Evolution du génome des Streptomyces : transfert horizontal et variabilité des extrémités chromosomiques

Frédéric Choulet

FACULTE DES SCIENCES & TECHNIQUES

U.F.R. Sciences & Techniques Biologiques
Ecole Doctorale Biologie Santé Environnement

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université Henri Poincaré, Nancy 1

en Génétique Moléculaire

par

Frédéric CHOULET

Evolution du génome des *Streptomyces* : transfert horizontal et variabilité des extrémités chromosomiques

Soutenue le 10 Novembre 2006

Membres du jury :

Rapporteurs :	Mme Cécile FAIRHEAD	Chargée de Recherche Institut Pasteur, Paris
	M. Eduardo ROCHA	Chargé de Recherche CNRS, Paris
Examinateurs :	M. Pascal SIMONET	Directeur de Recherche CNRS, Lyon (Président)
	M. Jean-Luc PERNODET	Directeur de Recherche CNRS, Orsay
	M. Bertrand AIGLE	Maître de Conférences, UHP, Nancy 1
	M. Bernard DECARIS	Professeur, UHP, Nancy 1
	M. Pierre LEBLOND	Professeur, UHP, Nancy 1 (Directeur de thèse)

Laboratoire de Génétique et Microbiologie – UMR INRA/UHP 1128 – IFR 110
Faculté des Sciences et Techniques – 54506 Vandœuvre-lès-Nancy

REMERCIEMENTS

Je tiens à remercier le Professeur *Bernard Decaris* 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. Merci de m'avoir fait confiance et de m'avoir soutenu dans mes choix.

Un immense merci au Professeur *Pierre Leblond*, mon directeur de thèse, pour m'avoir proposé de travailler sur la génomique de *S. ambofaciens* et de m'avoir accordé sa confiance pour mener à bien ce projet. Merci pour la qualité de son encadrement et la pertinence de ses remarques. Je le remercie également pour sa disponibilité quotidienne et pour son degré d'implication dans ces travaux. Merci enfin pour sa sympathie, sa simplicité et sa convivialité.

Je remercie vivement *Bertrand Aigle*, Maître de conférences, pour sa participation active à ce travail et ses précieux conseils. Merci d'avoir encadré ces travaux et d'avoir été aussi disponible et sympathique tout au long de ma thèse.

Merci à Madame *Cécile Fairhead*, Chargée de recherche à l'Institut Pasteur, et à Monsieur *Eduardo Rocha*, Chargé de recherche CNRS, pour me faire l'honneur d'être rapporteurs de ce travail de thèse.
Merci à Monsieur *Pascal Simonet*, Directeur de recherche CNRS pour avoir accepté de juger ce travail.
Merci à Monsieur *Jean-Luc Pernodet*, Directeur de recherche CNRS, pour sa participation à ce travail et pour ces très bons conseils concernant l'exploitation des données.

Ce travail est le fruit d'une collaboration entre le Laboratoire de Génétique et Microbiologie de Nancy, l'Institut de Génétique et Microbiologie d'Orsay et le Génoscope. Je tiens donc à remercier tous les acteurs de cette collaboration sans qui ce projet n'aurait pas abouti :
- Merci à *Jean-Luc Pernodet*, *Michel Guérineau*, *Claude Gerbaud* et *François-Xavier Francou* de l'Institut de Génétique et Microbiologie pour leur participation à ce travail.
- Merci à *Valérie Barbe*, *Sophie Mangenot* et *Chantal Truong* du Génoscope pour leur participation au séquençage du génome *S. ambofaciens*.
- Merci à *Frédéric Borges*, *Alexandre Gallois* et *Céline Fourrier* qui complètent l'équipe du laboratoire de Nancy ayant participée à ce travail.

Durant ma thèse, j'ai eu la chance de pouvoir profiter d'une collaboration établie entre le laboratoire de Génétique et Microbiologie de Nancy et le John Innes Centre de Norwich, UK. Je tiens donc à remercier *Bertrand Aigle* et *Keith Chater* pour m'avoir permis d'être un acteur de cette collaboration. Merci à *Keith Chater*, *Tobias Kieser* et *Govind Chandra* pour leur accueil et l'aide précieuse qu'ils m'ont apportée notamment au moment de commencer ces travaux. Merci tout particulièrement à Govind pour m'avoir initié aux méthodologies bioinformatiques.

Merci à tous les membres du laboratoire, passés et présents, que j'ai pu côtoyer. Difficile de citer tout le monde, alors merci à tous. J'ai été ravi de pouvoir travailler dans une ambiance aussi bonne.

Merci à tous mes amis. Merci à *Mélanie* pour la lecture et la critique du manuscrit.

Merci à mes parents d'avoir été présents tout au long de mon cursus et de m'avoir soutenu dans mes choix d'orientation. J'espère qu'ils liront ce manuscrit avec le plus grand intérêt et la plus grande attention.

Merci à *Géraldine* pour avoir rendu ces quelques années aussi belles. Merci d'avoir partagé les moments difficiles. Cette thèse n'aurait pas eu la même saveur sans elle.

SOMMAIRE

PREAMBULE

Depuis le séquençage complet du premier génome bactérien par le groupe de Craig Venter (TIGR) en 1995 (Fleischmann *et al.*, 1995), le nombre de séquences de génomes de procaryotes ne cesse d'augmenter (344 bactéries et 28 archées, en date du 31 août 2006). Cette abondance de données a révolutionné nos connaissances sur l'organisation et le contenu génétique des génomes bactériens. En l'absence de fossiles bactériens, l'analyse des génomes des espèces actuelles est l'unique moyen d'évaluer les caractères ancestraux et évolués et de reconstruire le passé évolutif. La génomique comparée permet de révéler les forces qui modèlent les génomes bactériens et d'approcher les voies d'évolution spécifiques qu'ont empruntées les bactéries pour assurer un succès évolutif sans égal.

Le paradigme du chromosome bactérien unique et circulaire a vécu. L'essor de la génomique a révélé une diversité en taille, composition, configuration et organisation des génomes. Les *Streptomyces*, bactéries sur lesquelles portent ces travaux et réflexions, présentent plusieurs caractéristiques génomiques extrêmes et originales chez les bactéries cultivables : une grande taille (8-11 Mb), une configuration linéaire et une composition en bases G+C de 71-73%. Elles sont également connues pour être affectées par des phénomènes d'instabilité génétique et chromosomique spectaculaires. D'un point de vue applicatif, ces bactéries synthétisent une grande diversité de métabolites secondaires présentant des intérêts en médecine, agro-alimentaire et biotechnologies.

Au début de ce travail, la séquence du génome de *Streptomyces coelicolor* (8,7 Mb), modèle d'étude de la différenciation bactérienne tant au niveau biochimique que morphologique, venait d'être publiée par Bentley *et al.*, 2002. Son analyse a corroboré les études préalables sur l'organisation très spécifique de l'information génétique, à savoir le confinement des gènes dits "essentiels" dans la région centrale du chromosome, et les gènes dits "accessoires" dans les régions terminales (appelées bras chromosomiques). La comparaison des génomes de *S. coelicolor* et de *Mycobacterium tuberculosis* (actinomycète apparenté à *Streptomyces* et agent de la tuberculose) avait également permis d'émettre l'hypothèse selon laquelle la linéarité et les régions terminales auraient été acquises lors d'un événement unique.

En 2000, une collaboration entre notre laboratoire (équipe de Pierre Leblond, Nancy), l'Institut de Génétique et Microbiologie (équipe de Jean-Luc Pernodet, Orsay) et le Génoscope (équipe de Valérie Barbe, Evry) a été initiée afin de séquencer et décrypter les régions terminales du chromosome de *Streptomyces ambofaciens* sur environ 3 Mb. La comparaison avec *S. coelicolor* et *Streptomyces avermitilis* (séquence publiée en 2003 par Ikeda *et al.*) a alors permis d'obtenir une vision dynamique de l'évolution des régions terminales spécifiques mais également des régions centrales conservées (malgré la stratégie de séquençage partiel retenue par le Génoscope).

Les questions posées concernaient alors l'organisation génétique, l'étendue et le niveau de spécificité des régions terminales. Dans le cadre d'autres travaux de notre équipe, la recherche de voies de biosynthèse de métabolites d'intérêt rentrait également dans les objectifs de cette analyse.

Puis, l'analyse s'est focalisée sur l'origine de l'organisation très spécifique du chromosome des *Streptomyces* (compartimentation génomique) ainsi que sur l'importance de la linéarité chromosomique dans la variabilité remarquable observée chez ces bactéries. La linéarité chromosomique, qui correspond à un caractère évolué, confère-t-elle un avantage en favorisant l'acquisition et le réarrangement de l'information génétique ?

L'introduction de ce manuscrit présente la diversité des génomes bactériens puis les contraintes et tolérances qui les modèlent.
Les résultats traitent, dans un premier temps, de la stratégie de séquençage du génome de *S. ambofaciens* et des méthodes bioinformatiques développées. Ils détaillent ensuite le décryptage des régions terminales en termes de composition et de codage (notamment ceux impliqués dans le métabolisme secondaire). Puis, ils se focalisent sur la comparaison de l'ensemble des séquences obtenues pour *S. ambofaciens* ATCC23877 avec les génomes séquencés de *S. coelicolor*, *S. avermitilis* et *S. scabies* (séquence accessible mais non annotée). Les mécanismes et forces qui dirigent le confinement de la variabilité aux extrémités du chromosome y sont débattus (article dans *Molecular Biology and Evolution*). Les régions terminales apparaissent comme des cibles privilégiées de transferts horizontaux et la taille des régions spécifiques terminales augmente avec la distance phylogénétique. Dans une dernière partie, la variabilité intraspécifique des répétitions terminales (~200 kb) est abordée. Les résultats aboutissent à la notion d'intégration et d'échanges de matériel génétique par l'intermédiaire des extrémités de réplicons linéaires (article dans *Journal of Bacteriology*).
La discussion propose des hypothèses quant aux mécanismes moléculaires associés à un gradient de fréquence croissante de réarrangements vers les extrémités chromosomiques. L'analyse est élargie à l'ensemble des génomes procaryotes séquencés.

Certains résultats sont présentés sous forme de résumés. Ils sont détaillés dans les publications suivantes qui ont été insérées dans les chapitres correspondants :

Publication n°1 :
Choulet F., Aigle B., Gallois A., Mangenot S., Gerbaud C., Truong C., Francou F.-X., Fourrier C., Guérineau M., Decaris B., Barbe V., Pernodet J.-L., Leblond P. Evolution of the terminal regions of the *Streptomyces* linear chromosome.
Molecular Biology and Evolution, Nov. 2006, Vol. 23, No. 11, sous presse.

Publication n°2 :
Choulet F., Gallois A., Aigle B., Mangenot S., Gerbaud C., Truong C., Francou F.-X., Borges F, Fourrier C., Guérineau M., Decaris B., Barbe V., Pernodet J.-L., Leblond P. Intraspecific variability of the terminal inverted repeats of the linear chromosome of *Streptomyces ambofaciens.*
Journal of Bacteriology, Sept. 2006, p. 6599-6610 Vol. 188, No. 18.

Ces travaux ont par ailleurs été exposés lors de différents congrès nationaux et internationaux :

Choulet *et al.* Horizontal transfer and evolution of the terminal regions of the linear chromosome of *Streptomyces.* ***Rencontres des Microbiologistes de l'INRA**, Dourdan, France, Juin 2006*

Choulet *et al.* Horizontal transfer and evolution of the terminal regions of the linear chromosome of *Streptomyces.* ***Journées Actinomycètes 2006**, Université Claude Bernard Lyon I, Lyon, France, Juin 2006*

Choulet *et al.* DNA rearrangements and integration of exogenous information in the terminal regions of the linear chromosome of *Streptomyces ambofaciens* revealed by comparative genomics, ***Prokagen 2005**, 2nd European Conference on Prokaryotic Genomes, University of Göttingen, Germany, Sep. 2005*

Choulet *et al.* Sequence annotation of the terminal regions of the chromosome of *Streptomyces ambofaciens.* Comparative genomics and evidence for an important interspecific variability, ***5èmes Journées Ouvertes de Biologie, Informatique et Mathématiques**, Montréal, Canada, Juin 2004, article JO72*

Choulet *et al.* Mise en évidence d'une importante variabilité entre les régions terminales des chromosomes de deux espèces phylogénétiquement proches : *Streptomyces ambofaciens* et *Streptomyces coelicolor*, ***Rencontre du club des Actinomycètes**, Centre d'Ingénierie des protéines, Université de Liège, Belgique, Mai 2004*

Choulet *et al.* Evidence for a High Heterogeneity Between the Terminal Regions of the Chromosomes of Two Phylogenetically Close Species, *Streptomyces ambofaciens* and *Streptomyces coelicolor* by Genome Comparison, ***Dissemination meeting**, Guilford, University of Surrey, UK, Juil 2003*

Choulet *et al.* Sequence analysis of the *Streptomyces ambofaciens* ATCC23877 large Terminal Inverted Repeats (210 kb) and evidence of high heterogeneity with the *Streptomyces coelicolor* A3(2) genome, ***Biology of streptomycetes and related actinomycetes**, University of Munster, Germany, Fév. 2003*

INTRODUCTION

INTRODUCTION

A. Organisation et diversité des génomes procaryotes

Les variations des différents paramètres qui caractérisent les génomes sont telles que le génome modèle n'existe pas. Une caractéristique reste, néanmoins, commune à tous les génomes procaryotes : la densité élevée en ADN codant. Ainsi, la variabilité des différents paramètres influe sur les séquences et le contenu en gènes chez les bactéries.

1. Diversité de taille des génomes

La taille des génomes est extrêmement variable dans les trois règnes du vivant. Chez les bactéries et les archées, la taille peut varier d'un facteur 20 parmi les génomes séquencés (Bentley et Parkhill, 2004 ; Ussery et Hallin, 2004a, b) (Fig.1). Le plus petit d'entre eux est celui d'une archée : *Nanoarchaeum equitans* avec 490.885 pb (536 CDS (Coding DNA Sequences) identifiées, (Waters *et al.*, 2003)). Cependant, le génome séquencé contenant le plus petit nombre de gènes prédits est celui d'une espèce bactérienne : *Mycoplasma genitalium* avec 476 CDS (580.076 pb ; (Fraser *et al.*, 1995)). A l'opposé, le plus grand génome bactérien séquencé est celui de la Protéobactérie alpha *Bradyrhizobium japonicum* avec 9.105.828 pb (parmi les archées : *Methanosarcina acetivorans*; 5.751.492 pb ; analyses réalisées à partir des génomes disponibles dans la banque EMBL (European Molecular Biology Laboratory, http://www.ebi.ac.uk/genomes/) en date du 10 mai 2006 : 312 souches de bactéries et 26 d'archées, Tableau annexe). Vu le nombre élevé de projets de séquençage en cours, ces valeurs extrêmes sont amenées à changer rapidement. Par exemple, le chromosome de *Streptomyces scabies* est disponible, mais non annoté, sur le site Internet du Centre Sanger (ftp://ftp.sanger.ac.uk/pub/pathogens/ssc/). Il s'agit du plus grand génome procaryote séquencé et disponible avec 10.148.695 pb.

