 box (GTT) motifs. Peak characteristics (*peak-variation* and width) are marked in the figure.

2.3.3. Peak extraction parameters. The parameters of extraction of the *iPeak-motifs* were selected to find the maximum number of SigR binding sites. Better detection of the SigR -35 box motif was obtained with a value of *peak-variation* set to 20% (Fig. 1). Table 1 shows the *iPeak-motifs* detected for *set-sco* with *peak-variation* set to 20%. All -35 boxes of the SigR binding sites are detected. The maximum width of the *iPeak-motifs* was fixed to 16 nucleotides. Less than 0.6% of the peaks were discarded by applying this threshold. Discarding these peaks noticeably improved the clustering process.

2.3.4. Clustering *iPeak-motifs*. Different similarity measures (local and global) were tested in the clustering. The Smith and Waterman local similarity (1981) gave the more homogeneous clusters. We empirically observed on the *set-sco* that the optimum number of sequences in a cluster was roughly 40–50. Deviation from this optimal cluster size skewed the consensus determination.

Table 1. Detection of SigR binding sites (-35 box: GGAAT) using HMM1 and HMM2

Gene_name	*Strand*	*Intergenic*	*HMM2*	*HMM1*
SCM1.15	1	√	AGT**GGAA**(t)	X
Trx A3, SCM1.18	1	√	GCGC**GGAAT**AC	CGC**GGAAT**ACC
SC6D7.18c, RbpA	1	√	CG**GGAAT**CTTT	X
RelA, SCL2.03	-1	√	CGGA**GGAAT**	A**GGAAT**C
7H2.09c	-1	√	ACAC**GGAAT**AG	AC**GGAAT**AGCG
7H2.11c	-1	√	TCCCG**GGAAT**GCC	CCG**GGAAT**GCC
6G10.34c	-1	√	TCCC**GGAAT**GAAT	CC**GGAAT**GAATC
8E4A.04c	-1	√	GCCG**GGAAT**GG	CCG**GGAAT**GGG
PepN	1	—	—	—
E20.23	1	√	CCGG**GGAAT**	X
E19A.11c	-1	—	—	—
ARNt 3226540	-1	√	CGCCG**GGAAT**AGG	CCG**GGAAT**AGGCT
E25.24c	-1	√	TGAGG**GGAAT**C	G**GGAAT**C
E87.13	1	√	TGA**GGAA**(t)	X
E22.04	1	√	CCGC**GGAAT**AG	CCGC**GGAAT**AGGTC
HrdD	-1	√	GTTG**GGAAT**TC	TTG**GGAAT**TCT
FoIE	-1	√	GCCC**GGAAT**GT	CC**GGAAT**GT
E9.22	1	√	CTG**GGAA**(t)	T**GGAAT**A
TrxB, H24.12c	-1	√	GCG**GGAAT**G	CG**GGAAT**GC
GuaB	1	√	GT**GGA**(at)	X
P8.26c	-1	√	AGCGTAG**GGAAT**GTT	**GGAAT**
MoeB	1	√	TCTC**GGAAT**GAAAAAG	X
23B6.11c	-1	—	—	—
SigR	1	√	GCCTG**GGAAT**G	X
RpmE3, 6G5.03	1	—	—	—
3D11.22	1	√	GCGGC**GGAAT**AGC	GC**GGAAT**AGCC
CinA	1	—	—	—
HfIX	1	√	CCG**GGAA**(t)	CCCCG**GGAAT**CTC
4B501c	-1	√	TCA**GGAAT**G	GTCA**GGAAT**GCGTC
6A5.08	1	√	CTCA**GGAAT**	X

—, not available, the -35 box of SigR binding site is located in coding sequence; X, not found.

With larger sizes, no clear consensus could be determined while in smaller clusters, the same consensus would appear in several clusters. During the Ward algorithm step, we used SigR SFBS as a marker. The process of merging two clusters, whose consensus motif contained GGAAT, is stopped when the motif consensus after merging does not contain GGAAT. The average size of the clusters was 50 sequences, and GGAAT still appeared in five clusters. In these five clusters, we observed that the motif GGAAT has a R'MES score higher than six, so this value was retained for assessing the exceptionality of five nucleotide motifs in further experiments. For a five letters motif, the R'MES threshold is 4.42 with 99.5% confidence (i.e., all motifs having score higher than this threshold are exceptional). This method assesses the validity of the consensus found by Multalin. These values were used in all further experiments, and represented the default parameters of the data mining. Extrapolating on the

entire genome, the 16,913 *iPeak*s DNA motifs were grouped into 360 clusters (i.e., approximately 47 sequences per cluster).

Table2. Comparison of the SigR binding sites (-35 box = GGAAT) recognition between HMM1 and HMM2 on the *set-sco*

	HMM2	*HMM1*
Sensitivity[a]	0.8	0.68
Specificity[b]	0.24	0.23

[a]Sensitivity = TP / (TP+FN)
[b]Specificity = TN / (FP+TN)
TP: true positive; FN: false negative; FP: false positive; TN: true negative

3. RESULTS AND DISCUSSION

Genomes of bacteria were obtained from the EMBL (Emmert et al., 1994) and were used to extract the raw sequence data used in this study as well as their corresponding annotation. DBTBS (Makita et al., 2004) was used for *B. subtilis* regulon analysis. The HMM2 core library (*CarottAge*: www.loria.fr/*jfmari/App/) was developed in C++, and included all the algorithms for modelling genomic sequences used in this study. A graphical interface was also developed in C++/XWindow to create, configure and display the model files, set the k*mer* reading methods, dump all the HMM physical topologies and screen their selected *a posteriori* probabilities alongside the sequences. The clustering software was developed in Java, and data mining and data formatting were done using Perl scripts. For additional results and data, see online Supplementary Material at www.liebertonline.com. All algorithms run on a P4 equivalent desktop computer, provided that the various parameters are set to reasonable values.

3.1. Identification of the dagA gene's promoters

The peak extraction algorithm was tested on the *dagA* gene. This gene was chosen, as it is not included in *set-sco* and it is not regulated by SigR. The *dagA* gene encodes the extracellular agarase precursor and is known to be controlled by four different promoters recognized by at least three (and probably four) different RNA polymerase holoenzymes (Buttner et al., 1988; Servin-Gonzalez et al., 1994). Using the HMM2 trained on the whole genome and the peak extraction algorithm—both tuned on *set-sco*—10 motifs could be extracted. Three of the previously known promoters were fully detected by our method. This includes *dagA*-p2 (ATTGTCA-N16-GTAGCATTC), *dagA*-p3 (GGAACTTT-N15-CTCTCGAAT), and *dagA*-p4 (TATAAGA); the motifs in brackets correspond to the previously identified binding sites. For *dagA*-p1 (TGGAGC-N18-TGGAATGA), only the -10 box (TGGAATGA) was found (for additional data, see *dagA analysis*; see online Supplementary Material at www.liebertonline.com). These results confirm the validity of the parameters tuned on the *set-sco*.

3.2. Selecting infrequent occurrences

Searching for under-represented TFBS consensuses can be done by selecting the R'MES score threshold to select infrequent occurrences (higher negative scores). The extracted *iPeak-motifs* exhibited the *BldN* binding consensus and identified the unique gene, annotated *bldM* (SCO4768), reported as a *BldN* target (Bibb et al., 2000).

3.3. Annotating SFBSs from S. coelicolor genome

This data mining method was next applied to the whole genome of *S. coelicolor*. We extracted 16,913 *iPeaks* DNA motifs grouped into 360 clusters. Each cluster had a significant motif sequence according to R'MES, and after filtering duplicated consensus, 357 motifs were obtained. So, 228 out of 357 motifs holding out R'MES score higher than six (see Section 2.3), were used for further analysis. Table 3 shows a partial list of cluster consensus motifs that contain a box part (-35 or -10) of already known SFBSs like SigB binding sites and WhiG binding sites.

The parameters *spacer* and l_{box2} were restricted to a range of values consistent with biological knowledge of the SFBSs. For all the 228 motifs, the number of possible consensuses is the product between the *spacer* and *length* of *box2*, which equals 8,280 possibilities for $3 \leq spacer \leq 5$ and $3 \leq l_{box2} \leq 5$. To reduce computation time from a couple of days to a few hours, the range of *spacer* and of the R'MES threshold was restricted ($14 \leq spacer \leq 20$). We found 6,236 *box1-spacer-box2* putative consensuses. Finally, applying the modeling SFBS motif algorithm (see Section 2.2), 812 potential TFBS consensuses were suggested and are available including known SFBS consensuses (for additional data, see online Supplementary Material at www.liebertonline.com).

Table 4 shows the main results of 812 predicted TFBSs:

- Five motifs that include or overlap the SigR consensus (GGAAT-N_{18}-GTT) were found.
- Three putative WhiG binding sites (Chater et al., 1989) were found. The extracted set of genes found downstream of these consensuses includes the two previously biologically determined targets of *whiG*, *whiI* (SCO6029) and *whiH* (SCO5819) (Ainsa et al., 1999; Ryding et al., 1998).
- A modified SigB consensus (ACGGTTT-N_{18}-TAC) was proposed. SigB is a general stress σ factor that has a binding site consensus ANGNNT-$N_{14\text{-}16}$-GGGTA (Cho et al., 2001).
- Two new consensuses for PTF2 were found. PTF2 is a regulatory motif proposed by Studholme et al. (2004) to be associated with the sigma factor regulation or expression.

Table 3. Partial list of cluster motifs defined from ClusterMean algorithm

No.	*TFBS box*	*Known Motif*	*Predicted Motif*	*R'MES score*
1. SigB	-35	ANGNNT	AAGATt	7,3824
1. SigB	-35	ANGNNT	ACGGCTt	8,4372
1. SigB	-35	ANGNNT	AGGCTt	8,4699
1. SigB	-35	ANGNNT	AGGATt	8,3462
1. SigB	-10	GGGTA	CGGGTa	8,2924
2. SigE	-10	TCTY	GCTCTt	6,628
2. SigE	-10	TCTY	ATCTT	7,9511
3. SigR	-35	GGAAT	CGGGAat	8,2738
3. SigR	-35	GGAAT	TCGGAAt	8,273
3. SigR	-35	GGAAT	GGGGAat	8,026
3. SigR	-35	GGAAT	GCGGAAt	8,2233
3. SigR	-35	GGAAT	AGGGAAt	8,2155
3. SigR	-10	GTT	GCGTTA	8,286
3. SigR	-10	GTT	aCCGTTt	8,2254
3. SigR	-10	GTT	CTGTTT	8,3584
4. WhiG	-35	TRVR	cCTGAAa	8,1572
4. WhiG	-35	TRVR	GCTGAa	8,201
4. WhiG	-35	TRVR	TGGATt	8,2126
4. WhiG	-35	TRVR	cTTGAAt	7,2117
4. WhiG	-35	TRVR	TGCGAA	8,2401
5. PTF1	-35	GAAC	GAACTt	6,332
5. PTF1	-10	GTTG	gTTGAa	7,901
6. PTF2	-35	TGGT	cTGGTAa	6,062
6. PTF2	-10	ACCA	ACCAAt	8,185
6. PTF2	-10	ACCA	ACCAT	8,038

Cluster motifs overlapping with known or putative TFBS -10 or -35 box motifs = 137 (total number of cluster motifs = 357).
R'MES stat = score of exceptionality (exceptionally frequent motifs will have high positive scores, whereas exceptionally rare motifs will have high negative scores).
PTF (1, 2, and 3): regulatory motif proposed by Studholme et al. (2004).
R = A/G; V = A/C/G; Y = C/T.

3.4. Advantages of the new data mining method

This method is innovative in its use of the second-order Markov model to capture stochastic dependencies in k*mers* (see Section 2.1). HMM2 shows good performance in modeling the DNA heterogeneities and possesses three main advantages:

- It successfully detects small motifs (5–16 bases), minimizing the elusion of important regulatory sequences.
- It efficiently detects statistically exceptional motifs in the genome (e.g., the sigma factor BldN binding site).

- It proves to be versatile. This is why this method could be applied to the detection of DNA motifs other than SFBSs. Some of our detected motifs correspond to regulatory motifs such as those proposed by Studholme et al. (2004). Also, the ribosome binding sites (RBS), which are translational signals, have also been detected during this analysis. The *S. coelicolor* genome contains 2,576 RBSs annotated by Bentley et al. (2002)—2,041 of them in intergenic regions. Among them, 236 (12.9%) were revealed by our HMM2. Further, the same background HMM2 was used to detect various DNA repeats in the *S. coelicolor* genome (Hergalant et al., 2002).

Table 4. Known TFBSs generated by the Data Mining algorithm

	Function/Involvement	*Consensus already defined*	*Consensus predicted*
R'MES score ≥ 6 $14 \leq spacer \leq 20$ $3 \leq l_{box2} \leq 5$			
WhiG	Early sporulation	**T**RVR-N_{14-16} - S**CCAG**NNW	GG**T**GAAT-N_{18}-**AG**T TGCGAA-N_{14}-CAG GCCGTAGG-N_{15}-GGCC
SigB	Osmotic stress response and aerial hyphae differentiation (sigma factor)	**A**N**G**NN**T**- N_{14-16} - **GGGTA**	**ACG**GT**T**T-N_{18}-**TA**C
SigR	Oxidative stress response (ECF sigma factor)	**GGAAT**-N_{18} -**GTT**	CG**GGAAT**- N_{18}-**GTT** TC**GGAAT**- N_{16}-TG**GT** GG**GGAAT**- N_{17}-G**GTT** CCC**GGAAT**- N_{17}-G**GT** AG**GGAAT**- N_{18}-**GTT**
R'MES score ≥ 4 $3 \leq spacer \leq 20$ $3 \leq l_{box2} \leq 4$			
PTF2 Studholme	Proposed to be associated with sigma factor regulation/expression	**TGAC**-N_{19} -**TGAC**	**TGAC**TTT-N_{16}-**TGA**

PTF2: regulatory motif proposed by Studholme et al. (2004).
N = A/C/G/T; S = G/C; R = A/G; V = A/C/G; W = A/T.

3.5. Parameters influencing the behaviour of the HMM2

3.5.1. Influence of the GC%. The genome of the *S. coelicolor* is GC rich (72.12%), and the intergenic regions present a slightly lower GC content (68.34%). In addition, the GC% of the extracted DNA motifs was 48.27%. Therefore, it may appear that the GC content may strongly influence our HMM2. To test this hypothesis, we classified the *iPeak-motifs* into 20 classes based on their GC%. Only one quarter of the *iPeak-motifs* showed a GC% in the 45–55 range, which includes the mean GC% (48.27%), and one third of them had a GC% scattered between the 40–45 and 55–60 range (for additional data, see distribution of *iPeak-motifs* GC content; see online Supplementary Material at www.liebertonline.com). Furthermore, when the GC content of the intergenic sequence was observed along with the

output probabilities, GC falls did not always generate peaks (data not shown). Thus, it can be concluded that the GC content does not have an exclusive influence on the model.

3.5.2. Influence of the pyrimidine/purine step and DNA repeats. DNA curvature plays a well characterized role in many protein/DNA interactions and also more precisely in transcriptional regulation mechanisms (Jauregui et al., 2003). It is also known that the flexibility of the DNA backbone is influenced by pyrimidine-purine steps (YR steps) (Bertrand et al., 1998). The percentage of YR steps determined within the intergenic regions of the genome (24.97%) and within the *iPeak-motifs* (23.58%) had a significant difference with a 95% confidence value. These results show that the YR step has an influence on the model (for additional data, see online Supplementary Material at www.liebertonline.com).

We also tested whether HMM2 was influenced by the presence of DNA repeats in the *iPeak-motifs*. Only 431 out of the 16,913 *iPeak-motifs* are found more than twice in the same intergenic region. Thus, HMM2 does not seem to be strongly influenced by local repeats of the same *iPeak-motif*. In contrast, we noticed that 8,520 *iPeak-motifs* (50.37%) were repeated at least once in another intergenic region. This result by itself is consistent and promising regarding the possibility of describing groups of co-regulated genes.

3.6. Preliminary application of the method to the bacterial model Bacillus subtilis

The above software suite was applied to the *B. subtilis* genome (Kunst et al., 1997). The main change to the model was in the HMM2± specification due to the GC skew of its genome (see Section 2.1). The GC% of *B. subtilis* (roughly 43%) is distinct to one's of *S. coelicolor* (72%), but as demonstrated in section 3.2.1, the GC content does not exclusively influence the behavior of the model. We carried out a preliminary analysis of the *B. subtilis* genome using the same parameters (HMM2 topology, clustering parameters, R'MES score, motif structure) than for *S. coelicolor* analysis. The HMM2 model (four states, 3-mors, *peak-variation* set to 20%) generated 19,507 *iPeaks*. The *iPeak-motifs* were clustered into 360 clusters. Among them, 191 clusters had a R'MES score higher than six and were further considered as significant motifs. Each of them was successively considered as *box1* (-35 box and -10 box) to find the second *box2* (-10 box and -35 box, respectively) by the SFBS consensus search algorithm in order to produce *box1-spacer-box2* consensuses. The length of the spacer was set between 14 and 25 nt and the length of the *box2* (l_{box2}) was set between three and five. Finally, the data mining process generated 2,317 and 2,342 putative TFBS motifs respectively depending if we used the significant motifs as *box1* or *box2*. The genes that are controlled by the biologically described SigD, E, F, G, H, W, and X sigma factors (Makita et al., 2004) were retrieved with a ratio ranging from 25% to 52% (for additional data, see table of *B. subtilis* preliminary results; see online Supplementary Material at www.liebertonline.com).

5. CONCLUSION

We have proposed a hybrid data mining method based on stochastic and combinatorial algorithms to find SFBSs. HMM2 was processed on the whole genome without prior assumptions about its genetic organization. As already shown in speech recognition, data mining and ecology, the modeling of the hidden process by a second-order Markov chain exhibited a good capability in representing short segments such as SFBSs and other DNA motifs involved a transcriptional regulation. The parameters of the data mining process were tuned on a subpart of the genome—the *set-sco*—defined as a biologically well characterized regulon (SigR). On this subset, the HMM2 outperform the HMM1. On the actinomycete *S. coelicolor* genome, some already known promoters (SFBS) were found that strengthen the validity of the method, and we suggest a list of putative promoters (TFBS). Furthermore, the preliminary results from *B. subtilis* are promising with respect to the generalization of the method to other bacterial genomes.

Analyse du génome de *Bacillus subtilis*

2.4 Compléments de résultats pour l'identification de SFBS putatifs sur *Bacillus subtilis*

L'analyse de *B. subtilis* a été réalisée en utilisant les mêmes paramètres définis pour l'analyse de *S. coelicolor* afin de déterminer si les paramètres trouvés étaient généralisables.
Le génome de *B. subtilis* est circulaire avec une taille de 4,2 Mpb et un contenu en base G et C de 43 %. Le génome de *B. subtilis* présente un biais compositionnel indiquant la richesse d'un brin en G par rapport au C : le GC skew. Ce biais influencerait le HMM2 (cf. section 1.3.1). La position du terminus de réplication a été identifiée (1 947 501 nt) en utilisant le logiciel genskew[4]. Cette prédiction du terminus de réplication est cohérente avec celle proposée précédemment (Kunst *et al.*, 1997 ; Franks *et al.*, 1995). Deux séquences génomiques ont été générées *in silico* afin de résoudre ce problème. La première séquence est la concaténation des brins sens du bras droit et gauche (brin orienté dans le sens de réplication) et la seconde séquence est la concaténation des brins complémentaires réverses. Ces deux séquences ont permis l'estimation des paramètres des modèles HMM2 (HMM2+ et HMM2-).
La même stratégie de recherche de SFBS appliquée à l'analyse du génome de *S. coelicolor* a été employée pour celle du génome de *B. subtilis* (annexes B, C et annexe Web A) et un total de 19 507 *iPeak-motifs* a été extrait en utilisant un seuil « *peak-variation* » de 20 %. L'algorithme de classification regroupe ces *iPeak-motifs* en 360 classes distinctes formant 360 consensus avec *MultAlin* (tableau 3). Après un regroupement des consensus identiques (348 consensus) et une sélection des consensus robustes avec *R'MES* (score > 6), il en résulte 191 consensus. La recherche de motif composite (*box1-séparateur-box2*) est réalisée en utilisant les 191 consensus robustes comme *box1* (boîte -35 puis boîte -10) afin de retrouver la *box2*. L'algorithme de recherche des SFBS utilise une longueur du séparateur compris entre 14 à 25 nucléotides et une longueur de la *box2* fixée entre 3 et 5. Dans un premier temps, les consensus robustes ont été assimilés à des *box1* (boîte -35) pour la recherche des *box2* (boîte -10), ce qui a révélé 10 879 SFBS putatifs. Les 10 meilleurs motifs (Top10) de chaque consensus représentent 2 317 SFBS putatifs. Dans un deuxième temps, les consensus robustes correspondent à des boîtes -10 pour trouver des boîtes -35, ce qui a défini 10 070 SFBS putatifs et la sélection des 10 meilleurs motifs pour chaque motif consensus a abouti à 2 342 SFBS putatifs. Un total de 4 659 SFBS putatifs a été extrait (tableau 3).
Il s'agit maintenant de comparer ces motifs avec des motifs caractérisés expérimentalement. Cette comparaison est effectuée au travers des données fournies par DBTBS (DataBase of Transcriptional regulation in *Bacillus subtilis*) (Sierro *et al.*, 2008).

[4] Genskew : http://webclu.bio.wzw.tum.de/genskew/genskew

Tableau 3 : Recherche de SFBS putatifs sur *B. subtilis*

	Box1 (box -35)	Box1 (box -10)
Nombre de classes (*MultAlin*)	348	348
Nombre de classes robustes (*R'MES*)	191	191
Nombre de SFBS putatifs	10,879	10,069
Nombre de SFBS putatifs (top10)	2,317	2,342

La base de données DBTBS (Sierro *et al.*, 2008) référence les informations concernant les TFBS chez *B. subtilis*. Ces données sont issues de données expérimentales dont la version utilisée (release 5, update 2008) comporte 1475 annotations de promoteurs. A partir des TFBS identifiés expérimentalement, le site propose une modélisation permettant la recherche de nouveaux TFBS putatifs par le calcul de matrices de poids-position (tableau 4) et une représentation découlant du calcul Weblogo (Crooks *et al.*, 2004).

Tableau 4 : Les SFBS des facteurs sigma de *B. subtilis* identifiés dans la DBTBS

Facteur sigma	Nombre d'unité de transcription[1]	Consensus (recensé sur DBTBS)[2]	Consensus connus[3]
SigA	362	TTGACA-N_{14}-tgnTATAAT	TTGACA-N_{17}-TATAAT
SigB	64	AGGTTT-N_{17}-GGGTAT	AGGTTT-N_{17}-GGGTAT
SigD	30	TAAA-N_{15}-GCCGATAT	TAAA-N_{15}-GCCGATAT
SigE	37	KMATATT-N_{14}-CATACA-T, GKMATRWY-N_{13}- MATACAMT	KMATAWW-N_{14}-CATACAMT
SigF	25	YGY(A/T)TA, GGMAWAMTA	GCATR-N_{15}-GGMRARMTW
SigG	55	GMATR, GG-CATXMTA	GMATR-N_{18}-CATWMTA
SigH	25	AGGTATT, GAATT	RVAGGAWWT-N_{14}-MGAAT
SigK	32	MACM-N_{16}-CATA---TA	AC-N_{17}-CATANNNTA
SigW	34	TGAAAC-N_{16}-CGTA	TGAAAC-N_{16}-CGTA
SigX	15	TGTAAC-N_{17}-CGAC	TGTAAC-N_{17}-CGAC
SigY	2	GAAC-N_{18}-GTC	GAAC-N_{18}-GTC

[1] Nombre d'unité de transcription recensé expérimentalement dans la base de données DBTBS.
[2] Nomenclature de IUPAC-IUB ; M = (A, C); (N ou X) = (A, G, C, T); R = (A, G); K = (G, T); Y = (C, T); W = (A, T); V = (G, A, C).
[3] Motifs expérimentalement définis par Haldenwang (1995) avec transformation au format IUPAC-IUB. Les égalités représentent les équivalences entre les annotations issues de Haldenwang (1995) qui sont entre parenthèses et celles de IUPAC-IUB ; (H) = M; (Z) = K; (X) = W; (W) = V.

La base de données DBTBS propose 11 consensus de SFBS. Ces consensus ne permettent pas de retrouver la totalité des SFBS identifiés expérimentalement. En effet, les consensus proposés par DBTBS ne reconnaissent qu'une partie des SFBS identifiés expérimentalement. Par exemple, le facteur sigma principal, σ^A, reconnait 362 sites et le consensus proposé par DBTBS est TTGACA-N_{14}-tgnTATAAT. La recherche de ce consensus dans les régions promotrices des gènes régulés ne donne qu'une cible unique. La représentation du consensus peut être plus flexible en utilisant la représentation Weblogo (TTGWMW-$N_{13\text{-}22}$-TAHAAT) où 29 séquences promotrices possèdent le motif. En élargissant au maximum la représentation avec Weblogo (TTGNNN-$N_{13\text{-}22}$-TANAAT), le nombre de motifs retrouvé est de 86, ce qui

reste encore insuffisant. Les résultats soulignent la flexibilité des sites de reconnaissance par σ^A.

La comparaison entre les consensus proposés par DBTBS et les TFBS putatifs prédits par notre méthode serait donc approximative. C'est pourquoi, nous avons entrepris la recherche des SFBS putatifs prédits par notre méthode dans les régions régulatrices, dénombrées dans DBTBS. Dans un premier temps, nous avons extrait les 681 régions promotrices reconnues par les onze facteurs sigma de la base de données DBTBS (tableau 5). Nous avons révélé que 377 des SFBS prédits par les HMM2 étaient retrouvés dans les régions promotrices des facteurs sigma.

Tableau 5 : SFBS identifiés dans les régions promotrices

Facteur sigma	Nombre de site de fixation caractérisés(1)	Séquence promotrice retrouvée par notre stratégie (2)	Ratio (2)/(1)	Nombre de SFBS putatifs bien localisés (3)	Ratio (3)/(1)
SigA	362	216	0.60	X	X
SigB	64	31	0.48	X	X
SigD	30	15	0.50	9	0.30
SigE	37	20	0.54	12	0.32
SigF	25	16	0.64	8	0.32
SigG	55	27	0.49	14	0.25
SigH	25	15	0.60	13	0.52
SigK	32	11	0.34	X	X
SigW	34	17	0.50	15	0.44
SigX	15	8	0.53	6	0.40
SigY	2	1	0.5	X	X
Sous-Total	221	118	0.53	77	0.35
Total	681	377	0.55	X	X

X : analyse non réalisée
La colonne 1 indique le nom des facteurs sigma. Les cases grisées sont les SFBS utilisés pour comparer les positions SFBS prédites par le HMM2 et les SFBS caractérisés dans DBTBS.
La colonne 2 dénombre le nombre de régulons retrouvés dans DBTBS pour chaque facteur sigma.
La colonne 3 correspond au nombre de SFBS putatifs retrouvés dans les régions régulatrices des facteurs sigma.
La colonne 5 énumère le nombre de SFBS putatifs concordant avec les positions réelles des SFBS analysés et identifiés dans DBTBS.

Afin de vérifier si les sites prédits correspondent aux sites expérimentalement décrits, nous avons localisé tous les SFBS putatifs par rapport aux SFBS réels. Les sites prédits ont été localisés pour sept des onze facteurs sigma identifiés (tableau 5). Ces sept facteurs sigma ont été analysés car le nombre de gènes régulés par ces facteurs sigma sont traitables manuellement. Les SFBS putatifs ont été comptabilisés lorsque leur boîte -35 et boîte -10 chevauchaient au moins deux nucléotides dans au moins une des boîtes -35 et -10 caractérisées. Un exemple d'identification des SFBS putatifs bien positionnés est présenté dans le cas de SigD (annexe D). Ainsi, sur les 221 unités transcriptionnelles analysées, 118

SFBS prédits sont localisés dans les régions promotrices dont 77 d'entre eux chevauchent au moins de deux nucléotides les régions régulatrices des 221 unités transcriptionnelles. Le nombre de SFBS identifiés est donc de 77 sur 221 soit un tiers. Dans le but de définir si les 41 SFBS prédits restants sont, soit des faux positifs (FP), soit des sites de régulations potentiels, un enrichissement de ces motifs prédits devrait être observé dans les régions intergéniques par rapport au reste du génome, une des caractéristiques des SFBS.
Le tableau 6 représente une moyenne de l'enrichissement calculé pour chaque SFBS potentiel qui sera comparée à celle trouvée pour le SFBS caractérisé expérimentalement.
Dans le cas du facteur sigma SigD, trente régulons ont été identifiés dans DBTBS et quinze SFBS prédits ont été localisés dans ces régions régulatrices. Parmi les quinze SFBS putatifs, neuf sont bien positionnés dans les SFBS caractérisés de la base DBTBS. Il reste alors six autres régions régulatrices où les SFBS putatifs sont localisés dans les régions intergéniques contenant les SFBS caractérisés mais ces SFBS putatifs ne sont pas bien positionnés (soit en partie, soit en totalité). Pour chaque région régulatrice, plusieurs SFBS putatifs peuvent être trouvés. Par exemple, la région régulatrice du gène *yjcP* est caractérisée par le site de reconnaissance (TAAT-N_{15}-CCCGATAT) et ce dernier est décrit par trois SFBS putatifs (ttt-N_{19}-TGGG ; aac-N_{22}-TGGG ; ata-N_{24}-TGGG). Une recherche du nombre d'occurrences des SFBS putatifs non positionnés correctement est effectuée sur tout le génome et seulement dans les régions intergéniques. Ceci permet de définir un ratio d'enrichissement dans les régions intergéniques. Cette procédure est réalisée sur l'ensemble des SFBS putatifs non positionnés correctement et une valeur moyenne du ratio est obtenue. Le tableau 6 présente les résultats obtenus pour les sept facteurs sigma analysés.
La comparaison (tableau 6) révèle une concordance entre les facteurs d'enrichissement des SFBS prédits et non identifiés et ceux identifiés expérimentalement (SigD, SigE, SigF, SigG, SigH). Certains de ces SFBS putatifs pourraient être des sites de régulations encore non identifiés.

Tableau 6 : Facteur d'enrichissement des SFBS potentiels

Facteur sigma	Nombre de SFBS potentiels non identifiés / Nombre de SFBS putatifs	Facteur moyen des enrichissements (intergénique/tout le génome) [déviation standard]	Facteur d'enrichissement pour les SFBS décrits expérimentalement
SigD	6/15	2.06 [1]	1.03
SigE	8/20	1.70 [1.52]	1.23
SigF	8/16	1.44 [0.86]	1.80
SigG	13/27	1.05 [0.34]	1.38
SigH	2/15	1.52 [0.55]	1.37
SigW	2/17	1.48 [0.73]	7.69
SigX	2/8	1.18 [1.16]	3.42

L'analyse du génome de *B. subtilis* avec notre stratégie de la fouille de données pour l'identification de SFBS est prometteuse. En appliquant les paramètres par défaut, au moins un tiers des SFBS décrits chez *B. subtilis* sont identifiés malgré l'utilisation de paramètres issus d'analyse d'un autre génome bactérien, *S. coelicolor* qui comporte des caractéristiques

compositionnelles différentes (GC%, GC skew). Cela conforte la généralisation possible de l'utilisation de cette stratégie.

Chapitre 3

Utilisation des HMM d'ordre 2 (M2M2) pour la détection de gènes acquis par transfert horizontal, HGT

INTRODUCTION 54
3.1 MÉCANISMES DE TRANSFERT HORIZONTAL DE GÈNES (HGT) 54
3.2 MAINTIEN DE L'INFORMATION GÉNÉTIQUE 57
3.3 AMÉLIORATION DES SÉQUENCES ACQUISES PAR TRANSFERT HORIZONTAL 58
3.4 MODÈLE D'ÉTUDE : *STREPTOCOCCUS THERMOPHILUS* 59
3.4.1 Le genre Streptococcus 59
3.4.2 Le « core » génome et le génome accessoire de Streptococcus thermophilus 61
3.4.3 Flux de gènes chez S. thermophilus 62
3.5 MÉTHODES DE DÉTECTION DES GÈNES EXOGÈNES 64
3.5.1 Méthodes phylogénétiques 64
3.5.2 Méthodes paramétriques 67

MATÉRIEL ET MÉTHODES 71
3.6 EXTRACTION DES POSITIONS ET DES RÉGIONS « ATYPIQUES » 71
3.6.1 Traitement des probabilités a posteriori 72
3.6.2 Détection et extraction des régions « atypiques » 73
3.7 CRÉATION D'UNE BASE DE DONNÉES 75
3.8 DÉFINITION D'UN NOUVEAU PARAMÈTRE : LA VALEUR DE DISPERSION 76
3.8.1 Réalisation d'un arbre phylogénétique des firmicutes 77
3.8.2 Calcul de la valeur de dispersion pour un gène 79

RÉSULTATS : ANALYSE DU GÉNOME DE *S. THERMOPHILUS* 81
3.9 ANALYSES PRÉLIMINAIRES DE SÉQUENCES CHIMÉRIQUES 81
3.10 ANALYSE GLOBALE DU GÉNOME DE *S. THERMOPHILUS* CNRZ1066 83
3.10.1 Extraction et analyse de gènes « atypiques » 83
3.10.2 Annotation des gènes « atypiques » 89
3.10.3 Distribution phylogénétique des gènes « atypiques » 91
3.10.4 Gènes d'adaptation chez S. thermophilus 99
3.10.5 Ilots génomiques S. thermophilus 100

CONCLUSION 104
3.11 CDS « ATYPIQUES » DE PETITE TAILLE ET PSEUDOGÈNES 104
3.12 ENRICHISSEMENT EN BASES A ET T DES SÉQUENCES « ATYPIQUES » 105

3.13 VALIDATION DU M2M2 COMME MÉTHODE D'IDENTIFICATION DE GÈNES ISSUS DU TRANSFERT HORIZONTAL 106
3.14 SCÉNARIO ÉVOLUTIF DU GÉNOME DE *S. THERMOPHILUS* 109

Introduction

Dans ce chapitre, nous allons présenter la deuxième application des HMM d'ordre 2 ; l'identification des séquences longues (1 à plusieurs kilobases) ayant des propriétés compositionnelles atypiques. Cette analyse sans *a priori* rend compte des propriétés statistiques de l'ADN sans présager de la nature de leur hétérogénéité et de leur signification biologique. Les séquences « atypiques » identifiées présentent donc des caractéristiques compositionnelles qui les distinguent du reste du génome. Elles peuvent correspondre à des séquences résultant d'événements de transfert horizontal de gènes ou encore à des séquences endogènes soumises à une pression de sélection induisant un biais compositionnel.
La première partie du chapitre correspond à une introduction s'intéressant aux différents mécanismes de transfert horizontal (HGT pour Horizontal Gene Transfer), aux mécanismes intervenant dans l'évolution du génome de *Streptococcus thermophilus* et aux méthodes de détection du HGT. La seconde partie du chapitre est une description détaillée de la stratégie de fouille de données développée pour l'identification et l'extraction des gènes « atypiques ». La troisième partie du chapitre présente les résultats obtenus sur des séquences chimériques tests ainsi que ceux obtenus pour l'analyse du génome complet de *S. thermophilus*. Les gènes « atypiques » extraits par le HMM d'ordre 2 (M2M2) sont ensuite analysés au travers de leur annotation fonctionnelle, de leur distribution au sein du phylum des firmicutes et de 47 souches de *S. thermophilus*.

3.1 Mécanismes de transfert horizontal de gènes (HGT)

Les bactéries se reproduisent de façon asexuée, ce qui engendre théoriquement un isolement génétique des individus présents à un temps donné dans un écosystème. Cependant, des mécanismes d'échanges d'information génétique (conjugaison, transformation, transduction) assurent non pas une homogénéisation du « pool » génétique (correspondant au patrimoine génétique) mais une diversification rapide des gènes « accessoires » des bactéries. Ces phénomènes, dits de parasexualité, permettent un brassage des gènes en absence de fécondation. Ils interviennent entre bactéries, dont les liens phylogénétiques sont variés, présentes dans un même écosystème. Ils ont été qualifiés de mécanismes de « transfert horizontal » par contraste à l'hérédité « verticale » qui s'opère de la cellule mère vers la cellule fille. Ces phénomènes, de part leurs mécanismes distincts, entraînent une hétérogénéité des séquences transférées en termes de fonctions codées et de longueur. De plus, l'existence de processus d'amélioration, c'est-à-dire une convergence de propriétés compositionnelles de la séquence exogène vers celle du génome hôte, rend plus complexe l'identification de ces séquences exogènes lors de l'analyse de la séquence du génome hôte.

Trois mécanismes majeurs de transfert horizontal de gènes ont été distingués (figure 15).

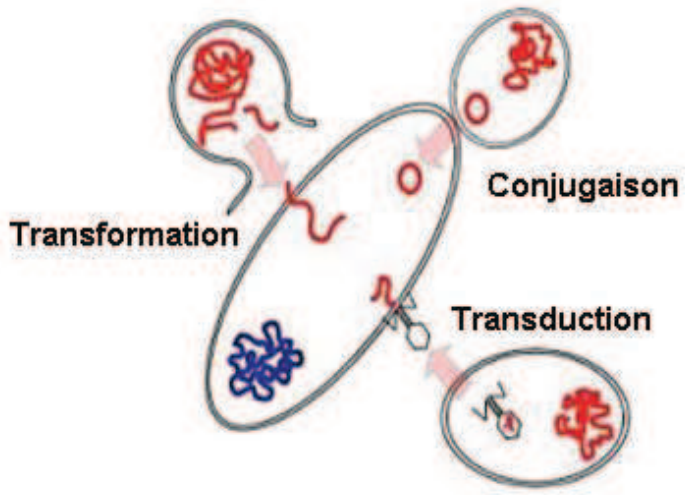


Figure 15 : Mécanismes de parasexualité chez les bactéries.

Transduction

La transduction requiert l'intervention de bactériophages comme vecteur d'ADN génomique (figure 15). La taille des génomes des bactériophages montre une variabilité importante (de quelques kb à plusieurs centaines de kb). Deux types de transduction peuvent être distingués : la transduction généralisée et la transduction spécialisée. La transduction généralisée résulte de l'encapsidation accidentelle d'un fragment d'ADN génomique de la cellule hôte dans une particule phagique. Au cycle d'infection suivant, cette séquence pourra être recombinée avec le génome de l'hôte. Dans ce phénomène, tous les gènes chromosomiques sont susceptibles d'être transférés, et la taille de la région transférée théorique est équivalente à la taille du génome phagique. La fréquence de formation de particules « transductrices » est le plus généralement de 1 à 2 %. Quant à la transduction spécialisée, elle est l'apanage des phages lysogènes, c'est-à-dire dont le génome s'intègre à l'ADN chromosomique de la bactérie hôte, et résulte de l'excision impropre de l'ADN du virus de son site chromosomique. Les quelques marqueurs adjacents au site d'insertion du virus peuvent être mobilisés avec le génome phagique. Lors du cycle d'infection suivant, le bactériophage transducteur s'intègre dans le génome du nouvel hôte et entraîne le transfert des gènes chromosomiques.

Le spectre de transfert par transduction est limité par le spectre d'hôte des bactériophages. En effet, les bactériophages reconnaissent un récepteur membranaire spécifique lors de l'étape d'adsorption (Sorensen *et al.*, 2005). En outre, les mécanismes d'encapsidation d'ADN chromosomique dans les particules phagiques peuvent être plus ou moins spécifiques et limitent voire interdisent la formation de particules transductrices (Friedman *et al.*, 1992 ; Smith et Feiss, 1995). Le nombre de gènes transférés (ou la taille de la séquence transférée) dépend de la taille du génome du bactériophage et de la nature de la transduction (spécialisée ou généralisée).

Transformation naturelle

La transformation naturelle implique le développement de l'état physiologique dit de « compétence » qui correspond à la capacité d'internaliser de l'ADN présent dans le microenvironnement cellulaire (Griffith, 1928 ; Avery *et al.,* 1944). Classiquement, l'ADN libre doit être sous forme double brin et son internalisation se fait sous forme simple brin. Le mode d'internalisation est légèrement différent entre les bactéries Gram positives et Gram négatives. Par exemple, dans le cas de *Streptococcus pneumoniae* (Gram positive) l'internalisation de l'ADN exogène est assistée par des complexes protéiques membranaires qui capturent l'ADN double brin et internalisent l'un des deux brins tout en dégradant l'autre (Méjean et Claverys, 1993). L'ADN internalisé peut alors s'intégrer par recombinaison (homologue ou illégitime). Le phénomène d'internalisation est le plus souvent non spécifique vis-à-vis de l'origine de la séquence d'ADN. Ainsi, le transfert peut avoir lieu entre bactéries extrêmement distantes d'un point de vue phylogénétique. La compétence est développée dans certaines conditions physiologiques. En effet, chez *Streptococcus mutans,* la fréquence de transformation est de 10 à 600 fois plus importante en biofilm qu'en croissance planctonique. Le développement de la compétence est lié à la perception de la densité cellulaire de la même espèce, phénomène aussi appelé quorum-sensing (Li *et al.*, 2001). En général, la taille de l'ADN internalisé est restreinte à quelques kilobases (Dubnau et Cirigliano, 1972). La transformation naturelle a été caractérisée dans plusieurs phyla bactériens, les firmicutes (*Bacillus subtilis, S. pneumoniae*), les protéobactéries (*Haemophilus influenzae, Neisseria gonorrhoeae, Helicobacter pylori*), les cyanobactéries (*Synechocystis spp*) mais également chez les archées (*Methanococcus voltae, Methanobacterium thermoautotrophicum*) (Johnsborg *et al.*, 2007).

Conjugaison

La conjugaison nécessite un contact physique entre les cellules donatrice et réceptrice. Le matériel génétique est transféré grâce à des éléments conjugatifs (plasmides, éléments mobiles). Ces éléments assurent la mise en place de la machinerie de conjugaison et de transfert de leur propre matériel génétique ainsi qu'éventuellement l'information génétique du génome de l'hôte. L'élément conjugatif code l'information nécessaire à l'établissement du pilus chez les bactéries Gram négatives ou de protéines induisant le contact cellulaire chez les bactéries Gram positives. La mobilisation d'information chromosomique peut être réalisée en *cis*, c'est-à-dire en association physique avec l'élément conjugatif. Le cas le plus documenté est celui du facteur de fertilité F d'*E. coli* avec le transfert de marqueurs selon un gradient à partir d'un site chromosomique (bactérie Hfr) ou par l'intermédiaire d'un plasmide F' (facteur F portant un fragment du chromosome). En conséquence, la quantité d'ADN transférée peut être considérable ; théoriquement la totalité du chromosome bactérien de la cellule donatrice peut être mobilisée par un transfert à partir d'une bactérie Hfr. Chez les bactéries où les marqueurs chromosomiques sont mobilisés en *trans,* comme chez les streptomycètes, l'élément conjugatif code une protéine présentant une identité importante avec les ADN translocases de type SpoIIIE qui pourrait transférer des fragments d'ADN chromosomiques de façon aspécifique (Wu *et al.*, 1995 ; Gunton *et al.*, 2007).

Les phénomènes de HGT au travers de la conjugaison s'effectuent entre espèces proches mais aussi éloignées grâce aux plasmides dits « à large spectre » qui permettent d'établir des transferts inter-espèces et inter-règnes. Ce type d'échange a été démontré par exemple entre la bactérie *E. coli* et la levure *Saccharomyces cerevisiae* (Heinemann et Sprague, 1989 ; Bates *et al.*, 1997).

3.2 Maintien de l'information génétique

L'ADN exogène après son transfert doit être capable de se maintenir dans la cellule. Certains éléments tels que les plasmides codent leur propre réplication, ce qui leur permet de se maintenir dans la cellule hôte. Néanmoins, dans la plupart des cas, les séquences nouvellement internalisées ne peuvent pas se répliquer de manière autonome et ne seront maintenues que si elles intègrent le chromosome bactérien. Cette intégration est réalisée par trois principaux mécanismes de recombinaison :

i) La recombinaison homologue est initiée entre séquences homologues de longueur variable suivant les organismes (quelques dizaines de pb chez *E. coli* et *B. subtilis*) (Watt *et al.*, 1985 ; Roberts et Cohan, 1993). La recombinaison homologue va avoir pour conséquence : (i) le remplacement allélique au locus chromosomique homologue, (ii) la recombinaison intragénique et la réassociation de fragments alléliques, (iii) l'insertion de séquences nouvelles dans un génome, par exemple par l'intermédiaire de séquences répétées dispersées comme les IS (Insertion Sequence) ou transposons.

ii) La recombinaison illégitime est initiée entre séquences peu ou pas homologues. Lorsqu'il ne s'agit pas d'erreurs réplicatives (glissement de la fourche de réplication), le réassortiment de séquences par recombinaison illégitime résulte de l'intervention de protéines de type Ku et d'une ligase ATP dépendante (Aravind et Koonin, 2001 ; Weller et Doherty, 2001 ; Pitcher *et al.*, 2007). Ces deux acteurs forment le mécanisme Non-Homologous End Joining (NHEJ), caractérisé initialement chez les eucaryotes, mais également plus récemment chez les bactéries. Environ 20 % des espèces bactériennes dont le génome a été séquencé possède des homologues des gènes codant les acteurs du NHEJ (Aravind et Koonin, 2001 ; Doherty *et al.*, 2001 ; Weller et Doherty, 2001).

iii) La recombinaison site-spécifique est l'apanage des bactériophages, des éléments conjugatifs intégrés (ICE) et des éléments dérivants (îlots génomiques). Elle fait intervenir une recombinase codée par l'élément catalysant la recombinaison entre deux sites, l'un chromosomique et l'autre présent dans l'élément. Il en résulte l'intégration de l'élément au chromosome bactérien. Ce mécanisme est majeur dans le processus d'acquisition d'information génétique exogène (Hacker *et al.*, 1997).

Autres acteurs de la mobilité des gènes et des séquences

Des éléments génétiques mobiles (IS ou transposons) sont également impliqués dans des phénomènes de transfert, mais le plus souvent indirectement, en stimulant la plasticité génomique. Ces éléments réarrangent le génome, stimulent la recombinaison et favorisent la mobilité intragénomique (chromosome vers plasmide et réciproquement). Les éléments

mobiles portés par les fragments d'ADN acquis par conjugaison, transformation et transduction peuvent ensuite promouvoir le maintien de ces fragments par transposition ou recombinaison avec le génome de l'hôte.
Dans certains cas, un élément de type transposon porte des fonctions conjugatives. Ces éléments ont généralement un spectre d'hôte large. Par exemple, le transposon conjugatif Tn*916* identifié chez *Enterococcus faecalis* est capable de se transférer vers divers genres bactériens : *Enterococcus*, *Streptococcus*, *Staphylococcus* et *Bacillus* (Franke et Clewell, 1981).

3.3 Amélioration des séquences acquises par transfert horizontal

Les caractéristiques des séquences acquises par transfert horizontal dépendent notamment des modes de transfert et de l'éloignement phylogénétique de la donneuse vis-à-vis de la réceptrice (cf. section 3.1.). La composition nucléotidique d'une séquence d'origine exogène dépend également du caractère plus ou moins ancien de l'acquisition. En effet, une séquence exogène aura tendance, au travers d'un processus d'accumulation de substitutions nucléotidiques, à converger vers les paramètres compositionnels du génome de l'organisme hôte. Ce phénomène est connu sous le nom de processus d'amélioration. De façon générale, les séquences subissent les forces mutationnelles d'une part, et les forces sélectives d'autre part. Les substitutions nucléotidiques résultent de phénomènes mutationnels aléatoires faisant apparaître de nouveaux allèles. La fixation et la dissémination de cet allèle au sein du patrimoine génétique de l'espèce dépend de facteurs sélectifs ; le nouvel allèle contribue t-il à l'adaptation de l'organisme à son milieu ? Ces deux forces contribuent aux caractéristiques compositionnelles des phases codantes. En effet, de part les contraintes imposées par le code génétique, la 3ème position des codons échappe pour l'essentiel à la contre-sélection des substitutions, ces dernières provoquant le remplacement d'un codon par un codon synonyme quant au codage des acides aminés (Sueoka, 1988). Les mutations à cette position sont donc neutres. En revanche, les substitutions aux positions 1 et 2 changent le codage, et sont donc soumises à la sélection appliquée sur la fonction codée. Ces substitutions informatives vont contribuer à l'établissement d'un usage particulier des codons, chaque organisme utilisant de façon préférentielle un ou plusieurs codons pour l'incorporation d'un acide aminé donné dans une chaîne polypeptidique. Par ailleurs, il est constaté, que bien que le code génétique permette de coder le même polypeptide à partir de deux séquences nucléotidiques dissemblables, la composition en acides aminés des protéomes est distincte entre les groupes bactériens indiquant que certains acides aminés sont plus utilisés que d'autres selon les organismes.
Si la combinaison des phénomènes de mutation et de sélection explique l'évolution nucléotidique dans les phases codantes, il n'explique pas la grande diversité des pourcentages en bases G et C des génomes bactériens. Cette diversité semble être associée au contexte environnemental de l'organisme ; les organismes à cycle de vie autonome possèdent un génome plus riche en bases G et C (Rocha et Danchin, 2002). En revanche, les organismes à

vie intracellulaire ou pathogènes obligatoires possèdent des génomes à faibles teneurs en bases G et C. Une corrélation a également été notée entre la taille du génome et son contenu en bases G et C ; plus le génome est grand plus le contenu en bases G et C est important, atteignant plus de 70 % chez les bactéries à grands génomes (supérieurs à 8 Mb).
Des hypothèses sur l'origine de ce biais ont été émises, qui reposent d'une part sur l'absence de certains mécanismes de réparation de l'ADN, notamment les glycosylases spécifiques ou encore le système de correction post-réplicative des mésappariements, chez les bactéries à petits génomes, et d'autre part sur l'exploitation des ressources énergétiques de l'environnement de l'organisme. L'exploitation des ressources énergétiques de l'environnement aboutit à l'incorporation préférentielle dans les génomes des bases A et T, moins coûteuses et plus disponibles, chez les organismes subissant des carences nutritionnelles (Rocha et Danchin, 2002). Ces hypothèses fondées soit sur l'établissement d'un biais mutationnel (perte de mécanismes de réparation), soit sur un phénomène sélectionniste (pression pour l'utilisation des ressources), expliqueraient la divergence du contenu en bases G et C dans les génomes bactériens.
En outre, l'hypothèse sélectionniste apporte par là-même une explication à un phénomène constaté mais non élucidé, à savoir une teneur plus faible en bases G et C des séquences issues du transfert horizontal telles que des plasmides, bactériophages et autres éléments génétiques mobiles. Les séquences exogènes seraient en compétition avec le reste du génome ; leur intégration et leur maintien dans le génome seraient favorisés par un abaissement de leur coût, et donc d'une meilleure exploitation des ressources cellulaires.

3.4 Modèle d'étude : *Streptococcus thermophilus*

Le genre *Streptococcus* regroupe des bactéries Gram positives et à bas pourcentage en bases G et C qui appartiennent au phylum des firmicutes. L'espèce *S. thermophilus* possède un génome d'environ 1,8 Mb. Son chromosome est circulaire et le contenu en bases G et C est d'environ 39 % en moyenne. A ce jour, les génomes de trois souches ont été entièrement séquencés et sont accessibles : ceux des souches CNRZ1066[5], LMG18311[5] et LMD-9[5].