Cette taille varie en fonction des événements de duplications, de transferts horizontaux et de pertes de gènes. Au sein d'un même phylum bactérien, la variation de la taille du génome peut être considérable. Par exemple, les Actinobactéries, auxquelles appartiennent les *Streptomyces*, regroupent des espèces dont le génome peut être parmi les plus petits (*Tropheryma whipplei*, 930 kb) ou, à l'inverse, parmi les plus grands connus (*Streptomyces*, entre 8 et 11 Mb ; Fig.1). Il en est de même pour les Protéobactéries. Etant donné l'éloignement phylogénétique de ces deux phyla, ces données suggèrent que les facteurs limitant la taille des génomes seraient communs à toutes les espèces. De ces constatations est né le concept de génome minimal qui définit le plus petit nombre de gènes requis pour la reproduction d'une cellule (Klasson et Andersson, 2004 ; Koonin, 2000). L'idée d'un génome minimal implique une niche écologique extrêmement riche. En effet, les organismes endosymbiotiques, qui vivent dans des conditions environnementales très riches et stables dans le temps, possèdent les plus petits génomes connus. Au contraire, les plus grands génomes bactériens connus sont retrouvés chez des organismes vivant de façon autonome dans des environnements complexes et changeants. C'est le cas des *Streptomyces* qui sont retrouvés principalement dans le sol. Cependant, cette règle ne semble pas s'appliquer aux pathogènes humains. En effet, on retrouve des

organismes à petits génomes comme certains streptocoques et d'autres avec de grands génomes comme *Bacillus anthracis* ou encore certaines espèces de *Salmonella*. Toutefois, les pathogènes humains possèdent très fréquemment un cycle de vie saprophyte, expliquant cette diversité de taille. Le génome minimal reste une notion sujette à controverses car sa définition est dépendante de celle d'un organisme vivant. Par exemple, le nombre de gènes utiles au développement de bactéries intracellulaires est très faible étant donné leur environnement très particulier. En 1989, Schmid *et al.* ont évalué à 200 le nombre de gènes essentiels chez *Escherichia coli* par obtention de mutants thermosensibles en croissance sur milieu riche (Schmid *et al.*, 1989). En 1996, ce nombre a été porté à 256 par génomique comparée (Mushegian et Koonin, 1996a). Par mutagenèse du chromosome de *Mycoplasma genitalium* (476 CDS prédites), ce nombre a d'abord été estimé entre 265 et 350 (Hutchison *et al.*, 1999) puis, par une technique améliorée, à 382 (Glass *et al.*, 2006) sans compter les gènes codant les ARN.

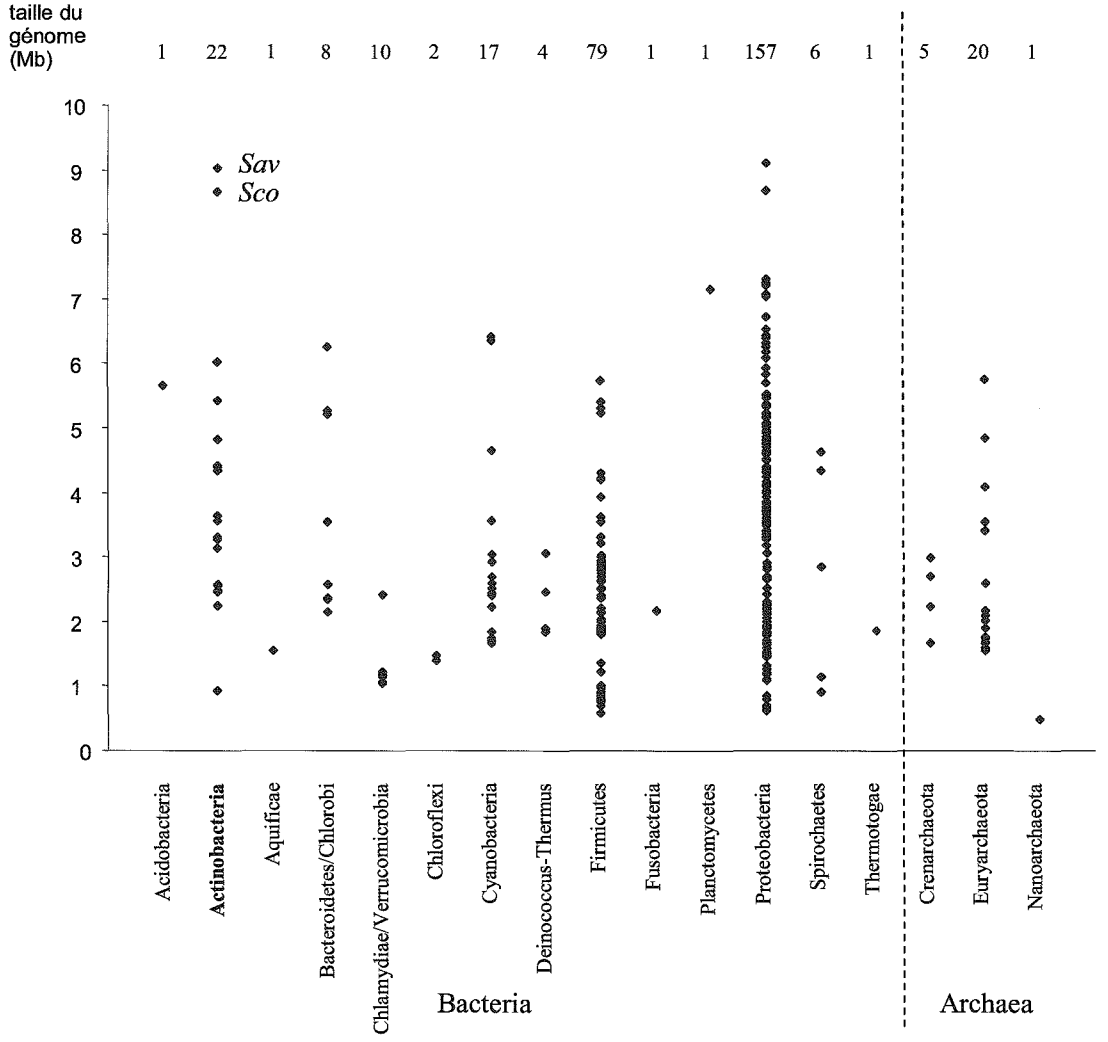


Figure 1 : Taille des génomes procaryotes séquencés pour chaque phylum.
A chaque génome séquencé correspond un point sur le graphe. Le nombre de génomes disponibles dans chaque phylum est indiqué au dessus de chaque nuage de points. L'analyse a été réalisée sur 312 génomes de bactéries et 26 génomes d'archées (ensemble des génomes disponibles en mai 2006) à l'aide de la base de données GENOMECOMP réalisée au cours de ce travail. *Sco* : *S. coelicolor*, *Sav* : *S. avermitilis*.

A l'opposé du concept de génome minimal, est-il possible de parler de génome maximal ? Cette notion apparaît nettement moins évidente à appréhender. Elle fait intervenir les facteurs limitant l'expansion des génomes à l'infini : le temps de génération et la vitesse de réplication, l'énergie nécessaire à la réplication et enfin la taille physique du génome et celle de la cellule.

Bien que la taille des génomes bactériens soit variable, la densité en gènes varie, quant à elle, extrêmement peu (Mira *et al.*, 2001). Elle est élevée : 86,0 +/- 5,7% en moyenne et s'étend de 50% chez *Mycobacterium leprae* à 96% chez *Candidatus Pelagibacter ubique* (calculé sur 338 génomes disponibles). Cette valeur révèle, par conséquent, une pression forte sur la taille des génomes qui s'oppose à l'accumulation de séquences d'ADN non fonctionnelles. L'augmentation de la taille du génome chez les bactéries s'accompagne donc d'une augmentation du nombre de gènes à la différence des eucaryotes. Théoriquement, l'augmentation du nombre de gènes chez une espèce bactérienne peut engendrer deux types de phénomènes : une augmentation du nombre de familles de protéines codées et/ou une augmentation du nombre de membres dans chaque famille.
Des analyses ont été réalisées sur 115 génomes procaryotiques (Konstantinidis et Tiedje, 2004) à l'aide de la base de données COG (Cluster of Orthologous Groups (Tatusov *et al.*, 2001)) et sur 56 génomes (Ranea *et al.*, 2004) en utilisant la base de données structurales CATH (Class Architecture Topology and Homologous superfamily (Orengo *et al.*, 2003)). Elles ont permis de distinguer deux grandes classes de familles de protéines : celle dont la fréquence est fortement corrélée à la variation de taille du génome et celle où aucune corrélation n'est observée. Parmi cette dernière figurent les protéines impliquées dans les mécanismes cellulaires tels que la traduction, la réplication/réparation de l'ADN, la division et la partition cellulaire ; leur nombre est sensiblement le même quelle que soit la taille du génome. En revanche, la première classe de familles regroupe des protéines impliquées dans la régulation, le métabolisme énergétique, la transduction du signal ou encore le métabolisme secondaire qui sont les fonctions en relation avec l'environnement (Fig. 2). Ainsi apparaît un lien fort entre la pression environnementale et la taille du génome.
Parmi les familles de protéines dont le nombre de représentants dans un génome est corrélé à sa taille, il est possible de distinguer celles dont le pourcentage augmente avec la taille du génome. C'est le cas des protéines impliquées dans les phénomènes de régulation (Ranea *et al.*, 2004). En effet, plus le nombre de gènes est grand, plus le pourcentage de régulateurs est élevé. Les génomes de *S. coelicolor* (et *S. avermitilis*) et de *Pseudomonas aeruginosa* (6.264.403 pb) sont de bons exemples puisque respectivement 12,3% et 9,4% de leur génome sont consacrés à la régulation (Bentley *et al.*, 2002 ; Ikeda *et al.*, 2003 ; Stover *et al.*, 2000). L'enrichissement en fonctions régulatrices et en métabolisme secondaire suggère que les grands génomes sont tolérés voire favorisés dans les environnements où les ressources sont rares mais diverses, comme le sont le sol (*Streptomyces*) et la rhizosphère (*Rhizobium*), et où la vitesse de croissance n'est pas un handicap pour le maintien dans une population.

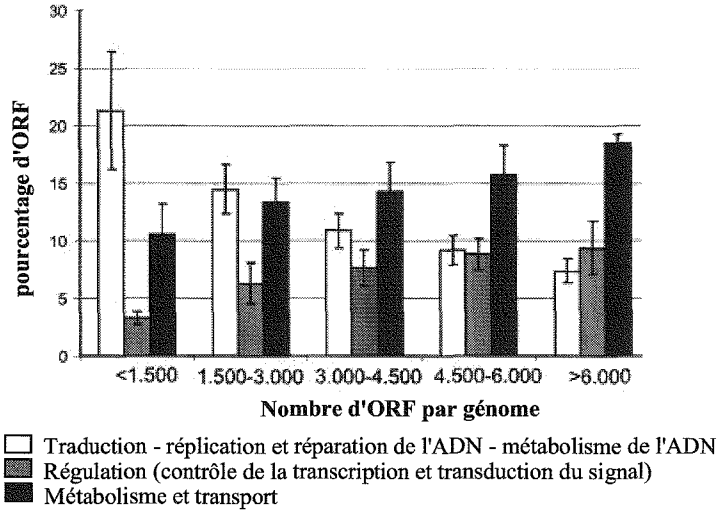


Figure 2 : Relation entre contenu en gènes et taille des génomes procaryotes.
L'histogramme représente les pourcentages d'ORF impliquées dans un même processus cellulaire (selon la classification par COG) dans un génome. L'évolution de ce pourcentage est corrélée à la taille du génome. Les écarts types sont représentés par les barres d'erreurs. D'après (Konstantinidis et Tiedje, 2004).

2. *Géométrie variable des génomes*

Bien que la plupart des bactéries connues possède un chromosome circulaire, des chromosomes linéaires ont toutefois été identifiés dans différents genres bactériens appartenant à trois phyla différents : *Streptomyces* et *Saccharopolyspora* (*Actinobacteria*, anciennement classé parmi les *Streptomyces* ; (Kieser *et al.*, 1992)), *Borrelia* (*Spirochaetes* ; (Ferdows et Barbour, 1989)) et *Agrobacterium* (*alpha-proteobacteria* ; (Allardet-Servent *et al.*, 1993)). Bien que des données expérimentales suggéraient la présence d'un chromosome linéaire chez *Coxiella* (Willems *et al.*, 1998), l'assemblage des séquences du chromosome de *Coxiella burnetii* a montré qu'il est circulaire (Seshadri *et al.*, 2003).

Les relations phylogénétiques entre ces bactéries montrent que la linéarité du chromosome est un caractère évolué apparu indépendamment dans différentes lignées et issu de l'ouverture d'un chromosome circulaire ancestral ((Volff et Altenbuchner, 2000), Fig. 3). Par ailleurs, l'événement inverse a pu être sélectionné en laboratoire chez *Streptomyces et Borrelia* (Ferdows *et al.*, 1996 ; Fischer *et al.*, 1997a ; Lin *et al.*, 1993). Malgré la viabilité des mutants possédant un chromosome circularisé, aucune souche ayant perdu la linéarité chromosomique n'a été isolée à l'état naturel. Certains plasmides peuvent se maintenir à la fois sous forme linéaire et circulaire. C'est le cas des plasmides pSLA2 (*Streptomyces lividans* ; (Chang et Cohen, 1994)) et pSCL (*Streptomyces clavuligerus* ; (Shiffman et Cohen, 1992)) isolés à l'état naturel sous forme linéaire mais dont les dérivés circulaires sont capables de se maintenir. Chez *Borrelia hermsii*, un plasmide de 180 kb est capable de se maintenir à l'état naturel, soit sous forme circulaire, soit linéaire (Ferdows *et al.*, 1996). Cette conversion est plus facilement envisageable chez les réplicons linéaires de *Borrelia* dont la réplication génère des molécules intermédiaires circulaires, ce qui ne semble pas être le cas chez *Streptomyces*. Cette différence résulte du mode de maintien des extrémités linéaires chez ces espèces qui ont adopté des stratégies totalement différentes, confirmant l'apparition indépendante de la linéarité chez les bactéries.

Les extrémités du chromosome (et des plasmides linéaires) des *Borrelia* sont des tiges-boucles fermées par une liaison covalente (Barbour et Garon, 1987 ; Casjens *et al.*, 1997 ; Tourand *et al.*,

2003). Les télomères des *Streptomyces* forment, quant à eux, des structures secondaires impliquant plusieurs tiges-boucles et sont liés de façon covalente à un complexe protéique composé des protéines Tap et Tpg (Qin et Cohen, 1998). *Agrobacterium tumefaciens* possède un chromosome circulaire qui dérive du chromosome ancestral et un chromosome linéaire qui dériverait d'un plasmide. Le mécanisme de maintien des extrémités est semblable à celui identifié chez *Borrelia* (Goodner *et al.*, 2001).

Le nombre de réplicons est également variable au sein des génomes bactériens. Ce nombre peut atteindre 21 plasmides (circulaires et linéaires) et 1 chromosome chez *Borrelia burgdorferi* (Casjens *et al.*, 2000). La distinction chromosome/plasmide peut être beaucoup plus problématique. Ce problème de classification résulte du flou qui entoure la notion d'essentialité. Pour de très nombreuses espèces pathogènes, certains plasmides sont porteurs de gènes essentiels au style de vie mais non essentiels en conditions de laboratoire, comme *Ralstonia solanacearum* (Salanoubat *et al.*, 2002) ou *Yersinia pestis* (Parkhill *et al.*, 2001b).
Malgré ces réserves, certaines espèces possèdent clairement deux chromosomes distincts comme *Vibrio cholerae* (Heidelberg *et al.*, 2000), *Deinococcus radiodurans* (White *et al.*, 1999), *Brucella melitensis* (DelVecchio *et al.*, 2002), *Burkolderia pseudomallei* (Songsivilai et Dharakul, 2000) et *Agrobacterium tumefaciens* (Goodner *et al.*, 2001). Des chromosomes supplémentaires semblent être issus de plasmides ayant acquis des fonctions essentielles. C'est le cas chez *V. cholerae*, *B. pseudomallei* et *A. tumefaciens*. En effet, la majorité des gènes essentiels sont portés par le chromosome n°1 mais des fonctions indispensables sont portées par le second réplicon (ex : un $tRNA^{ser}$ unique chez *B. pseudomallei*). De plus, les systèmes impliqués dans la partition et la réplication sont apparentés à des systèmes plasmidiques (Bentley et Parkhill, 2004).

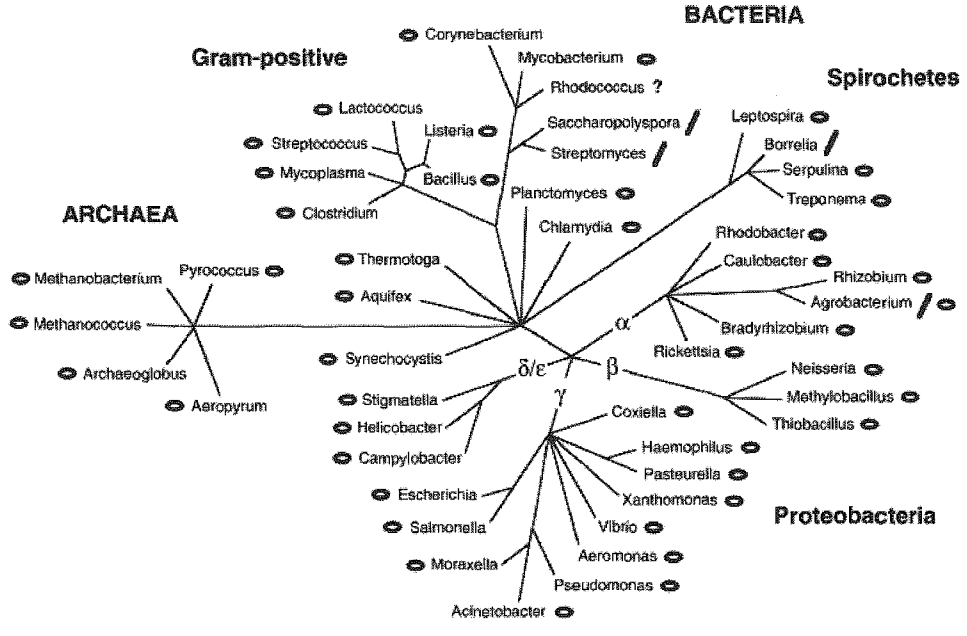


Figure 3 : Distribution des chromosomes linéaires (segments) et circulaires (cercles) chez les procaryotes. La situation n'est pas clarifiée chez *Rhodococcus*. D'après (Volff et Altenbuchner, 2000).

Chez *Streptomyces*, les souches naturelles connues possèdent toutes un chromosome unique et linéaire. De plus, certaines souches peuvent posséder un ou plusieurs plasmides circulaires et/ou linéaires. Cependant, une souche de *S. coelicolor* porteuse de deux chromosomes (7,2 Mb et 1,8 Mb) a été isolée en laboratoire (Yamasaki et Kinashi, 2004). Ils sont issus d'un événement de recombinaison illégitime entre le chromosome sauvage (8,7 Mb) et le plasmide linéaire SCP1 (0,356 Mb). L'événement s'est produit dans des séquences codantes (SCP1.136, hélicase putative ; SCO6388, fonction inconnue) et a engendré le passage des 1,6 Mb terminaux du bras chromosomique droit sauvage sur le plasmide SCP1. Les deux réplicons hybrides créés sont indispensables à la survie de ce clone indiquant que des gènes essentiels sont portés par les 1,6 Mb terminales du chromosome de *S. coelicolor*.

3. *Biais de composition en nucléotides*

a) Pourcentages en bases G+C et A+T

La composition en nucléotides d'un génome bactérien est soumise à des contraintes. Le pourcentage en bases G et C est, par exemple, un caractère peu variable au sein d'une même espèce bactérienne. Paradoxalement, la diversité des génomes bactériens est telle que ce pourcentage varie entre deux valeurs extrêmes : 26,5% chez l'endosymbiote obligatoire *Wigglesworthia glossinidia* (Akman *et al.*, 2002) et 74,9% chez *Anaeromyxobacter dehalogenans*. Etant données la forte densité en gènes codant des protéines chez les bactéries et les contraintes dues au codage des acides aminés, ces valeurs de pourcentage avoisinent le seuil au-delà duquel il n'est plus possible de coder des protéines. Pour preuve, le pourcentage en bases G et C est si élevé chez *Streptomyces* (72,1% chez *S. coelicolor*) que le biais à la troisième position des codons est, en moyenne, de 92,1% chez *S. coelicolor* (calculé sur toutes les CDS prédites). Cette particularité est par ailleurs utilisée afin d'identifier les bornes des phases codantes grâce au profil appelé "GC frameplot" (Ishikawa et Hotta, 1999). Etant donnée la stabilité du pourcentage en G/C (et A/T) au sein d'une espèce et étant donnée sa variabilité au sein du monde bactérien, cette mesure est très utilisée pour identifier les régions potentiellement acquises récemment par transfert horizontal.

Comment expliquer les écarts considérables de ce pourcentage entre différentes espèces bactériennes ? Tout d'abord, une corrélation positive a été observée entre la proportion de bases G/C et la taille du génome (étude sur 146 génomes (Bentley et Parkhill, 2004)). Par ailleurs, une corrélation peut aussi être établie entre ce paramètre et la niche écologique. Rocha et Danchin ont montré que les génomes des bactéries dépendantes d'un hôte, c'est-à-dire les endosymbiotes et les pathogènes obligatoires, tendent vers une richesse en bases A et T (Rocha et Danchin, 2002). De même, les plasmides, les phages et les séquences d'insertion (IS), pouvant être considérés comme des éléments "parasites" (ou symbiotiques) intragénomiques, montrent également un taux plus riche en bases A et T que le chromosome. Cette étude a été réalisée sur 52 génomes bactériens, 59 génomes de phages, 54 plasmides et 368 IS. La moyenne du pourcentage en bases G/C est de 38% pour les génomes des bactéries dépendant d'un hôte et de 49% pour les génomes de celles vivant de façon autonome (Fig. 4). Les phages et plasmides sont en moyenne respectivement 4% et 2,7% plus riches en bases A/T que leur génome hôte. De plus, cette différence n'est pas associée au type de gènes portés. En effet, la

comparaison de 245 gènes plasmidiques avec leur homologue chromosomique montre une richesse en bases A/T de +2% en moyenne.

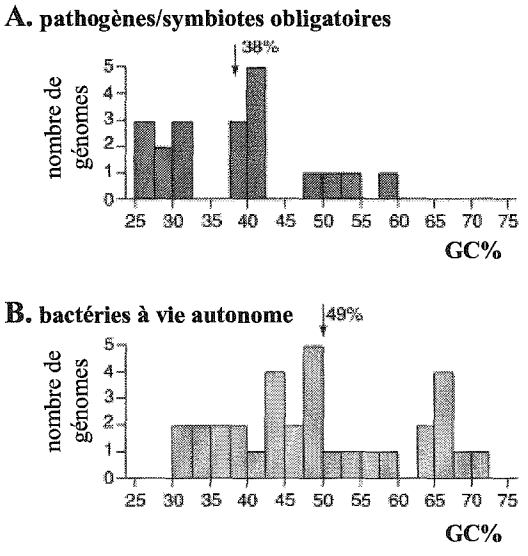


Figure 4 : Distribution du pourcentage en bases G+C chez des génomes d'organismes pathogènes (ou symbiotiques) obligatoires (A.) et d'organismes à vie autonome (B.). Les flèches indiquent la valeur moyenne. D'après (Rocha et Danchin, 2002).

Chez les bactéries à petit génome, les phases de réduction génomique ont engendré la perte de fonctions de réparation de l'ADN et de recombinaison (Moran, 2002). Dans de telles circonstances, les mutations ont pu s'accumuler et plus particulièrement la déamination de la cytosine en uracile (C→U, répliquée en thymine) qui est le mécanisme mutateur le plus fréquemment rencontré. La perte des capacités de réparation favoriserait donc l'enrichissement en bases A et T. De plus, la disponibilité plus limitée des molécules de GTP et CTP par rapport aux ATP et UTP/TTP serait un facteur déterminant qui oriente le biais vers une richesse en A/T (Rocha et Danchin, 2002). En effet, l'abondance du pool d'ATP dans la cellule s'explique par son rôle central dans le métabolisme énergétique. L'estimation de ces différents "pools" a été réalisée chez *E. coli* en croissance exponentielle et des différences notables ont pu être identifiées aussi bien concernant les pyrimidines triphosphates (ATP : 3,5 mM ; GTP : 1,9 mM) que les purines triphosphates (UTP : 2,0 mM ; CTP : 1,2 mM) (Danchin *et al.*, 1984).

Selon une hypothèse neutraliste, les différences observées des proportions en bases G/C par rapport à A/T résulteraient de biais mutationnels. Toutefois, Rocha et Danchin favorisent une hypothèse sélectionniste selon laquelle le biais en bases A/T des génomes des bactéries dépendantes d'un hôte résulterait d'une compétition pour les ressources. Ce biais serait sélectionné car il permettrait une meilleure exploitation des ressources de la cellule hôte. Le coût énergétique plus élevé que représente la synthèse des nucléotides GTP/CTP par rapport aux ATP/UTP et la disponibilité plus grande des ATP/UTP expliqueraient un tel avantage sélectif. Il apparaît très peu probable que chaque mutation G/C→A/T apporte un avantage sélectif qui permet sa fixation. Il est plus vraisemblable de penser que ce sont la ou les mutations engendrant ce biais mutationnel vers une richesse en A/T qui seraient sélectionnées. Dans des conditions de ressources limitées, la richesse en A/T deviendrait un avantage

sélectif. La réplication d'un génome plus riche en A/T serait ainsi favorisée. Cette hypothèse explique également le biais observé chez les phages, plasmides et IS qui sont en compétition avec le génome hôte pour leur maintien. Enfin, le taux généralement plus élevé en A/T des régions chromosomiques acquises par transfert horizontal, par exemple chez *Bacillus subtilis* (Moszer *et al.*, 1999), pourrait être expliqué par leur passage par des vecteurs de type phages ou plasmides.

Le biais de composition en nucléotides est plus prononcé dans les régions intergéniques et à la troisième base des codons, soumises de façon moindre aux pressions de sélection. Néanmoins, le biais touche toutes les positions des codons et affecte donc l'usage des acides aminés dans les protéines. Ce biais est visible par le calcul du point isoélectrique (pI) du protéome. Par exemple, chez *Buchnera aphidicola* (26,3% de bases G+C) le pI du protéome est très basique, 9,6 (*E. coli* : 7,2 ; *H. influenzae* : 7,3), dû notamment à une richesse particulière en lysine (codée par les triplets AAA et AAG ; (Shigenobu *et al.*, 2000)).
Considérant les fonctions des gènes, ce biais est plus marqué dans les gènes non essentiels. C'est le cas pour *Ureaplasma urealyticum* (Glass *et al.*, 2000) et *M. leprae* (Cole *et al.*, 2001) par exemple. La dérive en A/T s'oppose donc à la pression de sélection qui s'exerce sur les fonctions.

b) G/C skew

Un second type de biais de composition en nucléotides bien connu est appelé "G/C skew". Il est intimement lié au processus de la réplication. La majorité des génomes bactériens séquencés se caractérisent par une majorité de bases G par rapport à C sur le brin continu. Ainsi, la mesure de la proportion de bases G comparées aux bases C (G-C/G+C), ou inversement, permet de caractériser graphiquement un tel biais. Le profil correspondant se caractérise par une inversion ("shift") au niveau des sites d'initiation et de terminaison de la réplication, c'est-à-dire aux loci où le brin continu devient le brin discontinu et réciproquement. Ces deux transitions sont bien marquées pour le chromosome d'*E coli* K12 (Fig. 5). Ce biais est d'ailleurs utilisé pour déterminer les loci d'initiation et de terminaison de la réplication. Chez certaines espèces comme *Yersinia pestis* CO92, par exemple, des régions se caractérisent par un biais aberrant suggérant qu'elles dérivent, soit d'îlots génomiques nouvellement acquis, soit d'événements de recombinaison intragénomique (Fig. 5) (Parkhill *et al.*, 2001b). La présence de copies d'IS apparentées bornant ces trois régions est en faveur de la seconde hypothèse.
Cependant, certains réplicons échappent à cette règle. C'est notamment le cas des chromosomes linéaires des *Streptomyces* (Fig. 5). Chez *S. coelicolor* et *S. scabies* (mais contrairement à *S. avermitilis*), le brin continu présente un léger biais en bases C par rapport à G, c'est-à-dire la situation opposée à celle observée chez la majorité des chromosomes circulaires. Plusieurs inversions du biais sont toutefois observées, notamment au niveau de l'origine de réplication chez ces trois espèces. Cette particularité n'est pas une conséquence de la linéarité chromosomique étant donné le biais marqué retrouvé pour les chromosomes de *Borrelia* et *Agrobacterium tumefaciens*.

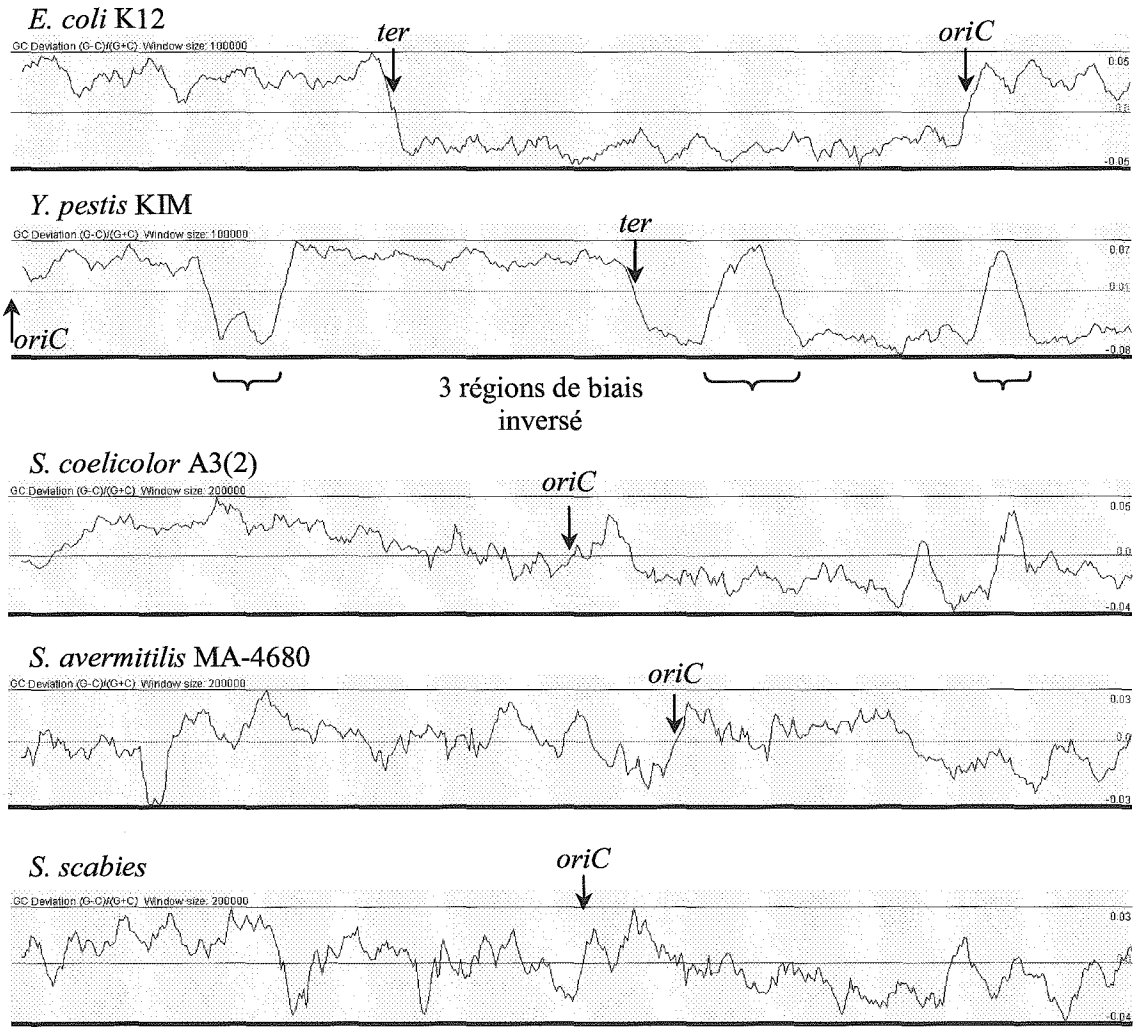


Figure 5 : Exemples de profils de GC skew [(G-C)/(G+C)].
Chez *E. coli* K12 et *Y. pestis,* comme chez la plupart des génomes séquencés, le brin continu se caractérise par un biais en bases G par rapport à C. Chez *Y. pestis*, trois régions présentant un biais inversé sont indiquées. Pour les chromosomes linéaires des trois *Streptomyces* dont le génome est totalement séquencé, aucun biais n'est clairement identifiable entre les deux brins. Cependant, chez *S. coelicolor* et *S. scabies* (mais pas chez *S. avermitilis*) le brin continu semble tendre vers un bais en bases C par rapport à G.