3.4.1 Le genre *Streptococcus*

L'espèce *S. thermophilus* est une des espèces de la division des Viridans. Au sein de cette division, toutes les espèces sont commensales excepté *S. thermophilus*. En effet, elles sont retrouvées, pour une majeure partie d'entre elles, dans les cavités buccales mais aussi les tractus intestinaux des mammifères. La division Viridans peut être divisée en cinq groupes majeurs : i) le groupe *mutans*, ii) le groupe *anginosus*, iii) le groupe *sanguinis*, iv) le groupe *mitis* et v) le groupe *salivarius* (Facklam, 2002). Les relations phylogénétiques du genre *Streptococcus* sont représentées par un arbre phylogénétique établi sur l'analyse du

[5] Les identifiants de ces génomes dans la base de données GENBANK (Benson *et al.*, 2008) sont : CP000024 (CNRZ1066), CP000023 (LMG18311) et CP000419 (LMD-9)

polymorphisme du gène *sodA* qui code la superoxyde dismutase manganèse dépendante (figure 16) (Poyart *et al.*, 1998).

Le génome de *S. thermophilus* contient 1915 séquences codantes (CDS pour Coding Sequence) pour la souche CNRZ1066 (selon l'annotation Bolotin *et al.*, 2004), 1889 CDS pour LMG18311 (Bolotin *et al.*, 2004) et 1710 CDS pour la souche LMD-9 (Makarova *et al.*, 2006). Un grand nombre d'entre elles (~ 80 %) sont présentes à la fois chez CNRZ1066 et LMG18311 et auraient des orthologues chez d'autres streptocoques. En effet, *S. thermophilus* est une espèce proche phylogénétiquement des deux autres espèces du groupe *salivarius* : *Streptococcus vestibularis* et *Streptococcus salivarius*, pathogènes opportunistes associés à des cas d'infections chez l'homme (endocardites et méningites) (Partridge, 2000 ; Idigoras *et al.*, 2001 ; Doyuk *et al.*, 2002 ; Conte *et al.*, 2005). Au contraire, *S. thermophilus* est la seule espèce du genre *Streptococcus* à être considérée comme une espèce propre à la consommation et reconnue comme organisme GRAS (Generally Recognized As Safe). Il a été proposé que l'émergence de *S. thermophilus* soit récente (3 000 à 30 000 ans) (Bolotin *et al.*, 2004) à partir de l'ancêtre commun qu'elle partage avec *S. salivarius*, ce qui est en concordance avec son utilisation massive dans le processus de fermentation des produits laitiers (estimée à environ 6 000 – 7 000 ans) (Fox, 1993). Au cours de ces fermentations, *S. thermophilus* est employée en co-culture avec *Lactococcus lactis* et *Lactobacillus delbrueckii* subsp *bulgaricus*. Ces espèces partagent donc le même écosystème.

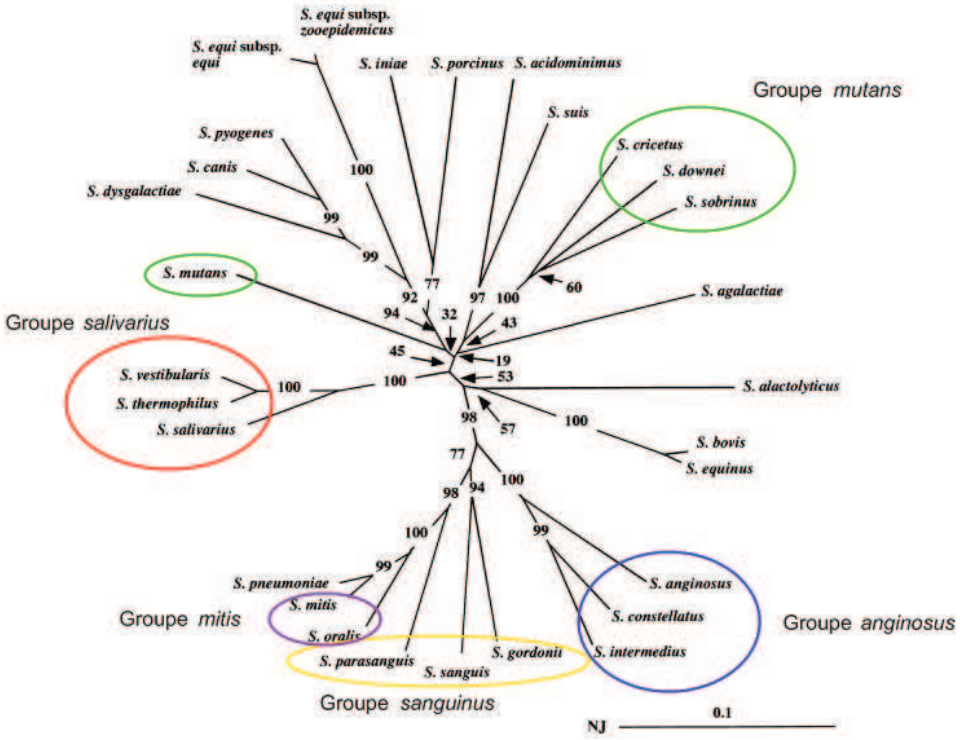


Figure 16 : Arbre phylogénétique basé sur l'analyse du polymorphisme du gène *sodA* chez les espèces du genre *Streptococcus* (Poyart *et al.*, 1998).

La construction de l'arbre est basée sur la méthode du plus proche voisin ou (Neighbor-Joining, NJ) dont le degré de confiance a été évalué par l'analyse de bootstrap. L'échelle correspond à la distance NJ qui représente 10 % de divergence nucléotidique. Les cinq groupes majeurs de la division des Viridans sont indiqués par des cercles de couleurs différentes.

3.4.2 Le « core » génome et le génome accessoire de *Streptococcus thermophilus*

Le « core » génome

Des analyses comparatives *in silico* de 26 génomes du genre *Streptococcus* permettent d'évaluer la taille du « core » génome, c'est-à-dire l'ensemble des gènes présents chez tous les streptocoques, à 600 gènes (Lefébure et Stanhope, 2007).

Une approche de la diversité génétique de l'espèce *S. thermophilus* a été réalisée par « MultiLocus Sequence Typing » (MLST) (Maiden *et al*., 1997). Cette analyse a porté sur huit gènes de ménage dans un échantillon de 27 souches de *S. thermophilus* (Delorme *et al*., 2007, 2008). Cette analyse a révélé un polymorphisme avec un pourcentage de divergence nucléotidique moyen de 0,19 %, qui est faible en comparaison avec ceux observés pour les espèces *S. salivarius* et *S. vestibularis*, phylogénétiquement proches de *S. thermophilus* (figure 16). En effet, ces espèces présentent un degré de divergence de 6,6 % parmi 22 souches de *S. salivarius* et de 3,6 % parmi 8 souches de *S. vestibularis*. L'homogénéité génétique de *S. thermophilus* révèle le caractère « clonal » de cette espèce. Ces données indiquent une émergence récente de l'espèce *S. thermophilus*.

Le génome accessoire

Le génome accessoire correspond à la partie variable du génome de la souche, de l'espèce ou du genre. Il regroupe des gènes participant à des phénomènes de contingence (réponse à l'environnement biotique ou abiotique, métabolisme secondaire, élément génétique mobile). Quant au « pan » génome, il est constitué de l'ensemble des gènes non redondants présents chez un organisme (souche, espèce ou genre), il correspond donc à chaque niveau phylogénétique, à la somme du « core » génome et du génome accessoire.

Chez les streptocoques, le « pan » génome s'élèverait à environ 6 000 gènes (Lefébure et Stanhope, 2007). Le « core » génome étant constitué de seulement 600 gènes, le génome accessoire constitue les 90 % restant du « pan » génome. Cela montre que les flux de gènes sont importants au sein de ce genre bactérien. Cette variabilité du génome accessoire a été décrite plus précisément pour l'espèce *S. thermophilus* par une analyse CGH (Comparative Genome Hybridization) portant sur 47 souches de *S. thermophilus*. Cette approche a permis d'estimer le « core » génome de l'espèce *S. thermophilus* à 1 271 gènes (Rasmussen *et al*., 2008), ce qui amène le génome variable à environ 700 gènes pour les trois souches entièrement séquencées. Cela révèle une grande quantité de gènes variables qui peuvent résulter de HGT récents ou de pertes d'information génétique. Des analyses *in silico* basées sur des biais compositionnels (GC%, composition en dinucléotides) ont proposé qu'un ensemble de 197 gènes ont été acquis par transfert horizontal dans les trois génomes de *S. thermophilus* entièrement séquencés et disponibles dans les bases de données (Liu *et al.*, 2009).

3.4.3 Flux de gènes chez *S. thermophilus*

Évolution par perte de gènes

L'évolution des génomes des bactéries de l'ordre des Lactobacillales (englobant les genres *Enterococcus*, *Streptococcus*, *Lactococcus*, *Leuconostoc* et *Lactobacillus*) se serait accompagnée d'une perte massive de gènes d'après Makarova et ses collègues (Makarova *et al.*, 2006) (figure 17). En effet, près de 25-30 % de gènes ont été perdus depuis l'émergence des Lactobacillales. *S. thermophilus* est l'organisme qui aurait subi la perte de gènes la plus importante : cette perte s'élève à 546 gènes depuis l'ancêtre commun de *S. thermophilus*, *L. lactis* ssp. *cremoris* et *L. lactis* ssp. *lactis*. Les auteurs mentionnent que malgré cette perte notable de gènes, la présence de nombreux pseudo-gènes suggère que le génome est encore dans un processus actif de réduction génomique. Chez les bactéries du groupe Lactobacillales, les pseudo-gènes peuvent représenter une proportion importante du génome comme par exemple, chez certaines bactéries présentes dans les produits laitiers telles que *S. thermophilus*, *Lactobacillus helveticus*, *Lactobacillus delbrueckii* subsp. *bulgaricus* qui possèdent respectivement 182, 217 et 533 pseudo-gènes. A l'inverse, les bactéries présentes dans l'intestin humain telles que *Lactobacillus salivarius* et *Lactobacillus gasseri* ont très peu de pseudo-gènes, respectivement 49 et 48 (Sullivan *et al.*, 2009).

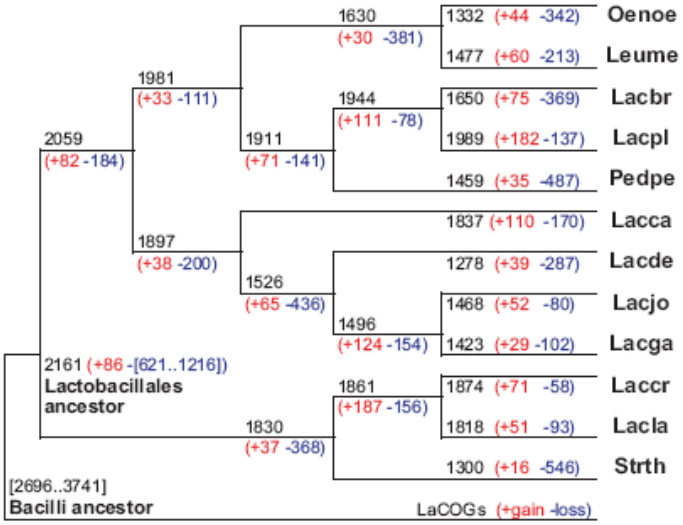


Figure 17 : Évolution des Lactobacillales par acquisition et perte de gènes (Makarova *et al.*, 2006).

L'arbre réalisé est enraciné par *Bacillus subtilis* (groupe externe). Les chiffres en rouge et en bleu indiquent respectivement un gain et une perte de gènes orthologues chez les Lactobacillales. Ces gènes sont définis selon la base de données COGs (Tatusov *et al.*, 2001). Oenoe, *Oenococcus oeni* ; Leume, *Leuconostoc mesenteroides* ; Lacbr, *Lactobacillus brevis* ; Lacpl, *Lactobacillus plantarum* ; Pedpe, *Pediococcus pentosaceus* ; Lacla, *Lactococcus lactis* ; Lacde, *Lactobacillus delbrueckii* ; Lacjo, *Lactobacillus johnsonii* ; Lacga, *Lactobacillus gasseri* ; Laccr, *Lactococcus lactis ssp. cremoris* ; Lacca, *Lactobacillus casei* ; Strth, *S. thermophilus*.

Les gènes de virulence (VRG pour Virulence Related Genes) présents chez les streptocoques pathogènes sont particulièrement affectés par le phénomène de perte de gènes chez *S. thermophilus* CNRZ1066 (Bolotin *et al.*, 2004). Au total, sur 50 VRG de

Streptococcus pyogenes SF370 et 118 VRG de *S. pneumoniae* TIGR4, seuls 8 (dont 1 pseudo-gène) et 58 (dont 7 pseudo-gènes) respectivement sont retrouvés chez *S. thermophilus* (tableau 7).

Tableau 7 : Identification de gènes homologues à des gènes de virulence de *S. pyogenes* SF370 (M1) et *S. pneumoniae* TIGR4 chez *S. thermophilus* CNZ1066 (Bolotin *et al.*, 2004)

	VRG	VRG chez *S. thermophilus*	VRG inactif chez *S. thermophilus*	% perte de VRG
***S. pyogenes* SF370 (M1)**				
VRG démontré	21	1	0	95,2
VRG putatif	29	7	1	75,9
Total	50	8	1	84,0
***S. pneumoniae* TIGR4**				
VRG démontré				
Adhérence	9	4	0	55,6
Mécanisme de défense	23	5	1	78,3
Métabolisme celullaire	35	20	2	42,9
Transport/protéases	29	19	3	34,5
Autre	22	10	1	45,5
Total	118	58	7	54,6

Acquisition de gènes par conjugaison ou transduction

S. thermophilus possède de nombreux éléments mobiles, notamment des séquences d'insertion (IS), qui peuvent favoriser l'acquisition de gènes en stimulant la plasticité du génome. Des événements de transfert d'IS ont été identifiés au sein de l'écosystème laitier (Guédon *et al.*, 1995). En effet, d'une part, certaines IS présentes dans le génome de *S. thermophilus* (IS*S1*, IS*981*, IS*911*) ont une grande similarité de séquences avec celles d'autres bactéries lactiques, et d'autre part, l'IS*1194* possède une forte identité avec des séquences de plasmide conjugatif de *S. pyogenes* (Bourgoin *et al.*, 1998). Ces IS ont probablement des origines diverses. Par exemple, IS*1191* est retrouvée chez des souches de *L. lactis* (Guédon *et al.*, 1995).

Récemment, des éléments mobiles plus complexes (ICE pour Integrative Conjugative Element et CIME pour *CIs* Mobilizable Element) ont été identifiés et participent à l'évolution et à la plasticité du génome. Ils sont notamment capables de mobiliser des gènes non impliqués dans le fonctionnement de l'élément (système de restriction modification, résistance aux métaux lourds). Les ICE sont des îlots génomiques qui sont capables de coder leur propre excision sous forme circulaire, leur transfert par conjugaison et leur intégration par recombinaison site-spécifique. Chez *S. thermophilus*, deux ICE, ICE*St1* et ICE*St3* ont été identifiés à l'état intégré dans l'extrémité 3' du gène *fba* codant une fructose-1,6-biphosphate aldolase et ces éléments semblent spécifiques de *S. thermophilus* (Burrus *et al.*, 2000 ; Pavlovic *et al.*, 2004). Les CIME, quant à eux, sont des éléments dérivants des ICE mais qui ont perdu leur module de conjugaison et/ou de recombinaison. Ces éléments, ICE et CIME,

ont une structure modulaire, dont les éléments constitutifs peuvent avoir des origines différentes (Pavlovic *et al.*, 2004). Des expériences réalisées au laboratoire ont montré que ICE*St1* et ICE*St3* se transfèrent de *S. thermophilus* CNRZ368 vers d'autres souches de *S. thermophilus* ou diverses espèces de bactéries lactiques alimentaires ou pathogènes comme *S. pyogenes* et *Enterococcus faecalis* (Bellanger *et al.*, 2009). D'autres faits mettent en évidence des événements de transfert entre *S. thermophilus* et des bactéries de l'ordre des Lactobacillales suggérant que certaines séquences d'origine exogène ont été acquises par conjugaison et transduction (Seizen *et al*. 2004). Une forte similarité d'identité de séquences a été notée entre une trentaine de gènes d'un prophage de *S. thermophilus* et des gènes d'un prophage de *L. lactis*. L'échange de gènes entre les deux génomes phagiques aurait eu lieu durant l'infection simultanée de *L. lactis* par ces deux phages. Le phage de *S. thermophilus* recombiné aurait été intégré dans le génome de *S. thermophilus* au cours d'une infection interspécifique. De façon similaire, le transfert d'un plasmide par transduction de *S. thermophilus* vers *L. lactis* a été récemment démontré (Ammann *et al*., 2008).

3.5 Méthodes de détection des gènes exogènes

L'identification des gènes exogènes peut se réaliser à travers deux approches distinctes ; phylogénétique et paramétrique. Les méthodes phylogénétiques ont l'avantage d'être robustes et les méthodes paramétriques sont rapides et plus faciles à mettre en œuvre.

3.5.1 Méthodes phylogénétiques

Le principe des méthodes phylogénétiques est de comparer l'évolution ou la distribution d'un gène candidat potentiellement acquis par transfert horizontal avec celle attendue d'une séquence de référence (arbre des espèces). Deux approches sont différenciées pour l'analyse des gènes exogènes à partir d'un arbre des espèces, l'une compare les arbres phylogénétiques pour déterminer des incohérences (arbres incongrus) entre l'arbre des espèces et l'arbre du gène et la seconde approche analyse une distribution du gène candidat potentiellement acquis par transfert horizontal directement dans l'arbre phylogénétique (distribution anormale). Ces deux approches nécessitent la construction d'un arbre phylogénétique robuste de référence qui constitue une étape clé. En effet, cette étape débute par le choix de la séquence de référence utilisée qui est important car cette séquence doit présenter un orthologue dans l'ensemble des espèces. Ensuite, un alignement multiple des séquences identifiées est obtenu en utilisant des algorithmes dédiés tels que ClustalW (Thompson *et al.*, 1994), Dialign (Morgenstern *et al.*, 1996), MultAlin (Corpet, 1988). Enfin, une reconstruction de l'arbre est réalisée suivant une méthode appropriée (calcul de distance, maximum de parcimonie et probabiliste) et elle est suivie de l'évaluation de sa robustesse (méthode bootstrap) (Efron, 1979 ; Felsenstein, 1985). Les approches de reconstruction se basent sur trois méthodes, i) les méthodes de calcul de distance (UPGMA et Neighbor-Joining) (Sneath & Sokal, 1973 ; Saitou et Nei, 1987), ii) les méthodes de maximum de parcimonie (Hennig, 1966 ; Felsenstein, 1978) et iii) les méthodes

probabilistes (Cavalli-Sforza et Edwards, 1967 ; Felsenstein, 1981). Les méthodes de calcul de distance se basent directement sur les distances calculées 2 à 2 puis regroupe les séquences. En revanche, les deux autres méthodes se basent directement sur les séquences où chaque position est considérée comme un caractère. Généralement, les méthodes basées sur la distance sont les plus utilisées grâce à leur simplicité et leur rapidité. Néanmoins, ces méthodes donnent lieu à une perte d'information portée par les séquences.
Des programmes phylogénétiques ont été développés permettant la construction des arbres phylogénétiques suivant une méthode spécifique ou parmi un choix multiple de méthodes et peuvent être semi automatisés ou automatisés tels que les programmes Pyphy (Sicheritz-Ponten et Andersson, 2001) ou PhyloGenie (Frickey et Lupas, 2004). Les programmes phylogénétiques sont listés sur le site web[6].

Arbres phylogénétiques incongrus

L'approche analysant les incongruences entre les arbres phylogénétiques compare l'arbre phylogénétique reconstruit à partir des orthologues d'un gène d'intérêt supposé acquis par transfert horizontal avec l'arbre des espèces reflétant l'évolution vraie des espèces. Une discordance entre l'arbre phylogénétique du gène candidat potentiellement acquis par transfert horizontal et l'arbre de référence suggère un HGT (Page et Charleston, 1997). Dans la figure 18, l'arbre phylogénétique du gène candidat est incongru car un orthologue de ce gène présent dans l'espèce D est plus proche de celui des espèces G et H que de ceux des espèces A, B et C normalement proposé par l'arbre des espèces. Cette discordance suggère un événement de transfert horizontal du gène candidat entre soit l'ancêtre commun des espèces G et H soit l'une des espèces G ou H vers l'espèce D.

Distribution phylogénétique anormale

L'approche analysant la distribution phylogénétique repose sur la recherche de séquences homologues au sein des différentes espèces (BLAST) (Altschul *et al.*, 1990). Elle consiste à déterminer la présence ou l'absence du gène candidat potentiellement acquis par HGT dans un groupe d'espèces ou dans toutes les espèces (Syvanen, 1994 ; Page et Charleston, 1997). La distribution des homologues du gène candidat dans l'arbre phylogénétique est comparée à celle de l'arbre des espèces. Un homologue du gène candidat trouvé dans deux espèces phylogénétiquement éloignées et absent des espèces les plus proches, révèle une distribution anormale qui présume que ce gène candidat a été soumis au HGT (figure 18.C).
Le programme DarkHorse (Podell et Gaasterland, 2007 ; Podell *et al.*, 2008) est une méthode d'analyse de la distribution phylogénétique basée sur la recherche d'homologue et le calcul d'un score LPI pour Lineage Probability Index. Il en résulte, une base de données de gènes candidats du transfert horizontal détectés au sein de 955 génomes bactériens. Le score LPI correspond à une distance phylogénétique du gène candidat avec les plus proches orthologues présents dans la base de données. Ces valeurs de LPI sont basées sur une analyse linguistique des informations taxonomiques des organismes proposées par la base de données taxonomiques du NCBI. Par exemple, la taxonomie de l'espèce *Escherichia* propose

[6] Liste des programmes sur http://evolution.genetics.washington.edu/phylip/software.html

l'appartenance à l'ensemble des taxons suivant Bacteria; Proteobacteria; Gamma-proteobacteria ; Enterobacteriales; Enterobacteriaceae. Ces informations taxonomiques sont ensuite pondérées. En effet, le programme attribut un poids faible pour les termes les plus représentés. Les termes « Bacteria », « Eukaryota » et « Viruses » ont les poids les plus faibles car ils désignent les niveaux taxonomiques les plus élevés. Les scores LPI bas suggèrent que le gène est un bon candidat pour être considéré comme potentiellement acquis par transfert horizontal.

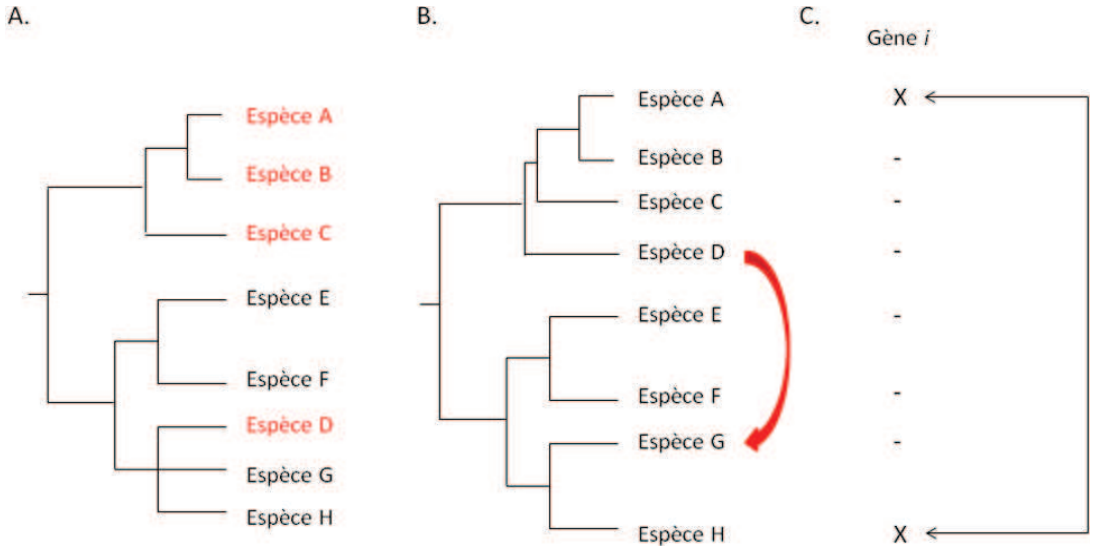


Figure 18 : Distribution phylogénétique de gènes acquis par HGT.
A. Un arbre phylogénétique construit à partir des orthologues du gène d'intérêt trouvés dans les différentes espèces. B. Un arbre d'espèce basé sur des relations de parenté entre une série d'espèces. La flèche rouge indique un phénomène de transfert produisant un arbre incongru basé sur le gène candidat. C. La présence (+) ou l'absence (-) d'un gène candidat (gène *i*) est indiquée dans chaque espèce. La présence du gène *i* dans seulement les espèces A et H phylogénétiquement éloignées supporte un phénomène de HGT entre ces deux espèces.

D'autres méthodes phylogénétiques appelées « réseau phylogénétique » ou « Network », utilisent une représentation des relations liées à l'évolution des espèces au travers d'un graphe (Bandelt *et al.*, 1999 ; Posada et Crandall, 2001 ; Gusfield *et al.*, 2007). Les réseaux phylogénétiques autres que les arbres phylogénétiques permettent de considérer les événements de recombinaison ou transfert. La figure 19.A décrit deux topologies d'arbres phylogénétiques possibles à partir des séquences analysées. La discrimination d'une topologie par rapport à l'autre n'est pas possible. En revanche, le réseau phylogénétique réticulé propose une topologie considérant les événements (recombinaison, HGT) par l'apparition à des nœuds de deux transitions entrantes (nœuds hybrides) (figure 19.B). L'arbre « parfait » est celui qui ne contient pas de recombinaison, c'est-à-dire où toutes les transitions arrivent directement dans les nœuds feuilles (extérieurs). L'implémentation et l'optimisation de méthodes telles que l'algorithme de décomposition sont nécessaires pour aboutir à cet arbre « parfait » (Huson et Bryant, 2005). Leur application est intéressante car elle permet une visualisation rapide des incongruités.

Les analyses phylogénétiques sont efficaces mais limitées. Le principal problème dans la reconstruction des arbres phylogénétiques est de différencier les phénomènes de duplication

suivis de pertes de l'une des deux copies du gène qui peuvent entrainer les mêmes incohérences de reconstruction qu'un arbre basé sur un gène acquis par transfert horizontal (Page et Charleston, 1997 ; Doolittle, 1999). D'autre part, la reconstruction des arbres phylogénétiques est difficile car elle doit s'effectuer sur des données représentatives (d'espèces et/ou de souches). La robustesse de l'arbre dépend également des données de départ et nécessite un jeu de données important d'orthologues (supposés). Dans le cas de l'analyse d'une distribution phylogénétique anormale, les paramètres de présence/d'absence d'un gène ou la recherche d'homologue dans une espèce est aussi arbitraire ; les seuils d'identité nucléotidique sont à affirmer.

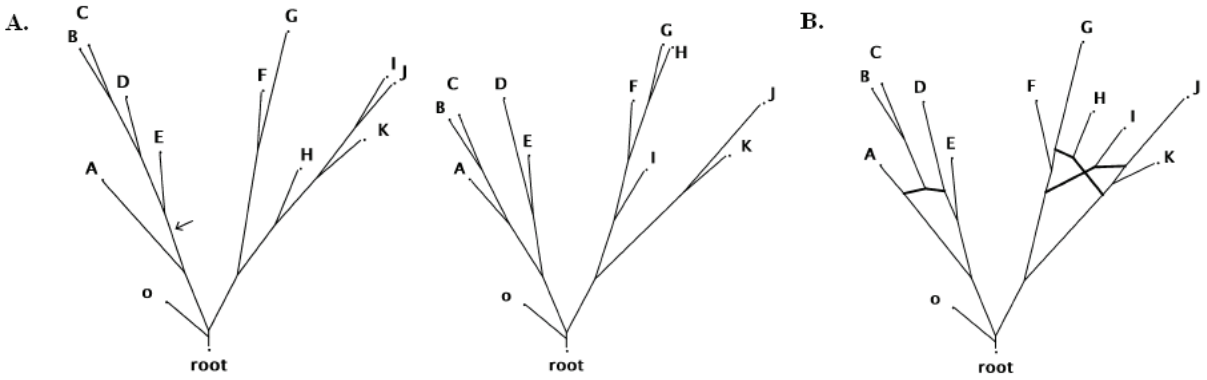


Figure 19 : Réseau phylogénétique (d'après Huson et Bryant, 2005).
A. Deux topologies d'arbres phylogénétiques possibles obtenues pour les séquences (a,b,c,d,e,f,g,h,i,j,k).
B. Représentation d'un réseau phylogénétique réticulé. Un nœud avec aucune transition entrante et deux sortantes représente le nœud racine. Un nœud avec une transition entrante et deux sortantes représente le nœud de spéciation. Un nœud avec deux transitions entrantes et une sortante représente un nœud hybride et un nœud avec une transition entrante et zéro sortante représente un nœud feuille qui correspond aux données observées (a,b,c,d,e,f,g,h,i,j,k).

3.5.2 Méthodes paramétriques

Les méthodes paramétriques tentent de détecter les biais compositionnels présents dans la séquence analysée. L'un de ces biais est le pourcentage en bases G et C. Le pourcentage en bases G et C varie de façon importante, entre 25 % à 75 %, entre espèces bactériennes. En revanche, cette composition est relativement homogène dans un génome (Muto et Osawa, 1987 ; Lawrence et Ochman, 1997). Le pourcentage en bases G et C se traduit dans les phases codantes par un biais de composition à la 3ème position des codons, et ceci d'autant plus fortement que le génome présente une composition fortement biaisée (soit vers des hauts ou des bas pourcentages en bases G et C) (Muto et Osawa, 1987 ; Sueoka, 1988). L'hypothèse sous-jacente des méthodes paramétriques est que ces biais constituent des signatures génomiques. Par conséquent, les séquences issues d'un transfert horizontal récent auront des caractéristiques compositionnelles différentes des gènes « résidents » de l'organisme hôte. Ainsi, l'identification de régions hétérogènes est possible au travers de la composition en nucléotides tels que le contenu en nucléotides G et C (Lawrence et Ochman, 1997, 1998 ; Ochman *et al.*, 2000), le biais d'utilisation des codons (Ikemura, 1985 ; Sharp, 1987 ; Karlin

et Burge, 1997 ; Karlin *et al.,* 1998), la fréquence des dinucléotides et des oligonucléotides (Karlin et Burge, 1997 ; Karlin *et al.,* 1997 ; Pride et Blaser, 2002 ; Fertil *et al.,* 2005 ; Dufraigne *et al.,* 2005). L'analyse de l'usage des codons a ainsi permis d'estimer que 16 % du génome *E. coli* aurait une origine exogène (Médigue *et al.*, 1991). Une autre analyse compositionnelle combinant des critères (GC% contenu dans les différentes positions avec le biais d'usage des codons) suggère que 17 % des gènes chez *E. coli* seraient issu d'un HGT (Lawrence et Ochman, 1997). Ces deux analyses attestent qu'une proportion de gènes exogènes dans le génome de *E. coli* peut être identifiée par des méthodes paramétriques. Il est alors possible de réaliser des analyses à grande échelle comme la base de données de gènes potentiellement acquis par HGT, HGT-DB (Garcia-Vallve *et al.,* 2003), qui a été réalisée au sein de 94 génomes de procaryotes. Les gènes potentiellement transférés ont été identifiés par l'analyse de biais compositionnels en bases G et C, l'usage des codons, l'usage des aminoacides et l'usage des codons synonymes. Quant au programme IslandPath (Hsiao *et al.*, 2003), il permet une visualisation des génomes sous forme de carte où chaque gène est représenté par un point et un code couleur qui permet de localiser les gènes de mobilité (IS). Cette représentation permet la détection et la localisation rapide d'îlots génomiques potentiels. Les HMM ont été aussi utilisés pour la prédiction de HGT (Nicolas *et al.*, 2002 ; Vernikos et Parkhill, 2005) et pour l'identification d'organismes donneurs potentiels (Waak *et al.*, 2006). Le programme, SIGI-HMM (Waack *et al.*, 2006) prédit des îlots génomiques et les donneurs potentiels sur la base de l'analyse de l'usage des codons (SIGI) (Merkl, 2004). En effet, l'usage des codons de chaque gène est comparé avec une table d'usage de codon établie sur un ensemble de gènes représentatifs des donneurs potentiels. Les auteurs ont généré une table d'usage des codons pour un ensemble de génomes, les gènes présentant une expression élevée ont été exclus de l'analyse. L'architecture du HMM utilisé est composée d'un état caché noté « natif » correspondant à l'usage des codons des gènes « résidents » du génome. Les états putatifs modélisent l'usage des codons des gènes « résidents » des autres génomes. Toutes les transitions sont permises afin de modéliser les structures mosaïques des HGT. La probabilité qu'un gène soit émis dans l'état « natif » est égale au produit des fréquences des codons synonymes dans l'état « natif » composant le gène. Dans le cas où le gène est considéré comme exogène, la probabilité d'émission du gène devrait être plus élevée pour l'un des autres états putatifs. L'algorithme Viterbi est utilisé pour déterminer le chemin le plus probable. Tous les gènes ayant une probabilité d'émission plus importante dans des états putatifs que dans l'état « natif » sont considérés comme des îlots génomiques et l'identité des états putatifs détermine le donneur potentiel pour chaque gène exogène. D'après les auteurs, l'analyse sur des génomes bactériens révèle un nombre moins important de HGT prédits que les autres méthodes utilisant les biais compositionnels et elle ne permet pas de différencier des gènes potentiels issus de HGT qui ont un usage des codons similaire aux gènes « résidents » du génome.

Vernikos et Parkhill (2005) ont développé le programme IVOMs qui identifie des biais compositionnels en utilisant la distribution des motifs d'ordre variable combinée à un HMM à deux états. IVOMs sélectionne des régions dont la fréquence de motifs de taille variable (1 à 8*mers*) est significativement différente de celle attendue dans tout le génome. Ensuite, les

auteurs ont employé un HMM à deux états cachés pour localiser de manière plus précise la position de transition entre séquence atypique et celle qui ne l'est pas. Les paramètres de ce modèle sont estimés sur l'ensemble des régions prédites comme atypiques et prolongées afin de contenir aussi des régions « natives » du génome. Ensuite, l'algorithme Viterbi est utilisé pour déterminer la suite d'états cachés la plus probable pour chaque région atypique allongée où la transition entre les deux états cachés représente la localisation de la transition entre ces deux types de régions. Cette méthode a été employée sur *Salmonella enterica* serovar *typhi* et a montré une meilleure sensibilité et spécificité pour l'identification de gènes acquis par HGT que la méthode de Tsirigos et Rigoustsos (2005a ; 2005b) ou celle de Garcia-Vallve et ses collègues (2003). Tsirigos et Rigoustsos (2005a ; 2005b) analysent la distribution de la fréquence des motifs de longueur fixe (8*mers*) observée par rapport à celle attendue dans le génome alors que Garcia-Vallve et ses collègues (2003) basent leur analyse sur les déviations de biais compositionnel (contenu en bases G et C, de l'usage des codons, des aminoacides, l'usage des codons synonymes).
Les méthodes telles que IVOMs (Vernikos et Parkhill, 2005) et SIGI-HMM (Waack *et al.*, 2006) utilisent des connaissances biologiques (usage des codons) ou un ensemble de séquences exogènes prédits pour estimer les paramètres d'un HMM. En revanche, Nicolas *et al.*, 2002 utilisent un HMM d'ordre 1 pour segmenter le génome de *B. subtilis* sans connaissance *a priori* du contenu génétique. Le modèle M1M2 indiquant une dépendance de la chaîne d'observations de 2 sur les nucléotides utilisés, se révèle être le plus approprié pour identifier des gènes exogènes. Ce modèle comporte 3 états cachés dont l'un d'entre eux identifie des gènes ayant un faible pourcentage en bases G et C et un usage des codons similaire à ceux des gènes acquis par HGT. En plus de ces biais compositionnels atypiques, ces gènes codent des fonctions connues pour être fréquemment transférées (ABC transporteur, enzymes impliquées dans la synthèse de la paroi cellulaire, résistance à des antibiotiques). Toutes les séquences de prophages intégrées caractérisées dans le chromosome de *B. subtilis* ont aussi été identifiées dans cet état caché.

En général, les méthodes phylogénétiques sont robustes mais elles nécessitent un jeu de données important alors que les méthodes paramétriques sont rapides. Ces dernières approches sont efficaces sauf dans le cas où les gènes issus d'un HGT ne présentent pas ou peu de biais. C'est le cas notamment, lorsque les séquences sont transférées entre des organismes proches (Koski *et al.*, 2001) ou lorsque le pourcentage en bases G et C de la séquence exogène est proche de celui du génome de la bactérie réceptrice. Ainsi, l'identification d'anciennes acquisitions de séquences par les méthodes d'analyse des biais compositionnels est rendue difficile à cause des phénomènes d'homogénéisation (cf. section 3.3).

Matériel et Méthodes

Notre approche de fouille de données est composée d'une étape d'extraction des régions « atypiques » à l'aide de différents types de HMM d'ordre 2, afin d'identifier des séquences longues (un à plusieurs kilobases) ayant des propriétés compositionnelles atypiques. Ces régions « atypiques » comportent des gènes « atypiques » dont la distribution au sein des firmicutes est établie par l'exploration d'une base de données développée pour cette étude. Cette distribution repose sur la recherche d'un homologue (présence, absence) dans les différents génomes des espèces composant cette base de données. Enfin, un nouvel indice numérique, correspondant à une valeur de dispersion, a été défini et caractérise comment une séquence « atypique » est distribuée au sein d'un arbre phylogénétique des firmicutes.
L'ensemble des programmes a été développé dans le langage Bioperl et leurs rôles sont décrits dans l'annexe E.

3.6 Extraction des positions et des régions « atypiques »

La figure 20 résume le processus d'extraction des positions et des régions « atypiques » développé. Ce processus débute par la génération, soit d'un modèle M2M2 considérant les nucléotides, soit d'un modèle M2M0 considérant les 3*mers*. Ensuite, une estimation de ces modèles est réalisée avec la séquence génomique de *S. thermophilus* CNRZ1066, sans *a priori* sur le contenu génétique, selon le processus décrit pour le logiciel *CarottAge* (cf. section 1.4.2) (figure 20.A, annexe E). Dans une seconde étape, des régions « atypiques » sont extraites sur la base d'une forte densité de positions dites déviantes (notées *pos_dev*). Deux procédures d'extraction des positions déviantes peuvent être mises en œuvre alternativement et ont été développées pour traiter les probabilités *a posteriori* générées par les modèles M2M0 et M2M2 (figure 20.B, annexe E).
Outre la différence de dépendances de la chaîne d'observations, les analyse basées sur les modèles M2M0 et M2M2 diffèrent par la caractéristique suivante : dans le cas de l'analyse utilisant un M2M0, les probabilités *a posteriori* correspondent aux probabilités *a posteriori* de rester dans un état caché donné, P_{ii}, alors que dans l'analyse employant un M2M2, les probabilités *a posteriori* correspondent aux probabilités *a posteriori* d'être dans un état caché donné, P_i (cf. section 1.3.3). Ces différences entre les deux modèles ont pour conséquence que les pics de probabilités *a posteriori* générés par le modèle M2M0 sont plus courts et nombreux alors que le modèle M2M2 produit des pics de probabilités plus lissés, comme le montre la figure 8. En conséquence de cette différence de profil, l'extraction des séquences « atypiques » suit une procédure différente de traitement des probabilités *a posteriori* selon le modèle utilisé.

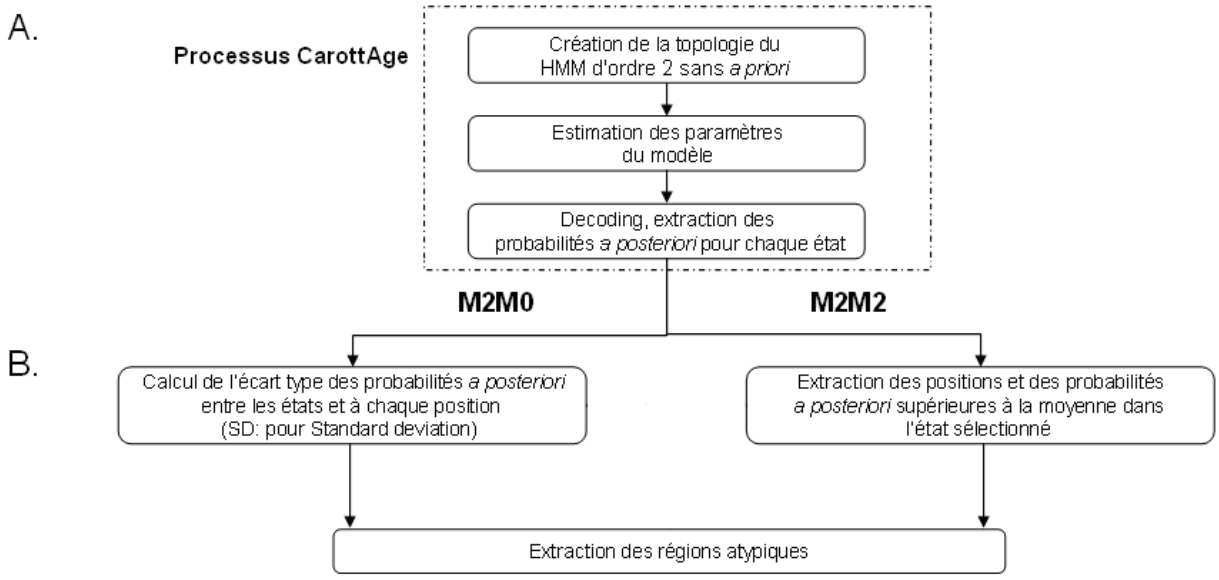


Figure 20 : Processus de fouille de données pour la recherche de régions « atypiques ».
A. Extraction des probabilités *a posteriori*.
B. Traitement des probabilités *a posteriori* et extraction des régions « atypiques ».

3.6.1 Traitement des probabilités *a posteriori*

Dans le cas d'un *decoding* sur un modèle M2M0, l'écart type (SD, standard deviation) entre les probabilités des différents états est calculé pour chaque index *j* (position d'un k*mer*), suivant la formule :

$$SD_j = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \tilde{x})^2}{(n-1)}}$$

Avec *n* le nombre d'états cachés, x_i la valeur de probabilité *a posteriori* de l'index *j* à chaque état caché *i* et $\tilde{x}$ la valeur moyenne des probabilités *a posteriori* pour chaque index *j* dans les *n* différents états cachés.

Ainsi, si une forte valeur de SD est observée, l'index correspondant est dit déviant. En effet, une forte valeur de SD indique qu'un état est fortement plus probable que les autres à cette position, révélant une atypicité compositionnelle locale par rapport au reste du génome. Une valeur de SD est considérée comme forte si elle dépasse un seuil fixé expérimentalement.

Dans le cas d'un *decoding* sur un modèle M2M2, l'extraction des séquences « atypiques » ne nécessite pas le calcul de SD et elle s'effectue directement sur la base de la forte probabilité d'être dans l'état caché le plus divergent des autres (appelé état D). Ainsi, les index qui ont une valeur de probabilité *a posteriori* supérieure à celle de la moyenne générale de l'état D sont extraits. L'état D est défini comme l'état caché le plus distant des autres. Cette distance est calculée en appliquant le calcul de la divergence de Kullback-Leibler (cf. section 1.3.3).

Quelle que soit la procédure d'extraction des positions déviantes (notées *pos_dev*, position ayant une valeur de SD élevée ou de probabilité *a posteriori* élevée dans l'état D), celle-ci est suivie d'une étape de sélection de régions qui présentent une forte densité de positions

déviantes (notées régions « atypiques »). Cette sélection repose notamment sur des contraintes de densité de positions déviantes et de longueur minimale des régions, de façon à sélectionner des atypicités compositionnelles compatibles avec l'échelle d'un gène ou d'un îlot génomique.

3.6.2 Détection et extraction des régions « atypiques »

L'identification automatiquement des régions contenant une forte densité de *pos_dev* est réalisée à l'aide d'une fenêtre coulissante. Chaque fenêtre doit vérifier au moins deux principaux paramètres pour être considérée comme des fenêtres candidates, le nombre de *pos_dev* dans la fenêtre courante doit être supérieur à un seuil (MinN) et le taux de *pos_dev* entre la fenêtre courante et la précédente doit être supérieur à un seuil (MinT). Ce dernier paramètre permet d'extraire des régions plus ou moins longues et denses en *pos_dev*. L'extraction des régions « atypiques » dépend des paramètres suivants :

- W : la taille de la fenêtre coulissante,
- S : le pas par lequel se déplace la fenêtre le long de la séquence,
- N : le nombre de *pos_dev* dans une fenêtre coulissante,
- MinN : le seuil minimum que N doit excéder pour que la fenêtre soit considérée,
- T : le taux de *pos_dev* dans une fenêtre par rapport à la fenêtre précédente,
- MinT : le seuil minimum que T doit excéder pour que la fenêtre soit considérée,
- L : la longueur des régions candidates (longueur cumulée des fenêtres considérées),
- MinL : le seuil minimum que L doit excéder pour que la région soit extraite.

L'algorithme d'extraction suit le processus décrit ci-dessous.

Algorithme : extraction des régions « atypiques »

1 : Calcul du nombre de *pos_dev* (N_1) dans la première fenêtre coulissante (W_1).
Si $N_1 \geq$ MinN, alors la fenêtre W_1 est candidate et l'on passe à l'étape 2.
Si $N_1 <$ MinN, alors la fenêtre W_1 n'est pas candidate. On déplace la fenêtre d'un pas S et l'on réitère l'étape 1 (jusqu'à ce que la fin de la séquence soit atteinte).

2 : Déplacement de la fenêtre d'un pas S et calcul du nombre de *pos_dev* (N_2) dans cette fenêtre courante W_2. Calcul du taux de *pos_dev* dans la fenêtre W_2 par rapport W_1 (T_1).
Si $T_1 \geq$ MinT et $N_2 \geq$ MinN, alors la fenêtre W_2 est candidate et l'étape 2 est réitérée jusqu'à ce que la fenêtre W_i ne soit pas candidate. Dans un tel cas, la fenêtre est déplacée d'un pas S et l'on reprend le processus à l'étape 1.

3 : Quand la totalité de la séquence a été parcourue, la longueur (L) des séquences couvertes par les fenêtres candidates chevauchantes est calculée.
Si L > MinL, alors cette séquence est considérée comme « atypique ». La région « atypique » extraite débute par la première *pos_dev* de la séquence « atypique » et elle se termine par la dernière *pos_dev* de la séquence « atypique ».

Au travers de l'exemple de la figure 21, l'extraction des régions appliquant les paramètres W = 400 nt, S = 300 nt, MinN = 5 positions, MinT = 0,2 et MinL = 1000 nt se déroule comme suit : sur la première fenêtre (W_1) de 400 nt, le nombre de *pos_dev* (N_1) est calculé. Ici, N_1 = 5, donc le seuil de MinN est atteint et la fenêtre coulisse d'un pas S de 300 nt. W_2 chevauche donc W_1 sur 100 nt. Dans cette nouvelle fenêtre, N_2 = 2 donc T_1 = 2/5 ce qui est supérieur au

seuil MinT de 0,2 ; en revanche, $N_2 = 2$, ce qui est inférieur à MinN. La fenêtre W_2 n'est donc pas candidate. Ainsi, la région candidate se réduit à W_1 qui a une taille de 400 nt. Elle est donc inférieure à la taille MinL de 1000 nt et ne sera donc pas extraite. La fenêtre coulisse de 300 nt. W_3 ne contient qu'une *pos_dev* et n'est donc pas non plus candidate. W_4 présente une valeur $N_4 = 5$ et atteint donc le seuil MinN. C'est également le cas de W_5 et W_6. De plus $T_4 = 1$, tout comme T_5, ce qui est supérieur au seuil MinT de 0,2. Ainsi, la région couverte par les fenêtres T_4, T_5 et T_6 répond aux seuils de densité (MinN) et d'homogénéité (MinT) requis. La taille (L) de cette région candidate est de 1000 nt, elle répond donc également au seuil MinL. La région atypique extraite de cette région candidate est plus petite puisqu'elle ne débute qu'à la première position *pos_dev* de la fenêtre W_4 pour s'achever qu'à la dernière position *pos_dev* de la fenêtre W_6.

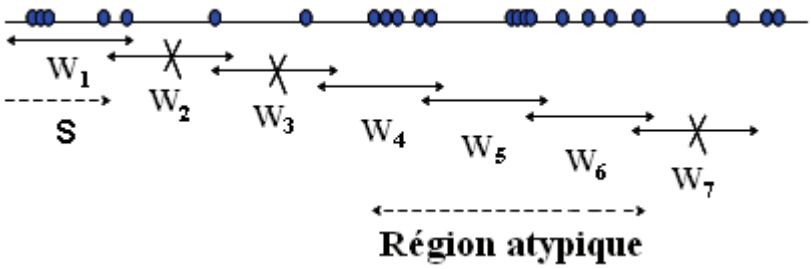


Figure 21 : Schématisation du processus d'extraction des régions atypiques.
Les cercles bleus symbolisent les *pos_dev*. Les paramètres suivants ont été appliqués : W = 400 nt, S = 300 nt, MinN = 5 positions, MinT = 0,2 et MinL = 1000 nt. Les croix correspondent aux fenêtres coulissantes qui seraient exclues car elles ne satisfont pas le paramètre MinN.

Une fois les régions « atypiques » extraites, il est possible de formater les résultats afin de les visualiser avec le logiciel *ARTEMIS* (Figure 22).

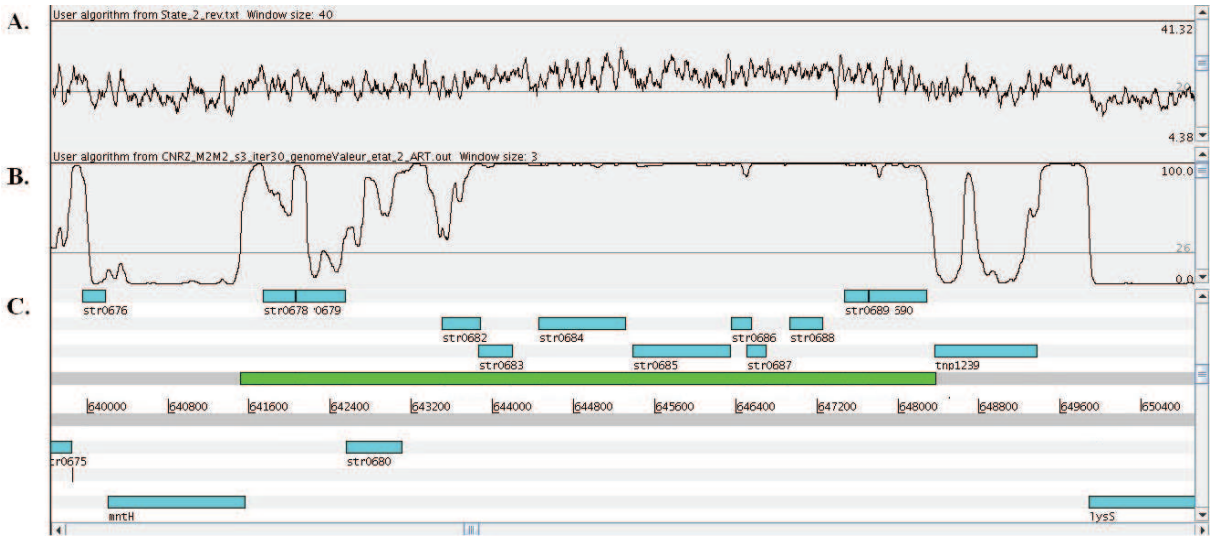


Figure 22 : Visualisation par *ARTEMIS* d'une région atypique.
Le logiciel *ARTEMIS* permet la visualisation de séquences génomiques et leur annotation. L'évolution des valeurs de SD (analyse du modèle M2M0) (A), les probabilités *a posteriori* de l'état D (analyse du modèle M2M2) (B) sont représentées le long de la séquence dont l'annotation est décrite en C. Les CDS et leur annotation sont symbolisées par des rectangles bleus et les régions « atypiques » ont été symbolisées par des rectangles verts.