Les raisons d'un tel biais de composition nucléotidique restent mal connues mais il semble que l'asymétrie de la machinerie de réplicon y contribue (Tillier et Collins, 2000b). Ce biais de composition s'oppose à la sélection des fonctions des protéines codées et pose problème quant à l'interprétation des relations phylogénétiques établies sur la base des comparaisons de séquences des gènes (Tillier et Collins, 2000b). Il est par ailleurs plus marqué aux positions neutres (McLean *et al.*, 1998).

c) Usage des codons

L'usage des codons est également biaisé. Une corrélation avec l'abondance en tRNA a été remarquée chez *E. coli* (Grantham *et al.*, 1981 ; Ikemura, 1981) et *Saccharomyces cerevisiae* (Ikemura, 1982) suggérant une adaptation des séquences des gènes à une meilleure efficacité de traduction. Les

séquences des gènes fortement exprimés dans un génome tendent à évoluer vers un usage des codons optimal, c'est-à-dire qui rend la traduction de l'ARMm par le ribosome plus efficace.
L'indice d'adaptation du code (CAI) permet de prédire, pour un gène donné, son niveau d'expression (Sharp et Li, 1987). Cet indice doit être calibré à partir de gènes dont la forte expression est prouvée ou fortement suspectée. En général, les gènes codant les protéines ribosomiques sont utilisés pour ce calibrage. L'usage des codons est donc une caractéristique génomique et, par conséquent, une barrière au maintien de gènes acquis par transfert horizontal. En effet, ce maintien implique que les gènes acquis soient exprimés pour conférer une fonction avantageuse.

d) Fréquence en dinucléotides

Une autre caractéristique structurant les génomes est la distribution des dinucléotides, notamment discutée par Karlin *et al.* (Karlin et Burge, 1995 ; Karlin, 1998, 2001). Une signature liée aux fréquences relatives des dinucléotides peut ainsi être définie pour chaque génome. Cette signature est une barrière aux échanges de gènes puisqu'elle pourrait notamment permettre de discriminer les séquences provenant d'autres organismes. Les raisons d'une telle signature sont mal connues. Néanmoins, il a été proposé qu'elle pourrait être le reflet de propriétés spécifiques d'espèces concernant les contraintes structurales liées à l'empilement des bases dans la molécule d'ADN, la modification, la réplication et la réparation de l'ADN (Karlin, 1998).

4. Biais de distribution des gènes sur le génome

La réplication a lieu de façon asymétrique. La synthèse coordonnée de deux brins en orientation antiparallèle implique des mécanismes différents pour chacun d'eux. La polymérisation d'un brin complémentaire se fait de façon continue (brin continu ; "leading strand") ou discontinue (brin discontinu ; "lagging strand"). De plus, elle induit un gradient entre les régions précocement et tardivement répliquées. Ainsi, de même que la composition en nucléotides (G/C skew), l'orientation et la localisation des gènes sur les réplichores peuvent être fortement contraintes par la réplication.

a) Rôle de la machinerie de réplication dans l'orientation des gènes

Les gènes tendent à être portés par le brin continu (McLean *et al.*, 1998). Il a tout d'abord été proposé qu'un tel biais résulterait d'une sélection qui minimiserait la fréquence de collisions entre les machineries de réplication et de transcription (Brewer, 1988), phénomènes qui peuvent provoquer des ralentissements de la fourche de réplication (Liu et Alberts, 1995) ainsi que des pauses et des arrêts de la transcription (French, 1992 ; Liu *et al.*, 1993). Par ailleurs, une forte corrélation a été mise en évidence entre le type de machinerie de réplication et un biais marqué d'orientation des gènes (Rocha, 2002). En effet, la présence d'une sous unité alpha de l'ADN polymérase III, appelée PolC, semble y jouer un rôle important. L'ADN polymérase III est composée de deux core-enzymes. Chez *E. coli*, une seule sous unité alpha catalytique est connue, DnaE. En revanche, *Bacillus subtilis* code deux sous unités alpha différentes ayant des rôles complémentaires, PolC et $DnaE_{BS}$, respectivement responsables de la synthèse des brins continu et discontinu (Dervyn *et al.*, 2001). Des orthologues de PolC sont retrouvés chez toutes les bactéries Gram positives à bas G+C et celles-ci présentent un biais

plus fort que les espèces dépourvues de PolC (Fig. 6). Les analyses ont été réalisées sur 64 chromosomes issus de 59 espèces différentes. Elles montrent une moyenne de 78% de gènes portés par le brin continu chez les organismes possédant PolC alors que le biais n'est que de 58% pour les autres. En revanche, le biais de composition en nucléotides ne semble pas être associé à la présence de PolC.

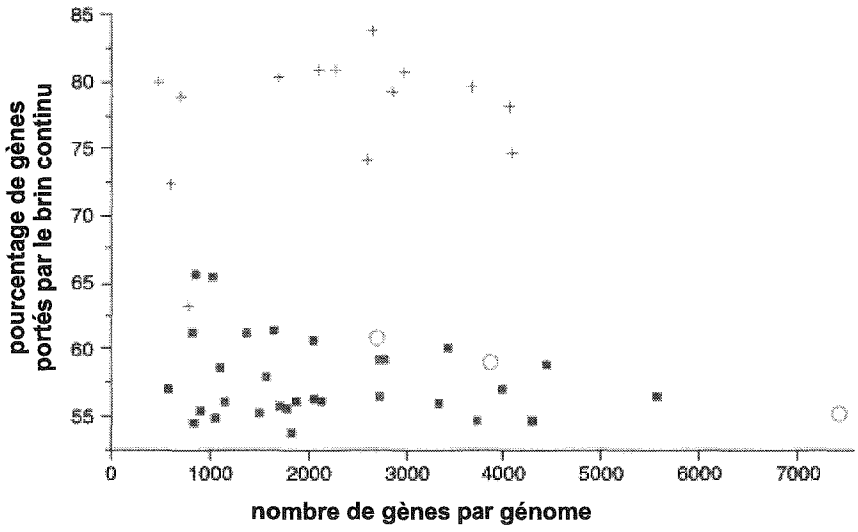


Figure 6 : Relation entre le nombre de gènes portés par le brin continu et le nombre total de gènes.
Les coordonnées calculées pour les génomes des bactéries Gram positive à bas G+C sont indiquées par des croix, ceux des bactéries Gram positive à haut G+C, par des cercles, et ceux des autres espèces bactériennes par des carrés.
D'après (Rocha, 2002).

b) Corrélation entre orientation et caractère essentiel des gènes

Une première hypothèse, selon laquelle l'orientation préférentielle des gènes sur le brin continu est le résultat d'une sélection en faveur d'une distribution des gènes fortement exprimés (notamment les opérons ribosomiques) sur le brin continu, a été proposée (Brewer, 1988). Elle a par la suite été nuancée. En effet, il a été montré que c'est le caractère essentiel des gènes qui sélectionne ce biais de distribution (Rocha et Danchin, 2003b). Une corrélation forte entre essentialité et distribution a tout d'abord été démontrée chez *E. coli* et *B. subtilis* pour lesquels de très nombreux résultats expérimentaux sont disponibles, permettant une estimation fiable du caractère essentiel des gènes (Rocha et Danchin, 2003b). Par la suite, ces résultats ont pu être étendus aux bactéries Gram positives à bas G+C, aux Protéobactéries gamma et enfin à tous les phyla disponibles (53 génomes ; (Rocha et Danchin, 2003a)).

Une inactivation systématique des gènes a permis d'identifier 271 gènes essentiels chez *B. subtilis* (Kobayashi *et al.*, 2003) et cet ensemble de gènes a ensuite été utilisé pour rechercher les orthologues chez les autres espèces. Le taux d'expression des gènes a par ailleurs été évalué par le calcul de l'index d'adaptation des codons (Sharp et Li, 1987) en vue d'étudier l'effet de l'expressivité des gènes sur leur distribution. Pour tous les génomes testés, les gènes essentiels sont plus fréquemment portés par le brin continu. Par exemple chez *B. subtilis*, ce biais atteint 94% pour les gènes essentiels alors qu'il n'est que de 74% pour les non essentiels. Pour la quasi-totalité des génomes testés, le niveau d'expression des gènes non essentiels n'influe pas sur leur orientation. Par ailleurs, la distribution des gènes essentiels sur le brin continu est un caractère conservé.

Minimiser les collisions entre les machineries de transcription et réplication revient donc à réduire le nombre de transcrits tronqués issus d'une transcription avortée. L'une des hypothèses proposées est que l'avortement de la transcription mène à synthétiser des peptides tronqués, dominants négatifs, qui rendent inactifs les complexes protéiques auxquels ils s'intègrent (Rocha et Danchin, 2003a). Cette situation serait donc fortement contre-sélectionnée. Néanmoins, l'intensité du biais est variable. La différence de composition du réplisome (PolC/DnaE), exposée précédemment, pourrait expliquer une partie de cette variabilité (Rocha, 2002).

c) Rôle de la réplication dans la position des gènes sur le chromosome : l'effet dose

Chez les bactéries à croissance rapide, plusieurs cycles de réplication du chromosome peuvent être initiés (en phase exponentielle) avant la séparation des chromatides sœurs et la division cellulaire. Ainsi, plus un gène est proche de l'origine de réplication, plus son nombre de copies par cellule peut être élevé et, par conséquent, plus son niveau d'expression pourra être décuplé : c'est l'effet dose associé à la réplication. Lors de la phase exponentielle, si un ou plusieurs gènes, dont l'expression est saturée, deviennent limitant pour la croissance, leur position sur le chromosome devient un paramètre soumis à une forte sélection. A plusieurs reprises, cette hypothèse a été évoquée pour expliquer les effets délétères de réarrangements génomiques provoquant le passage de gènes fortement exprimés de l'origine vers le terminus de réplication (ex : (Campo *et al.*, 2004 ; Segall *et al.*, 1988)).

La corrélation entre le niveau d'expression des gènes et leur position sur le chromosome a été étudiée par analyse de génomes (Couturier et Rocha, 2006). Existe-t-il une pression de sélection sur la position de certains gènes fortement exprimés ? Les auteurs ont défini les bactéries à croissance rapide comme celles montrant un rapport entre le temps de réplication du chromosome et le temps minimal de doublement de population supérieur à la moyenne (>0,5). Pour ces dernières, une telle corrélation a pu être mise en évidence pour les gènes impliqués dans la traduction et la transcription (ceux codant les ARNr, ARNt, les protéines ribosomiques et l'ARN polymérase). En revanche, la position relative des autres gènes fortement exprimés ne semble pas être influencée par un effet dose. De façon intéressante, les auteurs notent également un lien entre l'importance de l'effet dose et la stabilité des génomes (Couturier et Rocha, 2006). Pour les génomes montrant une organisation façonnée par l'effet dose, les réarrangements asymétriques auraient des effets plus marqués et seraient, par conséquent, contre sélectionnés de façon plus forte. L'effet dose s'impose donc, chez les espèces à croissance rapide, comme une contrainte à la plasticité génomique.

B. Evolution et dynamique des génomes

Les mécanismes moléculaires responsables de l'évolution sont : les mutations ponctuelles et la conversion génique qui modifient de façon graduelle l'information génétique, les réarrangements d'ADN endogène (sans perte ni acquisition de matériel) qui altèrent la structure du génome, les délétions et enfin l'acquisition d'information exogène. Ces deux derniers mécanismes contribuent majoritairement aux variations du "fitness" des organismes et à leur adaptation.

1. Expansion et contraction des génomes

La taille des génomes varie en proportion extrêmement importante, non seulement entre bactéries phylogénétiquement éloignées, mais également au sein de certaines espèces. Les modifications conséquentes sur le contenu en gènes et la rapidité avec laquelle ces changements s'opèrent expliquent la présence des bactéries dans toutes les niches écologiques.
L'expansion des génomes semble être en lien avec une adaptation à un environnement complexe et changeant. Les *Streptomyces* constitue un exemple d'expansion génomique. Alors que les espèces voisines de l'ordre des *Actinomycetales* présentent des génomes de taille allant de 2,5 à 6 Mb (Tableau annexe), tous les chromosomes caractérisés de *Streptomyces* contiennent plus de 8 Mb.

La contraction des génomes est souvent interprétée comme le résultat d'une adaptation à un environnement plus stable. Certains gènes devenant inutiles, leur perte peut être fixée par dérive voire sélectionnée. Ce type d'évolution peut être observé chez des organismes ayant changé de niche écologique récemment (Moran, 1996, 2002). C'est le cas du génome de *Buchnera aphidicola* (endosymbiote d'insecte) où des délétions de fragments d'ADN de grandes tailles (plusieurs dizaines de gènes) ont été fixées lors de la transition entre le mode de vie autonome et celui d'endosymbiote obligatoire (Moran et Mira, 2001).
L'accumulation de pseudogènes est la première étape de réduction des génomes avant leur délétion complète. Elle peut se produire par mutation(s) non sens, mutation(s) décalant le cadre de lecture (appelées "frameshifts" par la suite), troncature partielle de la phase codante, interruption de la phase codante (notamment par insertion d'IS) ou encore par réarrangements (ex : inversion dont une borne est localisée dans une séquence codante). De plus, des mutations n'affectant pas la séquence codante elle-même mais les signaux de régulation et d'expression (promoteurs, RBS, sites de fixation de régulateurs) peuvent également engendrer l'inactivation de gènes, bien que leur identification reste beaucoup plus problématique.
L'exemple le plus connu est celui de *Mycobacterium leprae* dont 41% des ORF sont décrits comme pseudogènes (1115/2720 CDS ; (Cole *et al.*, 2001)). Un autre cas, plus récemment découvert, mais tout aussi marquant, est le génome de *Sodalis glossinidius* (endosymbiote d'insecte) dans lequel 49% du chromosome est constitué d'ADN non codant (Toh *et al.*, 2006). Cet enrichissement en pseudogènes est également observé chez *Salmonella enterica* CT18 (211 pseudogènes sur 4600 CDS soit 4,6% ; (Parkhill *et al.*, 2001a)), chez *Shigella flexneri* (372 pseudogènes soit 8.1% (Wei *et al.*, 2003)) ou encore *Yersinia pestis* CO92 (150 pseudogènes soit 3.7% (Parkhill *et al.*, 2001b)). *Streptococcus thermophilus* n'est pas une espèce pathogène mais elle s'est adaptée à une niche

écologique particulière, le lait. De cette adaptation récente (~20000 ans) résulte la présence d'environ 10% de pseudogènes (Bolotin *et al.*, 2004).
Un cas spectaculaire de réduction du génome est observé chez *Bordetella pertusis* (Parkhill *et al.*, 2003). La comparaison avec *Bordetella bronchiseptica* montre qu'environ 20% de son génome ancestral aurait été perdu après adaptation à un hôte unique. Par ailleurs, son génome comporte encore 10% de pseudogènes détectables. Ces deux espèces sont extrêmement proches d'un point de vue phylogénétique (0,2% de divergence des ARNr 16S). Néanmoins le nombre de réarrangements génomiques fixés au cours de l'évolution depuis leur ancêtre commun a été estimé à 150. En revanche, le génome des espèces de *Mycoplasma* (de 580 kb pour *M. genitalium* à 1359 kb pour *M. penetrans*) est quasiment dépourvu de pseudogènes malgré une adaptation à un environnement plus stable accompagnée d'une réduction génomique.