Les analyses ultérieures peuvent porter soit sur les régions « atypiques », soit sur les gènes « atypiques ». Un gène est considéré comme « atypique » si au moins la moitié de sa longueur est incluse dans la région « atypique ».

3.7 Création d'une base de données

L'analyse de la distance des séquences « atypiques » au sein du phylum des firmicutes a été entreprise. Cette analyse de distribution nécessite l'implémentation d'une base de données contenant les séquences génomiques des organismes appartenant à ce phylum ainsi que la mise en place d'un programme de recherche de séquences homologues aux séquences « atypiques » dans cette base de données. Une autre analyse mise en place au cours de ce travail a porté sur le calcul d'une valeur de dispersion de chaque séquence « atypique » au sein des firmicutes. Cette valeur de dispersion prend en compte la divergence des différentes espèces qui composent la base de données par rapport à *S. thermophilus*.
Le traitement automatisé des séquences « atypiques » extraites (région ou gène « atypique ») se décompose en deux principales étapes décrites ci-dessous (figure 23).

Étape 1 : Importation et initialisation des bases de données
Au jour de la mise en place de la base de données des firmicutes, 126 génomes de souches bactériennes appartenant à 70 espèces différentes de firmicutes étaient accessibles. La liste des 126 organismes est donnée en annexe F. Ces génomes et leur annotation ont été importés et les séquences protéiques des CDS (Coding Sequence) ont alimenté une base de données formatée pour la recherche ultérieure d'homologie (notée *Firmicutes_prot*).
Parallèlement, une base de données répertoriant les caractéristiques de ces 126 organismes firmicutes a été réalisée. Les tables *Organisme* et *Souche* sont générées et alimentées et la structure des tables *Regions* et *Homologues* est générée et elles seront alimentées lors de l'étape 2 (figure 23.A).

Étape 2 : Traitement des régions « atypiques »
Dans cette étape, les tables *Regions* et *Homologues* sont alimentées. La table *Regions* est employée pour stocker les caractéristiques des séquences « atypiques » (identifiant de la région, longueur, GC%). Des recherches de gènes homologues aux séquences « atypiques » dans la base de données des firmicutes sont réalisées à l'aide du programme *BLASTX* (pour Basic Local Alignment Search Tool) (Altschul *et al.,* 1990). Le programme *BLASTX* recherche des identités ou similarités de séquence entre une séquence nucléotidique, qu'il traduit dans les six phases de lecture, et les séquences d'une base de données protéiques. S'il existe des séquences homologues aux séquences « atypiques » dans la base de données *Firmicutes_prot*, alors la table *Homologues* est utilisée pour stocker les informations pertinentes (fonction codée, organisme ayant une séquence homologue, longueur de l'homologie, pourcentage d'identité, e-value). Dans cette analyse, une séquence a été considérée comme homologue à une séquence « atypique » si l'alignement entre ces deux séquences révèle un pourcentage d'identité supérieur à 30 % et une e-value inférieure à 10^{-6}.

La e-value est l'espérence du nombre de sous-séquence dans la base de données avec un score plus grand que celui défini.

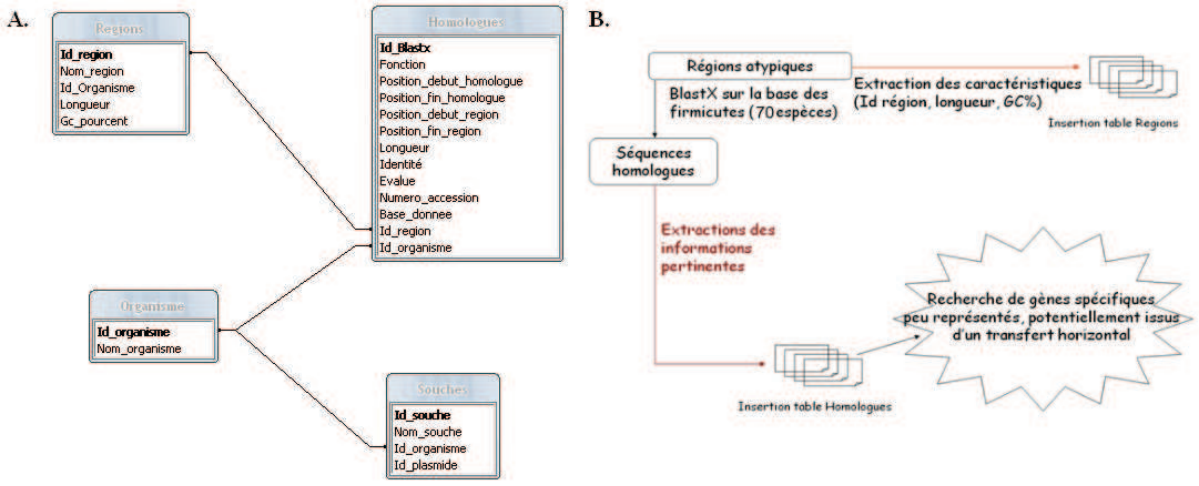


Figure 23 : Schéma relationnel de la base de données des firmicutes et du traitement des régions « atypiques ».

A. Schéma de la structure relationnelle de la base de données des firmicutes composée de quatre relations (*Regions*, *Homologues*, *Organismes* et *Souches*).
B. Schéma du traitement des séquences « atypiques ».

Une espèce est considérée comme ayant une séquence homologue à une séquence « atypique » si au moins l'une des souches de la base de données appartenant à cette espèce porte une séquence homologue. L'interrogation de la base de données des firmicutes permet d'extraire des séquences spécifiques ou peu représentées dans cette base de données (figure 23.B). Il est aussi possible d'extraire, pour une séquence « atypique » donnée, la distribution des homologues identifiés dans les différentes souches et/ou espèces de la base de données des firmicutes.

3.8 Définition d'un nouveau paramètre : la valeur de dispersion

L'objectif du calcul d'une valeur de dispersion est de déterminer comment une séquence « atypique » est distribuée au sein des génomes des firmicutes. Cette valeur indique comment les bactéries possédant ce gène sont localisées (proches ou éloignées) phylogénétiquement par rapport à *S. thermophilus* CNRZ1066.
Cependant, à notre connaissance, aucun arbre phylogénétique représentant toutes les espèces de la base de données des firmicutes n'existait. La réalisation d'un arbre phylogénétique des firmicutes a donc été réalisée et a été basée sur des données moléculaires des génomes de la base de données.

3.8.1 Réalisation d'un arbre phylogénétique des firmicutes

La qualité d'un arbre phylogénétique des espèces réside dans le choix de la séquence utilisée. Dans notre cas, la reconstruction de l'arbre phylogénétique s'est portée naturellement sur les gènes codant l'ARN ribosomique 16S.

L'étape d'extraction des gènes codant les ARNr 16S pose un certain nombre de difficultés car (i) une espèce peut être représentée par plusieurs souches dans la base de données des firmicutes, et plusieurs exemplaires peuvent être présents dans un même génome (ii) les gènes codant les ARNr 16S montrent une variabilité en taille entre les souches d'une espèce, voire dans certains cas entre les différents exemplaires d'ARNr 16S d'un génome. Ceci peut engendrer un alignement de séquences peu fiable et aboutir à un arbre peu robuste. Ainsi, pour surmonter ces deux difficultés, un gène d'ARNr 16S unique a été choisi comme représentant de l'espèce. La séquence d'ARNr 16S sélectionnée est celle qui, suite à une recherche d'homologie de séquences nucléotidiques (BLASTN) semble la plus représentative des séquences de l'espèce.

Les procédures de génération de l'arbre phylogénétique ont été effectuées sur la plateforme d'analyse de l'Institut Pasteur, Mobyle[7]. L'alignement des 70 séquences de gènes d'ARNr 16S représentatives des 70 espèces de la base de données a été réalisé par le programme *clustalw-multialign* qui implémente l'algorithme de clustalW (Thompson *et al.*, 1994) pour l'alignement multiple de séquences (Annexe Web B). Les paramètres utilisés sont ceux proposés par défaut :

- Utilisation d'une matrice d'identité,
- Pénalités d'ouverture de trous (10),
- Pénalités d'extension de trous (0,1),
- Pénalités de trous proches (8),
- Pourcentage d'identité maximum pour un alignement tardif (30 %).

Le programme *clustalw-multialign* produit un fichier contenant les séquences alignées mais aussi un fichier au format Newick[8] donnant la topologie de l'arbre avec des valeurs de distance d'évolution. Afin de reconstruire un arbre robuste des firmicutes, le programme Quicktree (Howe *et al.*, 2002) a été exécuté. Ce dernier implémente la méthode de distance *NJ* (pour Neighbor Joining) (Saitou et Nei, 1987) pour la reconstruction de l'arbre et il est possible d'utiliser comme modèle celui à deux paramètres de Kimura. Nous avons choisi cette méthode de distance pour la reconstruction de l'arbre phylogénétique des firmicutes car elle permet de tenir compte des taux de mutations différents dans chaque branche, ce qui n'est pas le cas pour la méthode *UPGMA* (pour Unweighted Pair Group Method with Arithmetic Mean) (Sneath et Sokal, 1973).

Nous avons aussi testé la construction une topologie d'arbre en utilisant la méthode du maximum de parcimonie. Cette méthode a l'avantage par rapport à celle de la méthode de distance NJ de considérer toute les positions de la séquence comme des sites informatifs. Néanmoins, l'analyse utilisant l'algorithme *dnapars* proposé dans la plateforme Mobyle n'a

[7] http://mobyle.pasteur.fr/cgi-bin/portal.py
[8] Newick format : http://evolution.genetics.washington.edu/phylip/newicktree.html

pas pu aboutir à l'obtention d'un arbre optimum mais cinq arbres ex-aequo (annexe Web C). Afin de ne pas induire de billet par rapport à l'arbre utilisé, nous avons choisi de poursuivre la contruction d'un arbre des espèces de firmicutes avec la méthode *NJ*.

Pour ce faire, le programme Quicktree a été employé aboutissant à un fichier au format Newick (Annexe Web D). L'arbre phylogénétique généré des 70 espèces appartenant au phylum des firmicutes (figure 24) est cohérent avec les arbres existants (arbre du genre *Streptococcus* ou arbre des lactocoques). Les distances entre tous les couples d'espèce ont été extraites en utilisant le logiciel PhyloDraw (Choi *et al.*, 2000) (Annexe G). Le calcul de distance entre deux séquences de souches (ou espèces) s'opère en additionnant toutes les branches visitées à partir d'une souche A jusqu'à atteindre la souche B.

Finalement, les distances entre la séquence de notre souche de référence, *S. thermophilus* CNRZ1066 et celle de toutes les autres souches représentant leur espèce ont été utilisées pour le calcul de la valeur de dispersion. Une espèce éloignée phylogénétiquement de *S. thermophilus* CNRZ1066 (*Staphylococcus saprophyticus*) a une distance plus grande que celle d'une espèce proche phylogénétiquement de *S. thermophilus* CNRZ1066 (*Streptococcus sanguinis*). Cependant, une espèce ayant une vitesse d'évolution rapide (par exemple, *Oenococcus oeni*) peut avoir une distance plus grande que celle d'une espèce phylogénétiquement éloignée de *S. thermophilus* CNRZ1066.

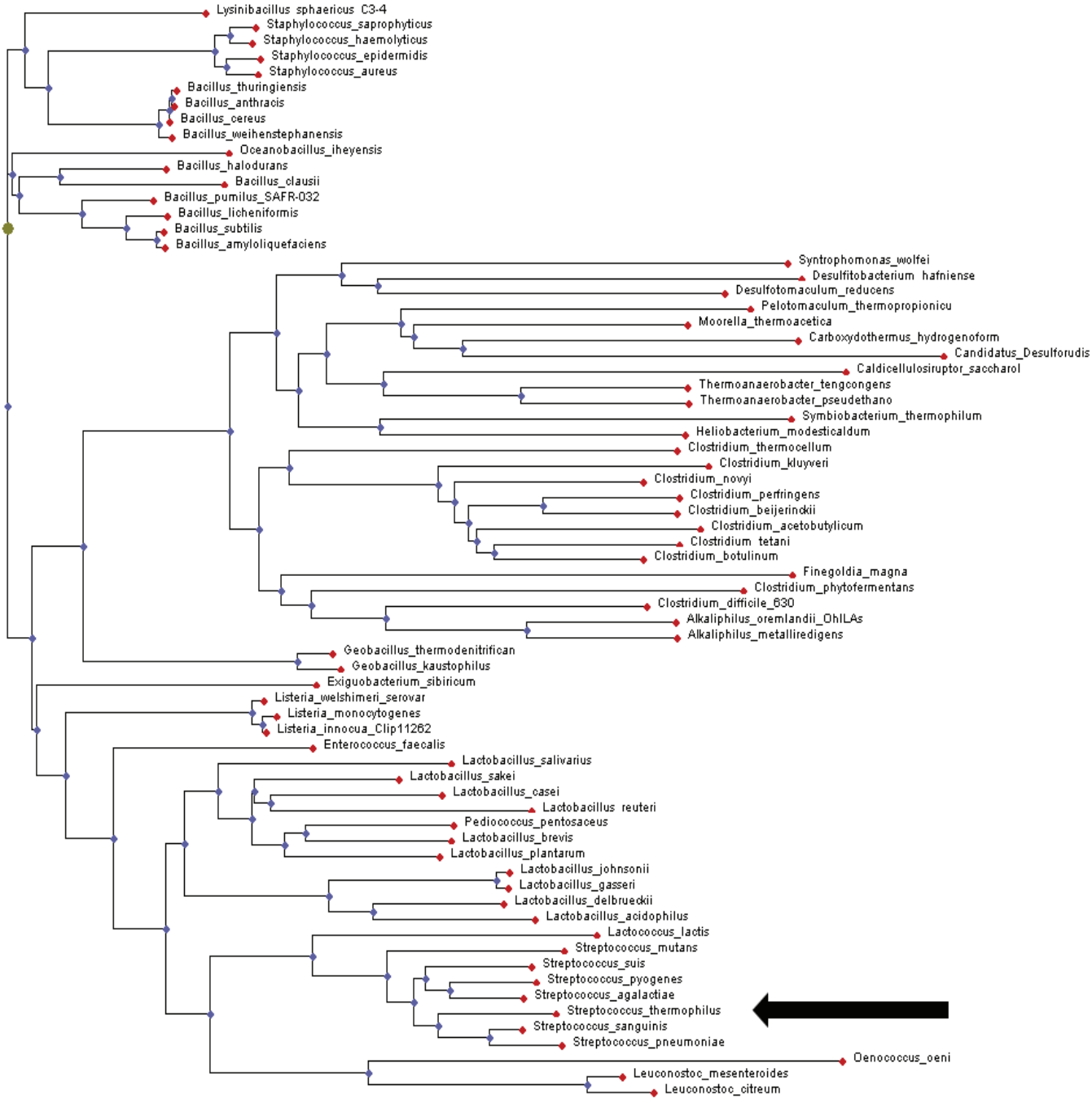


Figure 24 : Arbre phylogénétique des 70 espèces de la base de données des firmicutes basé sur les gènes d'ARNr 16S généré par le logiciel *PhyloDraw* (Choi *et al.*, 2000).
La méthode de reconstruction appliquée est le Neighbor Joining et le bootstrap est de 100 itérations.

3.8.2 Calcul de la valeur de dispersion pour un gène

La valeur de dispersion permet de rendre compte de la dispersion d'un gène au sein des firmicutes. En d'autres termes, il exprime numériquement l'éloignement des espèces présentant un homologue de ce gène par rapport à *S. thermophilus* CNRZ1066. Le nombre

d'espèces possédant un gène homologue est obtenu par l'interrogation de la table *Homologues* (cf. section 3.7) et l'information concernant la distance phylogénétique des espèces par rapport à *S. thermophilus* CNRZ1066 est fournie par les données de la matrice de distances des séquences (relatives à *S. thermophilus* CNRZ1066) calculées précédemment. Pour chaque gène « atypique » *j*, la valeur de dispersion correspond à un calcul d'écart type :

$$SD_j = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \tilde{x}\right)^2}{\left(n - 1\right)}}$$

Avec *n*, le nombre d'espèces possédant un gène homologue au gène « atypique », $\tilde{x}$, la valeur moyenne de la distance entre la séquence de *S. thermophilus* CNRZ1066 et les *n* espèces, possédant un gène homologue, et x_i, la valeur de distance entre la séquence de *S. thermophilus* CNRZ1066 et celle de l'espèce considérée *i*.

Ainsi, un gène qui a une valeur de dispersion élevée correspond à un gène dont des homologues sont retrouvés dans peu d'espèces, phylogénétiquement éloignées de *S. thermophilus*, de la base de données des firmicutes. Au contraire, un gène qui possède une valeur de dispersion faible correspond à un gène dont des homologues sont, soit présents dans peu d'espèces au sein des firmicutes mais proches phylogénétiquement de *S. thermophilus* CNRZ1066, soit largement distribués au sein des firmicutes.

Résultats : Analyse du génome de *S. thermophilus*

3.9 Analyses préliminaires de séquences chimériques

Afin de tester l'efficacité du modèle HMM d'ordre 2 (M2M0) pour la détection de séquences exogènes, une première analyse a été effectuée sur des séquences chimériques construites *in silico*. Pour cela, des séquences d'ADN chimériques ont été générées à partir de séquences d'éléments mobiles de *S. thermophilus*, en l'occurrence des éléments intégratifs et conjugatifs, ICE (pour Integrative and Conjugative Element) et des séquences apparentées CIME (pour *cis* mobilizable element). Les éléments ICE transfèrent, outre les gènes responsables de leur mobilité, des gènes de fonctions diverses inclus dans l'élément (ex. système de restriction-modification, résistance aux métaux lourds). La souche *S. thermophilus* CNRZ1066 est dépourvue d'élément au site *fba*, site spécifique d'intégration de ces éléments mobiles (Burrus *et al.,* 2000 ; Burrus *et al.,* 2001). Les chimères correspondent à l'insertion des éléments ICE*St1* (Genbank AJ278471), ICE*St3* (Genbank AJ586568), CIME302 (Genbank AJ586569), CIME19258 (Genbank AJ586571) et ΔCIME308 (Genbank AJ586570) au site d'insertion chromosomique *fba*. Ces éléments ont été choisis de par leur variabilité importante en termes de longueur et de contenu en bases G et C (Tableau 8). Ces éléments sont flanqués des sites d'attachement *attL* et *attR*, à l'exception de l'élément ΔCIME308 qui ne présente que le site *attR* (dans la partie droite). De plus, les éléments ICE*St1* et ICE*St3*, qui présentent un pourcentage en bases G et C moyen de 35 % et 36 % respectivement, sont caractérisés par la juxtaposition de deux régions de contenu en bases G et C distincts (31 % et 41 %).

Tableau 8 : Taille et composition des éléments ICE et CIME utilisés par rapport au génome de *S. thermophilus* CNRZ1066

	***S. thermophilus* CNRZ1066**	**ICE*St1***	**ICE*St3***	**CIME302**	**CIME19258**	**ΔCIME308**
Longueur (pb)	1 796 226	34733	28090	13148	14999	16832
Contenu en bases G et C (%)	39	35	36	30	34	36

Au sein des éléments ICE*St1* et ICE*St3*, le modèle de Markov d'ordre 2, estimé sur les génomes chimériques, a permis d'identifier plusieurs séquences du module de conjugaison : (i) la séquence *oriT* et (ii) une région incluant *orfA-D* (tableau 9 et figure 25). Le gène *orfD* code une protéine putative liant l'ATP/GTP et constitue un homologue d'*orf16* du transposon conjugatif Tn*916*. Le gène *orfC* code une protéine transmembranaire (homologue de *orf15* de Tn*916*). Quant à la fonction codée par *orfB*, elle est inconnue et le gène *orfA* code une protéine homologue de TraG/TrsG des plasmides conjugatifs de *Staphylococcus* pSK41 et pGO1. Ce module constitue la deuxième partie de l'élément dont le pourcentage en G+C (41 %) est proche de celui du génome de *S. thermophilus* CNRZ1066 (39,08 %). Quant à la

nature particulière du motif *oriT*, composé de séquences répétées inversées, elle pourrait avoir favorisé sa détection par le modèle M2M0.

Tableau 9 : Analyse des chimères *in silico*

	ICE*St1*	ICE*St3*	CIME302	CIME19258	ΔCIME308
Longueur (pb)	34733	28090	13148	14999	16832
Longueur détectée [a] (pb)	5498	6158	1626	3832	7236
M2M0 [b] (états-mers)	6-4	6-4	5-5	5-3	5-3

(a) Longueur cumulée des régions détectées par le M2M0. (b) Paramètres du M2M0 utilisés. Les régions « atypiques » ont été extraites automatiquement avec les paramètres suivants : W = 400 nt, S = 300 nt, MinT = 0,2, MinN = 20 et MinL = 2200 nt.

Dans les éléments CIME, le modèle détecte notamment *orf302A* (CIME302) codant une méthyltransférase putative pouvant appartenir à un système de restriction-modification. Au sein de l'élément CIME19258, la région détectée porte des gènes codant un système de transport du cadmium. Enfin, la moitié de l'élément ΔCIME308 est décelée ; cet élément contient notamment des homologues de gènes codant des systèmes de restriction-modification. De plus, la plupart des gènes de cet élément sont homologues à des gènes de *L. lactis*.

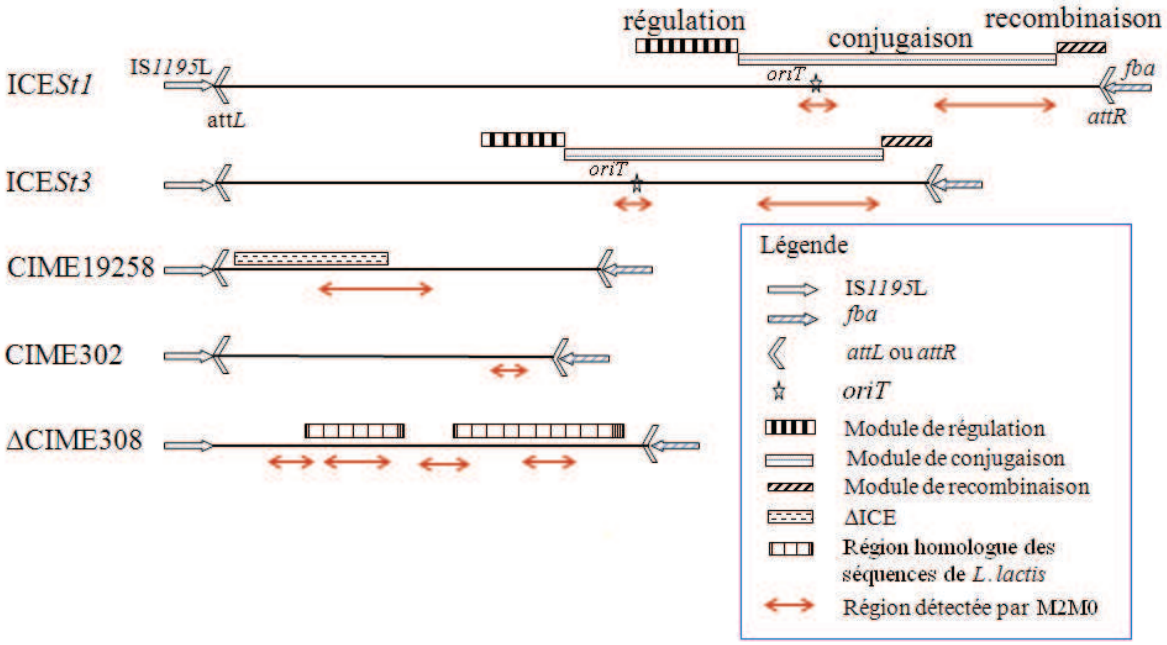


Figure 25: Localisation des régions identifiées par le HMM (M2M0) dans les éléments mobiles de *S. thermophilus*.

Ces tests *in silico* permettent de valider l'efficacité d'un modèle d'ordre 2 dans la détection de régions « atypiques ». Par conséquent, un modèle HMM d'ordre 2 a ensuite été appliqué à l'ensemble du génome de *S. thermophilus* CNRZ1066 et les régions décelées seront appelées « atypiques » tout au long du manuscrit.

3.10 Analyse globale du génome de *S. thermophilus* CNRZ1066

3.10.1 Extraction et analyse de gènes « atypiques »

La caractérisation des séquences et gènes décelés a été réalisée par les deux modèles HMM d'ordre 2 (M2M0 et M2M2). Nous avons choisi d'exposer l'analyse avec le modèle M2M2 permettant une comparaison ultérieure d'efficacité avec le modèle M1M2 pour la détection de gènes « atypiques » en fonction seulement de la dépendance de la suite d'états cachés. La répartition génomique des gènes « atypiques » détectés par le M2M2 le long du génome a été déterminée, et une analyse de la composition nucléotidique de ces séquences a été effectuée. Ces gènes ont ensuite été analysés au travers de leur annotation afin d'identifier des fonctions susceptibles d'avoir été transférées. L'analyse de leur distribution chez les firmicutes, phylum bactérien auquel appartiennent les streptocoques, a été entreprise afin d'évaluer le caractère variable de ces gènes au sein de ce phylum, et dans certains cas, d'interpréter cette variabilité en terme de gain (transfert horizontal) ou de perte au cours de l'évolution de ces bactéries. Enfin, la distribution de ces séquences au sein de 47 souches de *S. thermophilus* a également été déterminée grâce à des données de CGH qui nous ont été fournies par le groupe Chr. Hansen S/A.
Suite à l'analyse du génome de *S. thermophilus* CNRZ1066 par le modèle M2M2, l'extraction des régions « atypiques » a été accomplie avec les paramètres suivants : une fenêtre coulissante (W) de 400 nt, un pas (S) de 300 nt, un seuil minimum du nombre de positions ayant une probabilité *a posteriori* supérieure à la moyenne (MinN) de 20 et une taille minimale de région candidate fixée à 1000 nt.

Distribution des gènes « atypiques » le long du génome de S. thermophilus
Ce protocole d'analyse a conduit à considérer 146 régions « atypiques » représentant 17,7 % du génome de *S. thermophilus* CNRZ1066. Ces régions définissent un ensemble de 362 gènes « atypiques » dont la localisation le long du génome de *S. thermophilus* CNRZ1066 est représentée sur la figure 26. La région « atypique » la plus grande comporte 14 gènes. Quarante neuf régions sont constituées de 2 gènes et 41 sont des singlets.

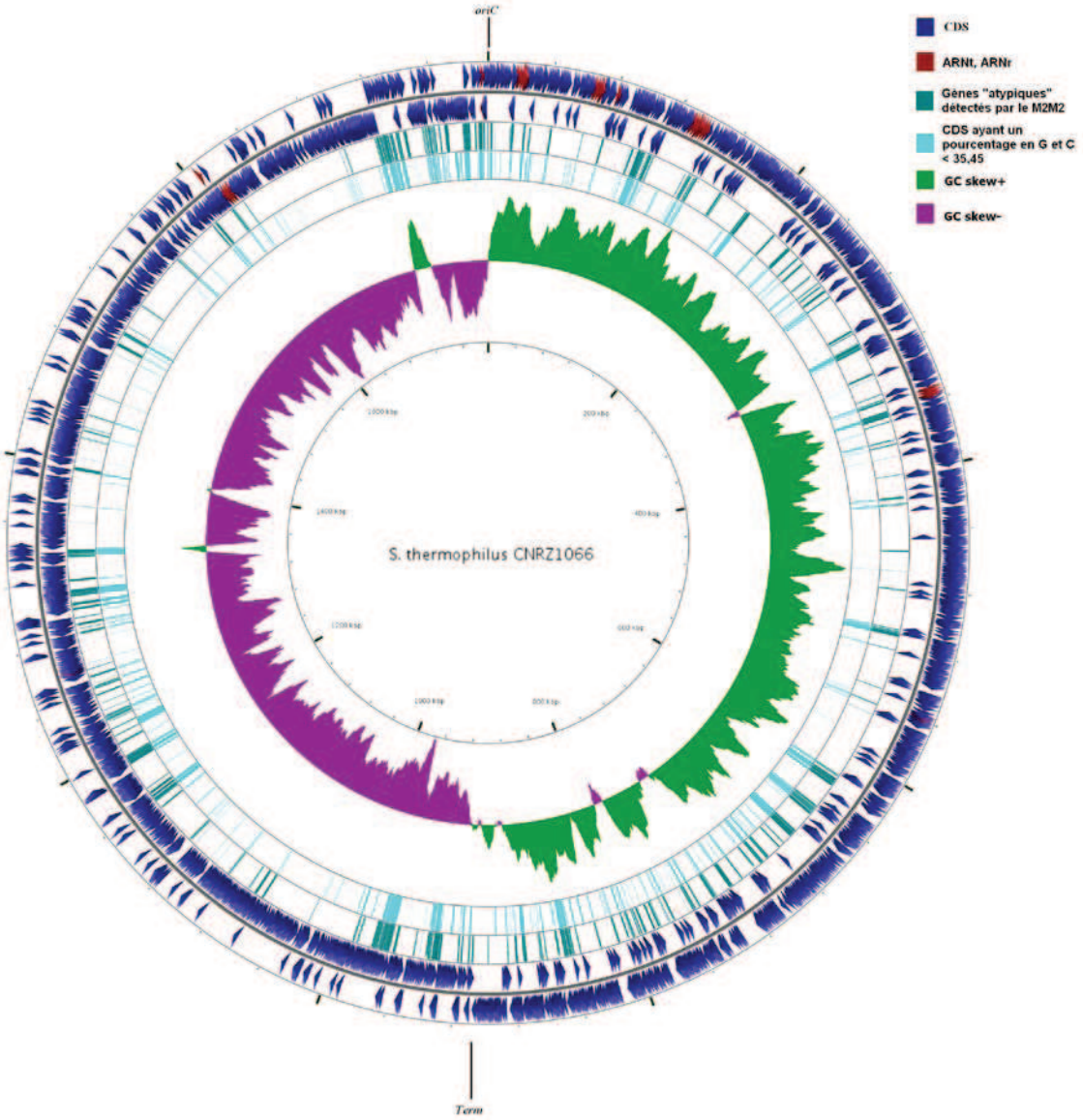


Figure 26 : Localisation des 362 gènes « atypiques » sur le génome de *S. thermophilus* CNRZ1066.

Les deux cercles extérieurs représentent les CDS sur chaque brin de l'ADN. Les traits verticaux verts symbolisent les 362 gènes « atypiques » détectés par le M2M2 et les traits verticaux bleus clairs indiquent les gènes ayant un contenu en bases G et C inférieur à 35,45 % (m-σ = 39,3-3,88). Le cercle le plus interne décrit l'évolution du GC skew le long du génome.

Il est possible de discerner des zones de répartition denses à proximité du terminus de réplication (*Term*) et à proximité de l'origine de réplication *oriC*. Au voisinage d'*oriC*, une région présente une inversion de GC skew. Compte tenu du fort biais observé sur le génome de *S. thermophilus*, cette inversion peut révéler, soit une insertion de séquences exogènes, soit une inversion. L'hypothèse d'une insertion de séquence exogène est renforcée par la présence de deux gènes annotés comme codant, une transposase tronquée et l'autre une intégrase/recombinase associée à un phage. Cinq autres régions de petite taille montrent aussi un GC skew inversé et trois d'entres elles sont identifiées comme « atypiques ».

Dans le cas de la région du terminus de la réplication, il a été rapporté par Daubin et Perrière (2003) que la plupart des génomes bactériens présentaient un appauvrissement du $(G+C)_3$

(contenu en bases G et C à la troisième position des codons) quelque soit le pourcentage en bases G et C moyen du génome. La région localisée entre les positions 600 000 et 1 000 000 du génome de *S. thermophilus* présente un pourcentage en GC de 37,82 %, approchant de celui du génome complet (39,08 %). Ce léger enrichissement en bases A et T ne permet pas d'expliquer la densité accrue des gènes « atypiques » dans cette région. Exceptées ces deux régions de forte densité, les autres gènes « atypiques » sont dispersés le long du génome.

Paramètres compositionnels des séquences « atypiques »

Afin de décrire les 362 gènes « atypiques », cinq caractéristiques ont été analysées : la taille des gènes extraits, leur composition en bases G et C, leur efficacité de traduction, la localisation cellulaire présumée des protéines codées et leur catégorisation fonctionnelle. Ces paramètres ont été comparés à ceux de deux ensembles de données : (i) à l'ensemble des 1915 CDS du génome, et (ii) l'ensemble des CDS du génome desquelles sont retranchées les gènes « atypiques », soit 1553 CDS. Ce dernier échantillonnage permet de mesurer l'impact des 362 gènes « atypiques » sur les biais analysés.

Longueur des gènes « atypiques »

La longueur moyenne des gènes « atypiques » est de 613 pb alors que celle des 1553 CDS non « atypiques » de *S. thermophilus* CNRZ1066 est de 829 pb. Comment expliquer cette différence de taille ? On observe que 21,3 % de ces gènes « atypiques » sont annotés comme « gènes tronqués » contre 15,7 % dans le reste du génome (tableau 10, figure 27.A). Toutefois la différence de taille est toujours notable entre les gènes non tronqués de ces deux ensembles de gènes (285 et 1309 gènes respectivement). Nous avons également envisagé que la faible taille des gènes « atypiques » résulte d'une proportion plus importante de gènes codant potentiellement des recombinases (portées par des IS) ou des gènes d'origine phagique. Cependant, seuls 7 gènes dont l'origine phagique est identifiée sont présents dans le jeu de gènes « atypiques », et bien que leur taille moyenne soit faible (643 nt), ils ne peuvent faire chuter significativement la taille moyenne. En revanche, les gènes codant des recombinases présentent une taille moyenne plus faible comme le montre l'analyse rapportée dans le tableau 10 et sont enrichis au sein des gènes « atypiques » (11 % contre 4,6 % dans le reste du génome). Toutefois, la taille moyenne des gènes « atypiques » retranchés des gènes codant des recombinases (635 nt) reste plus faible que dans le reste du génome (844 nt).

Il peut être noté que les gènes de fonction hypothétique sont de taille moyenne faible (~ 550 pb) et qu'ils sont présents en proportion plus importante dans les gènes « atypiques » que dans le reste du génome (44 % contre 26 %). Cependant, les gènes « atypiques » codant des protéines de fonction non hypothétique ont une taille moyenne de 670 pb qui est supérieure à la moyenne de 613 pb des gènes « atypiques » mais reste inférieure à la taille moyenne des autres CDS du génome (829 pb). Ainsi, aucune catégorie de gènes distinguée selon nos critères ne peut à elle seule expliquer la différence de taille moyenne entre les deux ensembles de gènes.

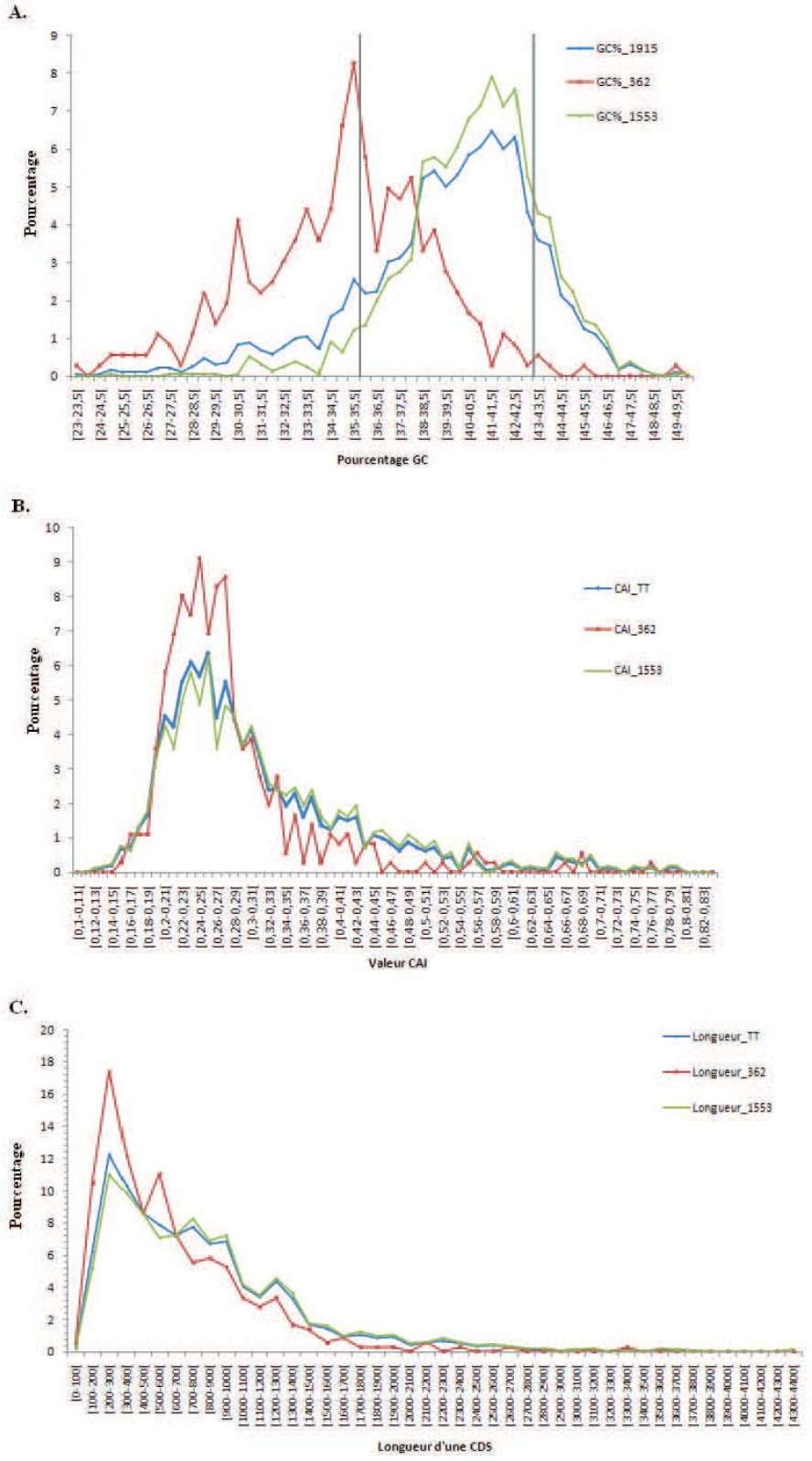


Figure 27 : Analyse compositionnelle des gènes « atypiques ».

A. Pourcentage en bases G et C. La zone comprise entre les traits verticaux gris indique l'intervalle m ± σ.
B. Valeur du CAI.
C. Longueur des CDS.
Des valeurs de pourcentage (nombre de gènes observés divisé par le nombre total de gènes dans chaque ensemble) sont utilisées afin de normaliser l'information car les jeux de données ont des tailles très différentes.

Tableau 10: Longueur (en pb) des 362 gènes « atypiques » et 1553 CDS du reste du génome

	Gènes « atypiques »	CDS du reste du génome
Nombre de gènes (1)	362 (613)	1553 (829)
Protéine tronquée (2)	77 (406)	244 (389)
Protéine non tronquée	285 (669)	1309 (911)
Proportion (2)/(1)	0,21	0,16
Fonction recombinase (4)	41 (445)	72 (527)
Fonction non recombinase	321 (635)	1481 (844)
Proportion (4)/(1)	0,11	0,046
Protéine hypothétique (3)	160 (543)	399 (553)
Protéine non hypothétique	202 (670)	1154 (925)
Proportion (3)/(1)	0,44	0,256

Le contenu en bases G et C des gènes « atypiques »

Le pourcentage moyen en bases G et C des phases codantes du génome de *S. thermophilus* CNRZ1066 est de 39,3 % ; cette valeur est proche de celle du génome global (39,08 %) tandis que celle des 362 gènes « atypiques » est de 34,7 %. Une hétérogénéité marquée de la composition des séquences « atypiques » est observée (figure 27.A). En effet, alors que 15 % des gènes totaux (287/1915) présentent un pourcentage en bases G et C inférieur à 35,42 % (c'est-à-dire m-σ = 39,3-3,88), 57% (206/362) des gènes « atypiques » présentent un pourcentage en bases G et C inférieur à cette même valeur. Il est à noter que parmi les 362 gènes, 151 (soit 41,7 %), présentent une composition en bases G et C moyenne (comprise entre m ± 3,88).

Cinq gènes « atypiques » possèdent un contenu en bases G et C supérieur à 43,18 % (*str0947*, *str1594*, *str1709*, *str1808*, *str1685*), parmi lesquels, un code une protéine ribosomique, deux codent des protéines hypothétiques, un autre code une protéine de transport (efflux) tronquée et le dernier code une bactériocine putative.

L'adaptation des codons des gènes « atypiques »

L'index d'adaptation des codons d'un gène (notée Codon Adaptation Index, CAI) est la somme des valeurs d'adaptation relative de chaque codon composant le gène ; valeurs comprises entre 1 si le codon est le plus fréquent et 0 si le codon n'est pas utilisé (Sharp et Li, 1987). La valeur d'adaptation relative d'un codon correspond au rapport entre la fréquence de ce codon parmi les codons synonymes et la fréquence du codon synonyme le plus fréquent. Plus la valeur de CAI d'un gène est élevée, plus celui-ci sera efficacement traduit.

La figure 27.B souligne une différence faible de la distribution des valeurs de CAI des gènes « atypiques » par rapport aux deux autres ensembles de gènes considérés. En effet, une diminution survient pour des valeurs de CAI supérieures à 0,34, ce qui indique que les gènes « atypiques » sont moins adaptés à une forte expression en proportion par rapport à la distribution des valeurs de CAI pour l'ensemble des CDS du génome. Le fort biais de la composition en bases G et C des gènes « atypiques » semble être à l'origine du biais de CAI.

Localisation cellulaire présumée des protéines codées

La prédiction de la localisation des produits des 362 gènes « atypiques » et des 1915 CDS du génome est basée sur la recherche d'une part d'hélices transmembranaires (à l'aide du logiciel TMHMM[9]) (Sonnhammer *et al.,* 1998) et d'autre part d'un signal d'excrétion (à l'aide du logiciel SignalP[10]) (Nielsen *et al.,* 1997). D'après ces prédictions, la proportion des protéines selon leur localisation cellulaire (ancrée dans la membranaire ou sécrétée) est similaire soit environ 23 % à 24 % pour l'ensemble des 362 gènes « atypiques » et pour l'ensemble des gènes du génome. Ce troisième critère ne permet pas de discriminer l'ensemble des 362 gènes par rapport au reste du génome (tableau 11).

Tableau 11 : Prédiction de la localisation cellulaire des protéines

	Protéines codées par les 1915 gènes du génome	Protéines codées par les 362 gènes «atypiques»
Protéines ancrées dans la membrane (%)	293 (22,5)	57 (22,1)
Protéines sécrétées (%)	28 (8,7)	3 (7,2)
Protéines présentant un peptide signal et au moins un domaine transmembranaire (%)	138 (7,2)	23 (6,4)
Total (%)	459 (24)	83 (22,9)

Catégorisation fonctionnelle des gènes « atypiques »

Les 362 gènes « atypiques » ont été ventilés dans les catégories fonctionnelles proposées par l'Institut TIGR (The Institute of Genomic Research)[11]. Cette distribution est donnée dans la figure 28. Il en ressort cinq principales catégories fonctionnelles qui recouvrent plus de 40 % des gènes « atypiques »: 'Enveloppe cellulaire' (11,60 %), 'Elément mobile et extrachromosomique' (10,22 %), 'Fonction de régulation' (7,73 %), 'Métabolisme de l'ADN' (6,35 %) et 'Processus cellulaire' (6,35 %). En comparaison, l'ensemble des gènes du génome de *S. thermophilus* CNRZ1066 se répartit différemment, les cinq principales catégories fonctionnelles étant les suivantes : 'Protéine de transport et de fixation' (11,49 %), 'Métabolisme énergétique' (8,10 %), 'Enveloppe cellulaire' (7,74 %), 'Synthèse de protéines' (6,83 %) et 'Fonction de régulation' (6,56 %). La distribution des 362 gènes « atypiques » dans les catégories fonctionnelles est en concordance avec une étude précédente de classification des gènes acquis par HGT provenant de 116 génomes de procaryotes (Nakamura *et al.,* 2004). En effet, d'après cette étude, les gènes ayant une origine exogène sont distribués principalement dans les quatre catégories fonctionnelles suivantes : 'Elément mobile et extrachromosomique', 'Enveloppe cellulaire', 'Fonction de régulation' et 'Processus cellulaire'.

Dans l'analyse des catégories fonctionnelles des gènes « atypiques », une catégorie fonctionnelle représente à elle seule 43 % des gènes « atypiques ». Il s'agit de la catégorie 'Fonction

[9] http://www.cbs.dtu.dk/services/TMHMM-2.0/
[10] http://www.cbs.dtu.dk/services/SignalP/
[11] TIGR database: http://www.tigr.org/db.shtml

inconnue'. Cette dernière correspond à trois catégories fonctionnelles fusionnées qui sont 'Protéine hypothétique', 'Protéine de fonction inconnue' et 'Protéine ayant une fonction encore non classée'. Dans la catégorie 'Protéine ayant une fonction encore non classée' se trouve des gènes codant des ABC transporteurs hypothétiques, des protéines de biosynthèse de lantibiotique et des protéines de biosynthèse d'exopolysaccharides qui sont connus pour être fréquemment échangés.

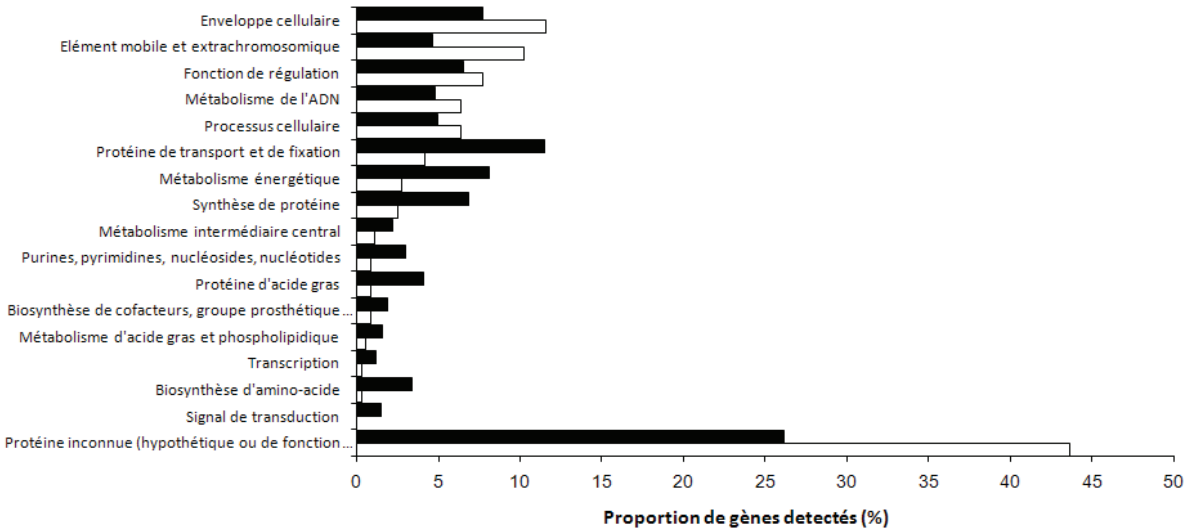


Figure 28 : Distribution des gènes dans les différentes catégories fonctionnelles chez *S. thermophilus* CNRZ1066.

Les barres noires représentent la proportion des gènes du génome dans les différentes catégories fonctionnelles. Les barres blanches indiquent la répartition des gènes « atypiques » dans ces mêmes catégories.

Cette catégorisation fonctionnelle montre que la distribution des fonctions des gènes « atypiques » est en accord avec leur potentielle origine exogène.

3.10.2 Annotation des gènes « atypiques »

L'annotation fonctionnelle peut suggérer, selon la fonction prédite, une susceptibilité plus ou moins forte aux transferts horizontaux. En effet, les gènes de résistance aux antibiotiques, les gènes codant des systèmes de restriction-modification ou encore les éléments génétiques mobiles sont réputés pour être fréquemment échangés. *S. thermophilus* CNRZ1066 possède 113 gènes codant des recombinases (annotés IS, intégrases/recombinases ou recombinases, putatives ou tronquées) qui représentent 5,9 % (113/1915) des CDS du génome. Quarante et un gènes « atypiques » codent des recombinases d'éléments mobiles (tableau 12). Ainsi, l'ensemble des 362 gènes « atypiques » présente un enrichissement d'un facteur 2 en « gènes de mobilité » (41/362) par rapport à l'ensemble du génome. Au total, 45 % des gènes « atypiques » ayant une fonction annotée (81/179) sont considérés comme ayant potentiellement une origine exogène (tableau 12).

Tableau 12 : Annotation fonctionnelle de gènes ou séquences suggérant une acquisition par HGT

Fonction ou séquence	Nombre de gènes « atypiques »
Recombinase	41
Biosynthèse ou résistance aux bactériocines	9
Système de restriction modification	9
Protéine associée à un phage	5
Système d'efflux multidrogues (ABC transporteur)	1
Biosynthèse d'exopolysaccharide	12
Ilot « exogène » de 17 kb (Bolotin *et al.*, 2004)*	4
Total	81

* Ilot exogène contenant 13 gènes présentant de fortes identités avec des séquences de *Lactococcus lactis* et *Lactobacillus bulgaricus*. Le M2M2 décèle 7 gènes de cet îlot dont un code une recombinase et deux codent un système de restriction modification qui sont inclus dans la catégorie « Recombinase » et « Système de restriction modification ».

Le cluster *eps* impliqué dans la biosynthèse d'exopolysaccharides présente une structure mosaïque qui est variable au sein de l'espèce *S. thermophilus* (Bourgoin *et al.*, 1999). Cette structure chimérique présente des traces d'événements de recombinaison intervenus entre séquences divergentes (de 10 à 50 %). Ce scénario d'évolution complexe implique l'acquisition par transfert horizontal de séquences de *L. lactis*. Le cluster *eps* de la souche CNRZ1066 est composé de 18 gènes dont 13 sont décelés par le modèle M2M2.

En plus de ces gènes qui codent des fonctions connues pour être soumises au HGT, d'autres régions ont été détectées comme ayant une origine exogène, c'est le cas du locus CRISPR et d'une région étiquetée comme « exogène » de 17 kb. Le locus CRISPR (Clustered Regulary Interspaced Short Palindromic Repeats) est composé de courtes séquences (21-48 pb) répétées et conservées. Ces séquences répétées sont espacées par des régions variables et de tailles similaires (entre 20 à 58 pb). Ces séparateurs sont supposés dériver de séquences de phages et fournissent un système de résistance à la bactérie contre les bactériophages (Barrangou *et al.*, 2007 ; Semenova *et al.*, 2009). Les CRISPR ont une structure conservée et répandue au sein des génomes bactériens. Les gènes *cas* sont toujours associés aux séquences CRISPR, et il est proposé que ces gènes codent notamment une nucléase qui permettrait une reconnaissance et un clivage spécifique d'une séquence cible (Barrangou *et al.*, 2007). Chez *S. thermophilus* (Horvath *et al.*, 2008), trois loci CRISPR ont été identifiés mais ne sont pas tous présents dans les trois souches dont les génome ont été entièrement séquencés et accessibles. En effet, le locus CRISPR3 est spécifique de LMD-9 alors que le locus CRISPR2 est retrouvé dans les trois génomes séquencés de *S. thermophilus* (figure 29).