Enfin, pour nuancer ces propos, il est nécessaire de préciser les précautions à prendre quant à l'interprétation du nombre de pseudogènes prédits dans les génomes ! Dans la plupart des cas, ce nombre reflète plus une limite du processus d'annotation qu'une réalité biologique. Cet aspect a particulièrement été bien discuté dans (Lerat et Ochman, 2004, 2005). La recherche ciblée de pseudogènes dans les génomes de différentes souches d'*E. coli* et *Shigella* a permis de mettre en évidence entre 98 et 168 pseudogènes non prédits initialement (Lerat et Ochman, 2004).

2. *Les flux de gènes*

Le transfert horizontal de gènes semble être d'ampleur minime chez les eucaryotes. Cependant, l'apparition des mitochondries et des chloroplastes dans ce règne est une forme de transfert horizontal. En revanche, son implication dans l'évolution des bactéries et des archées apparaît primordiale. La perte de gènes et leur acquisition par transfert sont les deux mécanismes prépondérants dans l'adaptation. Dans un contexte de contraintes fortes s'exerçant sur la taille des génomes, un équilibre s'établit entre acquisition et perte de gènes générant une compétition entre les gènes pour leur maintien au sein d'un génome. La coexistence de ces deux phénomènes entraîne un flux de gènes d'importance majeure dans la divergence entre bactéries.

a) Dynamique d'acquisition et de perte de gènes

Il existe trois mécanismes de transfert horizontal : la conjugaison, la transduction et la transformation naturelle. La grande majorité de l'ADN acquis ne se maintient pas dans la population et est ainsi perdue au fil des générations. Le maintien d'une séquence implique que l'ADN transféré puisse être répliqué par la cellule réceptrice afin d'être transmis aux générations suivantes. Le maintien peut donc intervenir, soit par réplication autonome de l'élément transféré, soit par son intégration dans un réplicon. Néanmoins, le maintien à long terme implique une pression de sélection s'exerçant sur les gènes acquis pour éviter leur perte par dérive génétique. En effet, l'ADN acquit est rarement fixé pour de multiples raisons. Il peut ne conférer aucune fonction. La ou les fonctions codées peuvent ne pas être exprimées dans le nouveau contexte. Toutes les fonctions nécessaires à une fonction biologique (ex : une voie métabolique) ne sont pas présentes. Enfin, la fonction nouvellement acquise ne contribue pas de façon significative au "fitness" de la bactérie. Ainsi, il est facile d'imaginer qu'une

infime part de l'ADN acquis se maintiendra. Dans ce cas, la cellule ayant vu son "fitness" augmenter, va émerger par compétition avec les cellules voisines et assurer la dissémination des gènes transférés dans l'ensemble de la nouvelle population. L'émergence d'une nouvelle lignée peut intervenir en réponse à une croissance accélérée de la cellule ayant acquis une nouvelle fonction ou encore par adaptation à un nouvel environnement.

Les pertes de gènes contribuent également de façon importante à l'évolution des génomes bactériens. Mira *et al.* suggèrent que les délétions se produisent à des fréquences plus élevées que les insertions dans les génomes bactériens et que ce biais est la force majeure de leur organisation et de leur densité en séquences codantes (Mira *et al.*, 2001). Les délétions peuvent affecter quelques nucléotides et ainsi créer des pseudogènes ; c'est notamment le cas chez *Rickettsia* (Andersson et Andersson, 1999, 2001). Elles peuvent aussi être de grande taille et entraîner l'élimination de plusieurs dizaines de gènes comme révélée par comparaison de génomes entre différentes souches de *M. tuberculosis* (Kato-Maeda *et al.*, 2001). Une preuve de ce biais en faveur des délétions a été apportée par la comparaison de séquences de pseudogènes avec leurs homologues fonctionnels (Mira *et al.*, 2001). Pour chacun des génomes testés, le nombre de délétions (de quelques nucléotides) détectées dans les pseudogènes est significativement plus élevé que celui des insertions.
Les délétions seraient majoritairement le fait de la dérive due à une sélection moins forte plutôt qu'à une sélection en faveur des petits génomes (sélection pour la perte de gènes). L'étude de la longueur des séquences intergéniques orthologues entre *E. coli* et *Buchnera aphidicola* a montré que leur taille ne diffère pas. Il n'y aurait donc pas de sélection de l'élimination de l'ADN non fonctionnel. Par ailleurs, la taille du génome n'est pas corrélée à la vitesse de croissance.
Le faible nombre de pseudogènes prédits dans la plupart des génomes bactériens a par ailleurs été interprété comme le résultat d'une sélection pour maintenir un taux de délétion élevé (Lawrence *et al.*, 2001). Les auteurs suggèrent que le flux d'éléments génétiques parasites (ex : transposons, bactériophages), potentiellement dangereux, est une des raisons pour lesquelles le maintien d'un taux de délétion élevé est sélectionné chez la majorité des bactéries.

b) L'établissement des flux de gènes

La conservation d'une séquence dans différents génomes résulte soit, du hasard (dérive génétique) soit, d'une pression de sélection, c'est-à-dire qu'une fonction se maintient dans la population si elle confère un avantage et contribue donc au "fitness" de l'organisme. Lawrence et Roth proposent une notion qualitative des fonctions de gènes du fait qu'il existe une graduation de la contribution d'un gène au "fitness" dans un environnement donné (Lawrence et Roth, 1999). Ainsi, les gènes peuvent être essentiels ou non essentiels à la croissance. Parmi ces derniers sont retrouvés les gènes de fonction importante, utile, marginale et enfin neutre (voire nuisible), c'est-à-dire qui ne contribuent pas au "fitness". La taille d'un génome bactérien étant fortement contrainte, son extension par acquisition d'ADN étranger est limitée. Ainsi, la perte et l'acquisition de gènes s'équilibrent au sein d'un génome définissant les flux de gènes. En schématisant la situation, tout matériel acquis et fixé entraîne la perte d'information génétique ayant ou non une fonction similaire présent dans le même génome (Fig. 7).

Les auteurs ont défini une valeur minimale de contribution au "fitness" (appelée valeur "s") en dessous de laquelle un gène ne sera pas maintenu par sélection naturelle et pourra être perdu par délétion ou mutation.

D'après cette théorie, la valeur seuil de contribution au "fitness" capable de conférer le maintien d'une séquence dans un génome est proportionnelle au taux de mutations chez un organisme. En effet, plus les mutations se produisent à haute fréquence, plus la valeur adaptative d'un gène doit être élevée pour conférer son maintien ; ainsi la valeur s est plus élevée dans un contexte de mutations fréquentes. Par ailleurs, cette valeur "s" est également fonction de la taille de la population. Quand celle-ci diminue, la fixation de perte de gènes par dérive est accrue. Au sein des populations plus petites, une valeur adaptative plus forte est requise pour se maintenir. Enfin, le taux de recombinaison intraspécifique influence l'efficacité avec laquelle la sélection agit sur les allèles mutés. En effet, plus ce taux de recombinaison augmente, plus la sélection peut éliminer efficacement l'accumulation d'allèles délétères et peut ainsi éviter le déclin selon le processus du rochet de Muller (Muller, 1964). Par conséquent, si le taux de recombinaison diminue, un gène devra contribuer de façon plus importante au "fitness" pour être maintenu.

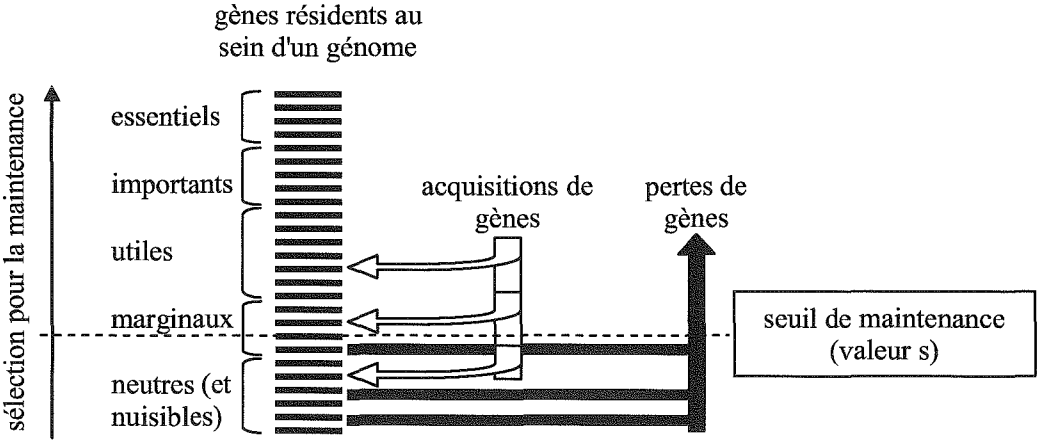


Figure 7 : Evolution des génomes par flux de gènes.
Chaque gène au sein du génome est représenté par un trait noir et est classé selon son importance. Les flux sont représentés par les flèches mimant les acquisitions de gènes par transfert horizontal et les pertes. Les gènes dont la valeur adaptative est inférieure au seuil de maintenance ne pourront pas être maintenus par sélection naturelle et seront perdus par délétion ou mutation. D'après (Lawrence et Roth, 1999).

De nombreux exemples confirment cette théorie : *E. coli* et les autres Entérobactéries ont un génome peu divergent en taille alors que de nombreux événements de transferts horizontaux ont été mis en évidence, expliquant la grande divergence du contenu en gènes chez ces espèces. De même, les chromosomes de *S. coelicolor* et de *S. avermitilis* présentent une taille très voisine (~4% de différence) mais respectivement 29% et 30% de leur CDS prédites sont spécifiques de l'une ou l'autre de ces espèces (Ikeda *et al.*, 2003).
L'équilibre établi entre pertes et acquisitions de gènes implique une compétition entre les gènes pour leur maintien au sein d'un génome.

Le nombre de gènes au sein d'un génome est également un facteur influençant la valeur s. Plus ce nombre est grand, plus la possibilité de contre-sélectionner les nombreux allèles délétères est difficile et donc, plus la valeur s augmente. Par conséquent, si les taux de recombinaison, taille de population et taux de mutation sont constants, la valeur minimale de contribution au "fitness" nécessaire pour se maintenir dans un génome est directement proportionnelle à la taille du génome.

c) Rôle des flux de gènes dans la spéciation

Le terme d'"espèce" procaryote est impropre étant donnée la confusion avec l'espèce eucaryote. Une espèce eucaryote est habituellement définie, de façon objective, comme l'ensemble des populations capables de se reproduire entre elles par voie sexuée en conditions naturelles. Ainsi, les échanges génétiques sont contraints par la barrière biologique de la reproduction. Chez les organismes à reproduction asexuée, la définition est subjective : elle se base sur des ressemblances de caractères. L'espèce procaryotique n'est définie que par comparaison avec une souche type, elle-même choisie de façon arbitraire. Toutes les souches présentant un certain niveau de ressemblance avec la souche type appartiendront à la même espèce.

La spéciation est le mécanisme par lequel une population ancestrale d'organismes (définie par leur capacité à exploiter une niche écologique particulière) évolue pour former deux populations exploitant deux niches écologiques différentes et recombinant plus souvent au sein de leur propre groupe qu'avec l'espèce sœur. Les flux de gènes entraînent constamment des changements de capacités métaboliques. Si un gène ou un ensemble de gènes récemment acquis altère le phénotype d'un organisme en lui permettant d'exploiter plus efficacement la niche écologique actuelle, il n'engendrera pas de séparation entre deux lignée (c'est-à-dire de spéciation) mais une sélection du génotype avantagé aura lieu. Si, en revanche, le ou les gènes acquis fournissent des capacités nouvelles pour la descendance, celle-ci pourra envahir une nouvelle niche écologique et le processus de spéciation pourra être initié.
L'accumulation de divergences génomiques entre la nouvelle lignée et les autres entraînent une diminution de l'efficacité de la recombinaison homologue et donc une barrière à l'échange de gènes (isolement génétique ; (Zahrt et Maloy, 1997)). La création de nouvelles fonctions implique, au départ, un mécanisme de duplication/divergence de l'information génétique au sein d'un organisme. Hormis le réassortiment de modules codant des domaines protéiques, la quantité d'événements mutationnels nécessaires et sélectionnables pour l'établissement d'une nouvelle fonction au sein d'un organisme est très élevée. La probabilité de succès est donc extrêmement faible. S'il n'est pas à l'origine de la création de nouvelles fonctions, le transfert horizontal permet leur dissémination à travers la population. De ce fait, même si les proportions du transfert horizontal et du mécanisme de duplication/divergence dans l'acquisition de nouvelles fonctions restent difficiles à appréhender, il est très probable que les transferts de gènes en soient responsables pour une grande partie. La plupart des fonctions connues aujourd'hui pourrait n'être apparue qu'une seule fois au cours de l'évolution. Par ailleurs, il est également envisageable que la plupart des gènes paralogues au sein d'un génome soit en fait issus du transfert horizontal de gènes homologues. Cependant, cette proportion dépend des fréquences de duplication propre à chaque génome. En effet, l'instabilité génomique constatée chez *Streptomyces* se traduit, par exemple, par des amplifications de segments d'ADN affectant des loci

portés par les régions terminales du chromosome linéaire (Leblond et Decaris, 1994). Par ailleurs, la formation de répétitions terminales inversées (TIR) chez *Streptomyces* est un mécanisme de duplication.

Le modèle d'évolution par flux de gènes (Fig. 7) s'applique parfaitement aux Entérobactéries et aux Actinobactéries qui possèdent de grands génomes et, par conséquent, une proportion importante de gènes non essentiels. C'est au niveau de ces gènes que les différences entraînent notamment la spéciation par l'exploitation de nouvelles niches écologiques, les gènes essentiels ne variant que très peu. En revanche, les flux de gènes peuvent affecter de façon très limitée l'évolution de certaines espèces tels que les endosymbiotes. Leur environnement reste constant et ces organismes sont très rarement en contact et/ou en compétition avec d'autres espèces. Cet isolement entraîne une diminution du nombre de gènes soumis à pression de sélection. Dans le cas particulier de *Mycobacterium leprae*, la présence de très nombreux pseudogènes montre une évolution régressive récente sans diminution de taille du génome.

Chez les bactéries, les gènes impliqués dans une même fonction sont très souvent organisés en "clusters" voire en opérons. Le transfert horizontal de gènes serait une des raisons pour lesquelles un tel regroupement est sélectionné. En effet, l'organisation en "clusters" de gènes augmente l'efficacité de transfert et donc le succès évolutif puisque le transfert d'une partie seulement des gènes impliqués dans une même fonction sera plus fortement contre-sélectionné. Pour aller plus loin, les gènes cotranscrits, c'est-à-dire les opérons, possèdent une probabilité de maintien accrue. En effet, les opérons arrivant dans un nouveau contexte et ne possédant qu'un seul promoteur à réguler, se maintiendront plus facilement que si chaque gène transféré possédait son propre promoteur.

Snel et Bork ont tenté d'établir une phylogénie non pas à partir des identités de séquences orthologues mais sur la base de la conservation du contenu en gènes parmi 13 espèces de microorganismes (Snel *et al.*, 1999). Les parentés entre espèces sont ainsi exprimées en pourcentage de gènes partagés. La phylogénie ainsi établie montre des similitudes remarquables avec celle définie à partir des similarités de séquences du gène codant l'ARNr 16S (Olsen *et al.*, 1994). Ainsi, le contenu en gènes partagés est corrélé à la phylogénie des organismes.

d) Mesure des flux de gènes

La mesure des flux de gènes implique l'identification des gènes étrangers au sein d'un génome, l'évaluation du temps depuis lequel ces gènes ont été acquis et l'estimation des pertes d'information. Au vu de l'absence de données fossiles chez les bactéries, la comparaison des génomes d'espèces actuelles reste la méthode de choix permettant d'accéder à ces informations. Le passé évolutif est alors reconstruit en respectant le principe d'un maximum de parcimonie. A titre d'exemple, la comparaison des génomes de deux espèces du genre de *Mycoplasma*, a révélé que le génome de *M. genitalium* (580 kb) est une sous partie de celui de *M. pneumoniae* (816 kb) (Herrmann et Reiner, 1998 ; Himmelreich *et al.*, 1996). Par ailleurs, le génome de l'archée sulfo-réductrice *Archeoglobus fulgidus* (2,18 Mb) est 25% plus grand que celui de l'espèce proche *Methanococcus jannashii* (1,66 Mb) qui est un méthanogène strict (Klenk *et al.*, 1997).

Une seconde méthode utilisée pour identifier les gènes nouvellement acquis consiste à analyser les déviations de certains paramètres par rapport à la signature du génome. En effet, chaque génome possède des caractéristiques propres : sa composition en bases G et C (ou A/T), sa fréquence en dinucléotides ou encore son usage des codons dans les séquences codantes. Cette signature est fonction de critères tels que les "pools" intracellulaires en désoxynucléosides triphosphate, la fréquence d'erreur de l'ADN polymérase, l'efficacité du système de réparation des mésappariements et les concentrations en différents types de tRNA.
L'émergence de cette signature du génome apparaît quand tout gène est soumis au même biais mutationnel pendant un temps donné. En effet, le taux de conversion des bases G/C en A/T peut être différent de celui des conversions des nucléotides A/T en G/C. Le contenu en bases G/C est dépendant du pourcentage global de ces bases pour le génome entier et également de l'usage des codons. Ainsi, les gènes étrangers nouvellement acquis présentent une signature typique des réplicons de la cellule donneuse qui peut être différente de celle du génome de hôte. L'acquisition de nouveaux gènes peut donc être estimée par cette approche. Si l'on considère une taille de génome constante, les pertes de gènes équivalent aux acquisitions. Il devient possible d'évaluer la dynamique des flux de gènes. Ainsi, la génomique comparée a permis de mettre en évidence que 15% du génome de *E. coli* est en fait composé de gènes acquis par transfert depuis sa séparation avec les *Salmonella*, il y a 100 millions d'années (Lawrence et Ochman, 1998). Les 755 gènes concernés par ces transferts auraient été acquis par 234 événements différents (234 loci non contigus).

La pression de mutation qui s'exerce sur les gènes nouvellement acquis tend à faire évoluer leur séquence vers la signature du génome. Ce processus a été appelé "amélioration des gènes" (Lawrence et Ochman, 1997). Chez *E. coli*, un tiers des gènes d'origine exogène montre des signes d'amélioration ce qui signifierait que la majorité des gènes repérables comme ayant été acquis par transfert horizontal l'ont été très récemment puisqu'ils n'ont pas encore subi d'amélioration. Ainsi, le taux d'acquisition d'ADN conférant un avantage sélectif chez *E. coli* a été estimé à environ 16 kb par million d'années (Lawrence et Ochman, 1998). En comparaison avec cette valeur, le taux de mutations ponctuelles a, quant à lui, été estimé à 22.000 substitutions par million d'années dont 90% affectent des positions synonymes. Les mutations ponctuelles contribueraient donc de façon moins déterminante que le transfert horizontal au "fitness" de l'organisme.

3. Conservation de l'ordre des gènes (GOC)

Les premières comparaisons de génomes d'espèces procaryotes ont montré que l'ordre des gènes est peu conservé (Dandekar *et al.*, 1998 ; Mushegian et Koonin, 1996b). Par exemple, à l'échelle chromosomique, aucune colinéarité n'avait été observée entre *E. coli* et *H. influenzae*. Cependant, la majorité des gènes communs entre espèces font partie de groupes de gènes, notamment des opérons, dont l'ordre est conservé. En 2000, Tillier et Collins ont montré que la majeure partie des réarrangements génomiques bouleversant l'ordre des gènes résulte d'inversions centrées autour de l'origine de réplication (Tillier et Collins, 2000a). Des comparaisons de génomes ont montré une fréquence plus élevée des translocations réciproques (impliquant plusieurs événements d'inversions

péricentriques) par rapport aux déplacements aléatoires des gènes sur le chromosome. Ces résultats suggèrent un rôle majeur de la réplication dans l'évolution de la structure des génomes procaryotes.

Le nombre de réarrangements fixés augmenterait de façon relativement constante au cours de la divergence des espèces (Suyama et Bork, 2001). La conservation de l'ordre des gènes serait donc un bon indicateur de l'évolution des espèces procaryotes. Cependant, certains génomes ne suivent pas cette règle : c'est le cas des *Chlamydia* et des *Mycoplasma* dont la fréquence d'interruption de synténie semble être beaucoup plus faible que chez les autres organismes testés. Cette particularité peut être associée à une absence de gène codant des protéines impliquées dans la recombinaison parmi la machinerie de réplication (ex : RecG, XerCD) chez ces espèces (Suyama et Bork, 2001).

Etant données les fréquences de remaniement de l'ordre des gènes, la synténie peut être interprétée comme le résultat d'une sélection positive favorisant le regroupement des gènes impliqués dans une même fonction. Par exemple, les paires de gènes dont l'ordre est conservé à travers l'évolution code des protéines qui pourraient interagir physiquement dans la cellule (Dandekar *et al.*, 1998 ; Overbeek *et al.*, 1999a). La recherche de synténie par génomique comparée a d'ailleurs été utilisée afin de prédire des couplages fonctionnels entre certains gènes (Overbeek *et al.*, 1999b). Le groupe de C. Medigue (Genoscope) a même implémenté la recherche de synténie pour la prédiction de fonctions au sein d'une plateforme d'annotation de génomes (MaGe) (Vallenet *et al.*, 2006).

Bien que les premières études de comparaison de génomes ont conclu à un niveau très faible de conservation de l'ordre des gènes, elles n'englobaient qu'un nombre restreint d'espèces qui, par ailleurs, sont très éloignées phylogénétiquement (Itoh *et al.*, 1999). Ces résultats suggéraient que les génomes subissent des remaniements qui bouleversent l'ordre des gènes. Les auteurs allaient jusqu'à proposer que le remaniement des génomes est virtuellement neutre à long terme. Cette vision de l'évolution a été quelque peu modérée par la suite. Une étude a été réalisée en 2006 sur 126 chromosomes bactériens afin d'évaluer l'évolution de la perte de synténie au cours du temps (Rocha, 2006). Le niveau de synténie entre deux génomes est calculé à l'aide d'un index, le GOC (gene order conservation), qui équivaut à la proportion de paires d'orthologues contigus dans les deux génomes comparés par rapport au nombre total d'orthologues. Le GOC correspond à la fréquence relative de réarrangements du matériel génétique partagé (transféré verticalement depuis l'ancêtre commun) entre deux espèces en s'affranchissant donc du transfert horizontal.

La diminution du GOC au cours du temps dépend de la présence d'une organisation en opéron : les paires de gènes contigus appartenant à un même opéron sont dissociées à une fréquence plus faible que celles appartenant à des opérons différents. Pour une distance évolutive estimée à 500 millions d'années, 80% des paires d'orthologues sont restées contiguës. Ainsi, la fixation de réarrangements d'information génétique endogène ne serait pas aussi fréquente qu'elle ne l'est apparue à travers les premières études de génomique comparée.
Il a également pu être montré qu'il n'existe pas de corrélation directe entre instabilité génomique et mode de vie de l'organisme. Cependant, les génomes des organismes à vie autonome tendent à être

plus instables que ceux des endomutualistes. Les *Streptomyces*, objet d'études au laboratoire, apparaissent parmi les plus instables dans cette étude (Rocha, 2006).

Les réarrangements affectant la structure des opérons se feraient selon un processus conservatif qui maintiendrait les gènes dans un même contexte fonctionnel (Lathe *et al.*, 2000). La conservation de l'organisation génétique se ferait à un niveau supérieur à celui des opérons (les "uber-operons"). Par exemple, le gène *tufA*, codant un facteur d'élongation de la traduction, est retrouvé dans différents environnements génétiques mais il est systématiquement associé à des gènes impliqués dans la traduction. Ainsi, les structures en opérons peuvent être rompues au cours de l'évolution mais seuls les réarrangements qui conservent un contexte fonctionnel seraient fixés. Cependant, certains opérons, comme l'opéron *rrn* (codant les ARN ribosomiques 16S, 23S et 5S), semblent récalcitrants aux remaniements. Cette particularité découle du fait que l'ARN est "maturé" afin de produire les trois ARNr fonctionnels. La structure de l'ARN précurseur doit donc être conservée afin de permettre sa maturation, ce qui n'est pas le cas des ARNm. Toutefois, chez *Pirellula* sp. strain 1 (*Planctomycetes*), le gène codant l'ARNr 16S est distant de 460 kb des deux autres (Glockner *et al.*, 2003).

C. Un désordre organisé !

L'évolution des génomes bactériens est régie par deux phénomènes contradictoires : d'un côté, les flux de gènes sont capables de générer une variabilité considérable du contenu en gènes au sein même d'une espèce, et d'un autre côté, les génomes sont soumis à des contraintes, résultant notamment de la réplication, qui impose une architecture fortement sélectionnée. Un génome bactérien peut donc être vu comme une structure en équilibre issu d'un conflit entre ordre et désordre.

1. Ampleur de la variabilité des génomes au sein des espèces

a) Pan-génome et génome core

Au sein des groupes monophylétiques définis comme des "espèces", la variabilité génomique peut atteindre des proportions très importantes. Les gènes portés par le génome d'une souche, c'est-à-dire d'un représentant d'une espèce, ne correspondent pas à la collection de gènes présents chez l'ensemble des membres d'une espèce. Le métagénome d'un clade est appelé "pan-génome" (pan signifie "entier" en grec) (Medini *et al.*, 2005). Il s'oppose au "génome core" qui définit, au contraire, l'ensemble des gènes partagés par tous les membres du groupe considéré.

Les comparaisons de génomes complets ont permis de prendre conscience non seulement de la variabilité existante au sein de certaines espèces mais également de l'absence de variabilité au sein d'autres espèces. Deux situations extrêmes sont bien décrites dans la littérature. Tout d'abord, la comparaison de trois souches d'*E. coli* a révélé que le génome core ne représente que 39,2% du pool de gènes identifiés (Welch *et al.*, 2002) (Fig. 8). A l'opposé, le génome des souches de *Buchnera aphidicola* montre une stabilité remarquable puisque aucun événement de transfert horizontal ou de réarrangement génomique ne semble s'être produit depuis les 50 derniers millions d'années (Tamas *et al.*, 2002) (Fig. 8). Ces différences flagrantes de vitesse de divergence reflètent des styles de vie bien distincts : *Buchnera* est une bactérie intracellulaire alors qu'*E. coli* vit de façon autonome, non dépendante d'un hôte.

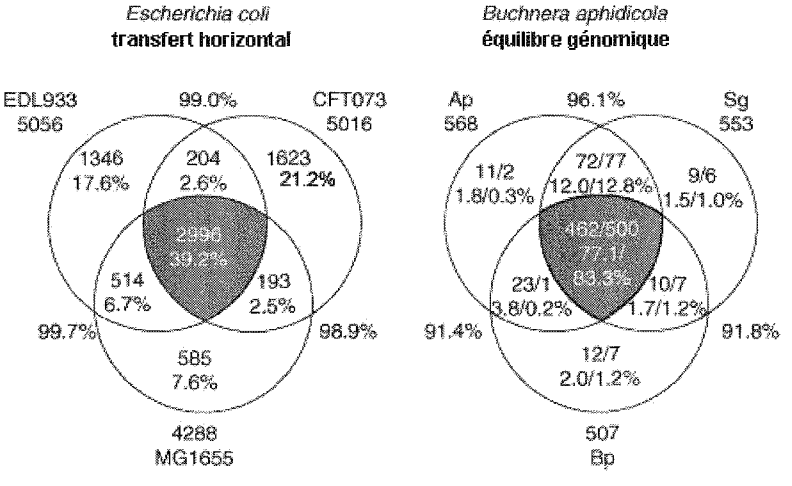


Figure 8 : Comparaison du contenu en gènes entre 3 souches chez *E. coli* et chez *Buchnera aphidicola*. Le génome core, c'est-à-dire les gènes partagés parmi les organismes comparés, est représenté en gris. Les valeurs inscrites en dehors des cercles représentent le nombre de gènes prédits dans chaque génome et le pourcentage d'identité du gène codant l'ARNr 23S.
Pour *Buchnera*, les valeurs tirées des analyses réalisées sans/avec les pseudogènes sont indiquées.
D'après (Lawrence et Hendrickson, 2005).

Une étude réalisée sur 116 génomes procaryotiques a estimé à 14% la proportion moyenne de gènes acquis récemment par transfert horizontal par génome (Nakamura *et al.*, 2004). La proportion la plus faible (0,5%) est observée pour *Buchnera* et la plus élevée (25,2%) pour l'archée *Methanosarcina acetivorans*. Par ailleurs, cette étude révèle une corrélation positive entre la taille du génome et la proportion de gènes acquis par transfert horizontal.

Streptococcus agalactiae (streptocoques du groupe B) constitue un exemple de situation intermédiaire où le séquençage de huit souches a révélé un génome "core" représentant environ 80% de chaque génome, le reste étant partiellement partagé voire spécifique de souche (Tettelin *et al.*, 2005). Le génome "core" compte 1806 gènes alors que 907 gènes constituent le génome dispensable. Des modèles mathématiques ont permis d'évaluer à 33 le nombre de gènes venant enrichir le métagénome pour chaque génome nouvellement séquencé. Ces modèles prédisent un pan-génome si vaste que des gènes spécifiques seraient encore découverts même après le séquençage de plusieurs centaines de génomes de cette espèce.

b) Estimation de la taille du génome core des procaryotes

La taille du génome core diminue à mesure que la diversité phylogénétique du groupe considéré augmente. Charlebois *et al.* ont imaginé des approches permettant d'estimer le nombre de gènes constituant le génome core parmi 130 génomes de bactéries et 17 d'archées (Charlebois et Doolittle, 2004). La première méthode consiste à rechercher les gènes orthologues conservés dans l'ensemble des représentants du groupe en utilisant BLASTP, l'orthologie entre deux séquences nécessitant un meilleur appariement réciproque entre les deux génomes considérés (reciprocal best match : RBM ou bidirectional best hit : BBH). La seconde est basée sur la conservation des noms de gènes dans l'annotation des génomes (consensus gene name : CGN).
Pour un génome donné, la moyenne du nombre de gènes ayant des orthologues dans l'ensemble des 146 autres génomes testés est de 14,8+/-2,6 gènes. En réunissant les différents ensembles de gènes ainsi partagés, les auteurs ont défini 30 gènes faisant partie du core par la première méthode. La deuxième approche (CGN) a permis d'ajouter 8 gènes à cette première estimation aboutissant ainsi à 38 gènes constituant l'ensemble des gènes quasi-ubiquitaires chez les procaryotes. Les gènes impliqués dans le processus de traduction seraient les seuls gènes ubiquitaires. La plupart des fonctions cellulaires générales (ex : métabolisme énergétique, synthèse de l'enveloppe cellulaire) peuvent donc être codées par différents types de gènes non homologues mais possédant des fonctions analogues dans le monde vivant, au contraire de la fonction de traduction qui implique une faible diversité de gènes tous homologues (Charlebois et Doolittle, 2004).

2. Contraintes s'opposant à l'établissement de la variabilité génomique

Les phénomènes de transferts horizontaux de grande ampleur (Lawrence et Ochman, 1998 ; Lawrence et Roth, 1999 ; Ochman *et al.*, 2000), peuvent donner l'idée fausse d'un chao régnant sur l'organisation du génome. En réalité, de nombreuses contraintes en faveur d'une architecture stable s'opposent à l'établissement du chao (pour revues (Lawrence et Hendrickson, 2005 ; Rocha, 2004b)).