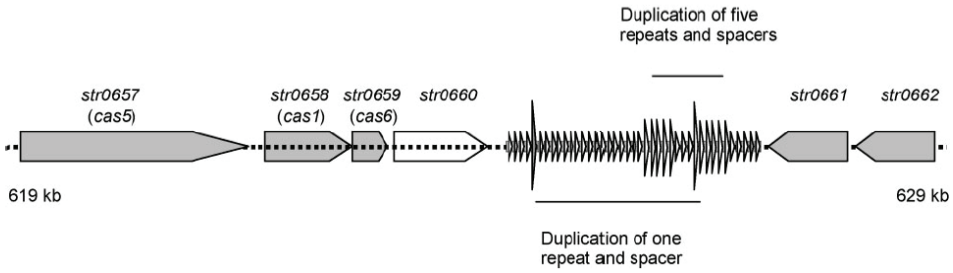


Figure 29 : Organisation du cluster CRISPR2 présent chez *S. thermophilus* CNRZ1066 (Bolotin *et al.*, 2005).
Les séquences répétées et espacées de séquences nucléotides sont représentées respectivement par des carrés et des triangles grisés.

En ce qui concerne l'îlot « exogène » de 17 kb, il est constitué de 23 gènes et est limité de part et d'autre par deux gènes qui codent chacun une dipeptidase tronquée d'après l'annotation (*str0834* et *str0858*). Cette région présente des séquences ayant plus de 90 % d'identité avec des séquences provenant de lactocoques (figure 30).

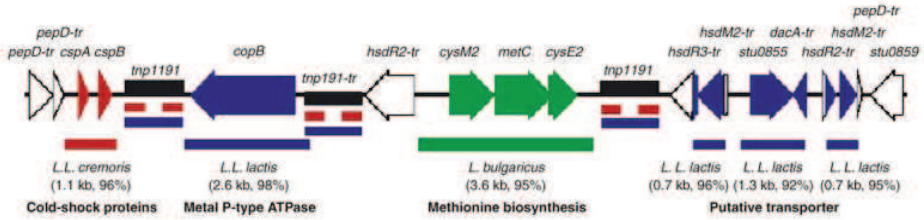


Figure 30 : Région exogène de 17 kb identifiée dans le génome de *S. thermophilus* (Bolotin *et al.*, 2004).

Ce jeu de 81 gènes « atypiques » ayant une fonction annotée sera noté *Annotation_HGT*.

3.10.3 Distribution phylogénétique des gènes « atypiques »

L'analyse phylogénétique est basée sur la reconstruction d'arbres phylogénétiques et une analyse de la distribution des gènes « atypiques » ; la reconstruction est fondée sur une séquence dont l'hérédité est verticale (et constitue l'arbre vrai) et l'analyse de distribution à partir d'une séquence sujette au transfert horizontal est réalisée par la recherche de séquences homologues dans l'arbre des espèces précédemment reconstruit. La distribution des gènes « atypiques » révèle des gènes supposés hérités par HGT. Les gènes accessoires qui sont des gènes fréquemment transférés sont aussi souvent remplacés et donc perdus au cours de l'évolution. L'étude de la distribution des gènes « atypiques » dans les groupes phylogénétiques que sont les firmicutes (phylum) et l'espèce *S. thermophilus* est susceptible d'apporter des éléments permettant d'inférer des événements de transfert horizontal.

Dans le phylum des firmicutes
Dans une première démarche « quantitative », la distribution des gènes « atypiques » a été déterminée en dénombrant les occurrences d'un gène donné dans le phylum des firmicutes (à

l'aide de la base de données constituée à cet effet) sans considérer les liens phylogénétiques entre espèces (cf. section 3.8). Pour cette approche quantitative, un seuil a été établi qui correspond au nombre d'espèces maximum présentant un gène homologue en dessous duquel un gène sera considéré comme sporadique. Si le nombre d'occurrences est inférieur au seuil, l'hypothèse d'une acquisition par transfert horizontal sera posée.
Afin de déterminer la valeur du seuil, trois jeux de contrôle de 362 gènes pris au hasard dans le génome de *S. thermophilus* CNRZ1066 ont été générés. Les occurrences des gènes de ces trois jeux ont été comptées au sein des firmicutes dans l'objectif de déterminer si les gènes « atypiques » (extraits par le modèle M2M2) présentaient une distribution particulière. Parmi les 362 gènes « atypiques », 79 (22 %) s'avèrent n'être présents que dans l'espèce *S. thermophilus* (figure 31) alors que pour les trois jeux de gènes tests, 20 gènes (5,5 %) en moyenne ne sont présents que chez *S. thermophilus*. Ainsi, les gènes « atypiques » sont enrichis en gènes spécifiques de l'espèce *S. thermophilus* par un facteur 4.
Par ailleurs, la comparaison de la distribution des gènes « atypiques » d'une part, et de la distribution moyenne des trois jeux tests d'autre part, permet de visualiser le seuil en dessous duquel nous pouvons considérer le nombre d'occurrences d'un gène comme compatible avec son acquisition par transfert. Le seuil choisi est de 7. En effet, il correspond au nombre d'espèce en dessous duquel les deux histogrammes divergent. Par ailleurs, une représentation cumulative des occurrences des 362 gènes « atypiques » en fonction de leur distribution au sein des firmicutes met en évidence une inflexion de la courbe au voisinage de 7 espèces (figure 32). Cela est cohérent avec le contenu de la base de données dans laquelle 7 génomes du genre *Streptocoques* sont représentés (*Streptococcus thermophilus*, *Streptococcus pneumoniae*, *Streptococcus sanguinis*, *Streptococcus agalactiae*, *Streptococcus suis*, *Streptococcus mutans*, *Streptococcus pyogenes*). Un exemple de fichier de sortie des résultats est proposé pour les gènes « atypiques » identifiés dans le cluster CRISPR (annexe H).

En appliquant ce seuil de 7, c'est-à-dire en sélectionnant parmi les 362 gènes « atypiques » ceux qui présentent des homologues dans au plus 7 espèces représentées dans la base de données des firmicutes, 189 apparaissent ainsi comme sporadiques (cet ensemble sera baptisé *Firmicutes_s7*). Cette proportion indique un enrichissement d'un facteur 5 en gènes à distribution sporadique dans l'ensemble des gènes « atypiques » par rapport aux jeux tests.

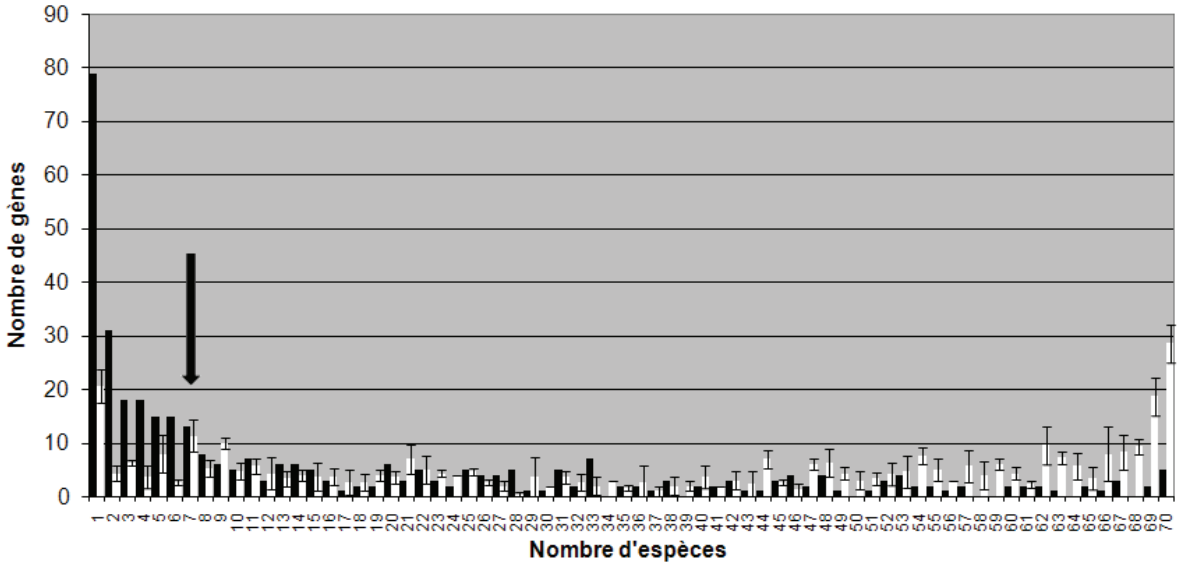


Figure 31 : Distribution des gènes « atypiques » chez les firmicutes.

L'histogramme blanc correspond à une moyenne des distributions des trois jeux tests (362 gènes choisis aléatoirement) et l'écart type est indiqué par des barres d'erreur. L'histogramme noir correspond à la distribution des 362 gènes « atypiques ».

La flèche indique le seuil de 7 choisi pour discriminer les gènes présentant une distribution sporadique.

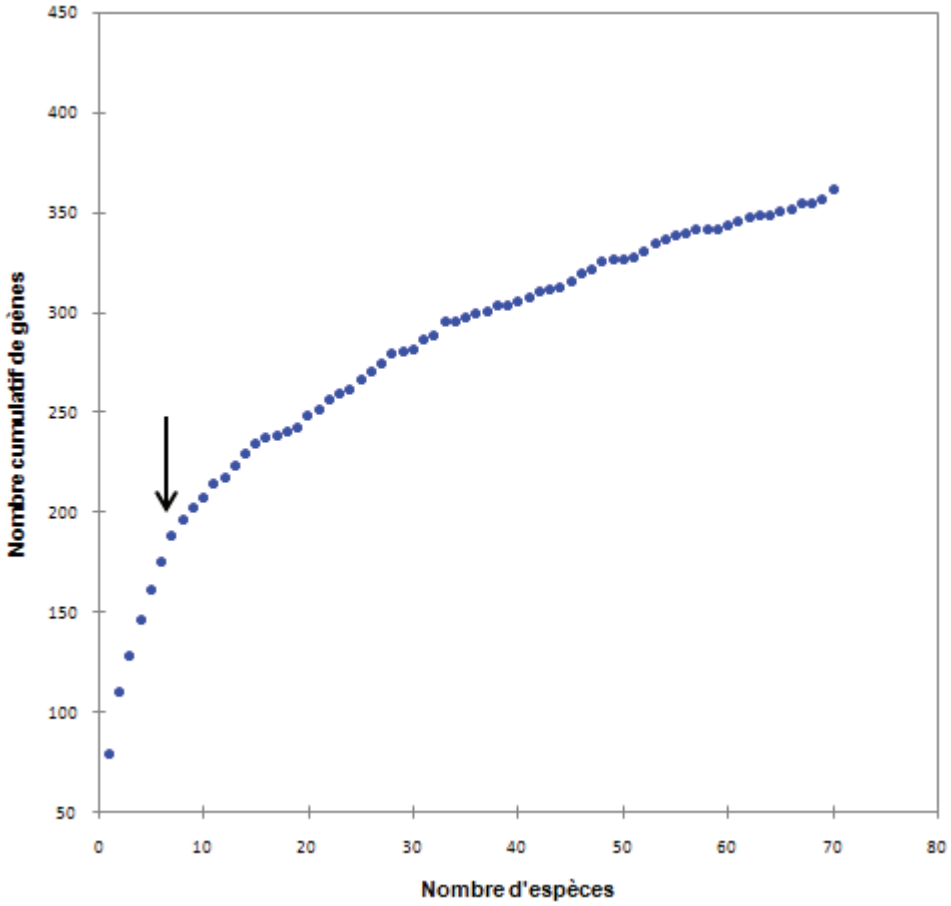


Figure 32 : Nombre cumulatif de gènes « atypiques » en fonction de leur occurrence chez les firmicutes.

La flèche indique le point d'inflexion. Les coordonnées de ce point correspond en abscisse au seuil de sept espèces et en ordonnées à 189 gènes « atypiques » présents dans au plus sept espèces de cette base de donnés.

Au sein de l'espèce S. thermophilus

La compagnie danoise Chr. Hansen S/A a entrepris une analyse comparative de 47 souches de *S. thermophilus* par la méthode de CGH (Comparative Genome Hybridization). La puce de référence comprend (i) les CDS du génome de la souche LMG18311, (ii) les CDS spécifiques des deux autres génomes entièrement séquencés (CNRZ1066 et LMD-9) ainsi que (iii) les CDS de l'espèce *S. thermophilus* provenant de la banque de données Genbank. Au total, 2250 CDS sont représentées chacune par un oligonucléotide. Une hybridation avec l'ADN génomique des 47 souches a été réalisée. Parmi les 1915 gènes de la souche CNRZ1066, 1782 ont générés des signaux exploitables. A partir des signaux obtenus, l'occurrence des gènes de la souche CNRZ1066 dans les 47 souches analysées a été établie. Celle-ci est comparée à la distribution des gènes « atypiques » (figure 33). Il apparaît que ces distributions sont quasi-superposables jusqu'au seuil de 15 souches. En effet, 73 CDS sont présentes dans moins de 15 souches et parmi elles, 54 (74 %) sont décelées par le M2M2. Ceci correspond à un enrichissement d'un facteur 4 en gènes variables au sein des gènes « atypiques » comparativement à l'ensemble des gènes du génome et un enrichissement d'un facteur 14 comparativement à l'ensemble des gènes non « atypiques ». Ces 54 gènes variables et « atypiques » sont répartis dans 14 segments du génome (Annexe I).

Le sous-ensemble de 73 gènes variables au sein de l'espèce *S. thermophilus* sera appelé *CGH-set*.

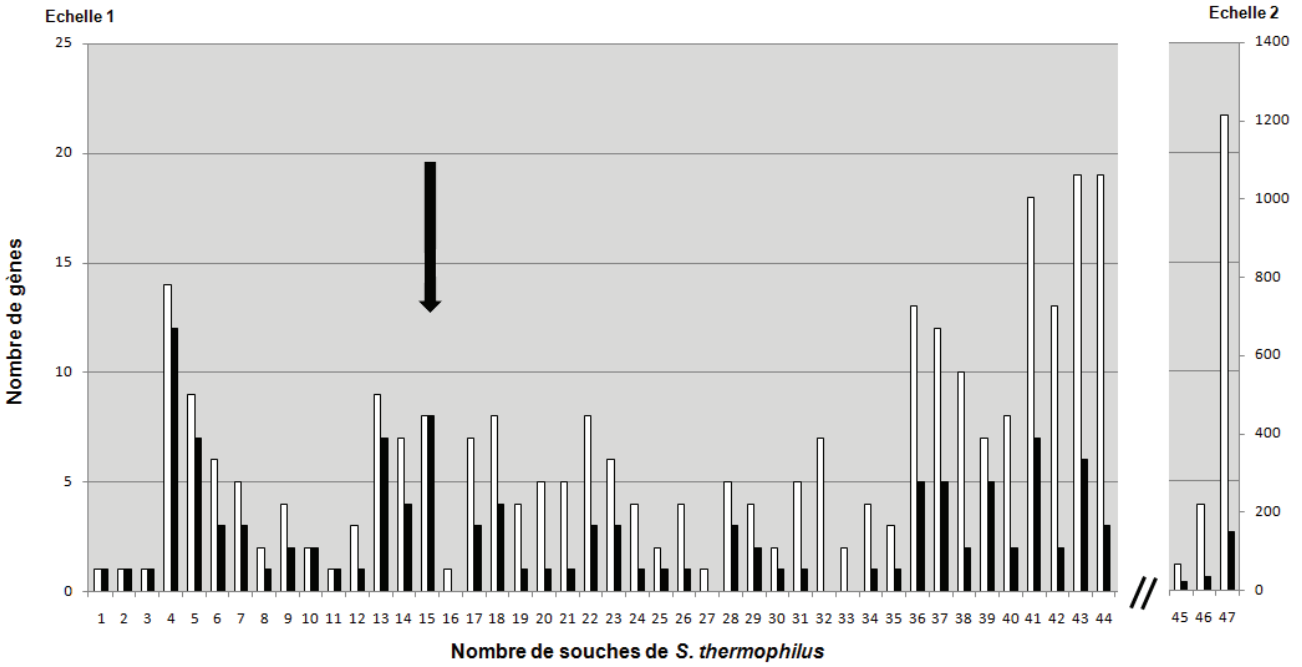


Figure 33 : Comparaison de la distribution des gènes variables et « atypiques » de *S. thermophilus* CNRZ1066 au sein des 47 souches utilisées dans les expériences CGH.

Distribution chez les 47 souches de *S. thermophilus* de l'ensemble des 1915 gènes et des 362 gènes « atypiques ». Seuls 1782 gènes parmi les 1915 et 324 parmi les 362 gènes « atypiques » ont donnés des signaux exploitables dans les expériences CGH.

Les histogrammes blancs et noirs représentent respectivement la distribution des 1782 gènes et des 324 gènes « atypiques ». L'axe des ordonnées représente le nombre de gènes. Deux échelles en ordonnées ont été générées afin d'apprécier la distribution de petit et de grand nombre de gènes de manière simultannée.

Indice de dispersion

Les informations de distribution d'un gène au sein des firmicutes et au sein des souches de *S. thermophilus* sont à exploiter de manière conjointe. Celles-ci renseignent sur le caractère spécifique d'espèce et/ou de souche du gène analysé et permettent d'émettre des hypothèses quant à une acquisition récente, c'est-à-dire postérieure, ou antérieure à la divergence de l'espèce *S. thermophilus*.

Une valeur de dispersion (cf. section 3.8) a été calculée pour chaque gène « atypique ». Cette valeur exprime l'éloignement par rapport à *S. thermophilus* des espèces chez lesquelles un homologue du gène « atypique » est présent (seuils appliqués : identité > 30 %, e-value < 10^{-6}). Ainsi, un gène présentant une valeur de dispersion élevée correspond à un gène dont des homologues sont identifiés dans peu d'espèces de la base de données des firmicutes, ces espèces sont également phylogénétiquement éloignées de *S. thermophilus*. A l'inverse, un gène possédant une valeur de dispersion faible correspond à un gène dont des homologues sont soit présents dans peu d'espèces au sein des firmicutes mais des espèces proches phylogénétiquement de *S. thermophilus*, soit largement distribués au sein des firmicutes. Ce paramètre permet par exemple de distinguer deux gènes présentant le même nombre d'homologues retrouvés dans le même nombre d'espèce, mais des espèces de liens phylogénétiques variés par rapport à *S. thermophilus*. Ainsi, un gène présent chez *S. thermophilus* et *S. mutans* aura une valeur de dispersion moins élevée qu'un gène présent chez *S. thermophilus* et *Bacillus subtilis*.

La figure 34.A représente la valeur de dispersion de chaque gène « atypique » contre sa dispersion au sein des 47 souches de *S. thermophilus* (résultats de CGH). Pour repère, la valeur de dispersion d'un gène présent dans les 7 espèces du genre *Streptococcus* présents dans la base de données des firmicutes est de 0,0234, et celle d'un gène présent dans les 70 espèces est de 0,0697. Une dispersion de 0 indique que le gène est présent seulement chez *S. thermophilus*.

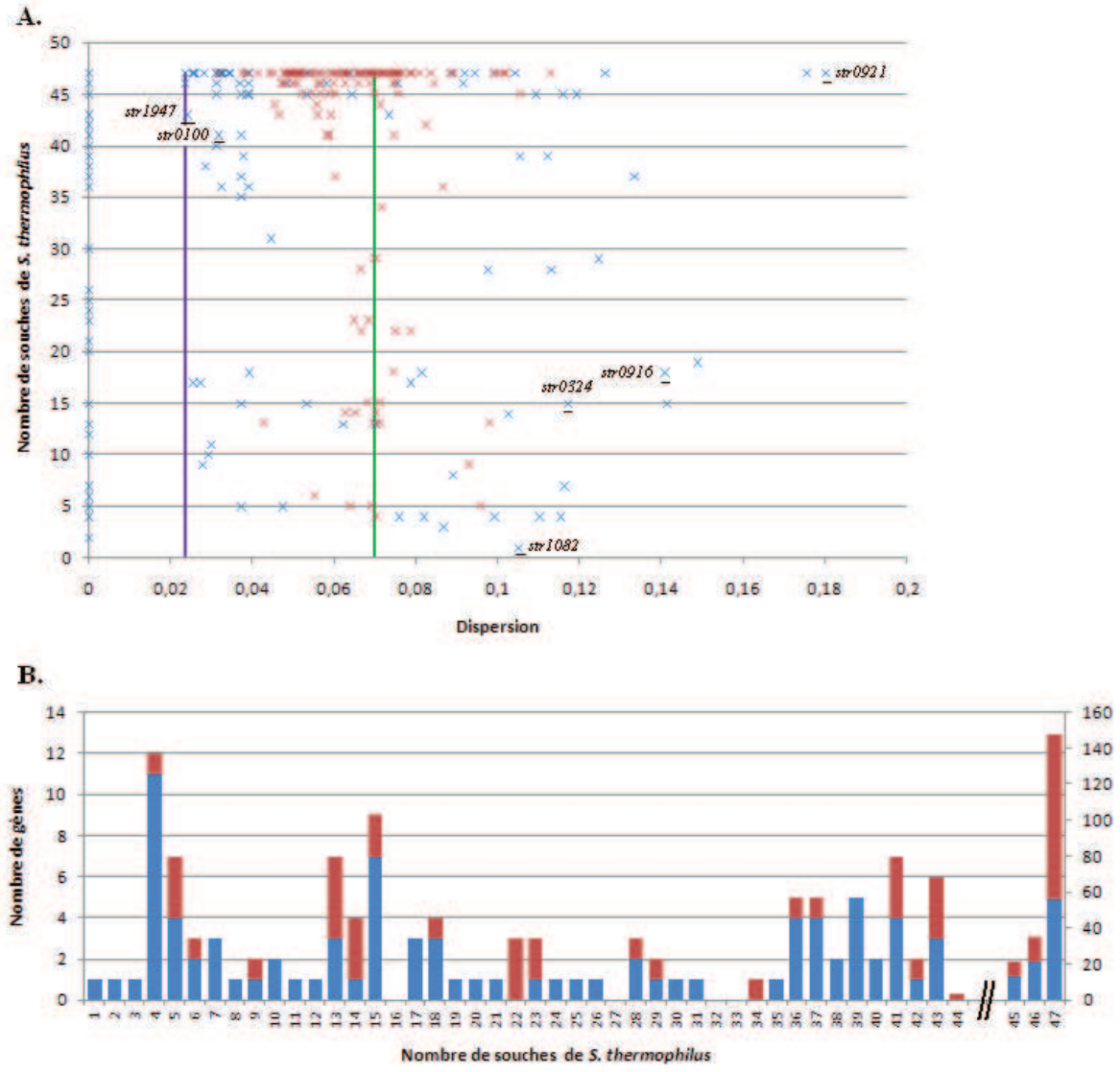


Figure 34 : Dispersion intra et inter spécifique des gènes « atypiques ».

A. Représentation graphique de la valeur de dispersion des 324 gènes « atypiques » en fonction de leur occurrence dans les souches de *S. thermophilus*. Les croix bleues symbolisent les gènes « atypiques » présents dans au plus 7 espèces au sein des firmicutes, les autres sont représentés par des croix rouges. Les croix bleues soulignées localisent les gènes dont la dispersion est commentée dans le texte. Le trait vertical violet représente la valeur de dispersion d'un gène présent dans les 7 espèces du genre *Streptococcus* (0,0234). Le trait vert symbolise la valeur de dispersion d'un gène présent dans les 70 espèces de firmicutes (0,0697).

B. Histogramme de la répartition des 324 gènes « atypiques » dans les 47 souches de *S. thermophilus*. Pour chaque barre, la partie bleue symbolise les gènes présents dans au plus 7 espèces et la partie rouge les gènes « atypiques » présents dans plus de 7 espèces.

Parmi les 189 gènes « atypiques » présents dans au plus 7 espèces de firmicutes (cf. section 3.10.3), 174 ont produit des données CGH exploitables et sont largement dispersés selon les deux axes de la figure (croix bleues). En revanche, les 150 gènes « atypiques » présents dans plus de 7 espèces de firmicutes (croix rouges) sont concentrés autour d'une valeur de dispersion de 0,0697 et sont majoritairement présents au sein des 47 souches de *S. thermophilus* (Figure 34.B). Cette valeur de dispersion de 0,0697 est proche de celle d'un gène présent dans toutes les espèces de firmicutes considérées dans notre analyse.

Parmi les gènes « atypiques » présents dans au plus 7 espèces de firmicutes qui présentent donc une distribution sporadique, le cas des gènes *str0921*, *str0916*, *str0324*, *str1082* est le plus extrême. Ils présentent une valeur de dispersion élevée (> 0,080). Ces gènes sont annotés respectivement comme codant :

- une activité agmatinase
- un régulateur transcriptionnel
- un ABC transporteur de peptides
- une glycosyl-transférase impliquée dans la biosynthèse d'exopolysaccharides.

Ces 4 gènes sont respectivement présents dans 47, 18, 15 et 1 des 47 souches de *S. thermophilus*.
A l'opposé de cette distribution, les gènes *str0100* et *str1947* ont une valeur de dispersion faible, respectivement de 0,0302 et 0,0230. Ces gènes sont annotés comme codant une protéine d'efflux de lantibiotique et un régulateur transcriptionnel de la famille MutR (*rgg*-like).

Dans la catégorie des gènes « atypiques » présents dans plus de 7 espèces de firmicutes, certains gènes, bien que présents dans plus de 7 espèces de firmicutes, présentent un indice de dispersion élevé, compatible avec l'hypothèse d'une acquisition récente. Par exemple, les gènes *str0264* et *str1084* (*epsF*) codent respectivement un transporteur de potassium et une protéine de biosynthèse d'exopolysaccharides. Le premier est présent dans les 47 souches de *S. thermophilus* et dans 14 espèces de firmicutes. Le second est présent dans 13 souches de *S. thermophilus* et dans 10 espèces de firmicutes. De plus les similarités les plus fortes obtenues pour *str1084* le sont avec les produits de gènes de firmicutes apparentées de façon très lointaine à *S. thermophilus* (dans l'ordre croissant de la e-value : *Pelotomaculum*, *Heliobacterium*, *Staphylococcus*, *Carboxydothermus*, *Lysinibacillus*, *Clostridium*, *Lactobacillus*, *Desulfotomaculum*, *Streptococcus pneumoniae*). Ceci est cohérent avec les fréquents événements de transferts horizontaux caractérisés pour ce type de locus.
Le gène *str1479* est un troisième exemple de gène de la seconde catégorie présentant une forte valeur d'indice de dispersion (0,095). Ce gène coderait une glycosyl-transférase impliquée dans la biosynthèse de composés pariétaux et serait présent dans 11 espèces de firmicutes diversement apparentées à *S. thermophilus*. Les identités de séquences obtenues permettent de distinguer deux niveaux d'identités : le premier avec des identités entre 43 % et 52 % et des valeurs d'e-value entre 10^{-75} et 10^{-92} ; le second est caractérisé par des valeurs d'identité autour de 30 % et d'e-value de 10^{-28}. Le premier niveau reflèterait une relation d'orthologie

alors que le second révèlerait une parologie. Ainsi, si le seuil d'identité permettant de conclure à la présence d'un gène chez une espèce donnée était relevé, ***str1479*** ne serait plus comptabilisé que chez ***S. sanguinis*** et ***S. suis***, et donc serait présent dans 3 espèces de firmicutes. L'indice de dispersion de gène chuterait fortement (0,030), et il ferait partie de la première catégorie (les gènes présents dans au plus 7 espèces de firmicutes) qui regroupent les gènes dont l'acquisition par transfert horizontal est probable. La présence de ce gène dans seulement 5 des 47 souches de ***S. thermophilus*** abonde dans ce sens. Ce cas illustre l'une des limites de l'approche concernant les familles de gènes. L'occurrence de gènes appartenant à une famille peut ainsi être surestimée. Le seuil d'identité nécessaire pour distinguer orthologues et paralogues dépendant de l'histoire évolutive de chaque séquence ; il ne peut donc être absolu et appliqué sans discernement à tous les gènes du génome.
A l'opposé de cette distribution, le gène ***str1668*** (alias ***cppA***) a une valeur de dispersion faible. Il code une protéine annotée « C3-degrading proteinase », activité présente chez ***S. pneumoniae*** où elle serait impliquée dans sa pathogénicité via la dégradation de la glycoprotéine C3 du complément, médiateur du processus d'inflammation locale chez l'hôte (Angel ***et al.***, 1994). Des homologues de ce gène sont identifiés dans les 7 espèces de streptocoques et ***Lactococcus lactis***, et les résultats de CGH semblent indiquer la présence de ce gène dans les 47 souches de ***S. thermophilus***. Ces éléments suggèrent une conservation depuis son acquisition datant de l'émergence du phylum des streptocoques ou peu avant. La présence de ce gène chez toutes les souches de ***S. thermophilus*** pose la question de la fonction de ce gène dans cette espèce non pathogène.
Ainsi, des cas de transferts horizontaux potentiels sont détectables même dans cette deuxième catégorie, mettant en évidence la stringence des paramètres appliqués dans l'analyse quantitative de dispersion.

En conclusion, l'indice de dispersion proposé dans notre analyse ne permet pas, seul, de valider l'hypothèse d'un transfert horizontal pour un gène « atypique » donné. Cependant, il permet de valider le caractère discriminant du seuil de 7 espèces appliqué pour étayer une hypothèse d'évolution verticale ou horizontale d'une séquence donnée. En outre, il est clair que le seuil de 7 est stringent, c'est-à-dire que si tous les gènes « atypiques » de la première catégorie (présents dans au plus 7 espèces) présentent un profil évolutif typique d'une acquisition par transfert horizontal, les gènes de la seconde catégorie ne doivent pas pour autant tous être considérés comme hérités de manière verticale. Ainsi les familles de gènes incluant, par exemple les gènes de synthèse des exopolysaccharides (glycosyl-transférases) ou encore les transporteurs (ABC transporteurs), présentent le plus souvent de nombreux paralogues au sein d'un même « pan » génome et leur analyse par notre méthode tend à les considérer comme largement distribués, et donc hérités de façon verticale. Il conviendrait donc pour ces gènes d'adapter le critère d'identité de séquences, de façon individuelle, pour conclure quant aux relations d'orthologie et de paralogie.

3.10.4 Gènes d'adaptation chez *S. thermophilus*

Parmi les 362 gènes « atypiques » identifiés dans le génome de *S. thermophilus*, 189 sont présents dans au plus 7 espèces de firmicutes. Parmi ces 189 gènes, 79 sont spécifiques de *S. thermophilus* (indice de dispersion 0) et présents entre 1 souche (CNRZ1066) et 47 souches de *S. thermophilus*. Parmi ces 79 gènes, 23 sont présents chez la totalité des souches testées. Ces 23 gènes pourraient définir un jeu de gènes d'adaptation de *S. thermophilus* au milieu lait. Afin de tester le caractère de « spécificité de niche » de ce jeu de gènes, l'espèce *Streptococcus salivarius* qui constitue l'espèce la plus étroitement apparentée à *S. thermophilus* a été introduite dans notre analyse. En effet, *S. salivarius* partage 99,6 % d'identité d'ARN 16S avec *S. thermophilus* alors que cette identité varie entre 93,7 % et 95,9 % pour les autres espèces de streptocoques dont le génome a été séquencé. Le génome de *S. salivarius*[12] a été rendu disponible récemment (NCBI, Avril 2009) et n'a de ce fait pas été intégré dans la base de données des firmicutes constituée pour ce travail. De plus, le génome de plusieurs autres espèces de firmicutes (*Streptococcus infantarius*, *S. gordonii*, *S. suis* 89/1591) a été introduit très récemment ou sous forme de shutgun dans les bases de données (respectivement 2008 2007, 2004). Ces génomes présentent des homologues pour 35 des 79 gènes (dont 15 des 23). Ainsi, l'hypothèse de spécificité d'espèce doit être restreinte aux 44 des 79 gènes pour lesquels soit aucune homologie n'est décelée (seuils appliqués : identité > 30 %, e-value < 10^{-6}) soit une homologie est décelée avec des espèces phylogénétiquement distantes (hors espèces appartenant au phylum des firmicutes). Parmi les 44 gènes spécifiques de l'espèce *S. thermophilus*, 8 sont présents dans la totalité des 47 souches de *S. thermophilus* et sont décrits dans le tableau 13.

Tableau 13 : Gènes « atypiques » et spécifiques de *S. thermophilus* et présents dans les 47 souches

CDS	Taille (aa)	Commentaire
str0091	130	Une homologie partielle est décelée avec la partie C-terminale d'une protéine de *S. infantarius*, *Blautia hansenii* et *Ruminococcus torques*. La protéine de *Blautia hansenii* est annotée comme codant le composant antitoxine d'un système toxine-antitoxine.
str0348	419	Cette CDS est annotée « peptidoglycan branched peptide synthesis protein, alanine adding enzyme ». Un homologue est identifié chez *Granulicatella elegans* (40 %, e-value = 10^{-84}), et une homologie partielle est décelée entre la partie C-terminale de *str0348* et une protéine de 183 acides aminés de *S. dysagalactiae*.
str1238	84	Cette CDS est spécifique et coderait une protéine de fonction inconnue. Une hélice transmembranaire est prédite.
str1388	82	Cette CDS et la précédente (*str1387*) constituent probablement un pseudogène : la région couvrant ces deux CDS présente une homologie avec un même gène de *Mesorhizobium opportunistum* (BlastX : 47 % d'identité protéique, e-value = 10^{-58})
str1391	119	CDS est spécifique.
str1709	78	CDS est spécifique.
str1743	142	CDS est spécifique. Trois hélices transmembranaires sont prédites.
str1746	63	Cette CDS est spécifique.

[12] Identifiant genbank : ACLO00000000

Parmi ces 8 gènes, le gène ***str0348*** est impliqué potentiellement dans la biosynthèse du peptidoglycane. La structure et la composition du peptidoglycane sont connues pour être variables d'une espèce à une autre. La spécificité de ce gène pourrait de ce fait être corrélée à celle de la composition de la paroi de *S. thermophilus*. En outre, parmi les 44 gènes spécifiques, on note le gène ***str1041***, lui aussi potentiellement impliqué dans la biosynthèse de la paroi, et le gène ***str1040*** qui le chevauche partiellement, suggérant fortement leur co-régulation. Le gène ***str0183*** compte lui aussi parmi ces 44 gènes et code une sérine-thréonine kinase putative. De façon intéressante, il appartient à une région « atypique » contenant également le gène ***str0182***, gène de la famille de régulateurs transcriptionnels *rgg*-like (famille MutR). Des analyses fonctionnelles réalisées au laboratoire indiquent que ce locus pourrait être impliqué dans la défense contre le stress oxydant et thermique (Fernandez ***et al.*** archives 2004 ; Henry R. et Leblond N., communication personnelle). Par ailleurs, un autre locus codant potentiellement un régulateur transcriptionnel *Rgg*-like présente sa plus forte homologie avec une protéine de *S. equi* (59 %, e-value 10^{-90}), alors que la meilleure homologie décelée avec les protéines codées par la souche LMD-9 n'est que de 29 % d'identité (e-value 10^{-14}), suggérant fortement qu'il s'agisse d'un paralogue. De plus, ce gène n'est présent que dans 6 des 47 souches, sur la base des données de CGH.
Enfin, on note, dans cette catégorie de 44 gènes, la présence de ***str1688*** codant la phéromone BlpC qui est impliquée chez la souche LMD-9 dans la régulation de la biosynthèse d'une bactériocine (Fontaine ***et al***., 2007). Le gène est présent chez les 3 souches dont le génome a été séquencé et il est également présent dans 40 des 47 souches analysées par CGH.

3.10.5 Ilots génomiques *S. thermophilus*

Se basant sur l'atypicité des régions (détection par le M2M2) ainsi que sur plusieurs autres caractéristiques dont (i) la composition en bases G et C, (ii) l'annotation fonctionnelle et (iii) les particularités structurales (présence de répétitions), des îlots génomiques se dégagent au sein de la séquence génomique de *S. thermophilus* CNRZ1066.
L'existence d'une séquence répétée de 139 nt a pu être mise en évidence. En totalité cette séquence répétée apparaît 10 fois dans le génome de la souche CNRZ1066. Dans huit des cas, ces occurrences sont localisées en amont d'une CDS annotée tnp*1239*. Dans les deux autres cas, ses répétitions sont en amont de CDS annotées « transposase tronquée ». Cette séquence pourrait donc faire partie de l'IS*1239*. Les gènes codant la transposase de ces éléments mobiles ne sont pas décelés par le M2M2. En revanche, il apparaît que la séquence répétée de 139 nt borne une région « atypique » dans 7 des 10 cas (figure 35). C'est le cas notamment :

- d'une région « atypique » couvrant les CDS ***str0097*** à ***str0104*** dont le pourcentage en GC est de 30,3 % et qui correspond à un IME (Integrative and Mobilizable Element) intégré au sein du gène *rpsI* contenant un locus de biosynthèse et d'efflux d'une bactériocine de type lantibiotique (figure 35.A),
- d'une région « atypique » couvrant les CDS ***str0678*** à ***str0690*** dont le pourcentage en GC est de 29,3 % (figure 35.B),

- d'une région « atypique » couvrant les CDS ***str1040*** à ***str1044*** et partiellement les deux gènes flanquants codant des transposases (figure 35.C). Cette région présente un pourcentage en bases G et C de 30 %. Le gène ***str1041*** serait impliqué dans la biosynthèse de la paroi. Les gènes ***str1040***, ***str1041*** et ***str1042*** sont spécifiques de *S. thermophilus*. Le gène ***str1043*** annoté tyrosyl-tRNA synthétase (***tyrSE***) serait un paralogue du gène ***tyrS***. Enfin, le gène ***str1044*** appartient à la famille ***rgg***-like. Il est présent dans 9 des 47 souches et il est absent de la souche LMD-9,
- d'une région « atypique » ne contenant que le gène ***str1668*** annoté ***cppA*** et codant potentiellement la protéase impliquée dans la dégradation de la glycoprotéine C3 du complément (figure 35.D). Il est à noter que ce gène présente un pourcentage en GC de 37,8 % proche de la composition moyenne du génome.

Une seconde séquence répétée de 105 nt a pu être identifiée en orientation directe, aux deux bornes de la région « atypique » contenant les CDS ***str1386*** à ***str1391***. De façon surprenante cette répétition est présente dans les régions codantes les plus distales. Le gène ***str1389*** est supposé coder une protéine d'efflux d'acides aminés (figure 36.A).
La région codant les exopolysaccharides (locus ***eps***) est très largement décelée par le M2M2 (figure 36.B). Ce locus a été décrit dans la littérature comme étant fortement sujet au transfert horizontal. Les trois régions « atypiques » recouvrant ce locus présentent un pourcentage en GC allant de 31,2 % à 36,5 %.

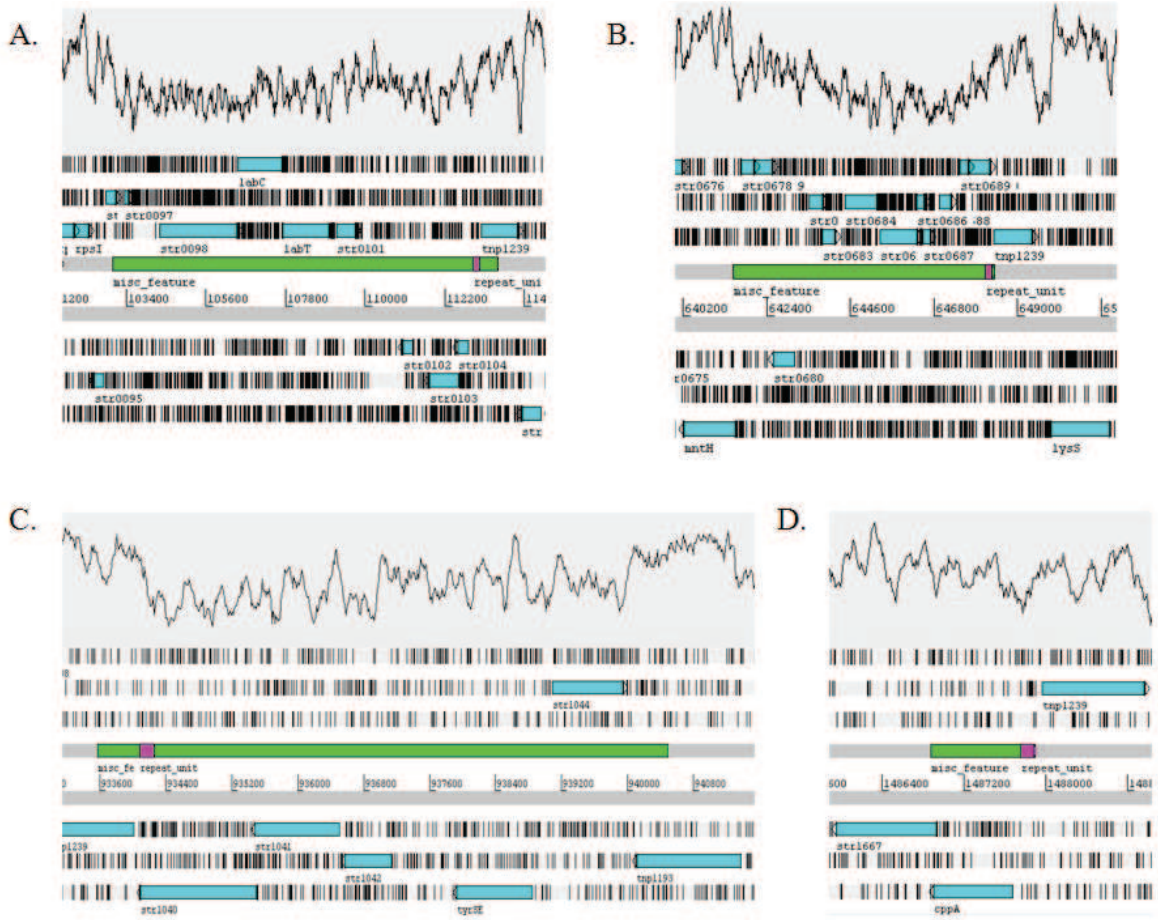


Figure 35 : Visualisation sous ARTEMIS de quatre régions « atypiques » bornées par une séquence répétée de 139 nt.

Le graphique en haut de chaque figure correspond à l'évolution du pourcentage en bases G et C le long de la région visualisée.

La partie inférieure de chaque figure indique l'annotation de la région où les rectangles bleus représentent les CDS sur les deux brins et les rectangles verts définissent les régions « atypiques ».

Les rectangles violets localisent la séquence répétée de 139 nt dans chaque région « atypique ».



Figure 36 : Visualisation sous ARTEMIS de deux régions « atypiques ».

Le graphique en haut de chaque figure correspond à l'évolution du pourcentage en bases G et C le long de la région visualisée.

La partie inférieure de chaque figure indique l'annotation de la région où les rectangles bleus représentent les CDS sur les deux brins et les rectangles verts définissent les régions « atypiques ».

A. Le rectangle jaune signale la séquence répétée de 105 nt dans la région « atypique » comprenant les CDS *str1386* à *str1391*.

B. Les trois régions « atypiques » recouvrent le locus *eps*, ils ont un pourcentage en bases G et C respectivement de 31,2 %, 34,5 % et 36,5 %.

Conclusion

3.11 CDS « atypiques » de petite taille et pseudogènes

Les 362 gènes « atypiques » identifiés par le M2M2 présentent une taille moyenne inférieure à celle des gènes de l'ensemble du génome. En effet, les 362 gènes ont une longueur moyenne de 613 pb contre 788 pb pour le reste du génome (retranché des 362 gènes décrits par le M2M2). Cette différence est notablement due à la présence de troncatures dans les gènes décelés par le M2M2. Il apparaît en effet que le M2M2 décèle préférentiellement des régions contenant des gènes de petite taille. Parmi les 79 gènes spécifiques de *S. thermophilus*, 21 sont annotés comme étant tronqués. Si l'on déduit ces troncatures de l'effectif, la taille moyenne passe de 542 pb à 635 bp. L'analyse au cas par cas des 79 gènes révèle également que pour 16 gènes additionnels, soit une erreur de détermination de la séquence, soit une erreur de son interprétation (par exemple, mauvaise identification du brin codant) mène à l'annotation de petites CDS. Dans plusieurs cas, l'introduction d'une insertion/délétion unique dans la séquence de la souche CNRZ1066 suffit à restaurer une CDS entière, apparaissant initialement en plusieurs fragments dans différentes phases de lecture. Dans d'autres cas, le M2M2 décèle de petites ORF consécutives qui se révèlent être les segments successifs d'une même ORF annotée chez une souche voisine. Lorsque cette CDS plus longue est présente chez les souches voisines, il est possible d'imaginer qu'une erreur de séquence se soit glissée dans le génome de la souche CNRZ1066. Ce type d'artéfact mènerait à la fois à surestimer le nombre de gènes décelés par le M2M2 et à sous-estimer leur taille. En revanche, lorsque la troncature est également pronostiquée dans une autre souche au moins, la présence d'un pseudogène est validée avec une sécurité plus grande. Le second type d'erreur concerne l'interprétation. Dans l'annotation génomique de la souche CNRZ1066 (Bolotin *et al.*, 2004), certaines ORF ont été annotées alors qu'elles n'ont pas été prises en compte dans les génomes des autres souches séquencées bien que la séquence correspondante soit présente. Dans ce cas de figure, la recherche d'homologues se faisant par BlastP, l'ORF est alors considérée comme spécifique. Ces ORF peuvent donc soit ne pas être fonctionnelles (et n'auraient pas dû être annotées dans le génome de CNRZ1066) soit être fonctionnelles et faussement considérées comme spécifiques par omission dans les autres génomes.

Par conséquent, si l'on retranche ces 16 CDS additionnelles, la taille moyenne des gènes décelés remonte à 744 pb. Outre ces artéfacts, il est à noter que les gènes de phages et que les gènes codant des recombinases (transposases d'IS) sont de taille inférieure à la taille moyenne des gènes de *S. thermophilus*.

L'enrichissement présumé en gènes exogènes et de mobilité au sein du jeu de 362 gènes décelés par le M2M2 pourrait influencer à la baisse, la taille moyenne des gènes de cet ensemble. Ainsi, le mécanisme d'internalisation et d'insertion de l'ADN exogène dans le génome de l'hôte pourrait lui-aussi participer à ce biais. Alors que la conjugaison bactérienne chez *E. coli* conduit à l'intégration de longs fragments génomiques (allant jusqu'à plusieurs centaines de kilobases) (Lloyd et Buckman, 1995), les mécanismes de transduction et de transformation naturelle pourraient, quant à eux, être plus limitants. En effet, la taille de

l'ADN transféré par transduction est le plus souvent très voisine de la taille du génome du bactériophage. De même, la transformation naturelle promeut le transfert de segment d'ADN de taille réduite à quelques kilobases (Claverys *et al.*, 2009).
Enfin, si la séquence exogène se maintient à court terme mais ne contribue pas à l'adaptation de l'organisme, elle ne subit plus de pression de sélection et accumule des mutations conduisant à la constitution d'un pseudogène. Ce dernier sera probablement, à terme, éliminé du génome par recombinaison. Les pseudogènes sont probablement plus fréquents dans des séquences d'origine exogène que dans des séquences endogènes. Ceci pourrait expliquer pour part l'enrichissement en pseudogènes du jeu de 362 gènes « atypiques ».
On peut également noter que le M2M2 a détecté 10 régions intergéniques de grande taille (*circa* 1 kb). Ces régions pourraient avoir été détectées en raison de la pauvreté en bases G et C des régions intergéniques comparativement aux régions codantes. Il peut également être envisagé qu'il ne s'agisse pas réellement de séquence intergénique du fait de la présence d'un gène non annoté (tronqué ou non). C'est le cas de la région « atypique » localisée entre les positions 1 031 882 et 1 033 116 où une CDS de 172 acides aminés peut être identifiée et présentant une identité de 94 % et 90 % avec des CDS de taille équivalente annotées chez les deux autres souches de *S. thermophilus*. Une recherche d'identité dans la base de données non redondante (*nr*) révèle qu'il s'agit d'une CDS spécifique de cette espèce.

3.12 Enrichissement en bases A et T des séquences « atypiques »

Les régions génomiques et gènes décelés par le M2M2 montrent un biais compositionnel important avec un pourcentage en GC faible. Les 362 gènes « atypiques » ont un pourcentage en bases GC moyen de 34,7 % contre 39,08 % pour l'ensemble des CDS du génome. Le principe du M2M2 est de fragmenter le génome en régions de composition distincte. Ici, ce sont des mots de trois lettres qui sont analysés. Il est évident qu'un biais compositionnel comme un enrichissement en bases G et C a un impact direct sur le biais de trinucléotides. Il n'est donc pas surprenant que le M2M2 « réagisse » à la lecture d'une région plus pauvre en G et C que la moyenne du génome. Par ailleurs, il a été constaté que les séquences d'éléments génétiques mobiles (plasmides, transposons, IS, phages, etc.) susceptibles de participer au flux de gènes présentent une composition en pourcentage en bases G et C inférieure de 1 % à 4 % au pourcentage du génome hôte (Rocha et Danchin, 2002). Il a été proposé récemment (Navarre *et al.*, 2007 ; Baños *et al.*, 2009) que le faible pourcentage en bases G et C des séquences acquises permettrait le contrôle différentiel de leur expression par la protéine histone-like H-NS comparativement aux gènes endogènes, évitant ainsi une contre sélection de cette séquence suite à son expression délétère. Les protéines de liaison à l'ADN (H-NS) sont très abondantes et sont impliquées dans la structuration et l'organisation des génomes bactériens en se liant à des séquences riches en A et T telles que les séquences issues du HGT (figure 37). Dans un premier temps, ces protéines inhiberaient l'expression des gènes exogènes et assureraient ainsi, une tolérance de l'organisme hôte vis-à-vis de ces dernières, laissant le temps aux processus d'amélioration d'opérer.

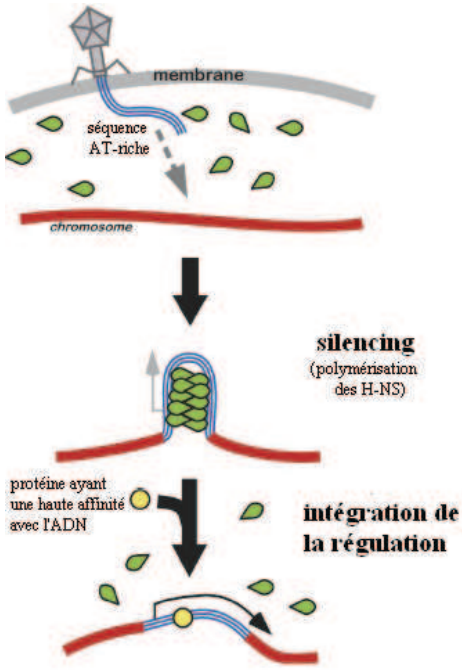


Figure 37 : Modèle de tolérance des gènes exogènes facilitée par les protéines H-NS (Navarre *et al.*, 2007).

Après intégration d'un fragment d'ADN exogène (ici par transduction), les protéines H-NS se lient à l'ADN empêchant l'expression de l'information contenue. Les protéines H-NS sont en compétition avec d'autres protéines qui pourraient favoriser l'expression de l'information exogène.

Ainsi, le faible pourcentage en bases G et C des gènes décelés par le M2M2 est en faveur d'un enrichissement en séquences exogènes. Toutefois, nous ne pouvons exclure que pour des raisons locales et/ou structurales des séquences endogènes présentent un biais compositionnel et génèrent de faux positifs.

Notons que parmi les 362 gènes « atypiques », 151 gènes ne présentent pas une composition significativement différente de la moyenne génomique. Le M2M2 n'est donc pas sensible qu'au seul enrichissement en nucléotides A et T mais probablement aussi aux biais compositionnels d'ordres supérieurs (di, tri, oligo-nucléotides).