Tout d'abord, le contenu en bases G+C (et A+T) et l'usage des codons constituent des barrières à l'expression des gènes nouvellement acquis et donc à leur maintien dans la population. De plus, la polarité du chromosome et les biais qui y sont associés sont autant de barrières supplémentaires au maintien de l'information génétique acquise et, par conséquent, de contraintes vis-à-vis de l'établissement de la diversité génomique.

a) Architecture et polarité du chromosome

Les biais déjà évoqués, notamment ceux associés à la réplication, façonnent l'architecture du chromosome et lui donnent une polarité ((Rocha, 2004a, b) ; Fig. 9). Il est possible de distinguer :

- la symétrie des réplichores (opposition entre origine et le terminus de réplication),
- les biais de composition en nucléotides,
- le biais de distribution préférentielle des gènes sur le brin continu,
- l'organisation de certains gènes en opérons ou en "clusters",
- l'effet dose des gènes fortement exprimés proches de l'origine de réplication,
- la polarité d'oligonucléotides associés à la résolution et la ségrégation chromosomiques (Hendrickson et Lawrence, 2006).

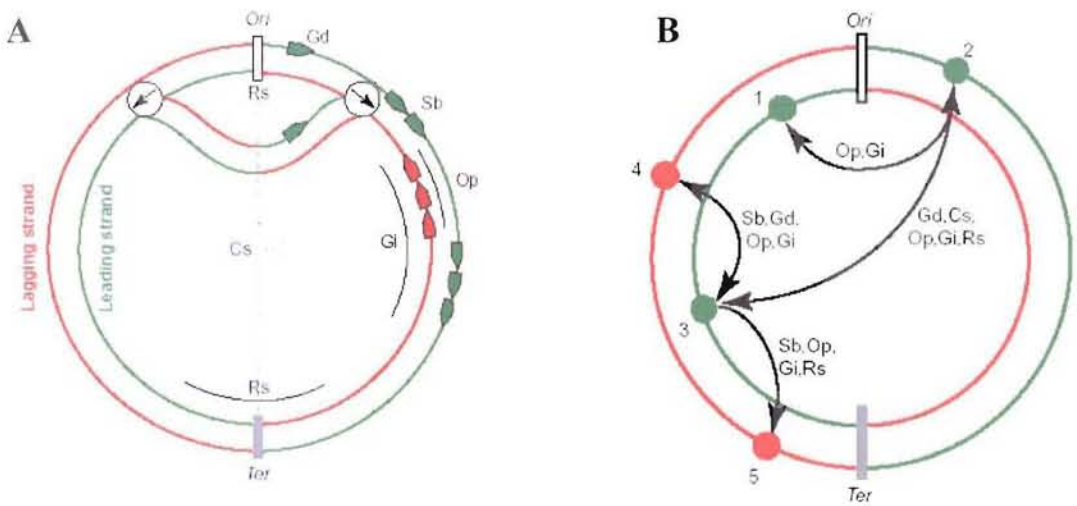


Figure 9 : **A.** Caractéristiques associées à la structuration des génomes bactériens.
Ori et Ter indiquent, respectivement, l'emplacement des origine et terminus de réplication et les flèches entourées indiquent le sens de déplacement des fourches de réplication. Les différents biais associés à l'organisation du chromosome sont schématisés : la symétrie du chromosome (Cs) qui résulte d'une sélection pour l'opposition de l'origine et du terminus de réplication, le biais de distribution des gènes, notamment essentiels, sur le brin continu (Sb), les opérons (Op) et les îlots génomiques (Gi), le biais de localisation préférentielle de gènes fortement exprimés proche de l'origine de réplication due à l'effet dose chez les bactéries à croissance rapide (Gd), le biais de distribution de certains oligonucléotides associés à la polarisation du chromosome et à la résolution des chromosomes répliqués (Rs). "Leading/Lagging strand" : brin continu/discontinu.
B. Conséquences des biais sur les inversions de segments chromosomiques. Les flèches représentent les bornes d'inversions et les caractéristiques génomiques (détaillées dans A.) s'opposant au maintien de chaque type d'inversion sont indiquées. D'après (Rocha, 2004b).

Les inversions non péricentriques sont fortement contre-sélectionnées car elles engendrent le passage de gènes essentiels du brin de synthèse continue vers le brin de synthèse discontinue (Rocha et Danchin, 2003b) et inverse la polarité du GC skew. Chez *Lactococcus lactis*, de grandes inversions bouleversant la symétrie du chromosome ont été induites expérimentalement (Campo *et al.*, 2004). Deux régions localisées autour de l'origine et d'un coté du terminus de réplication sont réfractaires aux réarrangements. L'orientation des gènes est fortement biaisée chez cette espèce (90% des gènes sont portés par le brin continu). Cependant, les événements qui inversent la polarité des deux brins mais n'affectent pas la taille des réplichores sont stables en conditions de laboratoire. Néanmoins, leur absence dans les souches naturelles indique qu'elles sont contre-sélectionnées à long terme. La génération artificielle d'inversions a également été testée chez *E. coli* et *Salmonella* (Rebollo *et al.*, 1988 ; Segall *et al.*, 1988). Elles ont révélé l'existence de régions interdisant les inversions du fait d'effets délétères au niveau fonctionnel (diminution du "fitness") et des régions où la recombinaison ne peut pas se produire pour des raisons d'accessibilité des séquences (Garcia-Russell *et al.*, 2004).

Les inversions les plus couramment observées sont les inversions conservatrices, c'est-à-dire celles qui ne bouleversent pas la symétrie des réplichores et la polarisation des deux brins. C'est le cas des inversions centrées autour de l'origine (et du terminus) de réplication. Elles peuvent se produire par recombinaison homologue entre séquences répétées comme les gènes d'ARNr et les IS. Ces inversions sont les plus fréquentes car elles auraient lieu au moment de la réplication (Tillier et Collins, 2000a). Pendant la réplication, les deux fourches sont approximativement équidistantes de l'origine de réplication et sont physiquement proches dans la cellule. Ainsi, les portions d'ADN sous forme simple-brin au passage des deux fourches seraient des cibles privilégiées de réarrangements expliquant une majorité d'inversions suivant l'axe de réplication.

Par ailleurs, la polarité du chromosome bactérien a été mise en évidence par l'identification d'oligonucléotides (octomères) présentant une distribution asymétrique dans les génomes bactériens (Hendrickson et Lawrence, 2006 ; Lawrence et Hendrickson, 2004, 2005). Ces octamères, appelés AIMS (architecture imparting sequence), s'accumulent au niveau du terminus de réplication (10% du chromosome) et sont préférentiellement portées (>70%) par le brin continu sur les deux réplichores. La présence de séquences AIMS a été recherchée et confirmée chez la quasi-totalité des chromosomes étudiés (Hendrickson et Lawrence, 2006). Les seules exceptions surviennent chez les petits génomes (<1 Mb) de bactéries intracellulaires et leur absence de détection reflète plus une limite de la méthode statistique (moins efficace avec les petits génomes) qu'une absence de AIMS chez ces organismes.

Ces séquences ne sont pas limitées aux régions intergéniques mais sont également retrouvées dans les régions codantes. De plus, leur abondance augmente de façon graduelle vers le terminus de réplication (exemples : *Pseudomonas aeruginosa* et *E. coli* ; Fig 10A-B). Le rôle biologique proposé est leur reconnaissance par des translocases de type FtsK capable de glisser le long du chromosome pour atteindre le terminus de réplication afin d'initier la séparation des chromosomes après réplication (Fig. 10C). Chez *E. coli*, après la réplication du chromosome, les termini sont tractés vers les cellules filles à travers le septum grâce à la protéine FtsK (Corre et Louarn, 2002, 2005 ; Ip *et al.*, 2003 ; Lau *et al.*, 2003 ; Pease *et al.*, 2005). FtsK participe à la résolution des dimères de chromosome avec les

recombinases XerC et XerD dont la cible est le site *dif* (Blakely *et al.*, 1993). La translocation du chromosome est rendue possible car FtsK pourrait reconnaître une AIMS, appelée Rag (RGNAGGGS), dont la distribution le long du chromosome est fortement biaisée (Bigot *et al.*, 2005 ; Levy *et al.*, 2005). Le biais de distribution de ces séquences augmente sensiblement au niveau du site *dif* puisque 97% de ces motifs sont présents sur le même brin dans les 640 kb encadrant ce locus (Corre et Louarn, 2002 ; Lobry et Louarn, 2003). Des AIMS ont par ailleurs été détectés dans les chromosomes linéaires bactériens (*S. coelicolor*, *A. tumefaciens* et *Borrelia*).

Chez *E. coli*, les AIMS ont été détectées non seulement dans les gènes ancestraux mais également dans les gènes acquis récemment par transfert (Fig 10A). Ce biais de distribution n'est pas dû aux biais de composition en nucléotides (biais mutationnels) (Daubin et Perriere, 2003) mais serait donc sélectionné pour sa fonction polarisatrice. Cependant, et de façon intéressante, seule la distribution non aléatoire de ces séquences semble soumise à une pression de sélection et non les séquences elles-mêmes. En effet, les positions des AIMS dans les gènes orthologues (entre taxons proches) ne sont pas conservées. Seules la distribution et l'abondance croissante vers le terminus sont partagées parmi les génomes étudiés (Hendrickson et Lawrence, 2006). Ainsi, la sélection des AIMS s'appliquerait au niveau des réplichores, c'est-à-dire à un niveau "plus élevé" que la pression de sélection qui s'exerce sur les fonctions des gènes.

La distribution de ces séquences aurait donc un impact fort sur l'évolution des génomes bactériens en constituant une barrière aux réarrangements, notamment aux inversions internes à un réplichore et au maintien de séquences acquises par transfert horizontal. En effet, les génomes d'espèces phylogénétiquement proches partagent souvent des AIMS similaires. Par exemple, GGGCAGGG est une AIMS (différente des sites Rag) chez toutes les Protéobactéries (Hendrickson et Lawrence, 2006). Ainsi, ces séquences polarisant le chromosome influent sur le succès évolutif des échanges de gènes en favorisant ceux impliquant des espèces proches et en les défavorisant fortement entre espèces éloignées (Lawrence et Hendrickson, 2004).

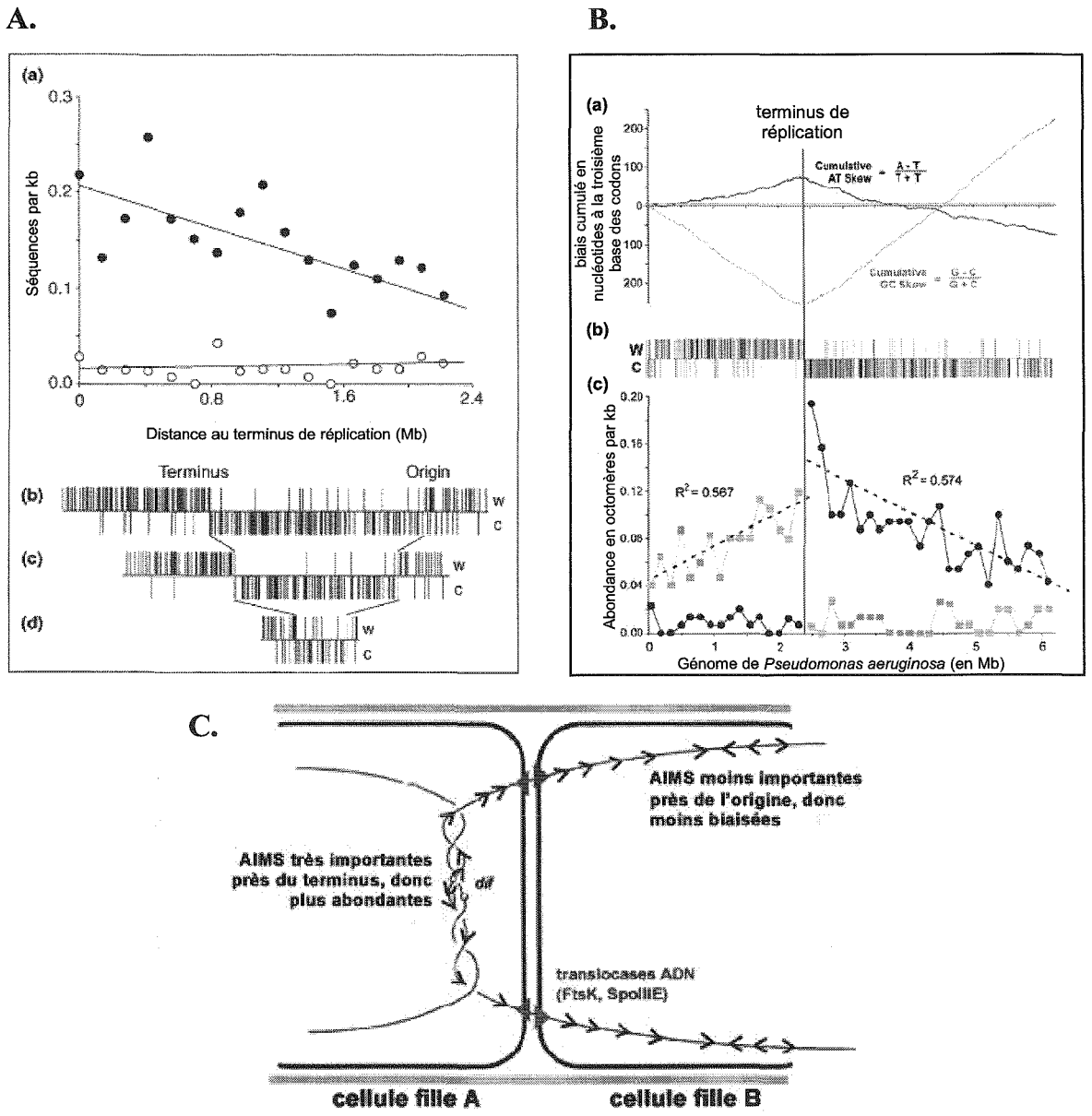


Figure 10 : Polarisation des réplicons

A. Biais de distribution d'oligonucléotides (AIMS) sur le chromosome d'*E. coli* K12. **a**) Abondance des quatre octomères (5'GGGGYAGG3' et 5'GGGYAGGG3') sur le brin continu (cercles pleins) et discontinu (cercles vides) du chromosome. Cette distribution est représentée pour le chromosome complet **b**), sur un chromosome virtuel constitué uniquement des gènes ancestraux d'*E. coli* (ceux retrouvés chez *Salmonella* et *Klebsiella* (**c**)) et sur un chromosome virtuel ne portant que les gènes spécifiques d'*E. coli* (**d**). La distribution biaisée de ces séquences parmi les gènes récemment acquis par *E. coli* reflète soit, l'existence préalable de ce biais parmi les séquences étrangères soit, une sélection pour son établissement rapide. Les traits verticaux sur les brins Watson (w) et Crick (c) indiquent la présence d'un octomère.

B. Séquences polarisant le chromosome de *Pseudomonas aeruginosa.* **a**) Les biais mutationnels permettent de positionner les origine et terminus de réplication. Les A/T skew et G/C skew ont été calculés uniquement à la troisième base des codons des gènes codant des protéines. **b**) Onze séquences octomériques sont retrouvées à une fréquence d'au moins 90% sur le brin continu (représentées par les traits verticaux). **c**) La distribution de trois d'entre elles GAGGAGGG, GGGAGGGG, et GGGTAGGG montre une augmentation significative et progressive vers le terminus de réplication qui n'est pas du au biais mutationnel. L'abondance de ces séquences est représentée pour les brins Watson (en noir) et Crick (en gris).