3.13 Validation du M2M2 comme méthode d'identification de gènes issus du transfert horizontal

Les modèles de Markov d'ordre 2 et le processus d'extraction automatique mis en place dans ce travail ont permis d'identifier 362 gènes « atypiques » dans le génome de *S. thermophilus* CNRZ1066 soit 18,9 % de l'ensemble des CDS du génome. L'analyse de l'annotation de ces gènes, de leur distribution au sein de l'espèce et plus largement du phylum bactérien des firmicutes a permis de proposer une origine exogène pour au moins 240 de ces gènes (figure 38).

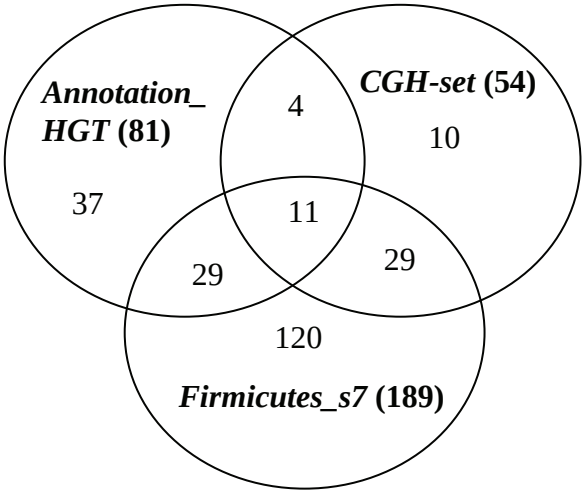


Figure 38 : Diagramme de Venn résumant l'analyse des gènes extraits par le M2M2 sur la base des critères de l'annotation (*annotation HGT*), de la variabilité au sein de l'espèce (*CGH-set*) et de la distribution au sein des espèces de firmicutes (*Firmicutes_s7*).

Ainsi, 54 gènes « atypiques » sont présents dans au plus 15 des 47 souches de *S. thermophilus* analysées en CGH. Ces gènes considérés comme variables au sein de l'espèce *S. thermophilus* ont plus probablement été acquis récemment (depuis l'émergence de l'espèce *S. thermophilus*) par un nombre d'espèces faible plutôt que perdus dans les nombreuses autres. Bien entendu, cette analyse est quantitative car elle ne tient pas compte des distances évolutives entre souches. Un facteur limitant l'impact de cette analyse est que l'on ignore si les souches analysées par CGH sont représentatives de la diversité de l'espèce. En effet, nous ne connaissons pas l'origine des souches retenues par notre partenaire dans cette étude, Chr. Hansen S/A. Parmi ces 54 gènes, 22 codent des protéines de fonction hypothétique et 7 correspondent à des protéines tronquées (Annexe Web E). Au sein de l'ensemble des gènes annotés, on compte 9 des 12 gènes *eps* codant la biosynthèse des expolysaccharides (locus *eps*) et 2 des 7 gènes de la famille *rgg*-like. Les autres gènes annotés sont, soit impliqués dans un système de restriction-modification (2), de transport (5), du métabolisme des sucres (3) et de biosynthèse de lantibiotique (1), soit apparentés à des protéines d'origine phagique (3).

Au sein des 362 gènes « atypiques », le critère de l'annotation fonctionnelle permet de distinguer une catégorie de 81 gènes « atypiques » (*Annotation_HGT*) dont l'annotation est compatible avec une fonction de contingence susceptible d'acquisition par transfert horizontal (tableau 12). Parmi eux, des gènes de résistance aux antibiotiques, des gènes codant des systèmes de restriction-modification, des éléments génétiques mobiles (incluant des IS, des bactériophages) ou encore des gènes de biosynthèse d'exopolysaccharides. Par ailleurs, l'analyse de la distribution de ces 362 gènes « atypiques » au sein des espèces de firmicutes, et sur la base de la présence d'un gène donné dans au plus 7 espèces de notre échantillon, a permis d'identifier 189 gènes dont la distribution est plus compatible avec un scénario évolutif de type perte/gain que de type transmission verticale. Ce seuil est par ailleurs en cohérence avec le nombre d'espèce de streptocoques présents dans notre échantillon

d'espèces de firmicutes. Parmi l'ensemble *Annotation_HGT*, 46 % (37 sur 81) ne sont distingués que sur la base de leur annotation. Cette proportion importante s'explique par le fait que parmi les 81 gènes, 41 ont une annotation correspondant à des transposases d'*IS* . Or ces séquences n'ont pas été utilisées comme sondes dans les expériences de CGH et leur distribution au sein de l'espèce n'a donc pas pu être étudiée. Enfin, les IS sont des éléments très largement répandus dans les génomes bactériens, et présentent des identités plus ou moins fortes de séquences protéiques entre leurs transposases. Aussi, comme pour les familles de gènes, les analyses menées sur les génomes de firmicutes à l'aide des seuils appliqués (identité et d'e-value d'une part, et nombre d'espèces dans les firmicutes d'autre part) n'ont pas permis de distinguer une distribution particulière de ces éléments mobiles.
Au final, onze gènes rencontrent les 3 critères de notre analyse (tableau 14).

Tableau 14 : Gènes « atypiques » rencontrant les 3 critères de l'analyse

CDS	Fonction putative de la protéine déduite
str0098	Protéine de biosynthèse de lantibiotique, tronquée
str0689	Protéine régulatrice d'un système de restriction-modification, putative
str0708	Sous unité d'un système de restriction-modification de type I
str0780	Protéine inconnue, origine phagique
str0781	Protéine de liaison à l'ADN d'origine phagique
str0782	Protéine inconnue, origine phagique
str1076	Protéine de biosynthèse d'exopolysaccharide
str1077	Protéine de biosynthèse d'exopolysaccharide
str1078	Protéine de biosynthèse d'exopolysaccharide
str1079	Protéine de biosynthèse d'exopolysaccharide
str1082	Protéine de biosynthèse d'exopolysaccharide

L'introduction d'un indice de dispersion a permis de montrer que les 189 gènes présentaient des valeurs de dispersion cohérentes avec une distribution sporadique, mais également que certains des 173 gènes restants (donc présents dans plus de 7 espèces de firmicutes) présentent néanmoins une distribution sporadique compatible avec une acquisition par transfert horizontal. Le seuil appliqué s'avère être strict mais assez significatif pour extraire des corrélations. L'application de ce seuil a permis d'identifier une dispersion sporadique des gènes codant les régulateurs de type Rgg. En effet, l'annotation fonctionnelle les classe comme des régulateurs de la famille MutR, ce qui ne suggère pas en soi que ces gènes soient sujets aux transferts horizontaux. De même, la distribution au sein des souches de *S. thermophilus* n'est pas non plus révélatrice puisque seuls deux sont présents dans le *CGH-set* et les 5 autres sont présents chez un nombre de souches très variable (entre 17 et 47). En fait, seuls deux, *str1511* et *str1044*, sont décelés par l'approche CGH étant respectivement présents dans 6 et 9 souches. La distribution au sein des firmicutes met en évidence que les 7 gènes *rgg*-like présentent une distribution sporadique (présents dans au plus 7 espèces de firmicutes) compatible avec une acquisition par transfert horizontal. L'analyse compositionnelle de ces gènes étaye également cette hypothèse puisqu'ils présentent des pourcentages en bases G et C variables et inférieurs à la moyenne du génome (de 26,5 % à 32,2 %). C'est vraisemblablement cette composition biaisée qui permet au modèle de déceler les 7 copies de gène *rgg*.

Il est intéressant de noter que la détection à l'aide des M2M2 est déterminante dans l'identification de certains gènes vraisemblablement issus d'événements de transfert, c'est-à-dire qu'ils ne le sont par aucun des autres critères utilisés ici. Par exemple, Bolotin et ses collègues (2004) ont décrit une région de 17 kb du génome de *S. thermophilus* contenant 20 gènes putatifs comme étant « exogènes » sur la base d'identités fortes avec des séquences de *L. lactis* et *L. bulgaricus*. Le M2M2 décrit 7 de ces 20 gènes (*str0837*, *str0838*, *str0841*, *str0855-0858*), et permet de révéler cette région (figure 30). Or :

- l'annotation fonctionnelle de ces gènes ne motive aucune hypothèse d'acquisition par transfert horizontal,
- seuls 3 de ces 20 gènes « atypiques » sont révélés par l'analyse de distribution chez les firmicutes,
- cette région est présente chez la plupart des 47 souches de *S. thermophilus* rendant sa détection impossible par l'analyse CGH.

En fait, l'acquisition de cette région est probablement intervenue de façon antérieure à l'émergence de l'ancêtre de *S. thermophilus*. Enfin comme précédemment avec les *IS*, les fonctions putatives codées par cette région (protéine de réponse au stress froid, perméase, peptidase, système de restriction-modification) sont retrouvées de très nombreuses espèces de firmicutes, et ces gènes présentent des homologues forts dans de nombreux génomes séquencés, même s'il ne s'agit que de paralogues.

3.14 Scénario évolutif du génome de *S. thermophilus*

L'espèce *S. thermophilus* est considérée comme une espèce clonale eu égard à son émergence récente, estimée entre 3 000 et 30 000 ans, concomitante du développement des pratiques agriculturales (Fox, 1993). Malgré cette émergence récente, le transfert horizontal est avéré chez cette espèce comme en témoignent les analyses *in silico* qui prédisent des transferts dans l'écosystème laitier et les approches *in vivo* qui identifient les mécanismes liés aux bactériophages et aux éléments conjugatifs (Bellanger *et al.*, 2009 ; Ammann *et al.*, 2008 ; Liu *et al.*, 2009). Nos analyses confortent le schéma évolutif de cet organisme. En effet notre méthode permet de prédire qu'au minimum 12,5 % (240) des gènes de *S. thermophilus* CNRZ1066 auraient une origine exogène. Il a en outre été rapporté au cours de l'annotation du génome qu'environ 10 % des séquences correspondaient à des pseudogènes (Bolotin *et al.*, 2004). Ainsi, l'émergence récente s'accompagne d'un processus d'évolution génomique rapide avec accumulation de mutations dans des gènes qui subissent une pression moindre dans le nouvel environnement, et par ailleurs une acquisition de fonctions nouvelles permettant d'exploiter au mieux ce nouvel environnement. Ainsi des fonctions phagiques, des fonctions de mobilité (par exemple, des transposases), des mécanismes de défenses (système restriction-modification, système de résistance aux bactériophages de type CRISPR), et des mécanismes de résistance aux antibiotiques ont pu être identifiés grâce à la stratégie employant les modèles de Markov cachés d'ordre 2 développée dans cette étude. De même, des gènes codant des fonctions identifiées, mais de contribution encore incertaine dans le

contexte environnemental, ont été décrits (gènes codant des exopolysaccharides, des ABC transporteurs ou des régulateurs transcriptionnels tels que Rgg).
Notre approche couplée à l'analyse de la distribution des gènes « atypiques » dans la phylogénie du phylum de firmicutes a permis d'identifier un sous ensemble de 79 gènes spécifiques de *S. thermophilus* dont 23 sont présents chez toutes les souches testées (données de CGH obtenues au travers d'une collaboration avec le groupe Chr. HANSEN S/A). Huit de ces gènes sont considérés comme spécifiques de la niche écologique « lait ». L'analyse a néanmoins été approfondie par la recherche de domaine fonctionnel conservé au travers de la base *Pfam*. Bien que la plupart des motifs ne présentent pas de score significatif, et avec toutes les précautions d'usage, 5 des 8 gènes pourraient participer à diverses fonctions de transport (peptide, malonate ou potassium), de synthèse de bactériocines, de biosynthèse de la paroi ou encore de système de restriction. Il a été démontré, par exemple chez *L. delbrueckii*, qu'une région interprétée comme issue du transfert horizontal codait un ABC transporteur impliqué dans le développement de la proto-coopération avec *S. thermophilus* (Guchte *et al.*, 2006). Ainsi, ces fonctions pourraient permettre une adaptation de *S. thermophilus* au milieu lait par exemple en permettant une utilisation optimisée des ressources nutritives disponibles dans le lait.

Le schéma évolutif de *S. thermophilus* apparaît donc plus clairement grâce à notre analyse. Les données génomiques montrent que le génome « core » est extrêmement conservé entre souches traduisant l'émergence récente de *S. thermophilus* à partir de *S. salivarius* (Schleifer *et al.*, 1991 ; Facklam, 2002). A côté de ce génome hérité verticalement, une proportion importante du génome subit une réduction génomique révélée par la formation de pseudogènes voués à l'élimination par contre-sélection et/ou recombinaison. Enfin, l'adaptation aux conditions biotiques (présence des organismes du milieu lait) comme abiotiques est promue par les nombreux gènes issus du transfert horizontal.

Chapitre 4

Discussion et perspectives

4.1 FOUILLE DE DONNÉES POUR LA RECHERCHE DES SITES DE RÉGULATION 111
4.1.1 Avantages de la méthode : grande flexibilité de recherche 112
4.1.2 Amélioration de la stratégie 115
4.2 FOUILLE DE DONNÉES DÉVELOPPÉE POUR L'ANALYSE DES GÈNES « ATYPIQUES » 117
4.2.1 Méthodes HMM vs. autres méthodologies de détection du HGT 117
4.2.2 Distribution des gènes dans un arbre phylogénétique : valeur de dispersion 119
4.3 LES HMM DU SECOND ORDRE POUR LA FOUILLE DE DONNÉES GÉNOMIQUES 121
4.4 DÉFINITION AUTOMATIQUE D'UNE TOPOLOGIE DE HMM D'ORDRE 2 123

Les deux stratégies fondées sur les HMM d'ordre 2 se révèlent efficaces pour l'identification de signaux de régulation et dans la détection de gènes atypiques. Nous allons exposer et discuter dans les deux premières parties, des avantages des analyses comparées à d'autres méthodes, des limites et des améliorations propres à chacune des stratégies employées. Dans une troisième partie, nous allons discuter des avantages et des inconvénients de l'utilisation des HMM d'ordre 2 dans l'analyse de séquences génomiques par une fouille de données. Enfin, la recherche de TFBS ou de gènes « atypiques », a nécessité la recherche expérimentale d'une topologie de HMM adéquate. En outre, nous proposons d'explorer une possibilité de recherche automatique d'une topologie de HMM.

4.1 Fouille de données pour la recherche des sites de régulation

L'originalité de la fouille est d'associer deux méthodes stochastiques (HMM2 et *R'MES*) pour l'identification de motifs et une méthode de classification pour regrouper les motifs identifiés afin de reconstruire des motifs composites robustes. La fouille de données sur le génome de *Streptomyces coelicolor* a permis de déceler des sites de fixation déjà caractérisés expérimentalement (σ^R, σ^{whiG}, σ^B) et de proposer de nouveaux SFBS putatifs. Les paramètres de la phase de recherche de motifs composites étaient ajustés à l'aide d'une étude effectuée sur le régulon σ^R (cf. section 2.3). Ces paramètres ont permis de cibler notre recherche, de comparer les différents modèles et de réduire le temps de calcul.

Ainsi, nous allons aborder dans un premier temps, les avantages de la stratégie d'analyse mise en œuvre et dans un second temps, nous allons suggérer des pistes d'optimisation de la stratégie.

4.1.1 Avantages de la méthode : grande flexibilité de recherche

Recherche de motifs rares (sous-représentés)

Durant l'étape de recherche de motifs composites SFBS, *R'MES* identifie des motifs exceptionnels et dans notre cas, nous avons sélectionné les motifs sur-représentés. Toutefois, il est possible d'extraire les motifs exceptionnellement sous représentés. Cette analyse pourrait révéler des motifs très peu fréquents comme cela est le cas pour le SFBS du facteur σ^{bldN}. Un *iPeak-motif* a été retrouvé dans la région promotrice du gène annoté *bldM* (SCO4768), ce dernier correspondant à une boîte définie par le SFBS du facteur σ^{bldN} (CGTAAC-N_{16}-CGTTGA). Le gène *bldM* est le seul gène décrit comme cible de ce facteur sigma (Bibb *et al.*, 2000). Cette donnée est étayée par l'absence du même motif consensus SFBS dans les régions intergéniques du génome.

Variation / ajustement sur la longueur du séparateur ou sur celle de la box2

Dans notre stratégie de modélisation, la reconstruction des motifs composites (*box1-séparateur-box2*) s'effectue en utilisant *R'MES*. Les paramètres de recherche sont donnés par l'utilisateur telles que la longueur du séparateur ou celle de la *box2*. Des paramètres stringents permettent de réduire l'espace de recherche et donc le temps de calcul lors de la recherche de motifs composites (Tableau 15). En effet, dans le premier cas où un motif recherché est plus ou moins caractérisé (intervalle de longueur, conservation d'un motif d'une boîte par rapport à l'autre, intervalle de longueur de la *box2*), il est judicieux d'utiliser des paramètres ajustés aux données.

Tableau 15 : Temps de calcul d'une recherche d'un motif composite à partir d'un consensus de classe

Longueur du séparateur	Longueur box2 ($l_{boîte10}$)	Temps de calcul en seconde (s)
[3-7]	[3-5]	29
[3-25]	[3-5]	119 (1 min 59s)
[3-25]	[3-7]	1000 (1h 10 min 40s)

Exemple de longueur de la box1 (l_{box1}) = 8 (cACCGCTt).

Dans le second cas où le motif recherché n'est pas bien caractérisé, l'exploration est possible avec des paramètres étendus. Cette exploration « à l'aveugle » est longue et fastidieuse du fait du grand nombre de possibilités à traiter. Pour effectuer une exploration optimale, il est recommandé de choisir le sens de recherche de la *box2* (figure 39). Dans l'analyse du génome de *S. coelicolor*, les motifs issus de la classification ont été considérés comme des *box1* correspondant aux motifs de la boîte -35. La procédure de reconstitution a recherché les *box2* représentant des motifs de la boîte -10 (figure 39, cas 1). Cette analyse a été à sens unique car

l'observation du consensus du SFBS de σ^R (GGAAT-N_{18}-GTT) caractérisé expérimentalement a révélé que le motif conservé de la boîte -35 était très conservé et de grande taille comparé à celui de la boîte -10. En revanche, dans l'analyse du génome de *B. subtilis*, nous avons pu considérer les *box1*, soit comme des motifs de la boîte -35, soit comme des motifs de la boîte -10 pour réaliser la recherche de la *box2* (figure 39, cas 1 et cas 2).

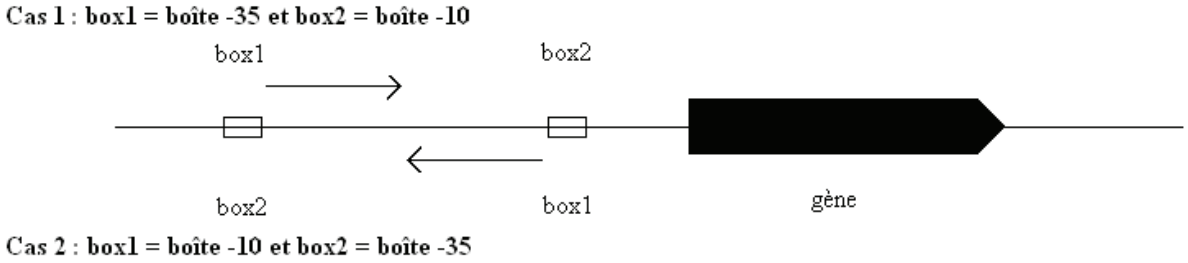


Figure 39 : Schéma d'extension des *iPeak-motifs*.
Les flèches représentent le sens de l'extension. Dans les deux cas, la *box1* (boîte -35 ou boîte -10) est fixée et la recherche de la *box2* (boîte -10 ou boîte -35) est réalisée dans les 50 pb adjacentes.

Possibilité de détecter des motifs différents des motifs composites SFBS

Possibilité de rechercher la boîte -10 étendue.

Le facteur σ^{70} est composé de quatre régions formées de sous-régions (figure 12) (Lonetto *et al.*, 1992). La région 2 ou plus précisément la région 2.4 interagit avec le consensus de la boîte -10. Une région localisée entre la région 2 et 3, notée 2.5 ou 3.0, pourrait interagir avec le motif 5'-GT-3' en amont de la boîte -10, appelée boîte -10 étendue. Une analyse des boîtes -10 étendues chez *B. subtilis* a été réalisée sur 236 promoteurs reconnus par σ^A. L'alignement de ces 236 promoteurs a révélé un motif conservé 5'-TRTG-3' en amont de la boîte -10 avec les nucléotides 5'-TG-3' (-15/-14) très conservés et les nucléotides 5'-TR-3' peu conservés (-17/-16) (Helmann *et al.*, 1995). La boîte -10 étendue contribuerait à l'activité optimale de nombreux promoteurs. Néanmoins, la région 3.0 des facteurs sigma est peu conservée voire absente chez certains facteurs sigma (Barne *et al.*, 1997).
Dans notre stratégie, la reconnaissance des boîtes -10 étendues serait possible et serait réalisée par le processus d'extraction des motifs complété par l'algorithme de classification. L'étape de recherche de motifs composites est alors optionnelle.

Possibilité de rechercher des sites de régulation autres que SFBS.

Notre stratégie appliquée à l'analyse de *S. coelicolor* a abouti à 812 motifs SFBS potentiels avec des paramètres de recherche limités (score *R'MES* supérieur à 6, valeur du séparateur entre 14 et 20 et la longueur de la *box2* comprise entre 3 et 5). Cette analyse a permis d'identifier des sites de fixation SFBS déjà bien caractérisés et des potentiels sites de régulation notamment PTF2, site de régulation proposé Studholme et ses collègues (2004).
L'ajustement des séparateurs permettrait d'identifier d'autres sites de régulation de facteur de transcription ; par exemple, le facteur de transcription CAP (Catabolism Activation Protein)

de classe II qui reconnait le consensus (5'-**TGTGA**TCTAGA**TCACA**-3') chez *E. coli*. Ce même site est aussi reconnu par un autre facteur de transcription de classe II, Crp chez *E. coli* (Sawers *et al.*, 1997 ; Busby et Ebright, 1999). Un homologue Crp est retrouvé chez *S. coelicolor* mais son TFBS n'a pas été identifié (Derouaux *et al.*, 2004). Dans notre analyse, parmi les 360 classes de consensus, 7 consensus de classes présentent une partie du consensus (tableau 16). Une recherche des motifs composites avec un séparateur compris entre 5 et 7 permettrait d'identifier des gènes régulés par ces facteurs de transcriptions.
De même, la reconnaissance des régions régulatrices inductibles lors de dommages de l'ADN serait possible en modulant la longueur du séparateur. Ahel et ses collègues (2002) ont déterminé une variation du promoteur RecA, induit lors de dommages aux UV à l'ADN, chez *Streptomyces* (5'-TTGTCAGTGGC-N_6-TAGggT-3'). Au travers de notre analyse chez *S. coelicolor*, 6 consensus de classe contiennent une partie de ces boîtes (tableau 16). Une recherche des motifs composites avec un séparateur compris entre 6 et 12 permettrait d'identifier des motifs complets dans les régions promotrices de *S. coelicolor* qui seraient induites lors de dommages aux UV.

Tableau 16 : Consensus de classes correspondant à des motifs régulateurs

	Motif	Classes	R'MES
Motif1	CRP (5'-**TCACA-3'**)		
	cACAAa	Class44_55	5.5233
	CACATT	Class118_64	8.2807
	CACATa	Class337_65	8.2043
	GA**TCA**	Class132_35	5.5097
	Dommage UV (5'-**TAGggT-3')**		
	aGGGTt	Class96_36	5.8916
	GGGTAg	Class319_32	2.4828
	CCCG**TAgg**	Class321_64	8.6451
Motif2	CRP (5'-**TGTGA-3'**)		
	tGTGAA	Class134_61	4.8237
	gtGAGAA	Class88_32	6.8526
	GTGAt	Class284_32	2.5110
	Dommage UV (5'- **TTGTCA -3')**		
	tGTCAtt	Class113_48	7.7660
	tCACTT	Class259_31	6.5259
	GTCTAg	Class183_51	7.5548

Dans la colonne motif, les caractères en gras dans les consensus de classe symbolisent les nucléotides présents dans le consensus caractérisé biologiquement. La colonne classe indique le numéro de classe où se trouve le motif ainsi que le nombre de séquences la composant. La colonne *R'MES* fournit le score *R'MES* associé à ces consensus de classe.

Possibilité de rechercher de TFBS dans d'autres génomes

En appliquant les paramètres d'analyse issus de l'analyse de *S. coelicolor*, de 25 % à 52 % des SFBS de huit facteurs sigma ont pu être identifiés correctement (cf. section 2.4, tableau 5). Ces deux bactéries, *B. subitilis* et *S. coelicolor* sont toutes les deux des bactéries Gram positives qui peuvent évoluer dans le sol mais celles-ci sont classées dans des phylums

différents et ont une composition génomique très distincte. Ainsi, l'application de la stratégie pour l'identification de SFBS chez *B. subitilis* est encourageante au vue d'une généralisation. Cette même stratégie employée dans l'analyse des génomes des organismes supérieurs est attirante mais plus complexe. Cette complexité est d'autant plus grande que leurs signaux de régulation font intervenir de multiples acteurs à des distances très longues jusqu'à 10 kb (*Saccharomyces cerevisiae*).

4.1.2 Amélioration de la stratégie

Étape de classification : regroupement de motifs similaires

L'étape de classification des *iPeak-motifs* est intégrée dans le processus de recherche de motifs composites. Cette étape est l'une des étapes clés qui permet de définir des motifs composites candidats et robustes. Le programme de classification (*ClassifMoyenne*) utilise un critère de distorsion pour regrouper les classes (représentant la dispersion d'un élément dans une classe). Cette distorsion nécessite un calcul de similarité entre deux séquences tel que l'algorithme d'alignement local de Smith et Waterman. Dans notre algorithme de classification, une matrice d'identité a été utilisée. Le nombre de classes optimum n'a pas pu être déterminé en utilisant les critères d'Akaike (AIC) ou de Rissanen (BIC). Ainsi, nous avons utilisé une séquence de référence pour déterminer ce nombre de classe à travers l'observation de l'évolution du regroupement des motifs de la boîte -35 de σ^R dans les différentes classes.

Le développement d'un modèle de classification robuste serait indispensable pour améliorer de manière qualitative la stratégie de recherche de motif composite. Une possibilité d'amélioration de ce modèle de classification serait d'utiliser une matrice spécifique au génome de *Streptomyces*. Cette dernière tiendrait compte du contenu en bases G et C et des événements de transition ou de transversion observés. Cette matrice serait donc adaptable au génome à analyser.

Regroupement des motifs composites

L'exécution du processus de fouille de données génère un résultat qui est présenté sous forme d'une liste de motifs composites avec leur nombre d'occurrences et les gènes possédant ce motif en amont de la région codante (Annexe Web F). Dans l'analyse du génome de *S. coelicolor*, 812 motifs SFBS putatifs ont été proposés.

L'analyse de ces motifs regroupés sous forme de consensus augmenterait considérablement la rapidité de recherche et/ou d'identification de régions régulatrices. Le regroupement de motifs induit une identification des gènes présentant ce motif dans leur région promotrice. Une hypothèse concernant leur co-régulation peut alors être envisagée. En outre, la détection d'un gène présent dans plusieurs groupes suggérait qu'il présente plusieurs motifs dans sa région promotrice et qu'il serait soumis à une régulation multiple comme pour le gène *dagA* (cf. section 2.3). L'identification de ce type de gène permettrait d'émettre des hypothèses quant à l'impact de son expression pour la bactérie. En effet, une régulation multiple indiquerait une régulation fine aboutissant à l'expression du gène à des différents moments pour la bactérie. L'identification des réseaux de régulation au sein de différents organismes rendrait possible

des comparaisons afin de déterminer des corrélations ou des évolutions distinctes de ces derniers.

Solutions éventuelles

Dans notre procédure, nous avons utilisé une étape de filtration simple qui consiste à sélectionner les dix premiers motifs composites ayant le plus grand nombre d'occurrences, appelés « Top10 » (cf. section 2.3). Si deux motifs composites sont strictement identiques avec le même nombre d'occurrences élevé mais avec une longueur de séparateur différente, les deux motifs sont comptabilisés comme un motif composite dans le Top10. Ce critère permet de considérer la variabilité de la taille des séparateurs pour un même motif composite. L'exploration de ces résultats peut être laborieuse. Une amélioration nécessaire serait de regrouper les motifs sous forme de consensus sans perdre l'information de la variabilité. C'est une étape délicate car elle nécessite un compromis entre l'aggrégation des données et la conservation d'information. La fusion ou une proposition de regroupement est indispensable pour les cas les plus simples de motifs composites similaires. Le tableau 17 présente un exemple de deux situations à traiter (comportant des motifs inclusifs ou peu divergents) parmi les 812 motifs SFBS potentiels retrouvés chez *S. coelicolor*.

Le premier cas représente un groupe de trois motifs TFBS putatifs qui ont une boîte -35 conservée et une variabilité pour la longueur de la boîte -10. Les motifs des boîtes -10 ont trois nucléotides en commun (TCA). Le troisième motif comporte un nucléotide T en plus par rapport au premier motif mais un en mois par rapport au deuxième motif. Le deuxième motif (GCGGCTT-N_{14}-tttca) inclut les deux autres. C'est une forme de redondance possible et un consensus peut être proposé sans trop de perte d'information. Le second cas est composé de motifs (boîte -35 et boîte -10) différents mais ils partagent la majorité des nucléotides qui composent leur motif. Ce sont des motifs peu divergents et le motif consensus peut être établi en fonction du motif ayant le moins de nucléotides.

Tableau 17 : Regroupement proposé pour quelques motifs SFBS similaires ou peu divergents dans l'ensemble des 812 SFBS potentiels

	Motifs composites redondants		**Motifs composites peu divergents**	
	Motif	**Taille du séparateur**	**Motif**	**Taille du séparateur**
	GCGGCTT-ttca	15	CGGGAAT-gtt	18
	GCGGCTT-tttca	14	GGGGAAT-ggt	17
	GCGGCTT-tca	16	GGGGAAT-ggtt	17
			CGGGAAT-cgtt	17
			CGGGAAT-gttg	18
			TCGGAAT-ctggt	15
			AGGGAAT-gtt	18
Proposition de consensus	GCGGCTT-14-ttTCA		NgGGAAT-15-ctSGTTg	

Les lettres en capital représentent des positions strictes et conservées alors que les lettres en minuscule indiquent une variabilité à cette position.

4.2 Fouille de données développée pour l'analyse des gènes « atypiques »

Nous avons présenté une méthode de fouille de données utilisant les modèles de Markov d'ordre 2 afin de déceler des hétérogénéités longues associées à des propriétés de séquences exogènes. Son application sur l'analyse de *S. thermophilus* a permis de détecter un ensemble de gènes « atypiques » dont l'origine exogène de certains d'entre eux est étayée par une analyse d'annotation fonctionnelle et une analyse de la distribution des homologues au sein des firmicutes (gènes spécifiques, valeur de dispersion). Dans cette partie, nous allons, tout d'abord, présenter une comparaison d'efficacité de notre approche par rapport à un modèle de Markov caché d'ordre 1 ainsi qu'une comparaison à différentes méthodes (HGT-DB, Islandpath, SIGI-HMM). La seconde comparaison permet de confronter la méthode développée avec des méthodes utilisant les critères de composition. Cette comparaison permet de mettre en évidence si la stratégie fondée sur le M2M2 identifie seulement les gènes « atypiques » par leurs caractéristiques compositionnelles connues. Ensuite, nous allons nous intéresser aux améliorations possibles dans le système d'analyse de la distribution phylogénétique.
Une interface graphique a été mise en place afin de former un pipeline d'analyse et de procurer une interface conviviale (annexe J).

4.2.1 Méthodes HMM vs. autres méthodologies de détection du HGT

Le jeu de données à comparer est issu des données de CGH, c'est-à-dire les 73 gènes de *S. thermophilus* CNRZ1066 présents dans au plus 16 souches (*CGH-set*). Ce jeu de données représente des gènes de *S. thermophilus* issu d'une acquisition récente dont les propriétés compositionnelles sont supposées être les plus distinguables car ils ont été le moins soumis aux mécanismes d'amélioration. Cet échantillon de gènes permet aussi de mesurer la capacité des autres méthodes à détecter ce que le HMM d'ordre 2 a identifié.

Comparaison des modèles de Markov d'ordre et de la dépendance des observations
Le M2M2 utilisé montre une meilleure sensibilité (0.74) que celle du M1M2 (0,64) alors que la spécificité du M1M2 est à peine plus grande (tableau 18). Le ratio (spécificité/sensibilité) proche de la valeur 1 avec des valeurs élevées de spécificité et de sensibilité indique un bon compromis entre la détection des vrais signaux (vrais positifs) et le bruit de fond (faux positifs). Ce ratio est légèrement meilleur pour le M2M2 (1,16) que pour le M1M2 (1,36). Il est intéressant de noter que le modèle M2M0 est supérieur sur tous les plans aux modèles M1M2 et M2M2. Le M2M0 identifie plus de gènes considérés comme exogènes et moins de candidats donc potentiellement moins de faux positifs. Cette identification est possible car elle est couplée à un algorithme d'extraction qui facilite la reconnaissance de régions denses en *pos_dev* (cf. section 3.6.2). Néanmoins, la présence de faux négatifs n'est pas gênant car

l'étape de classification permet de les filtrer. Nous donnons plus d'importance à la sensibilité c'est-à-dire à l'aptitude d'identifier des vrais motifs.

Tableau 18 : Sensibilité et spécificité des différents modèles de Markov sur les 73 gènes variables (*CGH-set*)

	M2M2	M1M2	M2M0
Nombre de gènes identifiés	362	313	277
Nombre de gènes présents dans *CGH-set*	54	47	58
Sensibilité[1]	0,73972603	0,64383562	0,79452055
Spécificité[2]	0,85674419	0,87381404	0,8937409
Ratio (Spécificité/Sensibilité)	1,15819121	1,35720053	1,12488079

[1]Calcul de la sensibilité : VP (Vrai Positif)/VP + FN (Faux Négatif).
[2]Calcul de la spécificité : VN (Vrai Négatif)/VN+FP (Faux positif). Les VN sont les 1915 CDS de l'ensemble du génome moins les 73 gènes de l'analyse CGH.

Méthodes HMM d'ordre 2 vs. autres méthodologies de détection du HGT

La comparaison s'est effectuée sur trois méthodes utilisant différents critères (composition, annotation fonctionnelle, usage des codons) :

- HGTDB (Garcia-Vallve *et al.*, 2003) est une base de données de gènes potentiellement acquis par transfert horizontal. L'identification de ces gènes est basée sur la déviation de leur composition en pourcentage de bases G et C, de leur usage des codons et des amino acides par rapport au génome.
- IslandPath (Hsiao *et al.*, 2003) permet une visualisation sous forme de carte avec des points qui représentnt les gènes ordonnés selon leur position sur le génome. Chaque gène est représenté avec un code couleur. Cette représentation met en relief les gènes exogènes ayant des caractéristiques compositionnelles (pourcentage de bases G et C) déviantes ou des gènes codant des éléments mobiles de type IS.
- SIGI-HMM (Waack *et al.*, 2006) prédit des îlots génomiques et les donneurs potentiels sur la base de l'analyse de l'usage des codons.

Le programme IslandViewer (Langille et Brinkman, 2009) a permis d'extraire les données issues d'IslandPath et de SIGI-HMM. En effet, IslandViewer utilise ces deux analyses afin d'effectuer des analyse comparatives, de permettre une visualisation et une localisation des données sur le génome. La sensibilité et la spécificité de détection des 73 gènes de l'analyse CGH (*CGH-set*) ont été calculées pour les quatre jeux de données (tableau 19).

Tableau 19 : Sensibilité et spécificité des méthodes pour les 73 gènes variables (*CGH-set*)

	M2M2	SIGI-HMM	IslandPath-DIMOB	HGT-DB
Nombre de gènes identifiés	362	60	126	151
Nombre de gènes présents dans *CGH-set*	54	22	8	36
Sensibilité[1]	0,73972603	0,30136986	0,10958904	0,49315068
Spécificité[2]	0,85674419	0,97978723	0,93979592	0,94123659
Ratio (Spécificité/ Sensibilité)	1,15819121	3,25111219	8,57563776	1,90861863

[1]Calcul de la sensibilité : VP (Vrai Positif)/VP + FN (Faux Négatif).
[2]Calcul de la spécificité : VN (Vrai Négatif)/VN+FP (Faux positif). Les VN sont les 1915 CDS de l'ensemble du génome moins les 73 gènes de l'analyse CGH.

Le M2M2 présente la meilleure sensibilité (0.74) alors que SIGI-HMM procure la meilleure spécificité (0.98) avec une sensibilité faible (0.30). Ainsi, SIGI-HMM génère peu de faux positifs mais il identifie peu de gènes considérés comme exogènes, il comporte de nombreux faux négatifs. Le meilleur modèle exprimant un ratio de spécificité et sensibilité proche de 1 est le modèle M2M2 par rapport aux autres méthodes pour l'identification de HGT à partir de l'ensemble *CGH-set*.

4.2.2 Distribution des gènes dans un arbre phylogénétique : valeur de dispersion

L'analyse de la distribution des gènes « atypiques » au sein des firmicutes a permis d'étayer l'hypothèse d'une acquisition par transfert horizontal d'une partie des gènes « atypiques ». La datation de cette acquisition peut être tentée dans certains cas. La valeur de dispersion nous a permis de distinguer plusieurs catégories (cf. section 3.10.3). Le calcul de cette valeur tient compte de la distance des séquences qui sépare les espèces de la base de données des firmicutes à notre génome de référence *S. thermophilus* CNRZ1066. Une valeur de dispersion faible indique que le gène présente soit un homologue dans peu d'espèces au sein des firmicutes mais des espèces proches phylogénétiquement de *S. thermophilus*, soit un homologue largement distribué au sein des firmicutes. Au contraire, une valeur de dispersion forte indique une distribution sporadique des homologues du gène d'intérêt. Ces homologues sont identifiés dans peu d'espèces de la base de données des firmicutes et ces espèces sont également phylogénétiquement éloignées de *S. thermophilus*. Dans les deux cas (valeur de dispersion faible ou forte), une interprétation sur l'acquisition par HGT peut être proposée ; soit l'acquisition peut être récente et donc ne concerner qu'un nombre faible d'espèces phylogénétiquement proches, soit l'acquisition peut être ancienne et avoir été suivie de nombreux événements de pertes.
L'utilisation de cette valeur de dispersion au sein des 47 souches de *S. thermophilus* permet d'affiner l'analyse :

i) des gènes spécifiques d'un petit nombre de *S. thermophilus* ;

ii) les gènes spécifiques de *S. thermophilus* et présents dans toutes les 47 souches de *S. thermophilus* ;

iii) les gènes présents dans au plus six autres espèces au sein des firmicutes.

Par exemple, l'analyse des 189 gènes atypiques de *S. thermophilus* présentant une distribution sporadique au sein des firmicutes suggère l'intervention de transferts en provenance d'espèces plus ou moins proches d'un point de vue évolutif ; ces espèces pouvant être présentes dans l'écosystème lait, par exemple *Lactococcus lactis*. Ainsi, un sous ensemble de 7 gènes présente une distribution remarquable : ils sont sporadiques au sein des firmicutes mais présents simultanément (avec la limite du critère d'identité retenu) chez des streptocoques dont *S. thermophilus* (majoritairement présents dans au plus 46 souches de *S. thermophilus*) ainsi que chez *Lactobacillus lactis*, qui constituent les principaux « co-locataires » du yaourt. Des échanges réciproques ont déjà été identifiés *in silico* entre ces deux partenaires ; en l'occurrence des séquences d'éléments génétiques mobiles (Guédon *et al.*, 1995 ; Burrus *et*

al., 2000 ; Pavlovic *et al.*, 2004). Le produit de quatre de ces 7 gènes présente une identité significative avec des protéines impliquées dans la biosynthèse de la paroi (acides téichoïques), la division cellulaire ou la régulation génique. Ces gènes pourraient donc être la trace d'événements de transferts entre les deux bactéries utilisées en co-culture pour la fabrication des yaourts bien que la directionnalité de ces transferts ne puisse être prédite. La coexistence de ces bactéries a sans doute potentialisé ces transferts ; il est également envisageable que cette coexistence ait pu alimenter la coévolution de ces deux organismes partenaires.
L'exploration de ces profils appuierait certains scénarios d'événements de HGT et nous renseignerait sur la datation des acquisitions. Est-ce une acquisition ancienne ou récente ? Dans le cas d'acquisition récente, la fonction codée par ces gènes a-t-elle un rôle dans l'adaptation de *S. thermophilus* à son milieu ? Ces axes de recherche contribueraient à la compréhension du schéma évolutif de *S. thermophilus.*

Les limites de l'approche

Pondération selon la représentativité des génomes dans la base de données

Dans notre analyse, un gène présent dans une souche est considéré comme présent dans l'espèce, ceci même s'il est absent des autres souches de l'espèce. Cette option entraîne vraisemblablement une sous-estimation du nombre d'événements de transfert. Le calcul d'un indice, qui reflète la distribution d'un gène dans un groupe de souches et/ou d'espèces dont le génome est séquencé, serait plus représentatif.

Recherche d'homologues dans des bases de données plus exhaustives

Un gène possédant une valeur de dispersion de 0 indique qu'il est spécifique de *S. thermophilus* au sein de la base de données des firmicutes. En général, ces gènes ont une fonction inconnue. Afin d'améliorer le processus d'identification, l'ajout d'une étape de recherche d'homologue de ces gènes dans la base de données non redondante (*nr*) ou dans des bases de données de domaines pourrait permettre d'identifier des fonctions ou/et des domaines fonctionnels particuliers (figure 40). Ce processus avait été réalisé de manière manuelle pour la recherche d'homologues des 23 gènes atypiques et spécifiques de *S. thermophilus*. Cette étape de recherche d'homologues dans la base de données non redondante (*nr*) permet d'élargir l'espace d'exploration à l'ensemble des gènes caractérisés dans les génomes non séquencés. Ces analyses complémentaires permettront d'étayer l'hypothèse d'événements HGT entre des espèces phylogénétiquement très éloignées et/ou d'attribuer une fonction hypothétique à ce gène.
De même, un gène ayant une valeur de dispersion proche de 0.06972 (cf. section 3.10.3) indique que ce gène est présent dans toutes les espèces de la base de données de firmicutes mais peut cependant avoir une origine exogène. Par exemple, le locus *eps* est très largement distribué mais des événements d'échanges de gènes ont été caractérisés chez de nombreuses souches (Bourgoin *et al.*, 1999 ; Minic *et al.*, 2007). Une analyse du meilleur homologue pourrait renseigner sur l'homologue le plus proche. Ainsi, le meilleur homologue ne serait pas forcément l'organisme le plus proche phylogénétiquement. Cette recherche sera un indicateur supplémentaire à intégrer dans une étape d'analyse de post-traitement.

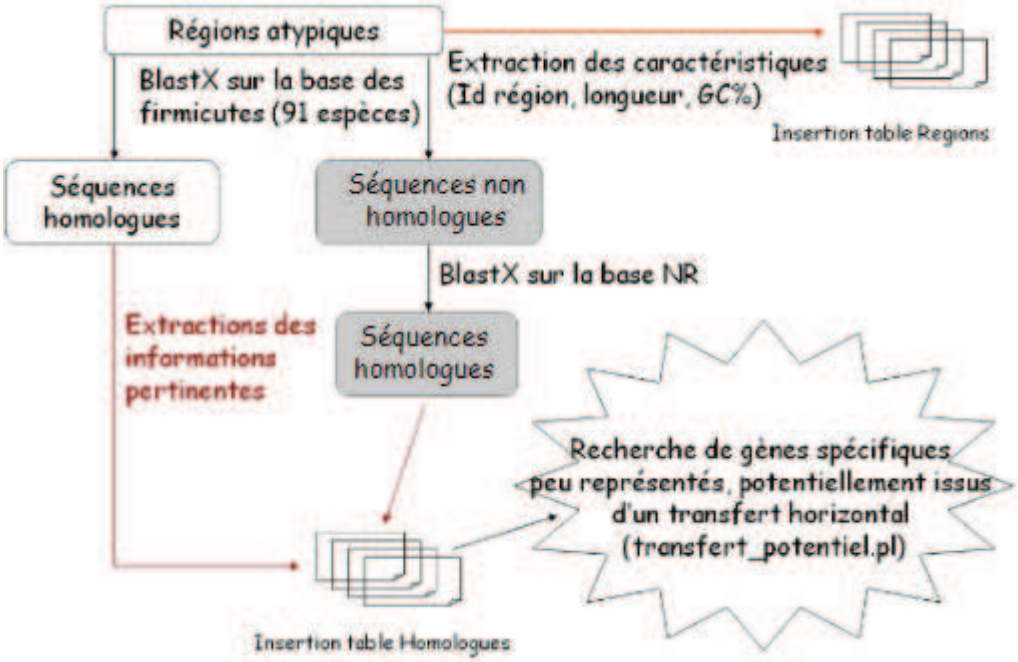


Figure 40 : Étape d'analyse de la distribution des gènes au sein des firmicutes.
Les rectangles grisés indiquent les étapes supplémentaires à automatiser par rapport à l'approche actuelle pour une analyse plus exhaustive.

4.3 Les HMM du second ordre pour la fouille de données génomiques

Les avantages

Les HMM d'ordre 2 employés se révèlent des outils efficaces pour l'extraction des probabilités *a posteriori* des états cachés dans l'identification de signaux de régulation ou dans la détection de gènes atypiques.

i) Nous les avons utilisés pour effectuer une fouille de données sur des données dont les structures ne sont pas connues. L'utilisation dans ce cas d'un apprentissage non supervisé a permis de déployer les HMM d'ordre 2 sur n'importe quelle analyse de séquence génomique sans traitement au préalable comme l'ont montré nos résultats sur deux génomes différents (*S. coelicolor* et *S. thermophilus*). Ce processus diffère de la majorité des analyses de séquences biologiques qui nécessitent un apprentissage supervisé sur un jeu de données ayant des structures connues telles que la recherche de gène (Krogh *et al.*, 1994 ; Kulp *et al.*, 1996 ; Yada et Hirosawa, 1996 ; Burge et Karlin, 1997), de domaine fonctionnel (Eddy, 1998 ; Sonnhammer et *al.*, 1998) et la recherche de promoteur (Jarmer *et al.*, 2001).

ii) Nous avons comparé deux types de HMM suivant les dépendances prises en compte entre observations (M2M0 et M2M2) en utilisant, soit des k*mers*, soit des nucléotides. Les premiers modèles, M2M0, présentent un intérêt pour la recherche de séquences courtes comme des promoteurs tandis que l'utilisation des M2M2 permet l'identification des gènes atypiques.

iii) Les HMM d'ordre 2 dans les stratégies d'analyses de données génomiques ont montré une meilleure sensibilité et spécificité d'identification des régions régulatrices et des

gènes potentiellement acquis par transfert horizontal par rapport aux HMM d'ordre 1 (HMM1 et M1M2). Cette comparaison s'est élargie, dans le cadre de l'identification des gènes exogènes, à d'autres méthodes d'identification de gènes exogènes. Il en ressort que les HMM d'ordre 2 sont plus efficaces pour le jeu de données étudié.

Les difficultés

La principale difficulté de l'utilisation des HMM dans notre cas, a été l'analyse des données brutes des probabilités *a posteriori* des états cachés qui nécessite un post-traitement efficace. L'intérêt de ce post-traitement était de diminuer les faux positifs. En effet, dans le cadre de l'identification des régions régulatrices, une première étape de filtration a consisté à sélectionner les motifs présents dans les régions intergéniques ou chevauchant une région intergénique et l'extrémité d'une région codante. Cette filtration a été appliquée car un enrichissement des *Peak-motifs* dans les régions intergéniques avait été observé (cf. section 2.3) et a permis de révéler les informations plus pertinentes. Néanmoins, cette étape de filtration a pour conséquence d'éliminer les motifs de régulations peu fréquents et présents dans des régions codantes. Une seconde étape importante du post-traitement est le développement d'une méthode combinatoire qui a permis de reconstituer des motifs composites significatifs.

Dans le cadre de l'identification de gènes « atypiques », un post-traitement a permis d'extraire des régions ayant une longueur supérieure à 1 kb avec un système de fenêtre coulissante (cf. section 3.6), ce qui permet d'éliminer des signaux courts.

Les caractéristiques compositionnelles

Les deux approches utilisant les HMM d'ordre 2 révèlent que ces derniers sont influencés par des séquences dont le contenu en bases G et C est faible par rapport à celui du génome moyen. Ces caractéristiques sont toutefois inhérentes aux types d'informations à identifier et aux propriétés du génome. En effet, les gènes exogènes ont des caractéristiques compositionnelles propres qui peuvent différer de celles du génome intégré. Ces caractéristiques peuvent être plus ou moins importantes selon la distance phylogénétique entre les espèces concernées. La stratégie fondée sur le M2M2 décèle une grande partie de gènes « atypiques » ayant une composition faible en bases G et C mais aussi certains gènes ayant un contenu proche de celui du génome étudié (cf. section 3.10). En ce qui concerne l'identification de motifs de régulation de composition faible en G et C par rapport à la composition moyenne des génomes (*S. coelicolor* ou *B. subtilis*), peut s'expliqué par la composition de ces motifs de régulation. Chez *S. coelicolor*, les consensus SFBS des différents facteurs sigma ont un contenu en bases G et C, pour les boîtes -35 et -10, relativement bas par rapport à celui du génome (72%) qui est très élevé même dans les régions intergéniques (68,34%) (cf. section 2.3, tableau 4). De même, les analyses sur le génome de *B. subtilis*, avec un contenu en bases G et C moyen de 43 %, ont permis de trouver et de localiser entre 25 à 52 % des SFBS analysés. Ces séquences régulatrices ont un contenu en bases G et C variable (cf. section 2.4, tableau 5). La stratégie fondée sur le HMM d'ordre 2 a identifié et correctement localisé 6 des 15 SFBS caractérisés expérimentalement. Le modèle

semble être influencé par la composition faible en G et C des éléments mais d'autres paramètres influencent aussi le modèle.

4.4 Définition automatique d'une topologie de HMM d'ordre 2

Le processus de génération d'un HMM d'ordre 2 implémenté (figure 6) consiste en la définition d'une topologie, puis à une estimation des paramètres et ensuite au calcul de la probabilité *a posteriori* d'un état particulier. Ce processus est automatisé excepté pour le choix de la topologie qui est guidé par la prise en compte de la distance entre les états cachés calculée grâce à la distance Kullback-Leibler.
L'étape manuelle du choix de la topologie est difficile pour un utilisateur non initié et nécessite souvent l'intervention d'un expert. La mise en place d'un système de fusion d'états automatique pourrait aider au choix de la topologie à générer. La fusion d'états cachés pourrait être guidée par la matrice de distance en fusionnant les états cachés les plus proches puis en définissant un état résultant grâce à un algorithme itératif analogue à l'algorithme de Ward. Cette stratégie conduit à mettre en œuvre beaucoup d'apprentissages de HMM qui sont des processus gourmands en temps. En ce sens, l'optimisation de ces HMM est un réel challenge.

Références bibliographiques

Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y. and De Moor, B. (2003) Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res* **31**(6): 1753-1764.

Ahel, I., Vujaklija, D., Mikoc, A. and Gamulin, V. (2002) Transcriptional analysis of the recA gene in Streptomyces rimosus: identification of the new type of promoter. *FEMS Microbiol Lett* **209**(1): 133-137.

Ainsa, J.A., Parry, H.D. and Chater, K.F. (1999) A response regulator-like protein that functions at an intermediate stage of sporulation in Streptomyces coelicolor A3(2). *Mol Microbiol*. **34**: 607-619.

Almiron-Roig, E., Mulholland, F., Gasson, M. J. and Griffin, A. M. (2000) The complete cps gene cluster from Streptococcus thermophilus NCFB 2393 involved in the biosynthesis of a new exopolysaccharide. *Microbiology* **146**: 2793-2802.

Altschul, S. F., Boguski, M. S., Gish, W. and Wootton, J. C. (1994) Issues in searching molecular sequence databases. *Nat Genet* **6**(2): 119-129.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* **215**(3): 403-410.

Ammann, A., Neve, H., Geis, A. and Heller, K.J. (2008) Plasmid transfer via transduction from Streptococcus thermophilus to Lactococcus lactis. *J Bacteriol* **190**(8): 3083-7

Amundsen, S. K., Neiman, A. M., Thibodeaux, S. M. and Smith, G. R. (1990) Genetic dissection of the biochemical activities of RecBCD enzyme. *Genetics* **126**(1): 25-40.

Angel C. S., Ruzek M., and Hostetter M.K. (1994) Degradation of C3 by Streptococcus pneumoniae. *J Infect Dis* **170**: 600-608.

Angert, E. R. (2005) Alternatives to binary fission in bacteria. *Nat Rev Microbiol* **3**(3): 214-224.

Apostolico, A., Bock, M. E., Lonardi, S. and Xu, X. (2000) Efficient detection of unusual words. *J Comput Biol* **7**(1-2): 71-94.

Aravind, L. and Koonin, E. V. (2001) Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res* **11**(8): 1365-1374.

Avery, O.T., MacLeod, C. and McCarty, M. (1944) Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. *Journal of Experimental Medicine* **79**(2): 137–158.

Azad, R. K. and Lawrence, J. G. (2005) Use of artificial genomes in assessing methods for atypical gene detection. *PLoS Comput Biol* **1**(6): e56.

Bailey, T. L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28-36.

Bailey, T. L. and Elkan, C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* **3**: 21-29.

Bandelt, H. J., Forster, P. and Rohl, A. (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**(1): 37-48.

Baños, R.C., Vivero, A., Aznar, S., García, J., Pons, M., Madrid, C. and Juárez, A. (2009) Differential regulation of horizontally acquired and core genome genes by the bacterial modulator H-NS. *PLoS Genet.* **5**(6): e1000513.

Barnard, A., Wolfe, A. and Busby, S. (2004) Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. *Curr Opin Microbiol* **7**(2): 102-108.

Barne, K. A., Bown, J. A., Busby, S. J. and Minchin, S. D. (1997) Region 2.5 of the Escherichia coli RNA polymerase sigma70 subunit is responsible for the recognition of the 'extended-10' motif at promoters. *EMBO J* **16**(13): 4034-4040.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. and Horvath, P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**(5819): 1709-1712.

Bateman, A. and Haft, D. H. (2002) HMM-based databases in InterPro. *Brief Bioinform* **3**(3): 236-245.

Bates, S., Cashmore, A. M. and Wilkins, B. M. (1998) IncP plasmids are unusually effective in mediating conjugation of Escherichia coli and Saccharomyces cerevisiae: involvement of the tra2 mating system. *J Bacteriol* **180**(24): 6538-6543.

Baum, L. E. (1970) An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequalities* **3**: 1-8.

Baum, L. E. and Egon, J. A. (1967) An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology. *Bull. Amer. Meteorology Soc.* **73**: 360-363.

Baum, L. E. and Petrie, T. (1966) Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Annals of Math. Statistics* **37**: 1,554-551,563.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970) A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Annals of Math. Statistics* **41**(1): 164-171.

Baum, L. E. and Sell, G. R. (1968) Growth Functions for Transformations on Manifolds. *Pacific J. Math* **27**(2): 211-227.

Bellanger, X., Roberts, A. P., Morel, C., Choulet, F., Pavlovic, G., Mullany, P., Decaris, B. and Guedon, G. (2009) Conjugative transfer of the integrative conjugative elements ICESt1 and ICESt3 from Streptococcus thermophilus. *J Bacteriol* **191**(8): 2764-2775.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. (2008) GenBank. *Nucleic Acids Res* **36**(Database issue): D25-30.

Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C. H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M. A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B. G., Parkhill, J. and Hopwood, D. A. (2002) Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**(6885): 141-147.

Bertrand, H.O., Ha-Duong, T., Fermandjian, S. and Hartmann, B. (1998) Flexibility of the B-DNA backbone: effects of local and neighbouring sequences on pyrimidine-purine steps. *Nucleic Acids Res* **26**(5): 1261-1267.

Besemer, J., Lomsadze, A., and 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 Res.* **29**, 2607- 2183.

Bianco, P. R., Tracy, R. B. and Kowalczykowski, S. C. (1998) DNA strand exchange proteins: a biochemical and physical comparison. *Front Biosci* **3**: D570-603.

Bibb, M. J., Molle, V. and Buttner, M. J. (2000) sigma(BldN), an extracytoplasmic function RNA polymerase sigma factor required for aerial mycelium formation in Streptomyces coelicolor A3(2). *J Bacteriol* **182**(16): 4606-4616.

Bize, L., Muri, F., Samson, F., Rodolphe, F., Ehrlich, S.D., Prum, B. and Bessières, P. (1999) Searching gene transfers on Bacillus Subtilis using Hidden Markov Models. *Recomb'99 Proceedings of the Third Annual International Conference on Computational Molecular Biology.*

Blanchette, M., and Tompa, M. (2002) Discovery of Regulatory elements by a Computational Method for Phylogenetic Footprinting. *Genome Research.* **12**(5): 739-48.

Bolotin, A., Quinquis, B., Renault, P., Sorokin, A., Ehrlich, S. D., Kulakauskas, S., Lapidus, A., Goltsman, E., Mazur, M., Pusch, G. D., Fonstein, M., Overbeek, R., Kyprides, N., Purnelle, B., Prozzi, D., Ngui, K., Masuy, D., Hancy, F., Burteau, S., Boutry, M., Delcour, J., Goffeau, A. and Hols, P. (2004) Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. *Nat Biotechnol* **22**(12): 1554-1558.

Bolotin, A., Quinquis, B., Sorokin, A. and Ehrlich, S. D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**: 2551-2561.

Borodovsky, M. and McIninch, J. (1993) GENMARK: parallel gene recognition for both DNA strands. *Computer & Chemistry* **17**(2): 123-134.

Bourgoin, F., Guedon, G., Gintz, B. and Decaris, B. (1998) Characterization of a novel insertion sequence, IS1194, in Streptococcus thermophilus. *Plasmid* **40**(1): 44-49.

Bourgoin, F., Pluvinet, A., Gintz, B., Decaris, B. and Guedon, G. (1999) Are horizontal transfers involved in the evolution of the Streptococcus thermophilus exopolysaccharide synthesis loci? *Gene* **233**(1-2): 151-161.

Broadbent, J. R., McMahon, C. J., Oberg, C. J. and Welker, D. L. (2001) Use of exopolysaccharide-producing cultures to improve the functionality of low fat cheese. *Dairy J.* **11**: 433–439.

Brown, K. L., Wood, S. and Buttner, M. J. (1992) Isolation and characterization of the major vegetative RNA polymerase of Streptomyces coelicolor A3(2); renaturation of a sigma subunit using GroEL. *Mol Microbiol* **6**(9): 1133-1139.

Buhler, J. and Tompa, M. (2001) Finding Motifs Using Random Projections. *Proc. RECOMB'01*: 69-76.

Burden, S., Lin, Y.-X. and Zhang, R. (2005) Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences. *Bioinformatics.* **21**(5): 601-607.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**(1): 78-94.

Burge, C. B. and Karlin, S. (1998) Finding the genes in genomic DNA. *Curr Opin Struct Biol* **8**(3): 346-354.

Burrus, V., Bontemps, C., Decaris, B. and Guedon, G. (2001) Characterization of a novel type II restriction-modification system, Sth368I, encoded by the integrative element ICESt1 of Streptococcus thermophilus CNRZ368. *Appl Environ Microbiol* **67**(4): 1522-1528.

Burrus, V., Pavlovic, G., Decaris, B. and Guedon, G. (2002) The ICESt1 element of Streptococcus thermophilus belongs to a large family of integrative and conjugative elements that exchange modules and change their specificity of integration. *Plasmid* **48**(2): 77-97.

Burrus, V., Roussel, Y., Decaris, B. and Guedon, G. (2000) Characterization of a novel integrative element, ICESt1, in the lactic acid bacterium Streptococcus thermophilus. *Appl Environ Microbiol* **66**(4): 1749-1753.

Busby, S. and Ebright, R. H. (1999) Transcription activation by catabolite activator protein (CAP). *J Mol Biol* **293**(2): 199-213.

Buttner, M. J., Chater, K. F. and Bibb, M. J. (1990) Cloning, disruption, and transcriptional analysis of three RNA polymerase sigma factor genes of Streptomyces coelicolor A3(2). *J Bacteriol* **172**(6): 3367-3378.

Buttner, M. J., Smith, A. M. and Bibb, M. J. (1988) At least three different RNA polymerase holoenzymes direct transcription of the agarase gene (dagA) of Streptomyces coelicolor A3(2). *Cell* **52**(4): 599-607.

Cappe, O. (2001). Ten years of HMMs.

Carvalho, A.M., Freitas, A.T., Oliveira, A.L. and Sagot, M.-F. (2004) Efficient Extraction of Structured Motifs Using Box-links. *Proceedings of the 11th Conference on String Processing and Information Retrieval (SPIRE 2004). Lecture Notes in Computer Science* **3246**: 267-268.

Casjens, S. (1998) The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* **32**: 339-377.

Cavalli-Sforza, L. L. and Edwards, A. W. (1967) Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet* **19**(3 Pt 1): 233-257.

Chater, K.F., Bruton, C.J., Plaskitt, K.A., Buttner, M.J., Méndez, C. and Helmann, J.D. (1989) The developmental fate of S. coelicolor hyphae depends upon a gene product homologous with the motility σ factor of B. subtilis. *Cell*. **59**: 133-143.

Chater, K. F. (1993) Genetics of differentiation in Streptomyces. *Annu Rev Microbiol* **47**: 685-713.

Chater, K. F. (2000) Developmental decisions during sporulation in the aerial mycelium in Streptomyces. *American Society for Microbiology Press* **Washington**(Brun YV, Shimkets LJ (eds) Prokaryotic development): 33–48.

Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L. and McAdams, H. H. (2004) Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci U S A* **101**(10): 3480-3485.

Cho, Y. H., Lee, E. J., Ahn, B. E. and Roe, J. H. (2001) SigB, an RNA polymerase sigma factor required for osmoprotection and proper differentiation of Streptomyces coelicolor. *Mol Microbiol* **42**(1): 205-214.

Choi, J. H., Jung, H. Y., Kim, H. S. and Cho, H. G. (2000) PhyloDraw: a phylogenetic tree drawing system. *Bioinformatics* **16**(11): 1056-1058.

Churchill, G. A. (1989) Stochastic models for heterogeneous DNA sequences. *Bull Math Biol* **51**(1): 79-94.

Claverys, JP., Martin, B. and Polard, P. (2009) The genetic transformation machinery: composition, localization, and mechanism. *FEMS Microbiol Rev* **33**(3):643-56.

Connolly, B., Parsons, C. A., Benson, F. E., Dunderdale, H. J., Sharples, G. J., Lloyd, R. G. and West, S. C. (1991) Resolution of Holliday junctions in vitro requires the Escherichia coli ruvC gene product. *Proc Natl Acad Sci U S A* **88**(14): 6063-6067.

Conte, A., Chinello, P., Civljak, R., Bellussi, A., Noto, P. and Petrosillo, N. (2006) Streptococcus salivarius meningitis and sphenoid sinus mucocele. Case report and literature review. *J Infect* **52**(1): e27-30.

Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16**(22): 10881-10890.

Crooks, G. E., Hon, G., Chandonia, J. M. and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res* **14**(6): 1188-1190.

Darling, A. C., Mau, B., Blattner, F. R. and Perna, N. T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**(7): 1394-1403.

Daubin, V., Lerat, E. and Perriere, G. (2003) The source of laterally transferred genes in bacterial genomes. *Genome Biol* **4**(9): R57.

Daubin, V. and Perriere, G. (2003) G+C3 structuring along the genome: a common feature in prokaryotes. *Mol Biol Evol* **20**(4): 471-483.

Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In: M Dayhoff, Editor, Atlas of Protein Sequence and Structure 5, National Biomedical Research Foundation, Washington, DC, pp. 345–358 suppl 3

de Saizieu, A., Gardes, C., Flint, N., Wagner, C., Kamber, M., Mitchell, T. J., Keck, W., Amrein, K. E. and Lange, R. (2000) Microarray-based identification of a novel Streptococcus pneumoniae regulon controlled by an autoinduced peptide. *J Bacteriol* **182**(17): 4696-4703.

Defrance, M. and Touzet, H. (2006) Predicting transcription factor binding sites using local over-representation and comparative genomics. *BMC Bioinformatics* **7**: 396.

Delcour, J., Ferain, T. and Hols, P. (2000) Advances in the genetics of thermophilic lactic acid bacteria. *Curr Opin Biotechnol* **11**(5): 497-504.

Della, M., Palmbos, P. L., Tseng, H. M., Tonkin, L. M., Daley, J. M., Topper, L. M., Pitcher, R. S., Tomkinson, A. E., Wilson, T. E. and Doherty, A. J. (2004) Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine. *Science* **306**(5696): 683-685.

Delorme, C. (2008) Safety assessment of dairy microorganisms: Streptococcus thermophilus. *Int J Food Microbiol* **126**(3): 274-277.

Delorme, C., Poyart, C., Ehrlich, S. D. and Renault, P. (2007) Extent of horizontal gene transfer in evolution of Streptococci of the salivarius group. *J Bacteriol* **189**(4): 1330-1341.

Demain, A. L. (1999) Pharmaceutically active secondary metabolites of microorganisms. *Appl Microbiol Biotechnol* **52**(4): 455-463.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via EM algorithm (with discussion). *J. of the Royal Stat. Soc. B* **39**: 1-38.

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G. and Fertil, B. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* **16**(10): 1391-1399.

Diaz-Lazcoz, Y., Aude, J. C., Nitschke, P., Chiapello, H., Landes-Devauchelle, C. and Risler, J. L. (1998) Evolution of genes, evolution of species: the case of aminoacyl-tRNA synthetases. *Mol Biol Evol* **15**(11): 1548-1561.

Didelot, X., Darling, A. and Falush, D. (2009) Inferring genomic flux in bacteria. *Genome Res* **19**(2): 306-317.

Doherty, A.J., Jackson, S.P. and Weller, G.R. (2001) Identification of bacterial homologues of the Ku DNA repair proteins. *FEBS Lett* **500**:186–88.

Dombroski, A. J., Walter, W. A. and Gross, C. A. (1993) Amino-terminal amino acids modulate sigma-factor DNA-binding activity. *Genes Dev* **7**(12A): 2446-2455.

Domelevo-Entfellner, J. B. and Gascuel, O. (2008). Une approche phylo-HMM pour la recherche de séquences. Jobim (Actes)**:** 133-139.

Doolittle, W. F. (1999) Phylogenetic classification and the universal tree. *Science* **284**(5423): 2124-2129.

Doyuk, E., Ormerod, O. J. and Bowler, I. C. (2002) Native valve endocarditis due to Streptococcus vestibularis and Streptococcus oralis. *J Infect* **45**(1): 39-41.

Dsouza, M., Larsen, N. and Overbeek, R. (1997) Searching for patterns in genomic data. *Trends Genet* **13**(12): 497-498.

Dubnau, D. and Cirigliano, C. (1972) Fate of transforming deoxyribonucleic acid after uptake by competent Bacillus subtilis: size and distribution of the integrated donor segments. *J Bacteriol* **111**(2): 488-494.

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A. and Deschavanne, P. (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res* **33**(1): e6.

Duggan, P. S., Chambers, P. A., Heritage, J. and Forbes, J. M. (2000) Survival of free DNA encoding antibiotic resistance from transgenic maize and the transformation activity of DNA in ovine saliva, ovine rumen fluid and silage effluent. *FEMS Microbiol Lett* **191**(1): 71-77.

Du Preez, J.A. (1998) Efficient training of high-order hidden Markov model using first-order representations. *Computer Speech & Language* **12**(1): 23-39.

Eddy, S. R. (1996) Hidden Markov models. *Curr Opin Struct Biol* **6**(3): 361-365.

Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics* **14**(9): 755-763.

Edgar, R. C. and Sjolander, K. (2003) Simultaneous sequence alignment and tree construction using hidden Markov models. *Pac Symp Biocomput*: 180-191.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *Ann. Statist* **7**: 1-26.

Emmert, D.B., Stoehr, P.J., Stoesser, G. and Cameron, G.N. (1994) The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Research.* **26**(1): 3445-3449.

Enright, M. C. and Spratt, B. G. (1999) Multilocus sequence typing. *Trends Microbiol* **7**(12): 482-487.

Eskin, E. and Pevzner, P. A. (2002) Finding composite regulatory patterns in DNA sequences. *Bioinformatics* **18 Suppl 1**: S354-363.

Eskin, E., Gelfand, M. and Pevzner. P.A. (2003) Genome-Wide Analysis of Bacterial Promoter Regions. In Pacific Symposium on Biocomputing **8**: 29–40.

Estrem, S. T., Ross, W., Gaal, T., Chen, Z. W., Niu, W., Ebright, R. H. and Gourse, R. L. (1999) Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. *Genes Dev.* **13**: 2134-2147.

Facklam, R. (2002) What happened to the streptococci: overview of taxonomic and nomenclature changes. *Clin Microbiol Rev* **15**(4): 613-630.

Feder, M. E. (2007) Evolvability of physiological and biochemical traits: evolutionary mechanisms including and beyond single-nucleotide mutation. *J Exp Biol* **210**(Pt 9): 1653-1660.

Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**(6): 368-376.

Felsenstein, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**: 783-791.

Ferdows, M. S. and Barbour, A. G. (1989) Megabase-sized linear DNA in the bacterium Borrelia burgdorferi, the Lyme disease agent. *Proc Natl Acad Sci U S A* **86**(15): 5969-5973.

Fertil, B., Massin, M., Lespinats, S., Devic, C., Dumee, P. and Giron, A. (2005) GENSTYLE: exploration and analysis of DNA sequences with genomic signature. *Nucleic Acids Res* **33**(Web Server issue): W512-515.

Fogel, G. B., Weekes, D. G., Varga, G., Dow, E. R., Harlow, H. B., Onyia, J. E. and Su, C. (2004) Discovery of sequence motifs related to coexpression of genes using evolutionary computation. *Nucleic Acids Res* **32**(13): 3826-3835.

Fontaine, L., Boutry, C., Guédon, E., Guillot, A., Ibrahim, M., Grossiord, B. and Hols, P. (2007) Quorum-Sensing Regulation of the Production of Blp Bacteriocins in Streptococcus thermophilus. *Journal of Bacteriology* **20**(189): 7195-7205.

Fox, P. F. (1993). Cheese: Chemistry, Physics and Microbiology. Chapman & Hall, London, UK.

Franke, A. E. and Clewell, D. B. (1981) Evidence for a chromosome-borne resistance transposon (Tn916) in Streptococcus faecalis that is capable of "conjugal" transfer in the absence of a conjugative plasmid. *J Bacteriol* **145**(1): 494-502.

Franks, A. H., Griffiths, A. A. and Wake, R. G. (1995) Identification and characterization of new DNA replication terminators in Bacillus subtilis. *Mol Microbiol* **17**(1): 13-23.

Fraser, C., Hanage, W. P. and Spratt, B. G. (2007) Recombination and the nature of bacterial speciation. *Science* **315**(5811): 476-480.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, R. D., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J. F., Dougherty, B. A., Bott, K. F., Hu, P. C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A., 3rd and Venter, J. C. (1995) The minimal gene complement of Mycoplasma genitalium. *Science* **270**(5235): 397-403.

Frickey, T. and Lupas, A. N. (2004) PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Res* **32**(17): 5231-5238.

Friedman, D. I. (1992) Interaction between bacteriophage lambda and its Escherichia coli host. *Curr Opin Genet Dev* **2**(5): 727-738.

Garcia-Vallve, S., Guzman, E., Montero, M. A. and Romeu, A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res* **31**(1): 187-189.

Germond, J. E., Delley, M., D'Amico, N. and Vincent, S. J. (2001) Heterologous expression and characterization of the exopolysaccharide from Streptococcus thermophilus Sfi39. *Eur J Biochem* **268**(19): 5149-5156.

Gordon, J. J., Towsey, M. W., Hogan, J. M., Mathews, S. A. and Timms, P. (2006) Improved prediction of bacterial transcription start sites. *Bioinformatics* **22**(2): 142-148.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res* **9**(1): r43-74.

Gregory, T. R. (2005) Synergy between sequence and size in large-scale genomics. *Nat Rev Genet* **6**(9): 699-708.

Griffith, F. (1928) The Significance of Pneumococcal Types. *The Journal of Hygiene* **27**(2): 113–159.

Grote, A., Hiller, K., Scheer, M., Munch, R., Nortemann, B., Hempel, D. C. and Jahn, D. (2005) JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res* **33**(Web Server issue): W526-531.

Grundy, W. N., Bailey, T. L., Elkan, C. P. and Baker, M. E. (1997) Meta-MEME: motif-based hidden Markov models of protein families. *Comput Appl Biosci* **13**(4): 397-406.

Guedon, G., Bourgoin, F., Pebay, M., Roussel, Y., Colmin, C., Simonet, J. M. and Decaris, B. (1995) Characterization and distribution of two insertion sequences, IS1191 and iso-IS981, in Streptococcus thermophilus: does intergeneric transfer of insertion sequences occur in lactic acid bacteria co-cultures? *Mol Microbiol* **16**(1): 69-78.

Gunton, J. E., Gilmour, M. W., Baptista, K. P., Lawley, T. D. and Taylor, D. E. (2007) Interaction between the co-inherited TraG coupling protein and the TraJ membrane-associated protein of the H-plasmid conjugative DNA transfer system resembles chromosomal DNA translocases. *Microbiology* **153**(Pt 2): 428-441.

Gusfield, D., Bansal, V., Bafna, V. and Song, Y. S. (2007) A decomposition theory for phylogenetic networks and incompatible characters. *J Comput Biol* **14**(10): 1247-1272.

Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T. and White, O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res* **29**(1): 41-43.

Hacker, J., Blum-Oehler, G., Muhldorfer, I. and Tschape, H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol* **23**, 1089–1097.

Haldenwang, W. G. (1995) The sigma factors of Bacillus subtilis. *Microbiol Rev* **59**(1): 1-30.

He, Y. (1988) Extended Viterbi algorithm for second-order hidden Markov process. *Proc. IEEE Int. Conf. Pattern Recognition*. **2**: 718–720.

Heinemann, J. A. and Sprague, G. F. (1989) Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* **340**: 205-209

Helmann, J. D. (1995) Compilation and analysis of Bacillus subtilis sigma A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucleic Acids Res* **23**(13): 2351-2360.

Henderson, J., Salzberg, S. and Fasman, K. H. (1997) Finding genes in DNA with a Hidden Markov Model. *J Comput Biol* **4**(2): 127-141.

Henikoff, S. and Henikoff, J. G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res* **19**(23): 6565-6572.

Henning, W. (1966) Phylogenetic systematics. *Urbana, IL: University of Illinois Press*.

Hergalant, S., Aigle, B., Decaris, B., Leblond, P. and Mari, J.-F. (2002) Intragenomic Reiterations Detection Using Hidden Markov Models. *Intelligent Systems for Molecular Biology (ISMB).*

Hernandez, D., Gras, R. and Appel, R. (2004) MoDEL: an efficient strategy for ungapped local multiple alignment. *Comput Biol Chem* **28**(2): 119-128.

Hiard, S., Maree, R., Colson, S., Hoskisson, P.A., Titgemeyer, F., Joris, B., van Wezel, G.P., Wehenkel, L. and Rigali, S. (2007) PREDetector: A new tool to identify regulatory elements in bacterial genomes. *Biochem Biophys Res Commun*. **357**(4): 861-4.

Hoebeke, M. and Schabtach, S. (2006) R'MES: Finding Exceptional Motifs. *User guide* **version 3**.

Hoffmeister, M. and Martin, W. (2003) Interspecific evolution: microbial symbiosis, endosymbiosis and gene transfer. *Environ Microbiol* **5**(8): 641-649.

Hols, P., Hancy, F., Fontaine, L., Grossiord, B., Prozzi, D., Leblond-Bourget, N., Decaris, B., Bolotin, A., Delorme, C., Dusko Ehrlich, S., Guedon, E., Monnet, V., Renault, P. and Kleerebezem, M. (2005) New insights in the molecular biology and physiology of Streptococcus thermophilus revealed by comparative genomics. *FEMS Microbiol Rev* **29**(3): 435-463.

Horvath, P., Coute-Monvoisin, A. C., Romero, D. A., Boyaval, P., Fremaux, C. and Barrangou, R. (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int J Food Microbiol* **131**(1): 62-70.

Horvath, P., Romero, D. A., Coute-Monvoisin, A. C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. and Barrangou, R. (2008) Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J Bacteriol* **190**(4): 1401-1412.

Howe, K., Bateman, A. and Durbin, R. (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* **18**(11): 1546-1547.

Hsiao, W., Wan, I., Jones, S. J. and Brinkman, F. S. (2003) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* **19**(3): 418-420.

Hu, J., Li, B. and Kihara, D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res* **33**(15): 4899-4913.

Hughes, J. D., Estep, P. W., Tavazoie, S. and Church, G. M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**(5): 1205-1214.

Huson, D. H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**(2): 254-267.

Idigoras, P., Valiente, A., Iglesias, L., Trieu-Cout, P. and Poyart, C. (2001) Meningitis due to Streptococcus salivarius. *J Clin Microbiol* **39**(8): 3017.

Ikeda, H., Aoki, K. and Naito, A. (1982) Illegitimate recombination mediated in vitro by DNA gyrase of Escherichia coli: structure of recombinant DNA molecules. *Proc Natl Acad Sci U S A* **79**(12): 3724-3728.

Ikeda, H., Shimizu, H., Ukita, T. and Kumagai, M. (1995) A novel assay for illegitimate recombination in Escherichia coli: stimulation of lambda bio transducing phage formation by ultra-violet light and its independence from RecA function. *Adv Biophys* **31**: 197-208.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. and Omura, S. (2003) Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol* **21**(5): 526-531.

Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**(1): 13-34.

Jacob, F. (1955) Transductioni of lysogeny in Esclwreiichiia coli. *Virology* **1**: 207-220.

Jacques, P.E., Rodrigue, S., Gaudreau, L., Goulet, J. and Brzezinski, R. (2006) Detection of prokaryotic promoters from the genomic distribution of hexanucleotide pairs. *BMC Bioinformatics.* **7**: 423.

Jain, R., Rivera, M. C., Moore, J. E. and Lake, J. A. (2002) Horizontal gene transfer in microbial genome evolution. *Theor Popul Biol* **61**(4): 489-495.

Jarmer, H., Larsen, T.S., Krogh, A., Saxild, H.H., Brunak, S. and Knudsen, S. (2001) Sigma A recognition sites in the Bacillus subtilis genome. *Microbiology* **147**:2417-24.

Jauregui, R., Abreu-Goodger, C., Moreno-Hagelsieb, G., Collado-Vides, J. and Merino, E. (2003) Conservation of DNA curvature signals in regulatory regions of prokaryotic genes. *Nucleic Acids Res* **31**(23): 6770-6777.

Jeffrey, H. J. (1990) Chaos game representation of gene structure. *Nucleic Acids Res* **18**(8): 2163-2170.

Jeon, Y. H., Yamazaki, T., Otomo, T., Ishihama, A. and Kyogoku, Y. (1997) Flexible linker in the RNA polymerase alpha subunit facilitates the independent motion of the C-terminal activator contact domain. *J Mol Biol* **267**(4): 953-962.

Johansson, O., Alkema, W., Wasserman, W. W. and Lagergren, J. (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics* **19 Suppl 1**: i169-176.

Johnsborg, O., Eldholm, V. and Havarstein, L.S. (2007) Natural genetic transformation: prevalence, mechanisms and function. *Research in Microbiology* **158** (10):767-778

Jordan, I. K., Makarova, K. S., Spouge, J. L., Wolf, Y. I. and Koonin, E. V. (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res* **11**(4): 555-565.

Kang, J. G., Hahn, M. Y., Ishihama, A. and Roe, J. H. (1997) Identification of sigma factors for growth phase-related promoter selectivity of RNA polymerases from Streptomyces coelicolor A3(2). *Nucleic Acids Res* **25**(13): 2566-2573.

Kanhere, A. and Bansal, M. (2005) A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics*. **6**(1).

Karlin, S. and Burge, C. (1997) Dinucleotide relative abundance extremes : a genomic signature. *Trends Genet.* **11**: 283-290.

Karlin, S., Mrazek, J. and Campbell, A. M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**(12): 3899-3913.

Karlin, S., Mrazek, J. and Campbell, A. M. (1998) Codon usages in different gene classes of the Escherichia coli genome. *Mol Microbiol* **29**(6): 1341-1355.

Karplus, K. and Hu, B. (2001) Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics* **17**(8): 713-720.

Karplus, K., Sjolander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L. and Sander, C. (1997) Predicting protein structure using hidden Markov models. *Proteins* **Suppl 1**: 134-139.

Kelemen, G. H., Brown, G. L., Kormanec, J., Potuckova, L., Chater, K. F. and Buttner, M. J. (1996) The positions of the sigma-factor genes, whiG and sigF, in the hierarchy controlling the development of spore chains in the aerial hyphae of Streptomyces coelicolor A3(2). *Mol Microbiol* **21**(3): 593-603.

Kenney, T. J. and Moran, C. P., Jr. (1991) Genetic evidence for interaction of sigma A with two promoters in Bacillus subtilis. *J Bacteriol* **173**(11): 3282-3290.

Kim, E. S., Song, J. Y., Kim, D. W., Chater, K. F. and Lee, K. J. (2008) A possible extended family of regulators of sigma factor activity in Streptomyces coelicolor. *J Bacteriol* **190**(22): 7559-7566.

Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* **217**(5129): 624-626.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**(2): 111-120.

Knight, R.D., Freeland, S.J. and Landweber, L.F. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol* **2**(4): RESEARCH0010.

Konstantinidis, K.T. and Tiedje, J.M. (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci U S A* **101**(9): 3160-3165.

Koski, L. B., Morton, R.A. and Golding, G.B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* **18**(3): 404-412.

Kowalczykowski, S.C., Dixon, D A., Eggleston, A.K., Lauder, S.D. and Rehrauer, W.M. (1994) Biochemistry of homologous recombination in Escherichia coli. *Microbiol Rev* **58**(3): 401-465.

Kowalczykowski, S.C., Dixon, D.A., Eggleston, A. K., Lauder, S.D. and Rehrauer, W. . (1994) Biochemistry of homologous recombination in Escherichia coli. *Microbiol Rev* **58**(3): 401-465.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K. and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235**(5): 1501-1531.

Krogh, A., Mian, I.S. and Haussler, D. (1994) A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res* **22**(22): 4768-4778.

Krogh, A. (1998) An Introduction to Hidden Markov Models for Biological Sequences. *Computational Methods in Molecular Biology*: 45-63.

Kullback, S. and Leibler, R A. (1951) On information and sufficiency. *Annals of Mathematical Statistic* **22**: 79-86.

Kulp, D., Haussler, D., Reese, M.G. and Eeckman, F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *Proc Int Conf Intell Syst Mol Biol* **4**: 134-142.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.K., Codani, J.J., Connerton, I.F., Danchin, A. et *al*. (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature* **390**(6657): 249-256.

Kurt, F., Caramori, T., Chen, Y.-F., Galizzi, A. and Helmann, J. D. (1995) Promoter architecture in the flagellar regulon of Bacillus subtilis: High-level expression of flagellin by the crD RNA polymerase requires an upstream promoter element. *Proc. Natl. Acad. Sci* **92**: 2582-2586.

Kuzminov, A., Schabtach, E. and Stahl, F.W. (1994) Chi sites in combination with RecA protein increase the survival of linear DNA in Escherichia coli by inactivating exoV activity of RecBCD nuclease. *EMBO J* **13**(12): 2764-2776.

Langille, M. G. and Brinkman, F. S. (2009) IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* **25**(5): 664-665.

Langille, M.G. and Brinkman, F.S. (2009) IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* **25**(5): 664-665.

Lawrence, C.E. and Reilly, A. A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **7**(1): 41-51.

Lawrence, C.E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**(5131): 208-214.

Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**(4): 383-397.

Lawrence, J.G. and Ochman, H. (1998) Molecular archaeology of the Escherichia coli genome. *Proc Natl Acad Sci U S A* **95**(16): 9413-9417.

Le Ber, F., Benoît, M., Schott, C., Mari, J.-F. and Mignolet, C. (2006) Studying Crop Sequences with CarrotAge, a HMM-Based Data Mining Software. *Ecological Modelling* **191**(1): 170-185.

Lee, E. J., Karoonuthaisiri, N., Kim, H. S., Park, J. H., Cha, C. J., Kao, C. M. and Roe, J. H. (2005) A master regulator sigmaB governs osmotic and oxidative response as well as differentiation via a network of sigma factors in Streptomyces coelicolor. *Mol Microbiol* **57**(5): 1252-1264.

Lefébure, T. and Stanhope, M. J. (2007) Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome Biol* **8**(5): R71.

Lesley, S.A. and Burgess, R.R. (1989) Characterization of the Escherichia coli transcription factor sigma 70: localization of a region involved in the interaction with core RNA polymerase. *Biochemistry* **28**(19): 7728-7734.

Li, H., Rhodius, V., Gross, C. and Siggia, E.D. (2002) Identification of the Binding Sites of Regulatory Proteins in Bacterial Genomes. *Proc Natl Acad Sci.* **99**(18): 11772-7.

Li, Y. H., Lau, P.C., Lee, J.H., Ellen, R.P. and Cvitkovitch, D.G. (2001) Natural genetic transformation of Streptococcus mutans growing in biofilms. *J Bacteriol* **183**(3): 897-908.

Liolios, K., Mavromatis, K., Tavernarakis, N. and Kyrpides, N. C. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **36**(Database issue): D475-479.

Liu, M., Siezen, R. J. and Nauta, A. (2009) In silico prediction of horizontal gene transfer events in Lactobacillus bulgaricus and Streptococcus thermophilus reveals protocooperation in yogurt manufacturing. *Appl Environ Microbiol* **75**(12): 4120-4129.

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

Lloyd, RG. and Buckman, C. (1995) Conjugational recombination in Escherichia coli: genetic analysis of recombinant formation in Hfr x F- crosses. *Genetics* **139**(3): 1123-48.

Lobocka, M. B., Rose, D. J., Plunkett, G., 3rd, Rusin, M., Samojedny, A., Lehnherr, H., Yarmolinsky, M. B. and Blattner, F. R. (2004) Genome of bacteriophage P1. *J Bacteriol* **186**(21): 7032-7068.

Lonetto, M., Gribskov, M. and Gross, C. A. (1992) The sigma 70 family: sequence conservation and evolutionary relationships. *J Bacteriol* **174**(12): 3843-3849.

Loots, G., Ovcharenko, I., Pachter, L., Dubchak, I. and Rubin, E. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Research* **12**(5): 832-9.

Maamar, H. and Dubnau, D. (2005) Bistability in the Bacillus subtilis K-state (competence) system requires a positive feedback loop. *Mol Microbiol* **56**(3): 615-624.

Maetschke, S. R., Towsey, M. W. and Hogan, J. M. (2006) Bacterial promoter modelling and prediction for E. coli and B. subtilis with Beagle. *Proceedings workshop on Intelligent systems for bioinformatics*: 9-13.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M. and Spratt, B. G. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**(6): 3140-3145.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., Polouchine, N., Shakhova, V., Grigoriev, I., Lou, Y., Rohksar, D., Lucas, S., Huang, K., Goodstein, D. M., Hawkins, T., Plengvidhya, V., Welker, D., Hughes, J., Goh, Y., Benson, A., Baldwin, K., Lee, J. H., Diaz-Muniz, I., Dosti, B., Smeianov, V., Wechter, W., Barabote, R., Lorca, G., Altermann, E., Barrangou, R., Ganesan, B., Xie, Y., Rawsthorne, H., Tamir, D., Parker, C., Breidt, F., Broadbent, J., Hutkins, R., O'Sullivan, D., Steele, J., Unlu, G., Saier, M., Klaenhammer, T., Richardson, P., Kozyavkin, S., Weimer, B. and Mills, D. (2006) Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A* **103**(42): 15611-15616.

Makita, Y., Nakao, M., Ogasawara, N. and Nakai, K. (2004) DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Res*. **32**: D75-77.

Mari, J.-F., Haton, J.-P. and Kriouile, A. (1997) Automaticword recognition based on second-order hidden Markov models. *IEEE Trans. Speech Audio Process* **5**: 22_25.

Mari, J.F. and Le Ber, F. (2006) Temporal and Spatial Data Mining with Second-Order Hidden Markov Models. *Soft Computing* **10**: 406-414

Marsan, L. and Sagot, M. F. (2000) Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J Comput Biol* **7**(3-4): 345-362.

Marsan, L. and Sagot, M. F. (2000) Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J Comput Biol* **7**(3-4): 345-362.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**(1): 374-378.

Mazurakova, V., Sevcikova, B., Rezuchova, B. and Kormanec, J. (2006) Cascade of sigma factors in streptomycetes: identification of a new extracytoplasmic function sigma factor sigmaJ that is under the control of the stress-response sigma factor sigmaH in Streptomyces coelicolor A3(2). *Arch Microbiol* **186**(6): 435-446.

McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V. and Lawrence, C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*. **29**: 774-82.

McGuire, A.M., Hughes, J.D. and Church, G.M. (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res*. **10**: 744–57

Medigue, C., Rouxel, T., Vigier, P., Henaut, A. and Danchin, A. (1991) Evidence for horizontal gene transfer in Escherichia coli speciation. *J Mol Biol* **222**(4): 851-856.

Mejean, V. and Claverys, J. P. (1993) DNA processing during entry in transformation of Streptococcus pneumoniae. *J Biol Chem* **268**(8): 5594-5599.

Merkl, R. (2004) SIGI: score-based identification of genomic islands. *BMC Bioinformatics* **5**: 22.

Minic, Z., Marie, C., Delorme, C., Faurie, J. M., Mercier, G., Ehrlich, D. and Renault, P. (2007) Control of EpsE, the phosphoglycosyltransferase initiating exopolysaccharide synthesis in Streptococcus thermophilus, by EpsD tyrosine kinase. *J Bacteriol* **189**(4): 1351-1357.

Mira, A., Ochman, H. and Moran, N. A. (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17**(10): 589-596.

Mirkin, B. G., Fenner, T. I., Galperin, M. Y. and Koonin, E. V. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* **3**: 2.

Mitchell, J. E., Zheng, D., Busby, S. J. and Minchin, S. D. (2003) Identification and analysis of 'extended -10' promoters in Escherichia coli. *Nucleic Acids Res* **31**(16): 4689-4695.

Mitrophanov, A. Y. and Borodovsky, M. (2006) Statistical significance in biological sequence analysis. *Brief Bioinform* **7**(1): 2-24.

Morgenstern, B., Dress, A. and Werner, T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc Natl Acad Sci U S A* **93**(22): 12098-12103.

Munch, R., Hiller, K., Barg, H., Heldt, D., Linz, S., Wingender, E. and Jahn, D. (2003) PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res* **31**(1): 266-269.

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

Muto, A. and Osawa, S. (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci U S A* **84**(1): 166-169.

Mwangi, M.M. and Siggia, E.D. (2003) Genome wide identification of regulatory motifs in Bacillus subtilis. *BMC Bioinformatics*. **16**: 4-18.

Naito, A., Naito, S. and Ikeda, H. (1984) Homology is not required for recombination mediated by DNA gyrase of Escherichia coli. *Mol Gen Genet* **193**(2): 238-243.

Nakamura, Y., Itoh, T., Matsuda, H. and Gojobori, T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36**(7): 760-766.

Narasimhan, C., LoCascio, P. and Uberbacher, E. (2003) Background rareness-based iterative multiple sequence alignment algorithm for regulatory element detection. *Bioinformatics*, **19**(15): 1952–1963.

Naughton, B. T., Fratkin, E., Batzoglou, S. and Brutlag, D. L. (2006) A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites. *Nucleic Acids Res* **34**(20): 5730-5739.

Navarre, W. W., McClelland, M., Libby, S. J. and Fang, F. C. (2007) Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. *Genes Dev* **21**(12): 1456-1471.

Nester, E. W. and Stocker, B. A. (1963) Biosynthetic Latency in Early Stages of Deoxyribonucleic Acidtransformation in Bacillus Subtilis. *J Bacteriol* **86**: 785-796.

Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S. D., Prum, B. and Bessieres, P. (2002) Mining Bacillus subtilis chromosome heterogeneities using hidden Markov models. *Nucleic Acids Res* **30**(6): 1418-1426.

Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Sys.* **8**: 581-599.

Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *Proc. Natl Acad. Sci. USA* **83**: 4-8.

Ochman, H., Lawrence, J. G. and Groisman, E. A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**(6784): 299-304.

Ochman, H. and Selander, R. K. (1984) Evidence for clonal population structure in Escherichia coli. *Proc Natl Acad Sci U S A* **81**(1): 198-201.

Ohnishi, Y., Ishikawa, J., Hara, H., Suzuki, H., Ikenoya, M., Ikeda, H., Yamashita, A., Hattori, M. and Horinouchi, S. (2008) Genome sequence of the streptomycin-producing microorganism Streptomyces griseus IFO 13350. *J Bacteriol* **190**(11): 4050-4060.

Omura, J. K. (1979) On the Viterbi Algorithm. *IEEE Trans. IT.* **Jan.**: 177-179.

Osada, R., Zaslavsky, E. and Singh, M. (2004) Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics* **20**: 3516-3525.

Omura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M., Takahashi, Y., Horikawa, H., Nakazawa, H., Osonoe, T., Kikuchi, H., Shiba, T., Sakaki, Y. and Hattori, M. (2001) Genome sequence of an industrial microorganism Streptomyces avermitilis: deducing the ability of producing secondary metabolites. *Proc Natl Acad Sci U S A* **98**(21): 12215-12220.

O'Sullivan, O., O'Callaghan, J., Sangrador-Vegas, A., McAuliffe, O., Slattery, L., Kaleta, P., Callanan, M., Fitzgerald, G. F., Ross, R. P. and Beresford, T. (2009) Comparative genomics of lactic acid bacteria reveals a niche-specific gene set. *BMC Microbiol* **9**: 50.

Roberts, A.P. and Mullany, P. (2000) Genetic basis of horizontal gene transfer among oral bacteria. *Periodontology* **42**: 36–46.

Page, R. D. and Charleston, M. A. (1997) From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol* **7**(2): 231-240.

Paget, M. S., Kang, J. G., Roe, J. H. and Buttner, M. J. (1998) sigmaR, an RNA polymerase sigma factor that modulates expression of the thioredoxin system in response to oxidative stress in Streptomyces coelicolor A3(2). *EMBO J* **17**(19): 5776-5782.

Paget, M. S., Chamberlin, L., Atrih, A., Foster, S. J. and Buttner, M. J. (1999) Evidence that the extracytoplasmic function sigma factor sigmaE is required for normal cell wall structure in Streptomyces coelicolor A3(2). *J Bacteriol* **181**(1): 204-211.

Paget, M. S., Leibovitz, E. and Buttner, M. J. (1999) A putative two-component signal transduction system regulates sigmaE, a sigma factor required for normal cell wall integrity in Streptomyces coelicolor A3(2). *Mol Microbiol* **33**(1): 97-107.

Paget, M. S., Molle, V., Cohen, G., Aharonowitz, Y. and Buttner, M. J. (2001) Defining the disulphide stress response in Streptomyces coelicolor A3(2): identification of the sigmaR regulon. *Mol Microbiol* **42**(4): 1007-1020.

Paget, M. S. B., Hong, H.-J., Bibb, M. J. and Buttner, M. J. (2002) The ECF sigma factors of Streptomyces coelicolor A3(2). In, Signals, Switches, Regulons and Cascades: Control of Bacterial Gene Expression. Hodgson, D.A. and Thomas, C.M. (eds). . *Cambridge University Press*: 105-125.

Partridge, S. M. (2000) Prosthetic valve endocarditis due to Streptococcus vestibularis. *J Infect* **41**(3): 284-285.

Pavesi, G., Mauri, G. and Pesole, G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* **17 Suppl 1**: S207-214.

Pavesi, G., Mauri, G. and Pesole, G. (2004) In silico representation and discovery of transcription factor binding sites **5**(3): 217-236.

Pavlovic, G., Burrus, V., Gintz, B., Decaris, B. and Guedon, G. (2004) Evolution of genomic islands by deletion and tandem accretion by site-specific recombination: ICESt1-related elements from Streptococcus thermophilus. *Microbiology* **150**: 759-774.

Pavlovic, G., Burrus, V., Toulmay, A., Choulet, F., Decaris, B. and Guedon, G. (2004) Characterization and evolution of a family of integrative and potentially conjugative or mobilizable elements from Streptococcus thermophilus. *Le Lait* **84**: 7-14.

Pearson, W. R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11**(3): 635-650.

Pesole, G., Prunella, N., Liuni, S., Attimonelli, M. and Saccone, C. (1992) WORDUP: an efficient algorithm for discovering statistically significant patterns in DNA sequences. *Nucleic Acids Res* **20**(11): 2871-2875.

Petersen, L., Larsen, T.S., Ussery, D.W., On, S.L. and Krogh, A. (2003) RpoD promoters in Campylobacter jejuni exhibit a strong periodic signal instead of a -35 box. *J Mol Biol.* **326**: 1361-72.

Pevzner, P. A., Borodovsky, M. and Mironov, A. A. (1989) Linguistics of nucleotide sequences. I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J Biomol Struct Dyn* **6**(5): 1013-1026.

Pevzner, P. A. and Sze, S. H. (2000) Combinatorial approaches to finding subtle signals in DNA sequences. *Proc Int Conf Intell Syst Mol Biol* **8**: 269-278.

Pitcher, R. S., Brissett, N. C. and Doherty, A. J. (2007) Nonhomologous end-joining in bacteria: a microbial perspective. *Annu Rev Microbiol* **61**: 259-282.

Pitcher, R. S., Wilson, T. E. and Doherty, A. J. (2005) New insights into NHEJ repair processes in prokaryotes. *Cell Cycle* **4**(5): 675-678.

Podell, S. and Gaasterland, T. (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol* **8**(2): R16.

Podell, S., Gaasterland, T. and Allen, E. E. (2008) A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. *BMC Bioinformatics* **9**: 419.

Posada, D. and Crandall, K. A. (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol Evol* **16**(1): 37-45.

Potuckova, L., Kelemen, G. H., Findlay, K. C., Lonetto, M. A., Buttner, M. J. and Kormanec, J. (1995) A new RNA polymerase sigma factor, sigma F, is required for the late stages of morphological differentiation in Streptomyces spp. *Mol Microbiol* **17**(1): 37-48.

Poyart, C., Quesne, G., Coulon, S., Berche, P. and Trieu-Cuot, P. (1998) Identification of streptococci to species level by sequencing the gene encoding the manganese-dependent superoxide dismutase. *J Clin Microbiol* **36**(1): 41-47.

Prakash, A., Blanchette, M., Sinha, S. and Tompa, M. (2004) Motif discovery in heterogeneous sequence data. *Pac Symp Biocomput*: 348-359.

Prangishvilia, D., Garrettb, R. A. and Kooninc, E. V. (2006) Evolutionary genomics of archaeal viruses: Unique viral genomes in the third domain of life This article is dedicated to the memory of Wolfram Zillig, the pioneer of the study of archaeal viruses. *Virus Research* **117**(1): 52-67.

Pride, D. T. and Blaser, M. J. (2002) Identification of horizontally acquired genetic elements in Helicobacter pylori and other prokaryotes using oligonucleotide difference analysis. *Genome Lett.* **1**: 2-15.

Puigbo, P., Bravo, I. G. and Garcia-Vallve, S. (2008) CAIcal: a combined set of tools to assess codon usage adaptation. *Biol Direct* **3**: 38.

Rabiner, L. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE* **77**(2).

Rajewsky, N., Socci, N.D., Zapotocky, M. and Siggia, E.D. (2002) The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res.* **12**: 298-308.

Rallu, F., Taillez, P., Ehrlich, S. D. and Renault, P. (2002) Common scheme of evolution between eps clusters of the food bacteria Streptococcus thermophilus and cps clusters of the pathogenic streptococci. *Proceedings of the 6th American Society of Microbiology Conference on Streptococcal Genetics*: 112.

Rangannan, V. and Bansal, M. (2007) Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability. *J Biosci.* **32**(5): 851-62.

Rasmussen, T. B., Danielsen, M., Valina, O., Garrigues, C., Johansen, E. and Pedersen, M. B. (2008) Streptococcus thermophilus core genome: comparative genome hybridization study of 47 strains. *Appl Environ Microbiol* **74**(15): 4703-4710.

Rigali, S., Titgemeyer, F., Barends, S., Mulder, S., Thomae, A. W., Hopwood, D. A. and van Wezel, G. P. (2008) Feast or famine: the global regulator DasR links nutrient stress to antibiotic production by Streptomyces. *EMBO Rep* **9**(7): 670-675.

Riley, M. and Labedan, B. (1997) Protein evolution viewed through Escherichia coli protein sequences: introducing the notion of a structural segment of homology, the module. *J Mol Biol* **268**(5): 857-868.

Roberts, M. S. and Cohan, F. M. (1993) The effect of DNA sequence divergence on sexual isolation in Bacillus. *Genetics* **134**(2): 401-408.

Robison, K., McGuire, A.M. and Church, G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. *J Mol Biol.* **284**: 241-54.

Robin, S., Daudin, J.-J., Richard, H., Sagot, M.-F. and Schbath, S. (2002) Occurrence probability of structured motifs in random sequences. *J Comp Biol.* **9**: 761-773.

Roca, A. I. and Cox, M. M. (1997) RecA protein: structure, function, and role in recombinational DNA repair. *Prog Nucleic Acid Res Mol Biol* **56**: 129-223.

Rocha, E. P. and Danchin, A. (2002) Base composition bias might result from competition for metabolic resources. *Trends Genet* **18**(6): 291-294.

Ross, W., Gosink, K. K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K. and Gourse, R. L. (1993) A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science* **262**(5138): 1407-1413.

Ruepp, A., Graml, W., Santos-Martinez, M. L., Koretke, K. K., Volker, C., Mewes, H. W., Frishman, D., Stocker, S., Lupas, A. N. and Baumeister, W. (2000) The genome sequence of the thermoacidophilic scavenger Thermoplasma acidophilum. *Nature* **407**(6803): 508-513.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A and Barrell, B. (2000) Artemis: sequence visualisation and annotation. *Bioinformatics*. **16**: 944-945.

Ryding, N.J., Kelemen, G.H., Whatling, C.A., Flärdh, K., Buttner, M.J. and Chater, K.F. (1998) A developmentally regulated gene encoding a repressor-like protein is essential for sporulation in Streptomyces coelicolor A3(2). *Mol Microbiol.* **29**: 343-357.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**(4): 406-425.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjolander, K., Underwood, R. C. and Haussler, D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res* **22**(23): 5112-5120.

Salzberg, S. L., Delcher, A. L., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **26**(2): 544-548.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**(Database issue): D91-94.

Sanger, F., Donelson, J. E., Coulson, A. R., Kossel, H. and Fischer, D. (1974) Determination of a nucleotide sequence in bacteriophage f1 DNA by primed synthesis with DNA polymerase. *J Mol Biol* **90**(2): 315-333.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**(12): 5463-5467.

Saunders, N. J., Boonmee, P., Peden, J. F. and Jarvis, S. A. (2005) Inter-species horizontal transfer resulting in core-genome and niche-adaptive variation within Helicobacter pylori. *BMC Genomics* **6**(1): 9.

Sawers, G., Kaiser, M., Sirko, A. and Freundlich, M. (1997) Transcriptional activation by FNR and CRP: reciprocity of binding-site recognition. *Mol Microbiol* **23**(4): 835-845.

Schilling, C. H., Held, L., Torre, M. and Saier, M. H., Jr. (2000) GRASP-DNA: a web application to screen prokaryotic genomes for specific DNA-binding sites and repeat motifs. *J Mol Microbiol Biotechnol* **2**(4): 495-500.

Schleifer, K. H., Ehrmann, M., Krusch, U. and Neve, H. (1991) Revival of the Species Streptococcus thermophilus (Ex Orla-Jensen, 1919). *Nom Rev. Syst. Appl. Microbiol* **14**: 386–388.

Schneider, T. D. and Stephens, R. M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**(20): 6097-6100.

Schultz, S. C., Shields, G. C. and Steitz, T. A. (1991) Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science* **253**(5023): 1001-1007.

Segal, E. and Sharan, R. (2005) A discriminative model for identifying spatial cis-regulatory modules. *J Comput Biol.* **12**: 822-834.

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N. and Whittam, T. S. (1986) Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl Environ Microbiol* **51**(5): 873-884.

Semenova, E., Nagornykh, M., Pyatnitskiy, M., Artamonova, I. and Severinov, K. (2009) Analysis of CRISPR system function in plant pathogen Xanthomonas oryzae. *FEMS microbiology letters.*

Servin-Gonzalez, L., Jensen, M. R., White, J. and Bibb, M. (1994) Transcriptional regulation of the four promoters of the agarase gene (dagA) of Streptomyces coelicolor A3(2). *Microbiology* **140(10)**: 2555-2565.

Sharp, P. M. and Li, W. H. (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* **24**(1-2): 28-38.

Sharp, P. M. and Li, W. H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**(3): 1281-1295.

Sharp, P. M., Tuohy, T. M. and Mosurski, K. R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14**(13): 5125-5143.

Shimizu, H., Yamaguchi, H., Ashizawa, Y., Kohno, Y., Asami, M., Kato, J.-I. and Ikeda, H. (1997) Short-homology-independent Illegitimate Recombination in Escherichia coli: Distinct Mechanism from Short-homology-dependent Illegitimate Recombination. *J. Mol. Biol.* **266**: 297-305.

Sicheritz-Ponten, T. and Andersson, S. G. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res* **29**(2): 545-552.

Siddharthan, R., Siggia, E. D. and van Nimwegen, E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* **1**(7): e67.

Siepel, A. and Haussler, D. (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol* **11**(2-3): 413-428.

Sierro, N., Makita, Y., de Hoon, M. and Nakai, K. (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res* **36**(Database issue): D93-96.

Siezen, R. J., Kuipers, O. P. and de Vos, W. M. (1996) Comparison of lantibiotic gene clusters and encoded proteins. *Antonie Van Leeuwenhoek* **69**(2): 171-184.

Siezen, R. J., van Enckevort, F. H., Kleerebezem, M. and Teusink, B. (2004) Genome data mining of lactic acid bacteria: the impact of bioinformatics. *Curr Opin Biotechnol* **15**(2): 105-115.

Sinha, S., Blanchette, M. and Tompa, M. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* **5**: 170.

Sinha, S. and Tompa, M. (2000) A statistical method for finding transcription factor binding sites. *Proc Int Conf Intell Syst Mol Biol* **8**: 344-354.

Smith, G.R. (1990) RecBCD enzyme. *Nucleic Acids and Molecular Biology* **4**: 78–98.

Smith, J.M., Dowson, C.G. and Spratt, B.G. (1991) Localized sex in bacteria. *Nature* **349**(6304): 29-31.

Smith, M.P. and Feiss, M. (1993) Sites and gene products involved in lambdoid phage DNA packaging. *J Bacteriol.* **175**(8): 2393-2399.

Smith, T.F., and Waterman, M.S. (1981) Identification of Common Molecular Subsequences. *Journal of Molecular Biology* **147**: 195-197.

Soding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**(7): 951-960.

Soliveri, J. A., Gomez, J., Bishai, W. R. and Chater, K. F. (2000) Multiple paralogous genes related to the Streptomyces coelicolor developmental regulatory gene whiB are present in Streptomyces and other actinomycetes. *Microbiology* **146**: 333-343.

Solomon, J. M. and Grossman, A. D. (1996) Who's competent and when: regulation of natural genetic competence in bacteria. *Trends Genet* **12**(4): 150-155.

Sonnhammer, E. L., Eddy, S. R. and Durbin, R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**(3): 405-420.

Sonnhammer, E. L., von Heijne, G. and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6**: 175-182.

Sorensen, S. J., Bailey, M., Hansen, L. H., Kroer, N. and Wuertz, S. (2005) Studying plasmid horizontal transfer in situ: a critical review. *Nat Rev Microbiol* **3**(9): 700-710.

Stewart, V., Landick, R. and Yanofsky, C. (1986) Rho-dependent transcription termination in the tryptophanase operon leader region of Escherichia coli K-12. *J Bacteriol* **166**(1): 217-223.

Stolcke, A. and Omohundro, S. (1993) Hidden Markov Model induction by bayesian model merging. *Neural Information Processing Systems* **5**: 11–18.

Stormo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics* **16**: 16-23.

Studholme, D. J., Bentley, S. D. and Kormanec, J. (2004) Bioinformatic identification of novel regulatory DNA sequence motifs in Streptomyces coelicolor. *BMC Microbiol* **4**: 14.

Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U S A* **85**(8): 2653-2657.

Syvanen, M. (1994) Horizontal gene transfer: evidence and possible consequences. *Annu Rev Genet* **28**: 237-261.

Takano, H., Obitsu, S., Beppu, T. and Ueda, K. (2005) Light-induced carotenogenesis in Streptomyces coelicolor A3(2): identification of an extracytoplasmic function sigma factor that directs photodependent transcription of the carotenoid biosynthesis gene cluster. *J Bacteriol* **187**(5): 1825-1832.

Tan, H. and Chater, K. F. (1993) Two developmentally controlled promoters of Streptomyces coelicolor A3(2) that resemble the major class of motility-related promoters in other bacteria. *J Bacteriol* **175**(4): 933-940.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. and Koonin, E. V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**(1): 22-28.

Than, C., Ruths, D. and Nakhleh, L. (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* **9**: 322.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P. and Moreau, Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17**: 1113-1122.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**(22): 4673-4680.

Thompson, J. D., Plewniak, F. and Poch, O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15**(1): 87-88.

Tompa, M. (1999) An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *Proc Int Conf Intell Syst Mol Biol*: 262-271.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. and Zhu, Z. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**(1): 137-144.

Touzain, F., Schbath, S., Debled-Rennesson, I., Aigle, B., Kucherov, G. and Leblond, P. (2008) SIGffRid: a tool to search for sigma factor binding sites in bacterial genomes using comparative approach and biologically driven statistics. *BMC Bioinformatics* **9**: 73.

Touzet, H. and Varre, J. S. (2007) Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol Biol* **2**: 15.

Towsey, M. W., Gordon, J. J. and Hogan, J. M. (2006) The prediction of bacterial transcription start sites using SVMs. *Int J Neural Syst* **16**(5): 363-370.

Tracy, R. B., Chedin, F. and Kowalczykowski, S. C. (1997) The recombination hot spot chi is embedded within islands of preferred DNA pairing sequences in the E. coli genome. *Cell* **90**(2): 205-206.

Tsirigos, A. and Rigoutsos, I. (2005) A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res* **33**(3): 922-933.

Tsirigos, A. and Rigoutsos, I. (2005) A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. *Nucleic Acids Res* **33**(12): 3699-3707.

Turner, K. M., Hanage, W. P., Fraser, C., Connor, T. R. and Spratt, B. G. (2007) Assessing the reliability of eBURST using simulated populations with known ancestry. *BMC Microbiol* **7**: 30.

Typas, A. and Hengge, R. (2005) Differential ability of σS and σ70 of Escherichia coli to utilize promoters containing half or full UP-element sites. *Mol. Microbiol.* **55**: 250-260.

Umezu, K., Chi, N. W. and Kolodner, R. D. (1993) Biochemical interaction of the Escherichia coli RecF, RecO, and RecR proteins with RecA protein and single-stranded DNA binding protein. *Proc Natl Acad Sci U S A* **90**(9): 3875-3879.

Umezu, K. and Kolodner, R. D. (1994) Protein interactions in genetic recombination in Escherichia coli. Interactions involving RecO and RecR overcome the inhibition of RecA by single-stranded DNA-binding protein. *J Biol Chem* **269**(47): 30005-30013.

Umezu, K., Nakayama, K. and Nakayama, H. (1990) Escherichia coli RecQ protein is a DNA helicase. *Proc Natl Acad Sci U S A* **87**(14): 5363-5367.

van de Guchte, M., Penaud, S., Grimaldi, C., Barbe, V., Bryson, K., Nicolas, P., Robert, C., Oztas, S., Mangenot, S., Couloux, A., Loux, V., Dervyn, R., Bossy, R., Bolotin, A., Batto, J. M., Walunas, T., Gibrat, J. F., Bessieres, P., Weissenbach, J., Ehrlich, S. D. and Maguin, E. (2006) The complete genome sequence of Lactobacillus bulgaricus reveals extensive and ongoing reductive evolution. *Proc Natl Acad Sci U S A* **103**(24): 9274-9279.

van Helden, J., Andre, B. and Collado-Vides, J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* **281**(5): 827-842.

van Helden, J., Rios, A. F. and Collado-Vides, J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* **28**(8): 1808-1818.

Vanet, A., Marsan, L., Labigne, A. and Sagot, M.F. (2000) Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori sigma(80) family of promoter signals. *J Mol Biol.* **297**: 335-53.

Vernikos, G. S. and Parkhill, J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics* **22**(18): 2196-2203.

Vierling, S., Weber, T., Wohlleben, W. and Muth, G. (2001) Evidence that an additional mutation is required to tolerate insertional inactivation of the Streptomyces lividans recA gene. *J Bacteriol* **183**(14): 4374-4381.

Viterbi, A. J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. IT* **13**: 260-269.

Volff, J. N. and Altenbuchner, J. (2000) A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol Lett* **186**(2): 143-150.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., Surovcik, K., Meinicke, P. and Merkl, R. (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* **7**: 142.

Wang, T., and Stormo, G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19**(18): 2369-80.

Ward, J. H. (1963) Hierarchical Grouping to optimize an objective function. *Journal of American Statistical Association* **58**: 236-244.

Ward, N. L., Challacombe, J. F., Janssen, P. H., Henrissat, B., Coutinho, P. M., Wu, M., Xie, G., Haft, D. H., Sait, M., Badger, J., Barabote, R. D., Bradley, B., Brettin, T. S., Brinkac, L. M., Bruce, D., Creasy, T., Daugherty, S. C., Davidsen, T. M., DeBoy, R. T., Detter, J. C., Dodson, R. J., Durkin, A. S., Ganapathy, A., Gwinn-Giglio, M., Han, C. S., Khouri, H., Kiss, H., Kothari, S. P., Madupu, R., Nelson, K. E., Nelson, W. C., Paulsen, I., Penn, K., Ren, Q., Rosovitz, M. J., Selengut, J. D., Shrivastava, S., Sullivan, S. A., Tapia, R., Thompson, L. S., Watkins, K. L., Yang, Q., Yu, C., Zafar, N., Zhou, L. and Kuske, C. R. (2009) Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils. *Appl Environ Microbiol* **75**(7): 2046-2056.

Watt, V. M., Ingles, C. J., Urdea, M. S. and Rutter, W. J. (1985) Homology requirements for recombination in Escherichia coli. *Proc Natl Acad Sci U S A* **82**(14): 4768-4772.

Wei, Z. and Jensen, S. T. (2006) GAME: detecting cis-regulatory elements using a genetic algorithm. *Bioinformatics* **22**(13): 1577-1584.

Weller, G. R. and Doherty, A. J. (2001) A family of DNA repair ligases in bacteria? *FEBS Lett* **505**(2): 340-342.

Weller, G. R., Kysela, B., Roy, R., Tonkin, L. M., Scanlan, E., Della, M., Devine, S. K., Day, J. P., Wilkinson, A., d'Adda di Fagagna, F., Devine, K. M., Bowater, R. P., Jeggo, P. A., Jackson, S. P. and Doherty, A. J. (2002) Identification of a DNA nonhomologous end-joining complex in bacteria. *Science* **297**(5587): 1686-1689.

West, S. C. (1996) DNA helicases: new breeds of translocating motors and molecular pumps. *Cell* **86**(2): 177-180.

West, S. C. (1996) The RuvABC proteins and Holliday junction processing in Escherichia coli. *J Bacteriol* **178**(5): 1237-1241.

Whitby, M. C., Ryder, L. and Lloyd, R. G. (1993) Reverse branch migration of Holliday junctions by RecG protein: a new mechanism for resolution of intermediates in recombination and DNA repair. *Cell* **75**(2): 341-350.

Whittam, T. S., Ochman, H. and Selander, R. K. (1983) Multilocus genetic structure in natural populations of Escherichia coli. *Proc Natl Acad Sci U S A* **80**(6): 1751-1755.

Workman, C. T. and Stormo, G. D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*: 467-478.

Wu, L. J., Lewis, P. J., Allmansberger, R., Hauser, P. M. and Errington, J. (1995) A conjugation-like mechanism for prespore chromosome partitioning during sporulation in Bacillus subtilis. *Genes Dev* **9**(11): 1316-1326.

Yada, T., Totoki, Y., Ishikawa, M., Asai, K. and Nakai, K. (1998) Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences. *Bioinformatics* **14**(4): 317-325.

Yang, Z. and Rannala, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol* **14**(7): 717-724.

Zaneveld, J. R., Nemergut, D. R. and Knight, R. (2008) Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. *Microbiology* **154**(Pt 1): 1-15.

Zhang, G. and Darst, S. A. (1998) Structure of the Escherichia coli RNA polymerase alpha subunit amino-terminal domain. *Science* **281**(5374): 262-266.

Zheng, J., Wu, J. and Sun, Z. (2003) An approach to identify over-represented cis-elements in related sequences. *Nucleic Acids Res* **31**(7): 1995-2005.

Annexe Web : Liste des fichiers accessibles par URL

Les fichiers sont stockés sous forme d'archive dans à l'URL :
http://caths.e-monsite.com/rubrique,these-c-eng-2010,320040.html

Le mot de passe pour désarchiver est : these2010HMM2

Chapitre 2 : Approche hybride stochastique et combinatoire pour la recherche et la modélisation de motifs de régulation génique

- Annexe Web A :
  - Etape de génération d'un HMM d'ordre 2 : Processus d'analyse type.

Chapitre 3 : Utilisation des HMM d'ordre 2 (M2M2) pour la détection de gènes acquis par transfert horizontal, HGT

- Annexe Web B :
  - Fichier d'alignement de séquence des 70 ARNr 16S extraits pour générer l'arbre phylogénétique.
- Annexe Web C :
  - Cinq arbres des espèces de firmicutes générés par la méthode du maximum de parcimonie.
- Annexe Web D :
  - Fichier au format Newick qui a permis de visualiser l'arbre phylogénétique.
- Annexe Web E :
  - Liste des 54 gènes variables du *CGH-set*

Chapitre 4 : Discussion et perspectives

- Annexe Web F (fichier excel) :
  - Liste des 812 motifs composites SFBS potentiels chez *S. coelicolor*.

Annexes

Annexe A

Processus de détermination de la spécificité et de la sensibilité des HMMs pour l'identification de la boîte -35 (GGAAT)

La spécificité et la sensibilité des HMM d'ordre 1 (HMM1) et d'ordre 2 (HMM2) sont calculées sur le jeu de données *set-sco*. L'identification de la boîte -35 (GGAAT) a été réalisée au niveau de la reconnaissance du motif (cf. section 2.3.1, tableau 1).
Dans le *set-sco,* il possible de retrouver 442 motifs GGAAT composé de 216 motifs pour les séquences du brin sens et 226 dans les séquences du brin anti-sens. Dans ces 442 motifs, seuls 25 représentent les vrais sites (VP) et donc les 417 autres motifs sont des VN.
En termes de spécificité les deux modèles sont équivalents pour le HMM2 (0.24) et le HMM1 (0.23) (cf. section 2.3.1, tableau 2). En revanche, la sensibilité du HMM2 est plus importante correspondant à 0.8 contre 0.68 pour le HMM1 et permettra de décrire l'ensemble du régulon. L'identification de tous les *iPeak-motifs* appartenant au régulon est essentielle même si le nombre de faux positifs (FP) augmente. Lors du processus de la fouille de données, les *iPeak-motifs* similaires sont regroupés et les FP seront disséminés dans les classes et ces derniers auront une probabilité plus faible de générer un consensus de classe.

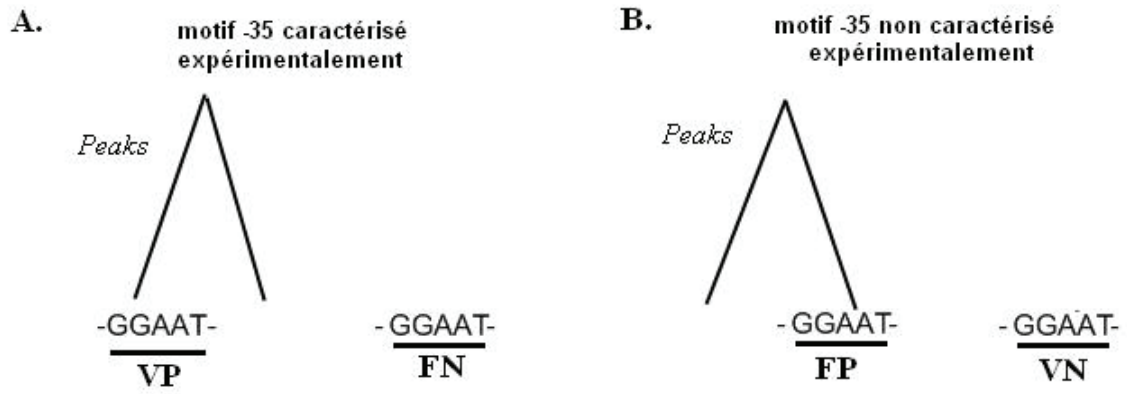


Figure 1 : Représentation de l'identification de la boîte -35 (GGAAT) du facteur σ^R.

VP (Vrai Positif), un vrai positif correspond à une boîte -35 caractérisée qui a été identifiée par les HMM.

FN (faux négatif), un faux négatif est défini par une boîte -35 caractérisée qui n'a pas été identifiée par les HMM.

FP (faux positif), un faux négatif représente une boîte -35 non caractérisée mais qui a été identifiée par les HMM.

VN (vrai négatif), un vrai négatif correspond à une boîte -35 non caractérisée et qui n'est pas identifié par le HMM.

A. Cette configuration représente le motif GGAAT correspondant à un site caractérisé expérimentalement du facteur σ^R. Si un *Peak* est observé sous un motif caractérisé alors il sera noté vrai positif (VP), c'est-à-dire que l'élément est identifié par les HMM alors qu'il le devait. Sinon si aucun *Peak* n'est observé sous un motif GGAAT caractérisé, c'est un FN.

B. Le motif GGAAT ne correspond pas à un site de reconnaissance du facteurσ^R c'est-à-dire que ce motif n'est pas localisé dans la région promotrice d'un gène régulé par ce facteur. Dans le cas où ce motif non caractérisé est identifié par les HMM au travers de l'apparition d'un pic de probabilité *a posteriori* correspondant à un *Peak*, alors ce motif sera considéré comme un FP. Dans l'autre cas, où le motif non caractérisé n'est pas identifié par les HMM, il est comptabilisé comme un VN.

Annexe B

Stratégie de recherche de SFBS : détail des étapes d'exécution des programmes

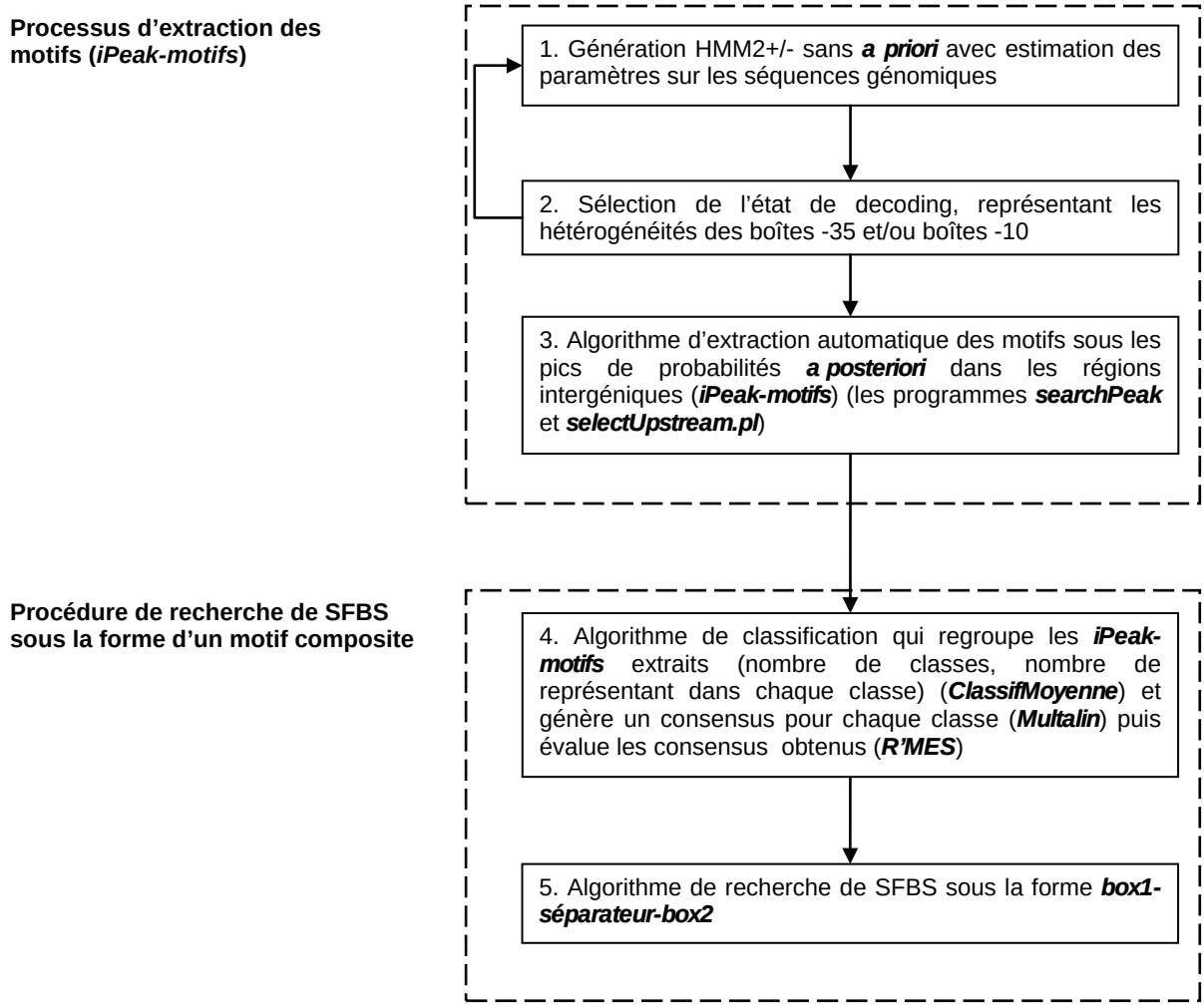
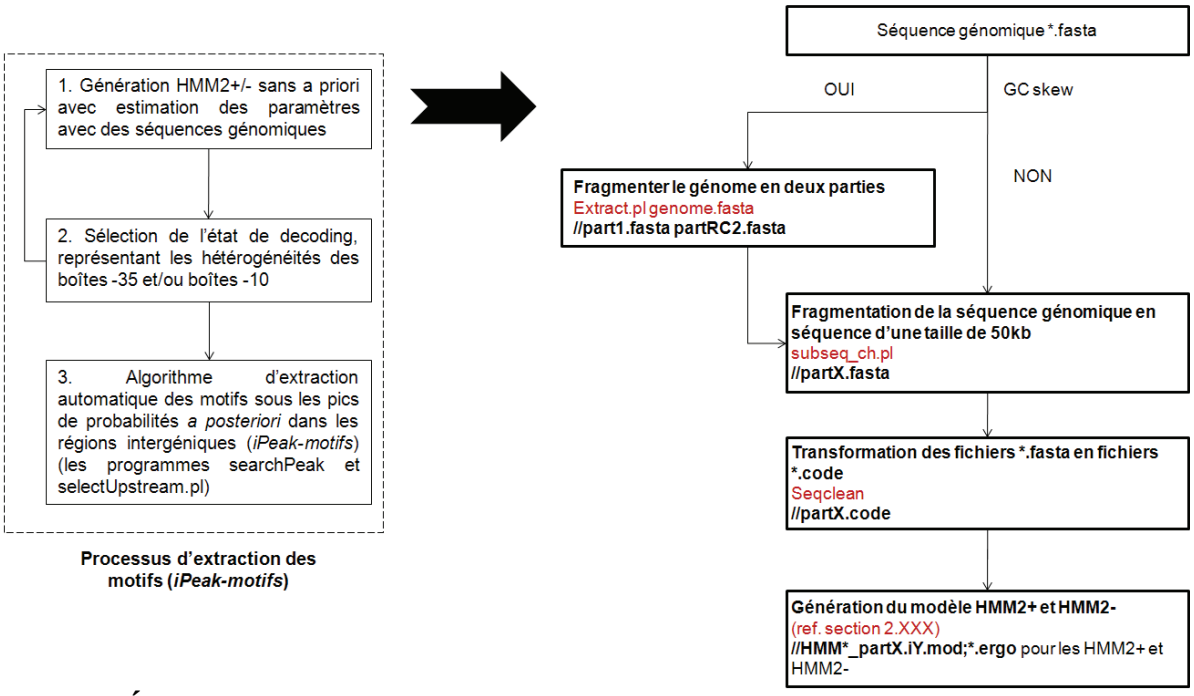
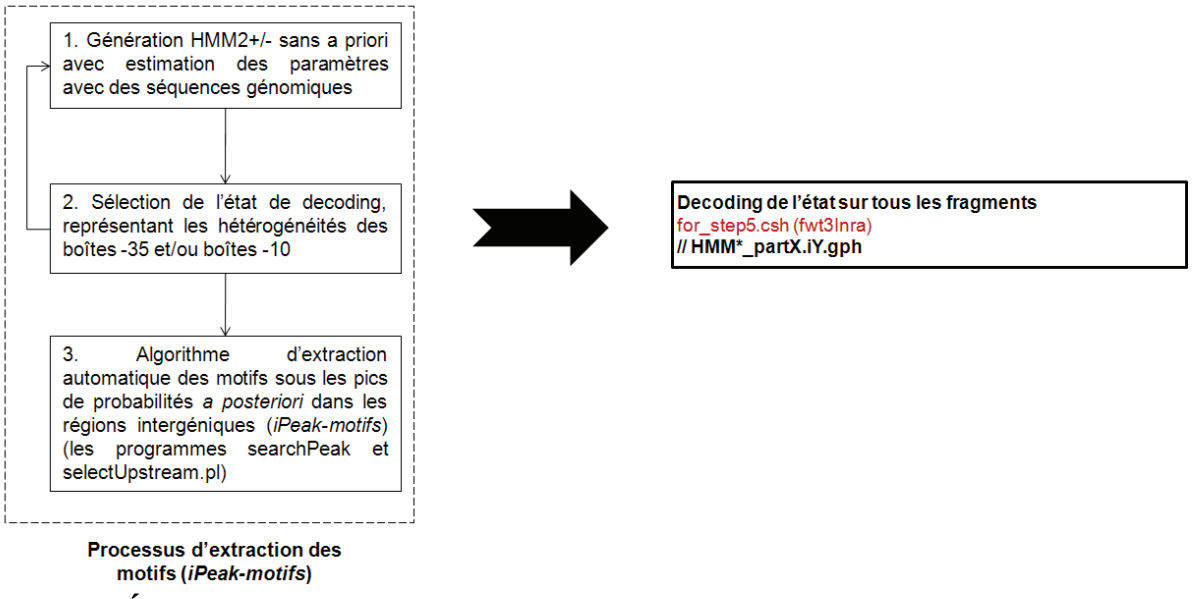


Figure 1 : Stratégie générale de recherche de SFBS



Processus d'extraction des motifs (*iPeak-motifs*)

Figure 2 : Étape 1



Processus d'extraction des motifs (*iPeak-motifs*)

Figure 3 : Étape 2

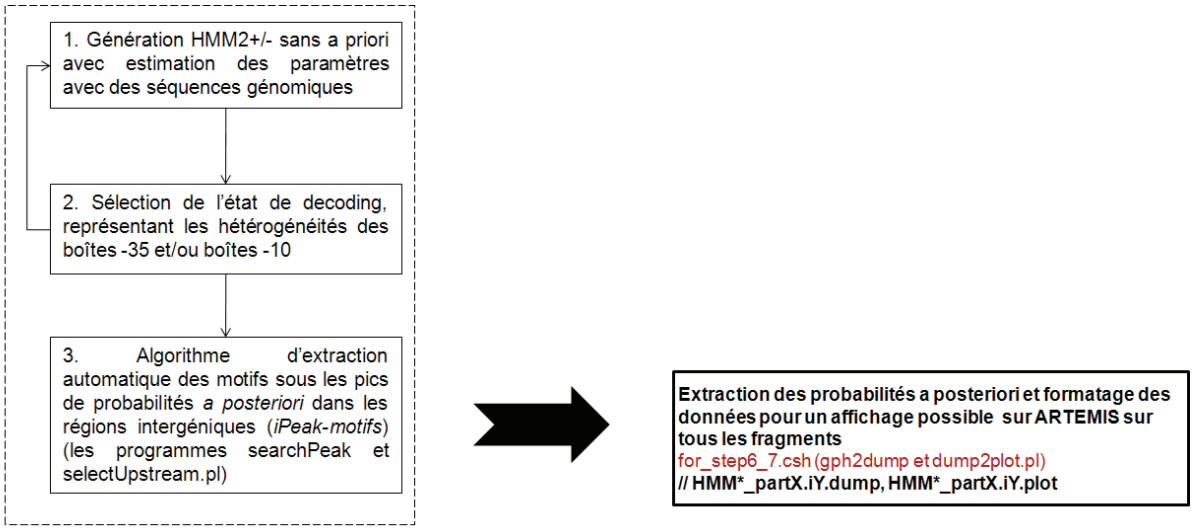


Figure 4 : Étape 3

Les programmes *searchPeaks.c* + *selectUpstream.pl* sont des algorithmes d'extraction automatique des pics (Algorithme 1, chapitre 2).

Le programme *searchPeaks.c* extrait les pics d'une longueur et d'une hauteur données qui dépendent des paramètres des HMM2 utilisés

Le programme *selectUpstream.pl* sélectionne les *iPeak-motifs* selon leur localisation (région codante ou intergénique) et leur longueur.

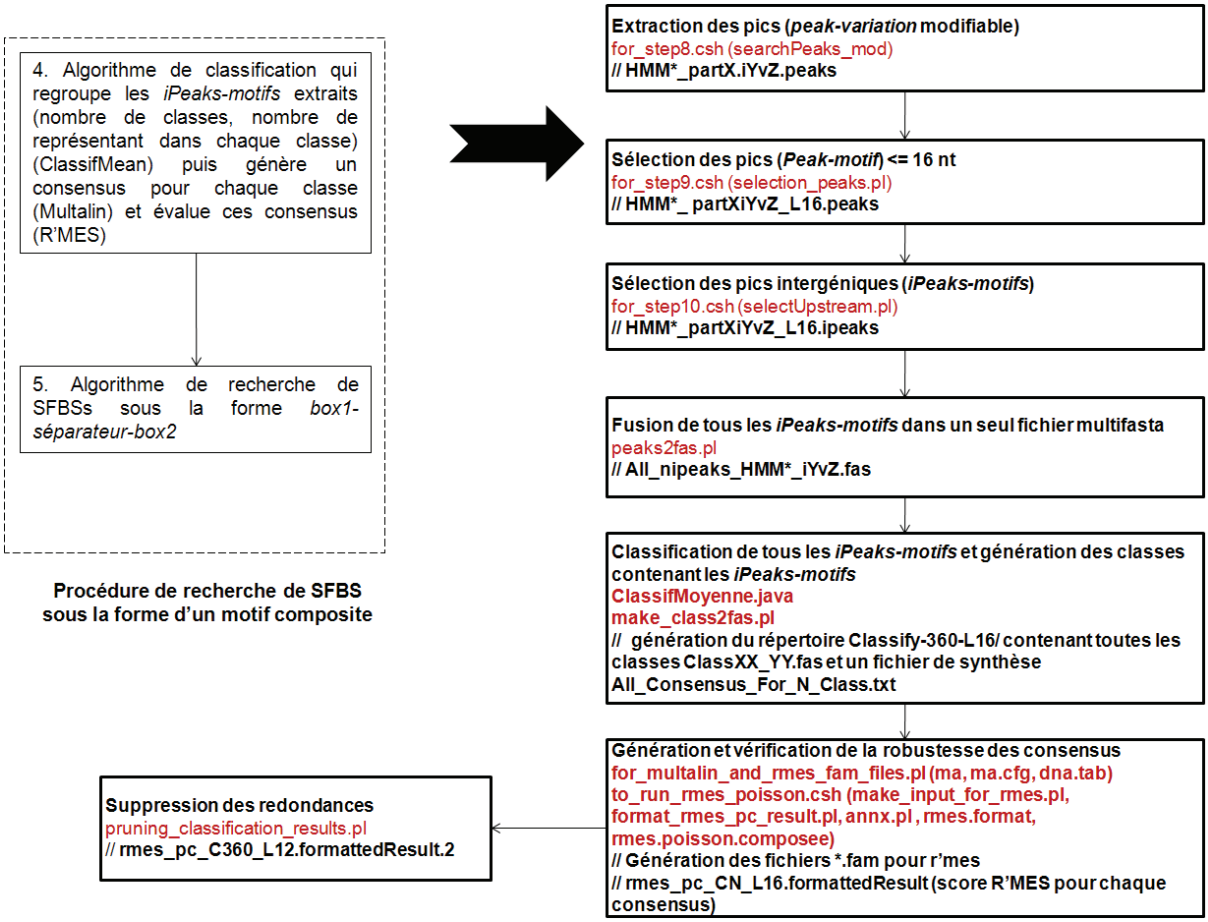


Figure 5 : Étape 4

L'algorithme de classification (*clusterMean*) parvient à regrouper efficacement un ensemble de séquences, ici les *iPeak-motifs*, en classes conduisant à des consensus robustes.

Le programme *ClassifMoyenne.java* regroupe les *iPeak-motifs*, en utilisant une méthode de classification hiérarchique établie par le calcul de distance moyenne entre les classes.

Le programme *MultAlin* (Corpet, 1988) recherche pour chaque classe, le consensus défini par les *iPeak-motifs* appartenant à la classe.

Le programme *R'MES* (Hoebeke et Schbath, 2006) donne un score statistique d'exceptionnalité pour chaque consensus de classe obtenu par *MultAlin*.

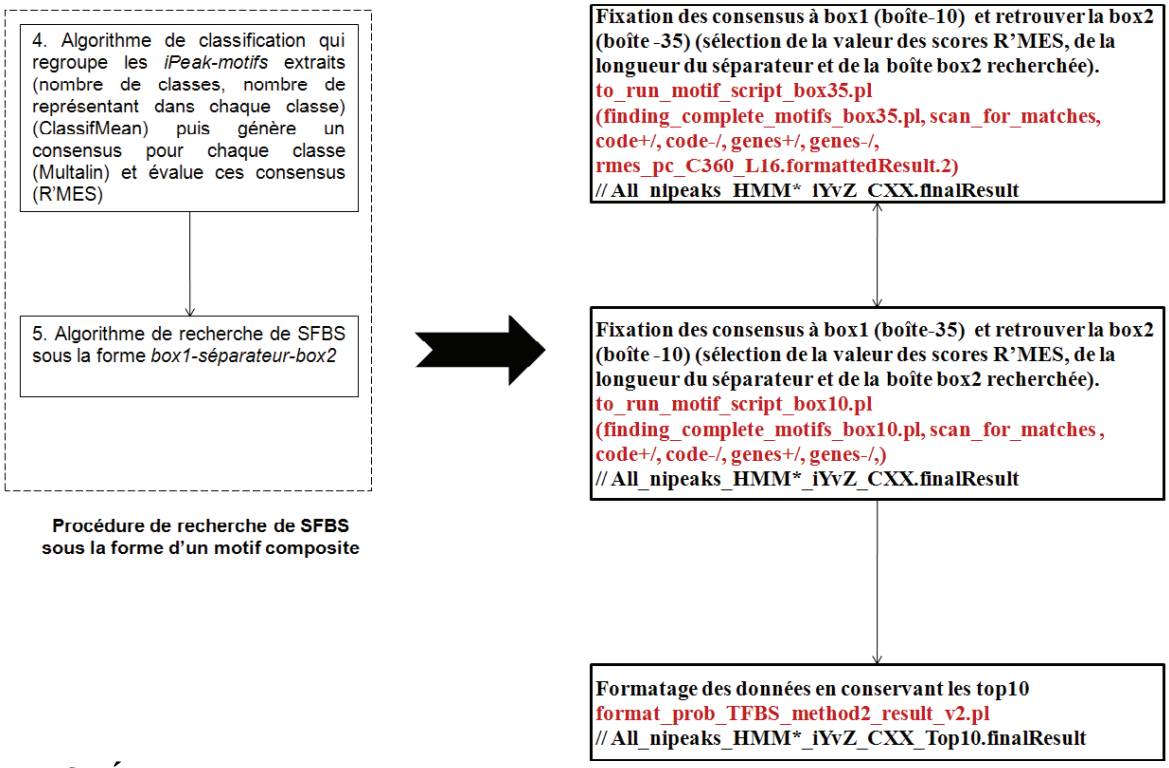


Figure 6 : Étape 5

Annexe C

Processus d'identification des potentiels SFBS de *B. subtilis*

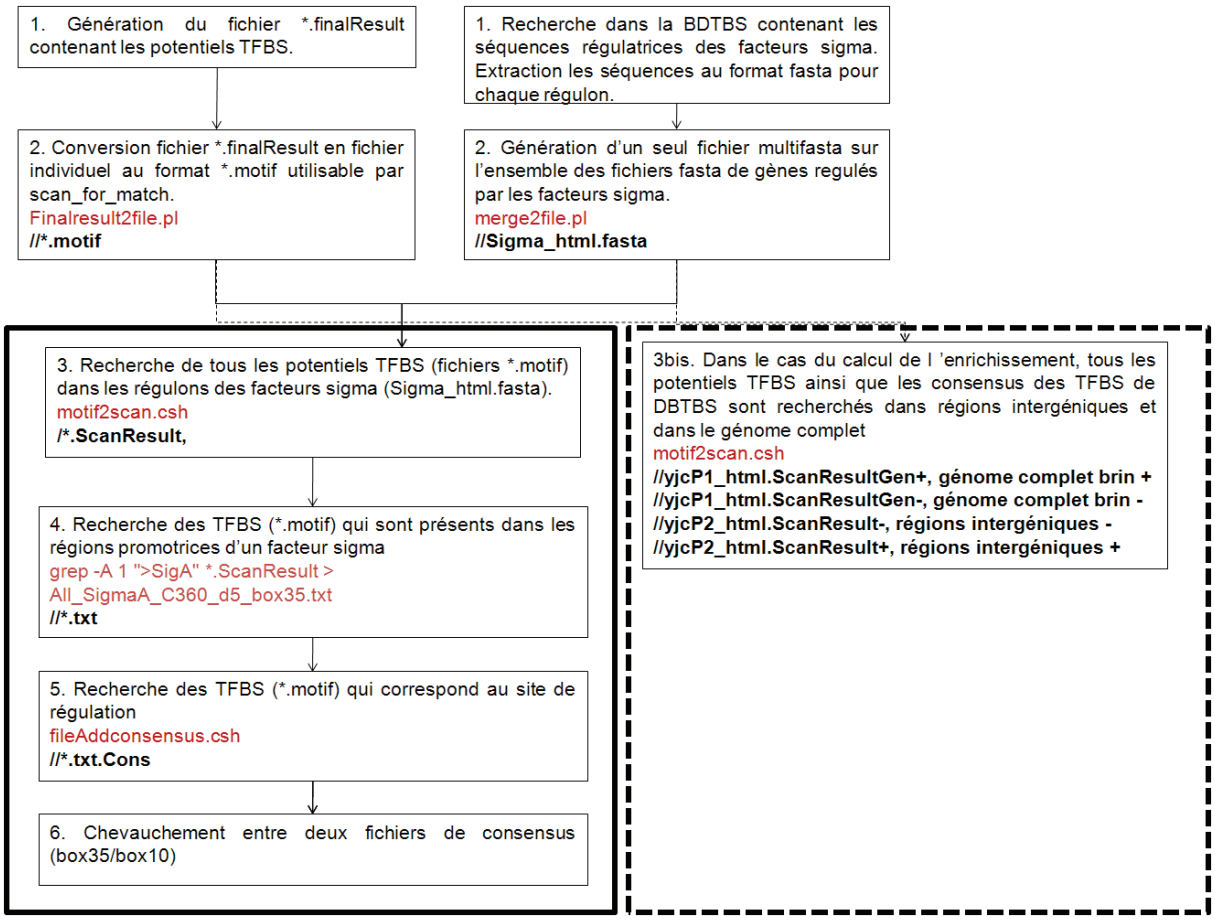


Figure : Étapes de l'analyse des SFBS putatifs et ceux proposés par DBTBS pour *B. subtilis*

Schéma d'analyse des programmes utilisés pour décrire la partie qui associe les SFBS potentiels aux SFBS réels et le seconde partie qui définie l'enrichissement des SFBS potentiels et des SFBS proposés dans DBTBS dans les régions intergéniques par rapport aux séquences codantes.

Annexe D

Localisation des SFBS putatifs par rapport aux positions réelles des motifs de régulation identifiés dans DBTBS : Exemple régulon SigD

	Putatifs SFBS	Régions promotrices du régulon SigD dans DBTBS
Fixation de la boîte -10	>SigD_flgB	
	cat 14...14 TTTCAAA	CATTTTTCTTCAAAAAGTTTCAAAAATGCCGAAAAGAAAGGAGAAAAAACAG
	aaa 21...21 AAGGAGA	CATTTTTCTTCAAAAAGTTTCAAAAATGCCGAAAAGAAAGGAGAAAAAACAG
	ttt 17...17 AAGGAGA	CATTTTTCTTCAAAAAGTTTCAAAAATGCCGAAAAGAAAGGAGAAAAAACAG
	ttc 16...16 AAGGAGA	CATTTTTCTTCAAAAAGTTTCAAAAATGCCGAAAAGAAAGGAGAAAAAACAG
	aaa 24...24 GGAGAA	CATTTTTCTTCAAAAAGTTTCAAAAATGCCGAAAAGAAAGGAGAAAAAACAG
	ttt 21...21 GCCGAA	CATTTTTCTTCAAAAAGTTTCAAAAATGCCGAAAAGAAAGGAGAAAAAACAG
	>SigD_lytF	
	aaa 20...20 ATATAAA	TAAAAAGGCAAAACGATGAGACAAATGTGCCGATATAAAAAAAGAAACAGCAAAT
	>SigD_mcpC	
	aaa 15...15 TTAGGA	TTTGATTATTCATTTTTCAAACGAAAGGGCCGATATAGAAATTAGGAGGGGA
	>SigD_motA	
	aaa 21...21 ATTAAC	AATGTCCCTAAAGTTCCGGGCACCAAAACCGATATTAACCATAGACA
	>SigD_ybdO	
	tat 16...16 GATAAAAA	TTTGAGGTTAATATATACATTATATTCGCCGATAAAAAAGAATAAGAGAGAATAC
	>SigD_yjcP	
	ttt 19...19 TGGGA	CAAATTGTTAATCGATAACTTTTCATTCCCGATATTAAAAATGGGAGT
	aac 22...22 TGGGA	CAAATTGTTAATCGATAACTTTTCATTCCCGATATTAAAAATGGGAGT
	ata 24...24 TGGGA	CAAATTGTTAATCGATAACTTTTCATTCCCGATATTAAAAATGGGAGT
	>SigD_ylqB	
	tat 24...24 TATGAA	AAATACGCTAAATATCATCCAATTTATTCCGATATTGTATATGAACCCTA
	atat 24...24 TATGAA	AAATACGCTAAATATCATCCAATTTATTCCGATATTGTATATGAACCCTA
	>SigD_yvyC	
	21...21 ATATAAA	ATATGTGTAATCTTATCTCGACTTAGTCGATATAAACGATAGATT
	ctt 15...15 ATATAAA	ATATGTGTAATCTTATCTCGACTTAGTCGATATAAACGATAGATT
	>SigD_yvyF	
	ttttt 15...15 CTAGA	AGAAGCTAAATGATTCTGTTTTTATGCCGATATAATCACTAGAAA
	ttt 15...15 CTAGAAA	AGAAGCTAAATGATTCTGTTTTTATGCCGATATAATCACTAGAAA
	ttt 16...16 CTAGAAA	AGAAGCTAAATGATTCTGTTTTTATGCCGATATAATCACTAGAAA
Fixation de la boîte -35	>SigD_degR	
	AAATAAA 16...16 TAT	CAAAATAGAAAAGAAAATAAAAATAAGCCGATATAACTATTGAACCAGA
	>SigD_epr	
	ATTAA 19...19 TAT	CCTTCTTACTATTAATCTCTCTTTTTTTCTCCGATATATATATCAAACATC
	>SigD_lytA	

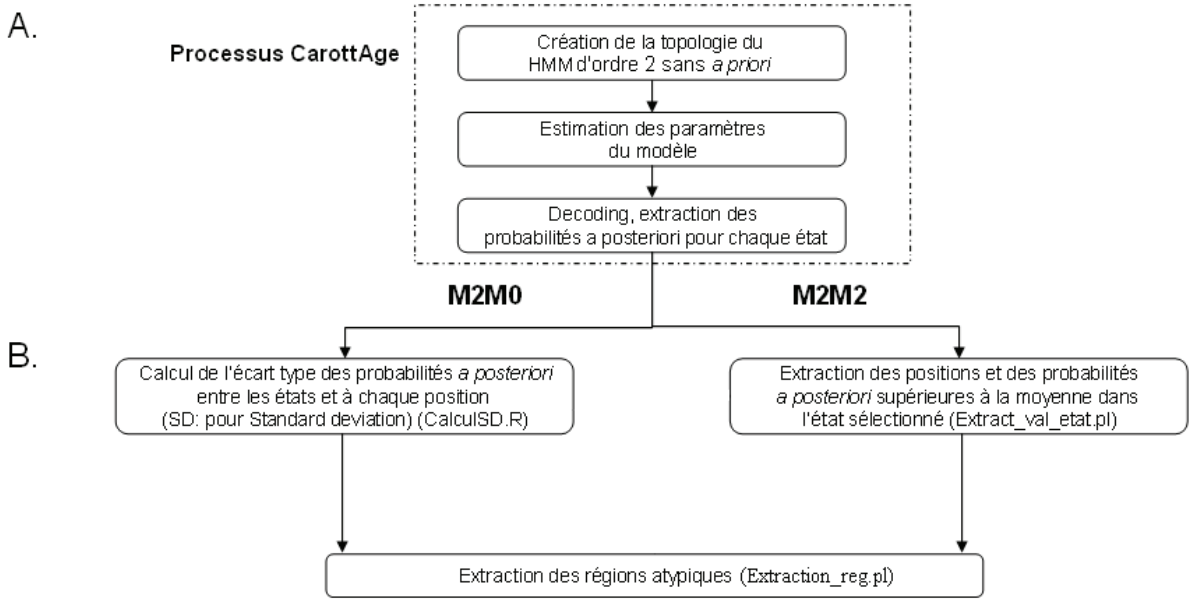
AATAG 20...20 AGG	TTTTTCTTTAAAATATAAAATAGGTTGACGATAAAATATAATGAGGTGA
ATATAAA 21...21 ATG	TTTTTCTTTAAAATATAAAATAGGTTGACGATAAAATATAATGAGGTGA
ATATAAA 21...21 ATGA	TTTTTCTTTAAAATATAAAATAGGTTGACGATAAAATATAATGAGGTGA
ATATAAA 22...22 TGA	TTTTTCTTTAAAATATAAAATAGGTTGACGATAAAATATAATTGAGGTGA
ATATAAA 16...16 ATA	TTTTTCTTTAAAATATAAAATAGGTTGACGATAAAATATAATGAGGTGA
>SigD_lytD	
AATGAA 20...20 AAA	TTCACCTATAAATGAAGCATGCGCGTGCCGATAAATAAACTAGTTTTATA
>SigD_yfmT	
ATAGA 17...17 TAT	TGAACCGATAGAAAAAATAGATTCGCCCATATTTTGATTTGCGGTTATAAAGGAG
>SigD_yoaH	
CTATAAA 23...23 ATG	TATCACTATAAAACTTGATAACACGTGTCGATATTATGGACATGAGC
CTATAAA 15...15 TCG	TATCACTATAAAACTTGATAACACGTGTCGATATTATGGACATGAGC
CTATAAA 19...19 TAT	TATCACTATAAAACTTGATAACACGTGTCGATATTATGGACATGAGC

Colonne 1 représente les SFBS putatifs proposés avec notre stratégie de fouille de données où la *box1* représente la boîte -10 ou la boîte -35 et l'algorithme recherche la *box2* représentant la boîte -35 ou la boîte -10 respectivement. Les lignes en jaune représentent les correspondances comptabilisées SFBS retrouvés à la bonne position.
Colonne 2 représente les régions promotrices du régulon SigD recensées dans DBTBS. Les motifs en rouge indiquent les SFBS localisés dans DBTBS. Les motifs soulignés localisent les SFBS putatifs dans les régions promotrices.

Annexe E

Description des programmes développés pour la fouille de données afin de détecter des régions atypiques

1. Programmes intervenant dans le processus d'extraction des positions et des régions atypiques



Le programme *CalculSD.R* calcule les valeurs de SD dans l'environnement R.

Le programme *Extract_val_etat.pl* extrait les index qui ont une valeur de probabilité *a posteriori* supérieure à celle de la moyenne générale de l'état D (état caché divergent).

Le programme *Extraction_reg.pl* permet d'identifier automatiquement des régions contenant une forte densité de *pos_dev* (position ayant une valeur de SD élevée ou de probabilité *a posteriori* élevée dans l'état D).

Une fois les régions atypiques extraites, il est possible de générer des fichiers de différents formats permettant l'analyse ultérieure de ces séquences.

Le programme *Comp_gene.pl* peut identifier et extraire l'ensemble des gènes contenu dans les régions atypiques. Il peut aussi générer un fichier contenant des informations de taille, de contenu en bases G et C de chaque région ou gène atypique.
Le programme *Extract_fasta.pl* génère un fichier multifasta contenant les séquences nucléotidiques de chaque région ou gène atypique.
Les programmes *RtoFT.pl* et *str2art.pl* permet de transformer un fichier contenant les positions déviantes ou contenant des gènes atypiques extraits vers un fichier visualisable par le logiciel *ARTEMIS*.

2. Programmes intervenant dans le processus de création des bases de données

Le programme *initialisation.pl* répertorie les caractéristiques de ces 126 organismes firmicutes. Ce programme génère et alimente les tables *Organisme* et *Souche* et génère la structure des tables *Regions* et *Homologues* qui seront par la suite alimentées.

Le programme *transfert_potentiel.pl* alimente les tables *Regions* et *Homologues* par la recherche de gènes homologues aux séquences atypiques dans la base de données des firmicutes.

Le programme *transfert_potentiel.pl* permet d'interroger la base de données des firmicutes par des requêtes multiples afin d'extraire des séquences spécifiques ou peu représentées dans cette base de données. Il est aussi possible d'extraire, pour une séquence atypique donnée, la distribution des homologues identifiés dans les différentes souches et/ou espèces de la base de données des firmicutes.

Annexe F

Liste des 126 souches bactériennes dans la base de données des firmicutes

Nom des souches	Projet ID	RefSeq_ID
Alkaliphilus metalliredigens QYMF	13006	NC_009633
Alkaliphilus oremlandii OhILAs	16083	NC_009922
Bacillus amyloliquefaciens FZB42	13403	NC_009725
Bacillus anthracis str. 'Ames Ancestor'	10784	NC_007530
Bacillus anthracis str. Ames	309	NC_003997
Bacillus anthracis str. Sterne	10878	NC_005945
Bacillus cereus ATCC 10987	74	NC_003909
Bacillus cereus ATCC 14579	384	NC_004722
Bacillus cereus E33L	12468	NC_006274
Bacillus cereus subsp. cytotoxis NVH 391-98	13624	NC_009674
Bacillus clausii KSM-K16	13291	NC_006582
Bacillus halodurans C-125	235	NC_002570
Bacillus licheniformis ATCC 14580	12388	NC_006270
Bacillus licheniformis ATCC 14580 DSM 13	13082	NC_006322
Bacillus pumilus SAFR-032	20391	NC_009848
Bacillus subtilis subsp. subtilis str. 168	76	NC_000964
Bacillus thuringiensis serovar konkukian str. 97-27	10877	NC_005957
Bacillus thuringiensis str. Al Hakam	18255	NC_008600
Bacillus weihenstephanensis KBAB4	13623	NC_010184
Caldicellulosiruptor saccharolyticus DSM *8903*	13466	NC_009437
Candidatus Desulforudis audaxviator MP104C	21047	NC_010424
Carboxydothermus hydrogenoformans Z-2901	253	NC_007503
Clostridium acetobutylicum ATCC 824	77	NC_003030
Clostridium beijerinckii NCIMB 8052	12637	NC_009617
Clostridium botulinum A str. ATCC 19397	19517	NC_009697
Clostridium botulinum A str. ATCC 3502	193	NC_009495
Clostridium botulinum A str. Hall	19521	NC_009698
Clostridium botulinum A3 str. Loch Maree	28507	NC_010520
Clostridium botulinum B str. Eklund 17B	28857	NC_010674
Clostridium botulinum B1 str. Okra	28505	NC_010516
Clostridium botulinum E3 str. Alaska E43	28855	NC_010723
Clostridium botulinum F str. Langeland	19519	NC_009699
Clostridium difficile 630	78	NC_009089
Clostridium kluyveri DSM 555	19065	NC_009706
Clostridium novyi NT	16820	NC_008593
Clostridium perfringens ATCC 13124	304	NC_008261
Clostridium perfringens str. 13	79	NC_003366
Clostridium phytofermentans ISDg	16184	NC_010001

Clostridium tetani E88	81	NC_004557
Clostridium thermocellum ATCC 27405	314	NC_009012
Desulfitobacterium hafniense Y51	16639	NC_007907
Desulfotomaculum reducens MI-1	13424	NC_009253
Enterococcus faecalis V583	70	NC_004668
Exiguobacterium sibiricum 255-15	10649	NC_010556
Finegoldia magna ATCC 29328	18981	NC_010376
Geobacillus kaustophilus HTA426	13233	NC_006510
Geobacillus thermodenitrificans NG80-2	18655	NC_009328
Heliobacterium modesticaldum Ice1	13427	NC_010337
Lactobacillus acidophilus NCFM	82	NC_006814
Lactobacillus brevis ATCC 367	404	NC_008497
Lactobacillus casei ATCC 334	402	NC_008526
Lactobacillus casei BL23	30359	NC_010999
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	16871	NC_008054
Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365	403	NC_008529
Lactobacillus gasseri ATCC 33323	84	NC_008530
Lactobacillus johnsonii NCC 533	9638	NC_005362
Lactobacillus plantarum WCFS1	356	NC_004567
Lactobacillus reuteri DSM 20016	15766	NC_009513
Lactobacillus reuteri JCM 1112	19011	NC_010609
Lactobacillus sakei subsp. sakei 23K	13435	NC_007576
Lactobacillus salivarius UCC118	13280	NC_007929
Lactococcus lactis subsp. cremoris MG1363	18797	NC_009004
Lactococcus lactis subsp. cremoris SK11	401	NC_008527
Lactococcus lactis subsp. lactis Il1403	72	NC_002662
Leuconostoc citreum KM20	16062	NC_010471
Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293	315	NC_008531
Listeria innocua Clip11262	86	NC_003212
Listeria monocytogenes EGD-e	276	NC_003210
Listeria monocytogenes str. 4b F2365	85	NC_002973
Listeria welshimeri serovar 6b str. SLCC5334	13443	NC_008555
Lysinibacillus sphaericus C3-41	19619	NC_010382
Moorella thermoacetica ATCC 39073	10648	NC_007644
Oceanobacillus iheyensis HTE831	284	NC_004193
Oenococcus oeni PSU-1	317	NC_008528
Pediococcus pentosaceus ATCC 25745	398	NC_008525
Pelotomaculum thermopropionicum SI	19023	NC_009454
Staphylococcus aureus RF122	63	NC_007622
Staphylococcus aureus subsp. aureus COL	238	NC_002951
Staphylococcus aureus subsp. aureus JH1	15758	NC_009632
Staphylococcus aureus subsp. aureus JH9	15757	NC_009487
Staphylococcus aureus subsp. aureus MRSA252	265	NC_002952
Staphylococcus aureus subsp. aureus MSSA476	266	NC_002953
Staphylococcus aureus subsp. aureus MW2	306	NC_003923
Staphylococcus aureus subsp. aureus Mu3	18509	NC_009782
Staphylococcus aureus subsp. aureus Mu50	263	NC_002758
Staphylococcus aureus subsp. aureus N315	264	NC_002745
Staphylococcus aureus subsp. aureus NCTC 8325	237	NC_007795
Staphylococcus aureus subsp. aureus USA300_FPR3757	16313	NC_007793
Staphylococcus aureus subsp. aureus USA300_TCH1516	19489	NC_010079
Staphylococcus aureus subsp. aureus str. Newman	18801	NC_009641
Staphylococcus epidermidis ATCC 12228	279	NC_004461

Staphylococcus epidermidis RP62A	64	NC_002976
Staphylococcus haemolyticus JCSC1435	12508	NC_007168
Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305	15596	NC_007350
Streptococcus agalactiae 2603V/R	330	NC_004116
Streptococcus agalactiae A909	326	NC_007432
Streptococcus agalactiae NEM316	334	NC_004368
Streptococcus mutans UA159	333	NC_004350
Streptococcus pneumoniae CGSP14	29179	NC_010582
Streptococcus pneumoniae D39	16374	NC_008533
Streptococcus pneumoniae G54	29047	NC_011072
Streptococcus pneumoniae Hungary19A-6	28035	NC_010380
Streptococcus pneumoniae R6	278	NC_003098
Streptococcus pneumoniae TIGR4	277	NC_003028
Streptococcus pyogenes M1 GAS	269	NC_002737
Streptococcus pyogenes MGAS10270	16364	NC_008022
Streptococcus pyogenes MGAS10394	12469	NC_006086
Streptococcus pyogenes MGAS10750	16366	NC_008024
Streptococcus pyogenes MGAS2096	16365	NC_008023
Streptococcus pyogenes MGAS315	311	NC_004070
Streptococcus pyogenes MGAS5005	13888	NC_007297
Streptococcus pyogenes MGAS6180	13887	NC_007296
Streptococcus pyogenes MGAS8232	286	NC_003485
Streptococcus pyogenes MGAS9429	16363	NC_008021
Streptococcus pyogenes SSI-1	301	NC_004606
Streptococcus pyogenes str. Manfredo	270	NC_009332
Streptococcus sanguinis SK36	13942	NC_009009
Streptococcus suis 05ZYH33	17153	NC_009442
Streptococcus suis 98HAH33	17155	NC_009443
Streptococcus thermophilus CNRZ1066	13163	NC_006449
Streptococcus thermophilus LMD-9	13773	NC_008532
Streptococcus thermophilus LMG 18311	13162	NC_006448
Symbiobacterium thermophilum IAM 14863	12994	NC_006177
Syntrophomonas wolfei subsp. wolfei *str.* Goettingen	13014	NC_008346
Thermoanaerobacter pseudethanolicus ATCC 33223	13901	NC_010321
Thermoanaerobacter tengcongensis MB4	249	NC_003869

Annexe G

Distances entre la séquence de *S. thermophilus* et celle des autres espèces de la base de données des firmicutes

Nom des espèces	Distances avec *S. thermophilus*
Streptococcus thermophilus	0
Streptococcus sanguinis	0,04336
Streptococcus pneumoniae	0,05012
Streptococcus agalactiae	0,05301
Streptococcus suis	0,05429
Streptococcus pyogenes	0,05532
Streptococcus mutans	0,07112
Lactococcus lactis	0,10657
Lactobacillus sakei	0,12597
Lactobacillus casei	0,13399
Lactobacillus plantarum	0,13441
Lactobacillus salivarius	0,13528
Lactobacillus brevis	0,13561
Bacillus pumilus	0,1368
Pediococcus pentosaceus	0,13725
Bacillus halodurans	0,13744
Bacillus licheniformis	0,13891
Bacillus subtilis	0,13892
Bacillus amyloliquefaciens	0,1392
Listeria welshimeri	0,14027
Listeria innocua	0,14077
Listeria monocytogenes	0,14244
Bacillus weihenstephanensis	0,14329
Bacillus cereus	0,14336
Lactobacillus delbrueckii	0,14453
Bacillus anthracis	0,14468
Bacillus thuringiensis	0,14474
Lactobacillus gasseri	0,14665
Lactobacillus johnsonii	0,14672
Enterococcus faecalis	0,14793
Bacillus clausii	0,14829
Lactobacillus reuteri	0,14904
Lysinibacillus sphaericus C3-4	0,14928
Lactobacillus acidophilus	0,14995
Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293	0,1505
Oceanobacillus	0,15174

Exiguobacterium sibiricum	0,15458
Leuconostoc	0,15573
Staphylococcus haemolyticus	0,15981
Staphylococcus saprophyticus	0,16071
Staphylococcus aureus	0,16086
Staphylococcus epidermidis	0,16094
Geobacillus sthermodenitrifican	0,16192
Geobacillus kaustophilus	0,16301
Oenococcus oeni	0,18471
Clostridium botulinum	0,18495
Clostridium novyi	0,18508
Clostridium difficile	0,18754
Alkaliphilus oremlandii OhILAs	0,18982
Alkaliphilus metalliredigens QYMF	0,19004
Clostridium tetani	0,19101
Clostridium beijerinckii	0,19102
Clostridium perfringens	0,19137
Clostridium thermocellum	0,19206
Heliobacterium modesticaldum Ice1	0,193
Moorella	0,19385
Clostridium acetobutylicum	0,19442
Clostridium kluyveri	0,19562
Thermoanaerobacter pseudethanolicus	0,19634
Thermoanaerobacter tengcongensis	0,19671
Desulfotomaculum reducens	0,20055
Clostridium phytofermentans	0,20131
Pelotomaculum thermopropionicum	0,20599
Symbiobacterium	0,20817
Finegoldia magna ATCC 29328	0,20993
Syntrophomonas wolfei	0,21091
Carboxydothermus hydrogenoformans	0,21143
Desulfitobacterium hafniense	0,21361
Caldicellulosiruptor saccharolyticus	0,21584
Candidatus Desulforudis audaxviator MP104C	0,23419

Annexe H

Exemple de résultat de sortie de l'analyse de distribution des gènes au sein des firmicutes : quatre gènes atypiques du locus CRISPR

Locus tag	Nombre d'espèces	Nom des espèces	Index des espèces	Fonction [espèce/souche]
str0657	1	*Streptococcus thermophilus*	1	hypothetical protein str0657 [*Streptococcus thermophilus* CNRZ1066]
str0658	7	*Staphylococcus epidermidis,Streptococcus pyogenes,Streptococcus agalactiae,Lactobacillus salivarius,Finegoldia magna ATCC 29328,Streptococcus mutans,Streptococcus thermophilus*	1,2,5,7,18,26,81	CRISPR-associated Cas1 family protein [*Staphylococcus epidermidis* RP62A], putative cytoplasmic protein [*Streptococcus pyogenes* MGAS2096], CRISPR-associated Cas1 family protein [*Streptococcus agalactiae* A909], CRISPR-associated protein [*Lactobacillus salivarius* UCC118],hypothetical protein FMG_0059 [*Finegoldia magna* ATCC 29328],hypothetical protein SMU.1404c [*Streptococcus mutans* UA159], hypothetical protein str0658 [*Streptococcus thermophilus* CNRZ1066]
str0659	7	*Lactobacillus casei,Finegoldia magna ATCC 29328,Streptococcus mutans,Lactobacillus salivarius,Streptococcus pyogenes,Streptococcus agalactiae,Streptococcus thermophilus*	1,2,5,7,11,18,81	CRISPR-associated protein [*Lactobacillus casei*],hypothetical protein FMG_0060 [*Finegoldia magna* ATCC 29328],hypothetical protein SMU.1403c [*Streptococcus mutans* UA159],CRISPR-associated protein [*Lactobacillus salivarius* UCC118],hypothetical cytosolic protein [*Streptococcus pyogenes* MGAS10750],CRISPR-associated Cas2 family protein [*Streptococcus agalactiae* A909], hypothetical protein str0659 [*Streptococcus thermophilus* CNRZ1066]
str0660	1	*Streptococcus thermophilus*	1	hypothetical protein str0660 [*Streptococcus thermophilus* CNRZ1066]

Un exemple de sortie est présenté pour quatre gènes atypiques retrouvés chez *S. thermophilus* CNRZ1066. Ces gènes appartiennent au locus CRISPR et auraient une origine exogène (Bolotin *et al.*, 2005).

Annexe I

Analyse de gènes atypiques (M2M2) qui sont variables (*CGH-set*) et/ou spécifiques

Tableau 1 : 54 gènes atypiques (M2M2) et variables (*CGH-set*) 17575
Tableau 2 : 73 gènes atypiques et spécifiques regroupés en 13 régions 176

Tableau 1 : 54 gènes atypiques (M2M2) et variables (*CGH-set*)

Numéro de la région	Nom	Fonction annotée
Région 1	str0098	lantibiotic biosynthesis protein, truncated
Région 2	str0323	hypothetical protein str0323
	str0324	peptide ABC transporter ATP binding/permease protein, putative
Région 3	str0682	hypothetical protein str0682
	str0683	hypothetical protein str0683
	str0684	hypothetical protein str0684
	str0685	hypothetical protein str0685
	str0686	hypothetical protein str0686
	str0687	hypothetical protein str0687
	str0688	hypothetical protein str0688
	str0689	restriction-modification system regulatory protein, putative
	str0690	hypothetical protein str0690
Région 4	str0706	hypothetical protein str0706
	str0708	type I restriction-modification system specificty subunit
	str0709	hypothetical protein str0709
Région 5	str0780	unknown protein, phage associated
	str0781	HTH DNA binding protein, phage associated
	str0782	unknown protein, phage associated
Région 6	str0908	ABC transporter substrate binding protein
Région 7	str0914	hypothetical protein str0914
Région 8	str1037	hypothetical protein str1037
	str1038	hypothetical protein str1038
	str1040	hypothetical protein str1040
	str1042	hypothetical protein str1042
	str1043	tyrosyl-tRNA synthetase E

	str1044	positive transcriptional regulator MutR family
Région 9	str1053	hypothetical protein str1053
Région 10	str1076	exopolysaccharide biosynthesis protein
	str1077	exopolysaccharide polymerization protein
	str1078	exopolysaccharide gene claster protein
	str1079	exopolysaccharide polymerization protein
	str1080	exopolysaccharide biosynthesis protein, sugar transferase
	str1081	exopolysaccharide biosynthesis protein, acetyltransferase
	str1082	exopolysaccharide biosynthesis protein, glycosyltransferase
	str1084	exopolysaccharide biosynthesis protein
	str1085	exopolysaccharide biosynthesis protein
Région 11	str1345	adenylosuccinate lyase, putative
	str1346	amino acid (glutamine) ABC transporter ATP-binding protein
	str1347	ABC transporter amino acid permease protein
	str1348	hypothetical protein str1348
	str1349	hypothetical protein str1349
	str1350	putative aspartate aminotransferase, truncated
	str1351	putative aspartate aminotransferase, truncated
	str1352	glycerol dehydrogenase, truncated
	str1353	glycerol dehydrogenase, truncated
Région 12	str1374	hypothetical protein str1374
Région 13	str1474	glycosyltransferase, putative teichoic acid biosynthesis protein
	str1477	glycosyl transferase, truncated
	str1478	glycosyl transferase, truncated
	str1479	glycosyl transferase
Région 14	str1508	hypothetical protein str1508
	str1511	positive transcriptional regulator MutR family
	str1512	hypothetical protein str1512
	str1513	ABC transporter ATP binding/permease protein

Tableau 2 : 73 gènes atypiques et spécifiques regroupés en 13 régions

Numéro de la région	Nom	Fonction annotée
Région 1	str0098	lantibiotic biosynthesis protein, truncated
	str0099	lantibiotic biosynthesis protein
	str0102	integrase/recombinase, phage integrase family, truncated
	str0103	hypothetical protein str0103
	str0104	hypothetical protein str0104
Région 2	str0182	positive transcriptional regulator MutR family
	str0183	conserved hypothetical protein, putative protein kinase
Région 3	str0679	MutT/nudix family protein

	str0680	hypothetical protein str0680
	str0682	hypothetical protein str0682
	str0683	hypothetical protein str0683
	str0684	hypothetical protein str0684
	str0685	hypothetical protein str0685
	str0686	hypothetical protein str0686
	str0687	hypothetical protein str0687
	str0688	hypothetical protein str0688
	str0689	restriction-modification system regulatory protein, putative
	str0690	hypothetical protein str0690
Région 4	str0706	hypothetical protein str0706
	str0707	hypothetical protein str0707
	str0708	type I restriction-modification system specificty subunit
	str0709	hypothetical protein str0709
Région 5	str0908	ABC transporter substrate binding protein
	str0912	hypothetical protein str0912
	str0913	hypothetical protein str0913
	str0914	hypothetical protein str0914
	str0916	transcriptional regulator
Région 6	str1014	maltose operon transcriptional repressor, putative
	str1018	maltose/maltodextrin ABC transporter / tnp1167 hybrid protein, truncated
Région 7	str1037	hypothetical protein str1037
	str1038	hypothetical protein str1038
	str1039	IS1239 transposase
	str1040	hypothetical protein str1040
	str1041	UDP-N-acetylglucosamine enolpyruvyl transferase, putative
	str1042	hypothetical protein str1042
	str1043	tyrosyl-tRNA synthetase E
	str1044	positive transcriptional regulator MutR family
Région 8	str1076	exopolysaccharide biosynthesis protein
	str1077	exopolysaccharide polymerization protein
	str1078	exopolysaccharide gene claster protein
	str1079	exopolysaccharide polymerization protein
	str1080	exopolysaccharide biosynthesis protein, sugar transferase
	str1081	exopolysaccharide biosynthesis protein, acetyltransferase
	str1082	exopolysaccharide biosynthesis protein, glycosyltransferase
	str1084	exopolysaccharide biosynthesis protein
	str1085	exopolysaccharide biosynthesis protein
Région 9	str1206	truncated IS1239 transposase
Région 10	str1308	biotin synthase
	str1309	proline/glycine betaine ABC transporter

	str1310	proline/glycine betaine ABC transporter ATP-binding protein
	str1311	proline/glycine betaine ABC transporter membrane-spanning protein
	str1313	histidine ammonia-lyase
	str1315	urocanate hydratase
Région 11	str1345	adenylosuccinate lyase, putative
	str1346	amino acid (glutamine) ABC transporter ATP-binding protein
	str1347	ABC transporter amino acid permease protein
	str1348	hypothetical protein str1348
	str1349	hypothetical protein str1349
	str1350	putative aspartate aminotransferase, truncated
	str1351	putative aspartate aminotransferase, truncated
	str1352	glycerol dehydrogenase, truncated
	str1353	glycerol dehydrogenase, truncated
Région 12	str1474	glycosyltransferase, putative teichoic acid biosynthesis protein
	str1475	hypothetical protein str1475
	str1476	glycosyl transferase
	str1477	glycosyl transferase, truncated
	str1478	glycosyl transferase, truncated
	str1479	glycosyl transferase
Région 13	str1509	hypothetical protein str1509
	str1510	hypothetical protein str1510
	str1511	positive transcriptional regulator MutR family
	str1512	hypothetical protein str1512
	str1513	ABC transporter ATP binding/permease protein

Annexe J

Développement d'une interface d'analyse web

Dans l'objectif d'effectuer des analyses à grande échelle, nous sommes en cours de développement d'une interface utilisation web sécurisée.
L'interface est développée avec des scripts CGI/PERL (Common Gateway Interface) qui procurent un interfaçage web. Tous les programmes sont exécutés sur le serveur. Les utilisateurs peuvent effectuer des analyses à plusieurs niveaux soit une analyse intégrale pour la détection de HGT avec la construction d'un HMM d'ordre 2 soit une analyse phylogénétique au sein des firmicutes avec des données formatées.

La page Web se compose d'un entête, d'un panneau latéral gauche et d'un panneau central (figure 1). Le panneau latéral gauche permet une navigation sur le choix des différentes analyses. Le premier choix est soit de réaliser des blast locaux soit d'effectuer une analyse de données afin d'identifier de potentiels gènes exogènes dans une séquence génomique en utilisant les HMM d'ordre 2 (HMM2 modèle). Le panneau central informe des différents paramètres utilisateurs à modifier (des paramètres par défaut sont préétablis).
Le « HMM2 model » est composé de deux analyses distinctes dont l'une est la génération du modèle HMM d'ordre 2 pour l'extraction des gènes atypiques et l'autre est l'analyse de la distribution des gènes au sein de la base de données des firmicutes.



Figure 1 : Page d'accueil.

Dans le cas d'une analyse intégrale, nous débutons par la création d'un modèle HMM d'ordre 2 en cliquant sur « HMM2 selection » au niveau du panneau latéral gauche, sous la partie « HMM2 generation » (figure 2). A ce niveau le choix de la topologie optimale est à l'appréciation de l'utilisateur. Comme l'indique le panneau central (figure 2), les paramètres sont :

i) le choix de l'ordre du modèle de Markov d'ordre 2 (M2M0 ou M2M2) ;

ii) l'alphabet utilisé (nucléotide, dinucléotides, …, hexanucléotides) ;

iii) le nombre d'itérations pour l'estimation des paramètres ;

iii) le fichier fasta comportant la séquence génomique ;

iv) la présence d'un GC skew prédominant et dans ce cas, il faut renseigner la position où se situe le point d'inflexion.

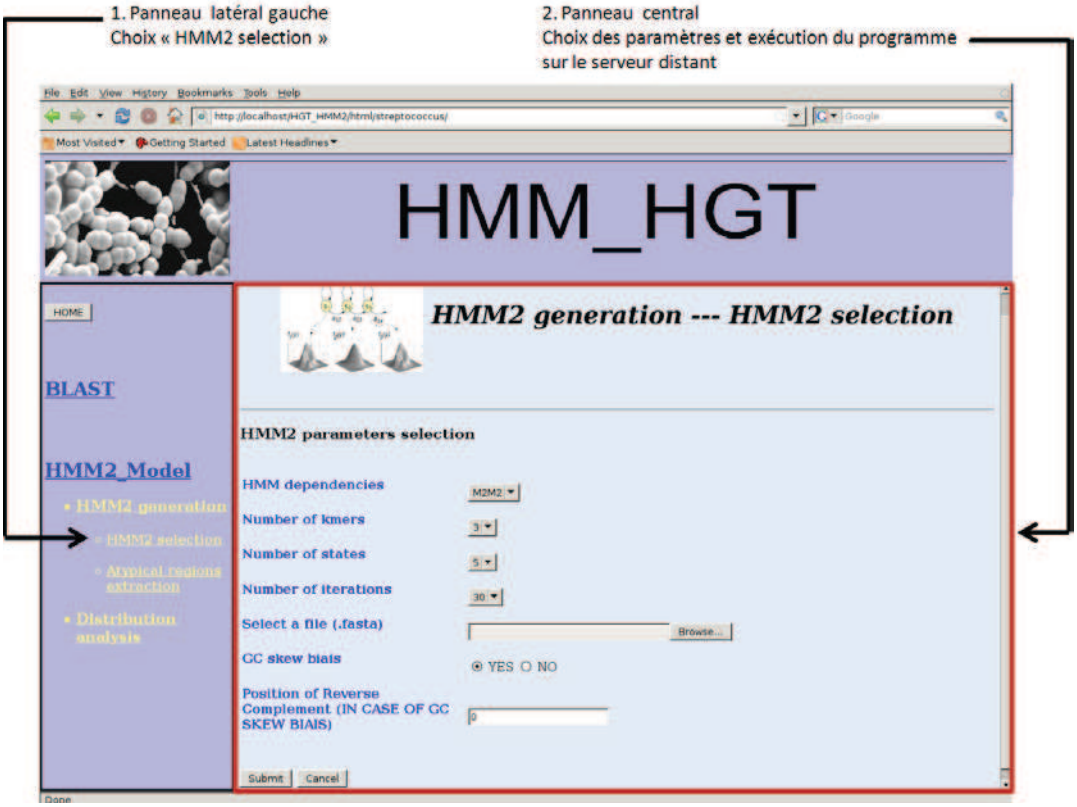


Figure 2 : Génération du modèle HMM d'ordre 2 et estimation de ses paramètres.

A ce stade, le modèle HMM d'ordre 2 est construit selon la topologie choisie par l'utilisateur, les paramètres ont été estimés avec la séquence génomique et les probabilités *a posteriori* d'être dans chaque état caché à chaque position sont fournies dans un fichier **.dump* (à télécharger sur le serveur). La topologie du HMM d'ordre 2 ainsi que les probabilités de transition entre les états cachés et d'émission des symboles dans chaque état caché sont aussi téléchargeables mais cette opération nécessite un programme de visualisation spécifique (*gviewmod* à compiler sur la plateforme utilisateur) (cf. Annexe Web A).

L'extraction des régions atypiques est réalisable par l'interface en cliquant sur « Atypical regions extraction » au niveau du panneau latéral gauche, sous la partie « HMM2 generation » (figure 3). Il faut renseigner un certain nombre de données :

i) le fichier de séquence au format genbank, les annotations des CDS sont utilisées pour extraire les gènes atypiques contenus dans les régions atypiques ;

ii) l'état caché susceptible de capturer les hétérogénéités ;

iii) l'indication de la position où se trouve l'inflexion en présence de GC skew prédominant ;

iv) la taille de la fenêtre coulissante pour l'extraction des *pos_dev* (position déviante, cf. section 3.6) ;

v) le paramètre T correspondant au taux de *pos_dev* dans une fenêtre par rapport à la fenêtre précédente ;

vi) le paramètre L pour la longueur des fenêtres candidates.

Les données de iv) à vi) sont les paramètres d'extraction des régions nécessaires pour les considérer comme des régions atypiques (cf. la section 3.6).

A ce niveau, l'exécution des programmes va permettre l'extraction des régions atypiques ainsi que celle des gènes atypiques. Deux fichiers sont générés et sont téléchargeables par l'utilisateur. Le premier est un fichier d'information qui comporte une liste de la valeur de GC% et de longueur pour chaque gène atypique. Le second fichier est un fichier au format multifasta des séquences nucléotidiques des gènes atypiques. Ces deux fichiers sont aussi générés pour les régions atypiques.

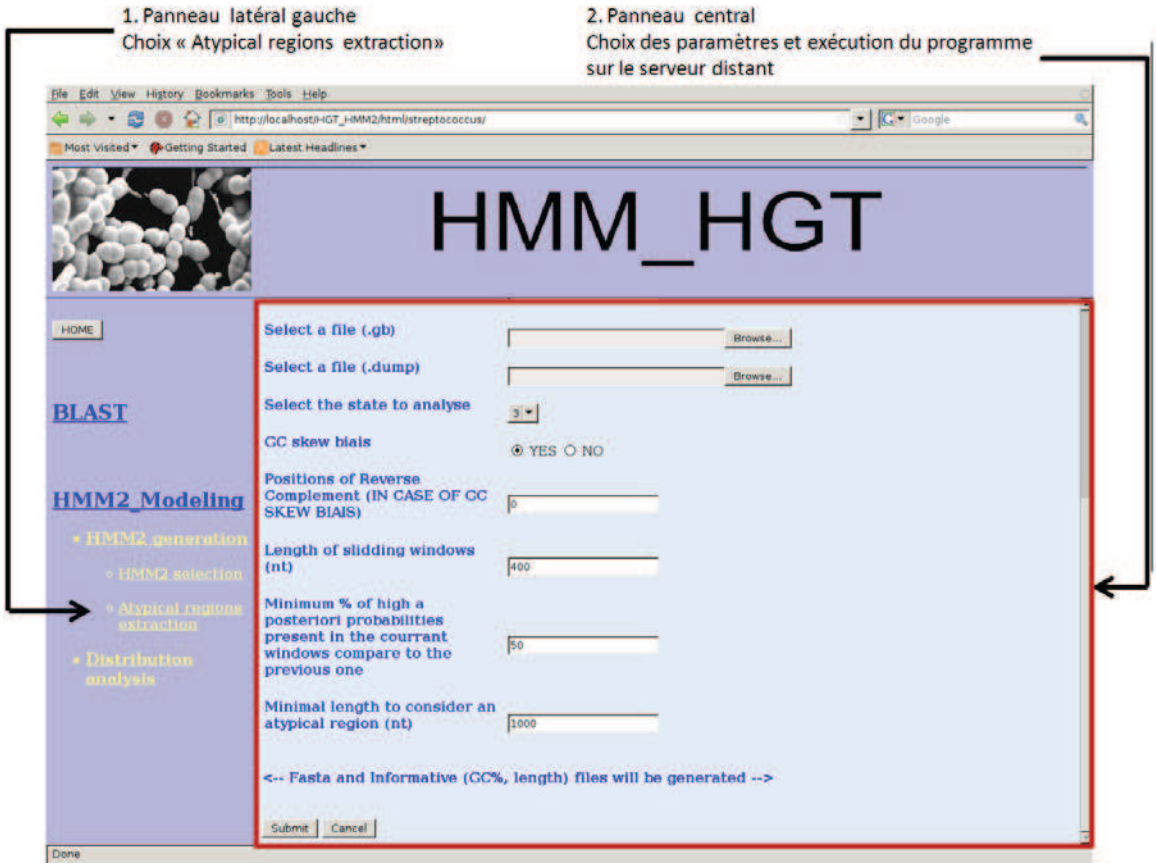


Figure 3 : Extraction des régions et gènes atypiques.

Les étapes de génération du HMM d'ordre 2 et l'extraction des gènes atypiques ont permis de détecter des gènes atypiques. Dans cette dernière étape, l'identification des gènes atypiques comme issus de HGT est déterminée par l'analyse de la distribution de ces gènes atypiques au sein des Firmicutes. Cette étape est accessible par l'interface en cliquant sur « Distribution analysis » au niveau du panneau latéral gauche, sous la partie « HMM2 model » (figure 4). L'exécution des analyses de distribution est indépendante de l'analyse d'extraction des régions atypiques. L'utilisateur peut fournir son propre jeu de données et dans ce cas, l'interface propose de formater les données pour les analyser (figure 5). En effet, celles-ci doivent mimer la structure des résultats obtenus par l'analyse d'extraction des régions et gènes atypiques. Dans ce cas, l'utilisateur fournit un fichier au format genbank (annotation des CDS) ainsi que la liste des gènes à analyser avec une structure simple. Pour chaque gène, il suffit de débuter la ligne du fichier par le signe « > » suivi du nom du gène (locus tag).

Enfin, l'utilisateur indique le nom de la souche au complet où les analyses de distribution seront effectuées.
L'utilisateur pourra télécharger les fichiers générés tels qu'ils seraient fournis par l'analyse d'extraction des régions/gènes atypiques.

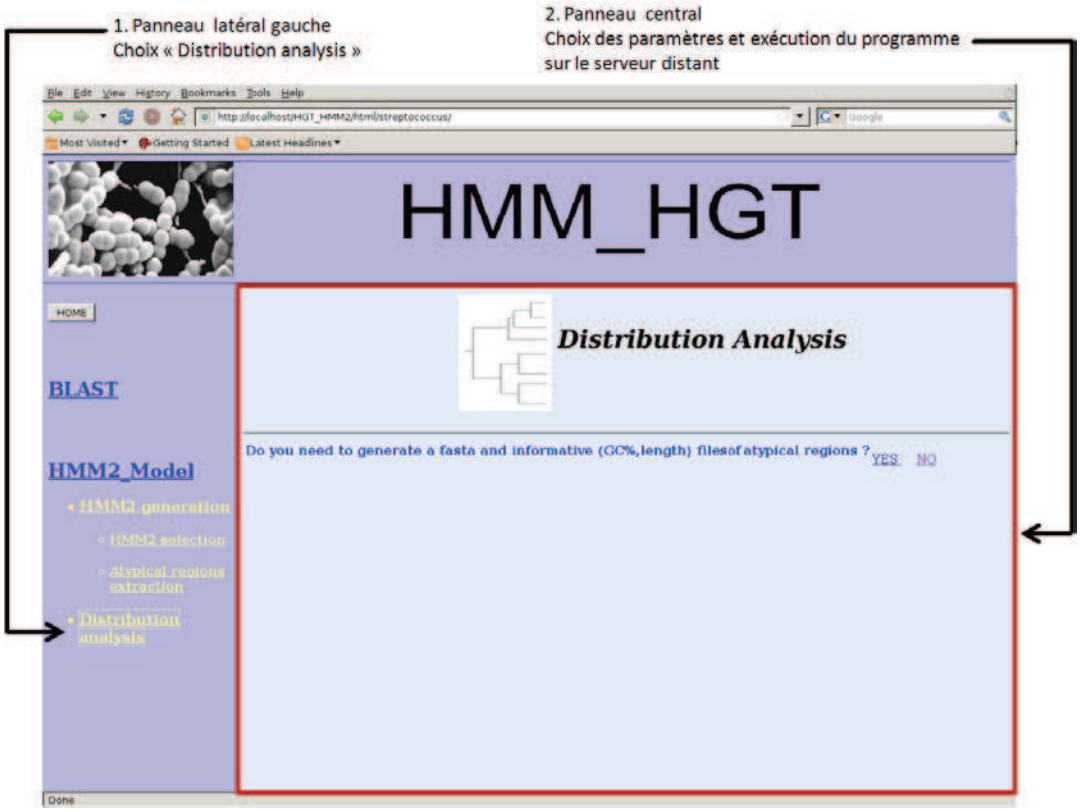


Figure 4 : Analyse de la distribution au sein de la base de données des firmicutes.

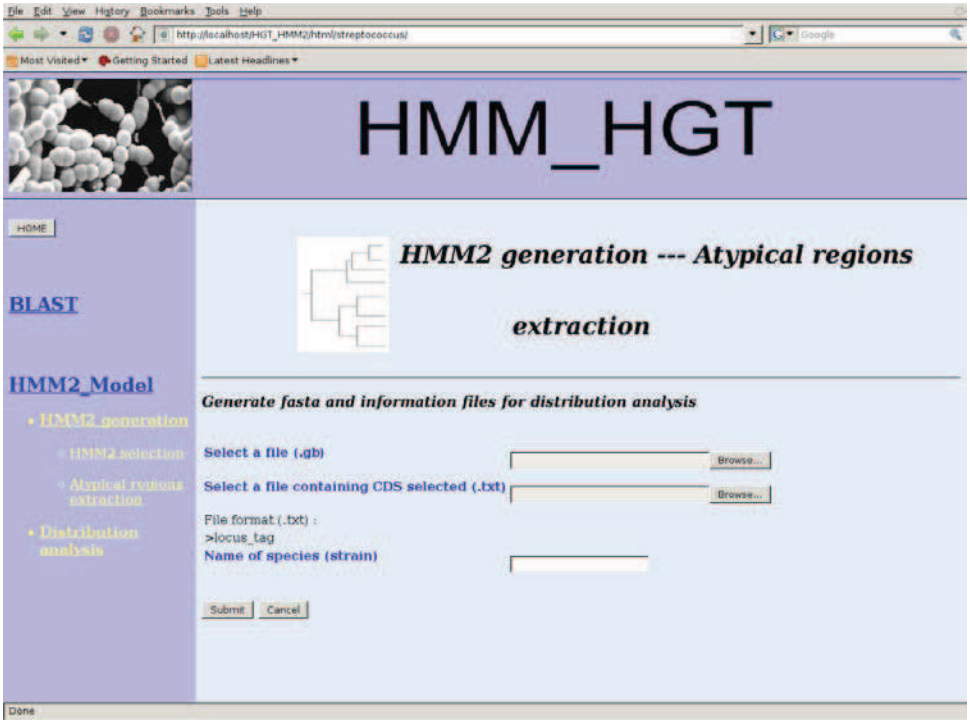


Figure 5 : Formatage des données utilisateur sous la forme de données issues de l'analyse d'extraction de régions et gènes atypiques.

L'analyse de la distribution débute par la vérification que l'utilisateur possède les données au format requis (figure 4, cadre 2) puis la sélection du 'no' entraîne l'ouverture de la page d'analyse de la distribution phylogénétique (figure 6). L'analyse de la distribution nécessite :

i) les deux fichiers générés précédemment soit par l'extraction des régions/gènes atypiques soit par l'étape de substitut (figure 3, figure 5) ;

ii) le nom de la souche (complet) où les gènes atypiques sont à identifier ;

iii) la sélection de la base de données.

La base contient actuellement 126 génomes de Firmicutes et pourrait être incrémentée. De plus, la possibilité de faire une recherche sur d'autres bases de données locales est envisageable. L'exécution du programme va générer un fichier au format texte (*.txt) qui fournit pour chaque gène atypique (annexe G) :

- Son identifiant (locus tag).
- Le nombre d'espèces où un homologue a été retrouvé.
- La liste des espèces ayant un gène homologue retrouvé.
- La liste des index des espèces dans la base de données. Cette information est importante lors du calcul de la valeur de dispersion (cf. section 3.8).
- La liste des fonctions des gènes homologues retrouvés au sein des Firmicutes.

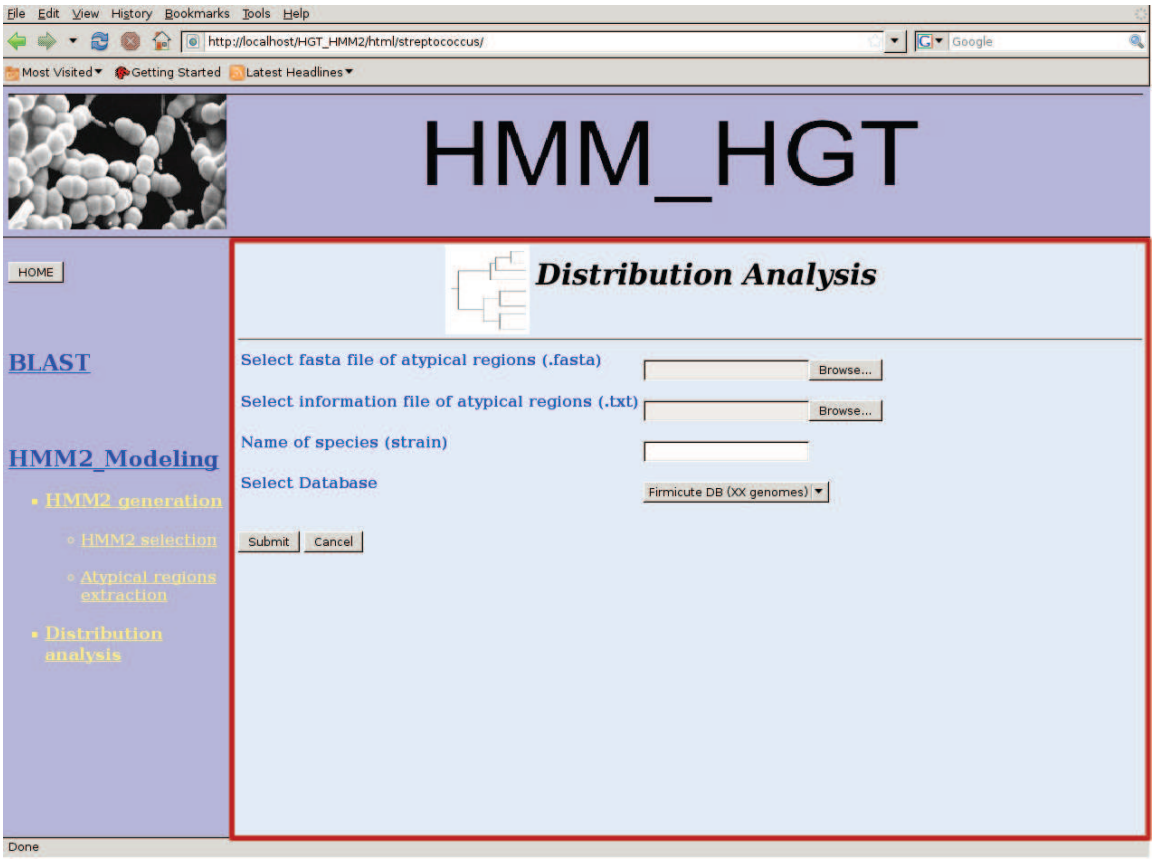


Figure 6 : Analyse de la distribution des gènes atypiques au sein de la base de données des Firmicutes.
XX représente le nombre de génome à ce jour, la base de données comporte 126 génomes de Firmicutes.

Accessibilité

La mise à disposition n'est pas encore possible mais elle est en cours de réalisation pour diverses raisons. Notamment, la manipulation des modèles pour les non initiés est complexe, une amélioration par fusion des états serait utile (cf. section 4.4). De plus, la mise en place d'une telle plateforme requiert un gros système de stockage (génération de base de données) et une administration des comptes afin de s'assurer de la confidentialité des données.

Une première possibilité serait la création de compte utilisateur c'est-à-dire d'un espace de travail (ou workspace). L'utilisateur pourrait s'identifier afin de récupérer ses résultats. Mais ce processus exige l'administration de comptes complexes (ajout, suppression de comptes, gestion de l'espace disque). Une seconde possibilité qui serait plus rapide dans la mise en œuvre est l'envoi d'un mail avec un lien sur les données stockées sur une partie du serveur spécifique.

Par ailleurs, un grand espace disque est nécessaire pour le stockage des données car à chaque analyse une base de données de comparaison est générée entre l'organisme de référence (donné par l'utilisateur) et la base de données des génomes des firmicutes. Cela correspond à une occupation de l'espace disque qui peut devenir très limitant pour des analyses de grands génomes (plusieurs milliers de octets). Pour finir, il serait indispensable de créer un module de mise à jour automatique qui permettrait d'effectuer des analyses avec les données les plus complètes et les plus récentes.

Abstract

Second-order Hidden Markov Models (HMM2) are stochastic processes with a high efficiency in exploring bacterial genome sequences. Different types of HMM2 (M1M2, M2M2, M2M0) combined to combinatorial methods were developed in a new approach to discriminate genomic regions without *a priori* knowledge on their genetic content. This approach was applied on two bacterial models in order to validate its achievements: *Streptomyces coelicolor* and *Streptococcus thermophilus.*
These bacterial species exhibit distinct genomic traits (base composition, global genome size) in relation with their ecological niche: soil for *S. coelicolor* and dairy products for *S. thermophilus*. In *S. coelicolor*, a first HMM2 architecture allowed the detection of short discrete DNA heterogeneities (5-16 nucleotides in size), mostly localized in intergenic regions. The application of the method on a biologically known gene set, the SigR regulon (involved in oxidative stress response), proved the efficiency in identifying bacterial promoters. *S. coelicolor* shows a complex regulatory network (up to 12% of the genes may be involved in gene regulation) with more than 60 sigma factors, involved in initiation of transcription. A classification method coupled to a searching algorithm (i.e. *R'MES*) was developed to automatically extract the *box1-spacer-box2* composite DNA motifs, structure corresponding to the typical bacterial promoter -35/-10 boxes. Among the 814 DNA motifs described for the whole *S. coelicolor* genome, those of sigma factors (σ^{B}, σ^{WhiG}) could be retrieved from the crude data. We could show that this method could be generalized by applying it successfully in a preliminary attempt to the genome of *Bacillus subtilis*.

The Hidden Markov Models (HMM2) were also shown to be efficient to discriminate genome regions on a larger scale, i.e. the size of a bacterial gene. These new models were applied to *S. thermophilus* genome and preliminary analyses revealed that a large part of the tackled genes had functions reported to be prone to horizontal gene transfer, HGT (transporters involved in antibiotic resistance, restriction-modification systems, exopolysaccharide biosynthesis, mobile genetic elements). In order to confirm these early predictions, the distribution of these atypical genes in phylogenetic reconstructions of the firmicutes group to which the streptococci belong was established. A phylogenetic reconstruction of the *S. thermophilus* species (47 strains) was also achieved from data collected in the frame of the collaboration with the Chr HANSEN compagny (Danemark). Vertical heredity (from ancestor to progeny) versus horizontal heredity (gene exchange between bacteria within the same ecosystem) can be inferred from the distribution of a gene through a phylogenetic tree retracing the species past. A broad or at contrary a sporadic distribution provides a rough dating of the acquisition of a gene. Among the 362 extracted genes (corresponding to 18% of the global *S. thermophilus* CNRZ1066 content), two-thirds (240 genes) show a distribution consistent with an HGT origin. The distribution of these genes at different phylogenetic levels (i.e. firmicutes and *S. thermophilus* species) allowed defining a set of 23 genes specific to the *S. thermophilus* species which could turn out to be niche specific. These genes could have provided adaptation of the bacteria to its environment. Hence, it is now clearly admitted that *S. thermophilus* has very recently emerged (*circa* 3-30.000 years) from an ancestor of *Streptococcus salivarius*. Its emergence would have been concomitant to the development of dairy processes, and would have hale *S. thermophilus* to adapt to a new ecological niche (i.e. yoghurts and cheeses). While *S. thermophilus* was reported as a species showing a clonal genetic structure with a very low genetic heterogeneity, these results suggest that HGT is very frequent in this species allowing an unexpected diversification of the accessory genome. The evolutionary schema of the bacterial species appears typical of a recent and quick adaptation to a new ecosystem.
The interfaced or source code algorithms together with the whole data are available on the URL of "bio-informatique en lorraine". This whole work constitutes an original data mining approach combining stochastic and data analysis methods.

Keywords: Bioinformatics, data mining, second order hidden Markov model, transcriptional factor binding site, stochastic and combinatorial approach, horizontal gene transfer, *Streptomyces coelicolor*, *Streptococcus thermophilus.*

Résumé

Les modèles de Markov d'ordre 2 (HMM2) sont des modèles stochastiques qui ont démontré leur efficacité dans l'exploration de séquences génomiques. Cette thèse explore l'intérêt de modèles de différents types (M1M2, M2M2, M2M0) ainsi que leur couplage à des méthodes combinatoires pour segmenter les génomes bactériens sans connaissances *a priori* du contenu génétique. Ces approches ont été appliquées à deux modèles bactériens afin d'en valider la robustesse : *Streptomyces coelicolor* et *Streptococcus thermophilus*. Ces espèces bactériennes présentent des caractéristiques génomiques très distinctes (composition, taille du génome) en lien avec leur écosystème spécifique : le sol pour les *S. coelicolor* et le milieu lait pour *S. thermophilus*.

Chez *Streptomyces coelicolor*, un premier système de fouille de données articulé autour de HMM2 a permis d'extraire des hétérogénéités discrètes (longueur 5-16 nucléotides) des régions intergéniques. Appliqué au régulon SigR (impliqué dans la réponse au stress oxydant), le modèle M2M0 a démontré son efficacité pour l'identification de promoteurs bactériens. *S. coelicolor* présente un réseau de régulation génique complexe (pas moins de 12% des gènes coderaient des régulateurs) dont au moins 60 facteurs transcriptionnels sigma. Afin d'automatiser le processus d'extraction et de recherche de promoteurs bactériens, une recherche de motifs composites *box1-séparateur-box2* a alors été développée à l'aide du logiciel *R'MES* et d'une méthode originale de classification des hétérogénéités décelées par les HMM2. Parmi les 814 motifs régulateurs potentiels décrits chez *S. coelicolor*, ceux de deux facteurs sigma (σ^{B} et σ^{WhiG}) ont ainsi pu être décrits. La généralisation de cette méthode à d'autres génomes bactériens a été montrée sur le modèle *Bacillus subtilis*.

Une méthode de fouille similaire a également permis de segmenter le génome en unités de tailles plus importantes, à l'échelle du gène. Ces nouveaux modèles stochastiques de type M2M2 ont été appliqués au génome de la bactérie *Streptococcus thermophilus*. Les analyses préliminaires ont permis de conclure qu'une partie des séquences extraites correspondait à des fonctions connues pour être soumises au transfert horizontal (transporteurs spécifiques impliqués dans la résistance aux antibiotiques, systèmes de restriction-modification, synthèse d'exopolysaccharides, éléments génétiques mobiles). Afin de confirmer que ces modèles constituent des méthodes d'identification de séquences exogènes, la distribution des gènes extraits dans l'arbre phylogénétique des firmicutes, groupe bactérien auquel appartiennent les streptocoques, a été établie. La phylogénie de l'espèce *S. thermophilus* a été également reconstruite à partir de données portant sur un échantillon de 47 souches au travers d'une collaboration développée avec le groupe Chr HANSEN S/A (Danemark). L'hérédité verticale (ascendant vers descendant) ou horizontale (issue d'échanges entre bactéries présentes dans le même écosystème) peut être inférée selon la distribution d'un gène dans la phylogénie. Ainsi, une distribution large ou au contraire sporadique permet d'interpréter l'acquisition d'un gène dans le patrimoine bactérien comme étant plus ou moins ancienne. Sur les 362 gènes extraits (représentant 18% du génome de *S. thermophilus* CNRZ1066) par les HMM2, les deux-tiers (240) présentent une distribution compatible avec une acquisition par transfert horizontal. L'analyse de la distribution de ces gènes à différentes profondeurs phylogénétiques a permis d'identifier un échantillon de 23 gènes spécifiques de l'espèce *S. thermophilus*. Ces gènes pourraient avoir participé à l'adaptation de la bactérie à son environnement. En effet, il est maintenant admis que *S. thermophilus* a émergé très récemment (*circa* 3-30.000 ans) à partir d'un ancêtre de *Streptococcus salivarius*. Cette émergence, concomitante au développement des pratiques de fermentation laitière, contraint *S. thermophilus* à s'adapter à une nouvelle niche écologique. Alors que *S. thermophilus* a été décrite comme une espèce à évolution clonale, c'est-à-dire sans hétérogénéité génétique, ces travaux suggèrent que le transfert horizontal est très fréquent chez cette espèce diversifiant fortement le génome accessoire. Ainsi, le schéma évolutif de cette espèce apparaît caractéristique d'une évolution récente et rapide vers un nouvel écosystème.

Les algorithmes interfacés ou leur code ainsi que les données générées dans les deux modèles bactériens seront accessibles sur le site bioinformatique en lorraine. L'ensemble constitue une approche originale en fouille de données génomiques combinant des méthodes stochastiques et d'analyse de données.

Mots-clés : Bioinformatique, fouille de données, modèle de Markov du second ordre, site de fixation des facteurs de transcription, approche stochastique et combinatoire, transfert horizontal de gènes, *Streptomyces coelicolor, Streptococcus thermophilus.*