C. Mécanisme de translocation des chromosomes nouvellement répliqués à travers le septum au moment de la division cellulaire. Les translocases de type FtsK/SpoIIIE, enchâssées dans la membrane au niveau du septum, pompent l'ADN dans les cellules filles en reconnaissant des séquences AIMS polarisant le chromosome. D'après (Lawrence et Hendrickson, 2004, 2005).

b) Contraintes liées à l'organisation en domaines structuraux

L'organisation contrainte des réplicons peut non seulement être mise en évidence *in silico*, mais également *in vivo*. Le surenroulement de la molécule d'ADN influence l'expression des gènes et la mobilité des éléments transférables (Manna *et al.*, 2004). Par ailleurs, la transcription influence la diffusion du surenroulement le long du génome (Deng *et al.*, 2004). Il a également été démontré que la densité de surenroulement est soumise à la sélection (Crozat *et al.*, 2005). Certaines mutations affectant des gènes responsables du surenroulement (*fis* et *topA*) ont été sélectionnées parmi des populations d'*E. coli*. Ces mutants voient leur "fitness" augmenté en condition de compétition pour les substrats.

La structuration physique du chromosome joue également un rôle sur le taux de réarrangements génomiques en contraignant l'accessibilité des séquences homologues capables de recombiner (Garcia-Russell *et al.*, 2004). Le chromosome est en effet organisé en domaines surenroulés. En plus d'une telle organisation régio-spécifique, la position et les mouvements du chromosome pendant le cycle cellulaire sont hautement régulés.
Par microscopie à fluorescence, il a été montré chez *E. coli* (Boccard *et al.*, 2005), *B. subtilis* (Teleman *et al.*, 1998) et *Caulobacter crescentus* (Viollier et Shapiro, 2004 ; Viollier *et al.*, 2004) que le chromosome possède une organisation spatiale dans le cytoplasme contrôlée qui influence l'expression des gènes et la recombinaison. Dans chaque cas, le chromosome adopte une structure en cercle dans la cellule. Les différents segments d'ADN chromosomique seraient donc organisés dans l'ordre de la carte génétique. Chez *E. coli*, le chromosome s'organise en quatre macrodomaines (~1 Mb) et deux régions moins structurées (Valens *et al.*, 2004). Deux de ces macrodomaines correspondent aux régions incluant l'origine (domaine Ori) et le terminus de réplication (domaine Ter). Ils ont tout d'abord été définis par FISH (fluorescence *in situ* hydridization) (Niki *et al.*, 2000) puis confirmés par des expériences mesurant l'efficacité de recombinaison site-spécifique entre deux sites *att* dispersés le long du chromosome (Valens *et al.*, 2004). Les deux autres macrodomaines flanquent le domaine Ter alors que les deux régions présentant une organisation non structurée flanquent le domaine Ori (Fig. 11). Ces expériences ont permis de suggérer que deux loci localisés dans deux macrodomaines différents ne pourraient pas interagir. Cette séquestration des macrodomaines les uns par rapport aux autres résulterait de la localisation cellulaire contrôlée des macrodomaines au cours du cycle cellulaire.

Chez *Streptomyces*, peu de choses sont décrites sur la structuration du chromosome linéaire dans le mycélium et dans les spores. Les protéines terminales liées aux télomères sont capables de s'associer *in vivo*. L'ADN chromosomique adopterait donc une conformation circulaire (Yang et Losick, 2001). Par ailleurs, les compartiments des hyphes végétatifs contiennent plusieurs copies, entre 4 et 8, du chromosome non condensé (Jakimowicz *et al.*, 2005). En revanche, peu avant la septation des hyphes aériens qui requiert FtsZ (Schwedock *et al.*, 1997), la partition permet un positionnement non aléatoire des chromosomes condensés au centre des préspores (Jakimowicz *et al.*, 2005). Ces résultats suggèrent donc qu'une structuration du chromosome existe chez *Streptomyces*.

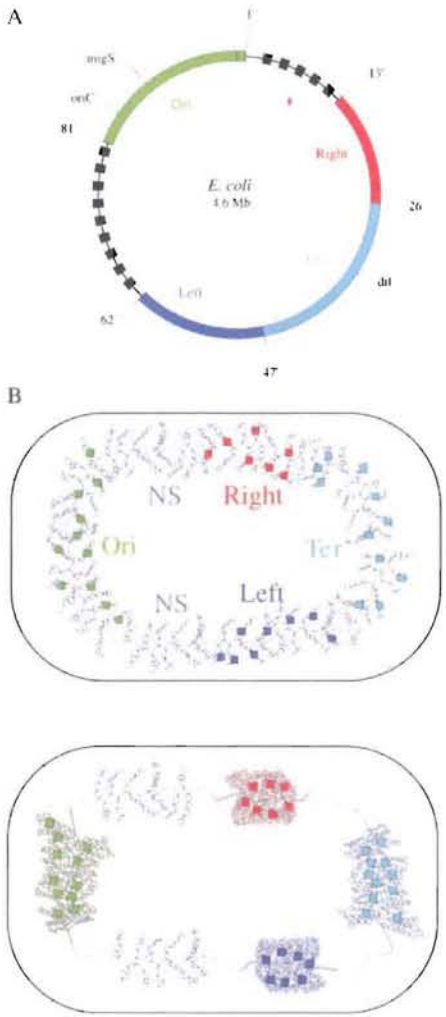


Figure 11 : Organisation structurale du chromosome en macrodomaines.
A. Représentation schématique des macrodomaines (barres colorées) et domaines moins structurés (traits noirs interrompus) chez *E. coli*. Les positions de l'origine de réplication (*oriC*), des sites *dif* et *migS* (impliqué dans la ségrégation chromosomique) sont indiquées.
B. Modèle de structuration du chromosome chez *E. coli*. Le chromosome s'organise dans la cellule selon un cercle composé de quatre macrodomaines (Ori, Ter, Left et Right) et deux domaines moins structurés (NS). La fixation de facteurs inconnus, ayant une localisation spatiale spécifique dans la cellule, sur des sites portés par les macrodomaines concentrerait ces régions en domaines structuraux. Cette séquestration inhibe les interactions entre loci portés par des macrodomaines différents alors que les régions non structurées présentent plus de flexibilité. D'après (Boccard *et al.*, 2005).

c) Propension variable des gènes au transfert horizontal

Tous les gènes ne sont pas transférés (avec succès évolutif) à la même fréquence. La dispersion des gènes à travers les espèces est reliée à la fonction biologique (Nakamura *et al.*, 2004). Un biais concernant les fonctions biologiques des gènes transférés est observé en faveur de trois grandes catégories fonctionnelles : enveloppe cellulaire (13,8%), régulation (11,0%) et processus cellulaires (10,0%). Parmi les gènes classés dans la catégorie "enveloppe cellulaire", ceux impliqués dans la structure de la surface cellulaire et ceux impliqués dans le métabolisme des polysaccharides (et lipopolysaccharides) de surface semblent être plus fréquemment transférés. Parmi ceux impliqués dans les processus cellulaires, des gènes reliés à la pathogénicité, la compétence et le métabolisme secondaire sont retrouvés. Enfin, parmi les gènes transférés, ceux impliqués dans la régulation sont fortement représentés chez les bactéries du sol comme les *Streptomyces* (Nakamura *et al.*, 2004). Etant donnée la corrélation positive entre la proportion de gènes acquis par transfert horizontal et la taille du génome, ces résultats rejoignent les conclusions de (Konstantinidis et Tiedje, 2004) sur l'enrichissement des grands génomes en fonctions régulatrices.

Les gènes impliqués dans les fonctions dites "informationnelles" (transcription, traduction et réplication) (Rivera *et al.*, 1998) sont, au contraire, rarement transférés. De plus, certains gènes de fonctions dites "opérationnelles" (Rivera *et al.*, 1998) sont également peu transférés. C'est le cas des gènes impliqués dans la synthèse des acides aminés et des nucléotides, par exemple (Nakamura *et al.*, 2004).
Ainsi, les gènes les plus fréquemment transférés sont les gènes non essentiels, aussi appelés gènes de contingence. Ceux-ci sont capables de fournir un avantage immédiat à la cellule réceptrice par l'apport d'une ou plusieurs fonctions nouvelles. Bien que certains gènes soient récalcitrants au transfert, les gènes du génome core peuvent tout de même être transférés ou remplacés. Le transfert de gènes orthologues, y compris les gènes d'ARN ribosomiques, est possible et a été démontré chez un certain nombre d'espèces de bactéries et d'archées. Ainsi, un opéron d'ARNr entier a été acquis horizontalement chez *Thermomonospora chromogena* (Yap *et al.*, 1999), le gène de l'ARNr 16S est mosaïque chez *Rhizobium galegae* (Eardly *et al.*, 2005) et une hétérogénéité intragénomique marquée de ce gène est détectée dans le genre *Haloarcula* (jusqu'à 5% de divergence entre différentes copies) (Boucher *et al.*, 2004).

La propension variable au transfert est également le résultat d'une mobilisation non aléatoire des séquences. En effet, les gènes portés par les îlots génomiques mobiles (et mobilisables) sont plus fréquemment transférés (Dobrindt *et al.*, 2004). Par exemple, le transposon conjugatif CTn*scr94* détecté chez *Salmonella senftenberg* mesure environ 100 kb alors que la fonction sélectionnée semble être la fermentation du sucrose codée par un "cluster" de 5 kb seulement (Hochhut *et al.*, 1997). Ainsi, 95% des gènes ayant été transférés n'apportent pas d'avantage détectable mais sont acquis car mobilisés par un élément mobile.

3. *Confinement de la variabilité et compartimentation des génomes*

L'évolution de chacun des gènes au sein d'un génome n'est pas isolée mais dépend du contexte génomique. L'équilibre entre ordre et désordre se traduit, chez certaines espèces, par un confinement de la variabilité. Plus précisément, certaines régions concentrent l'essentiel de la variabilité alors que d'autres restent très stables.

a) Plasticité génomique accrue au niveau du terminus de réplication

La région entourant le terminus de réplication des chromosomes bactériens circulaires serait la région où la sélection naturelle qui s'exerce sur les fonctions des gènes est la plus forte (Lawrence et Hendrickson, 2004). A chaque cycle de réplication, le terminus est la cible d'événements de recombinaison requis pour la séparation des chromosomes (recombinaison site-spécifique au site *dif* grâce aux recombinases XerCD ; cf. paragraphe précédent). Cette recombinaison pourrait favoriser les remaniements en augmentant, par exemple, la fréquence des cassures double-brin.
Le transposon *Tn*7 s'insère de façon préférentielle dans la région entourant le terminus de réplication chez *E. coli* (Peters et Craig, 2000). Les auteurs ont pu montrer que l'affinité particulière du transposon pour cette région du chromosome s'explique par une fréquence d'apparition de cassures double-brin

plus élevée. En effet, l'induction artificielle de telles cassures à différents loci chromosomiques montre une insertion préférentielle de *Tn*7 au niveau de ces loci.

Par comparaison de génomes d'espèces proches (20 génomes comparés uniquement par dot plots), Suyama et Bork suggèrent que les réarrangements seraient plus fréquemment fixés dans la région entourant le terminus de réplication chez certaines espèces : *Mycobacterium leprae* et *tuberculosis*, *Vibrio cholerae* (chromosome 1), *E. coli*, *Chlamydia pneumoniae* et *trachomatis* et les espèces d'archées *Pyrococcus horikoshii* et *abyssi* (Suyama et Bork, 2001).
Chez *E. coli*, la fréquence d'excision d'un prophage (lambda cI857) est plus élevée au niveau du terminus de réplication qu'aux autres loci chromosomiques (Corre *et al.*, 1997 ; Louarn *et al.*, 1991). Ce phénomène a été appelé "hyper-recombinaison" terminale (Louarn *et al.*, 1994). La présence du prophage, dans une certaine orientation, entraîne une instabilité génomique au niveau du terminus de réplication qui s'explique par une inhibition de la résolution des dimères de chromosomes au site *dif* (Corre *et al.*, 2000). Il a été montré plus tard que cette inhibition est due à une perturbation de la polarité du chromosome reconnue par FtsK (Corre et Louarn, 2002).
Le phage CTX, codant le principal facteur de virulence chez *V. cholerae*, s'intègre par recombinaison site-spécifique au terminus de réplication. Ce phage ne code aucune recombinase mais utilise les recombinases XerC et XerD de l'hôte, dont le rôle est d'effectuer la résolution des dimères de chromosome (Huber et Waldor, 2002). Le site d'intégration de CTX (*att*B) est similaire au site *dif* défini chez *E. coli*.
D'autres phages semblent avoir adopté cette stratégie d'intégration : f237 (Iida *et al.*, 2002), CUS-1, CUS-2, phiLf (Lin *et al.*, 2001), Cf16-v1 (Dai *et al.*, 1988), Xfphif1. L'avantage d'un tel mode de maintien réside probablement dans le fait que les recombinases XerC, XerD et le site *dif* sont largement répandus et conservés dans le monde bactérien. De même, l'îlot génomique GGI de *Neisseria gonorrhoeae*, acquis par transfert horizontal, s'intègre par recombinaison site-spécifique au site *dif* (Hamilton *et al.*, 2005).

b) Confinement de la variabilité et plasmides

Les spirochètes du genre *Borrelia*, agents responsables de la maladie de Lyme, possèdent un chromosome linéaire (911 kb chez la souche *B. burgdorferi* B31 (Fraser *et al.*, 1997)) et le plus grand nombre de plasmides décrit chez une bactérie : 21 dont 12 linéaires et 9 circulaires. L'ensemble de l'ADN plasmidique représente 611 kb soit 40% du génome (Casjens *et al.*, 2000).
La variabilité et l'instabilité du génome des *Borrelia* sont principalement le fait des plasmides. En effet, 92% des gènes plasmidiques prédits (670 CDS) sont spécifiques des *Borrelia*. La plupart de ces plasmides sont apparentés et leur distribution à travers les différentes espèces de *Borrelia* diffère. Ils peuvent être perdus, pour la plupart d'entre eux, en conditions de laboratoire mais certains sont impliqués dans la virulence. De nombreux échanges d'ADN entre ces plasmides, notamment entre séquences télomériques (Huang *et al.*, 2004), se sont produits au cours de l'évolution (Casjens *et al.*, 1997 ; Casjens *et al.*, 2000). La famille de plasmides circulaires cp32 compte 7 membres de taille et de séquence très proches. Par ailleurs, chez *B. burgdorferi* B31, le plasmide linéaire lp56 est issu de l'intégration d'un plasmide circulaire de la famille cp32 dans un plasmide linéaire. La conversion d'un plasmide linéaire en plasmide circulaire a également été montrée (Ferdows *et al.*, 1996). L'évolution

rapide de ces plasmides chez *Borrelia* semble être en relation avec une abondance en ADN répété en tandem favorisant les interactions entre réplicons (Casjens *et al.*, 2000). Elle est également soutenue par une faible densité en ADN codant (<70%) par rapport à celle observée sur le chromosome et par la présence d'un grand nombre de petites ORF et d'ORF tronquées (20% de pseudogènes), leur fréquence pouvant atteindre plus de 50% chez certains plasmides (Casjens *et al.*, 2000). Le caractère recombinogène des extrémités d'ADN pourrait expliquer l'instabilité des plasmides linéaires chez *Borrelia* (Casjens, 1999). Par ailleurs, des transferts horizontaux de plasmides entre souches de *B. burgdorferi* ont été mis en évidence (Qiu *et al.*, 2004).

c) Réarrangements et variabilité des régions subtélomériques chez les eucaryotes

La linéarité du chromosome étant une exception chez les espèces procaryotes cultivables, les études concernant l'impact de la linéarité sur l'instabilité génomique se limitent à *Streptomyces* et *Borrelia*. Cependant, l'évolution des chromosomes linéaires a été largement étudiée chez les eucaryotes et notamment chez la levure (pour revue (Dujon *et al.*, 2004)). Le projet "Genolevures", consistant en un séquençage à faible taux de couverture (0,2 x et 0,4 x) de 13 espèces d'hémiascomycètes, a permis une exploration des mécanismes moléculaires régissant l'évolution de leur génome par comparaison avec le génome complètement séquencé de *S. cerevisiae* ((Souciet *et al.*, 2000) et articles associés). Le modèle d'évolution proposé pour les génomes de levures est basé sur la formation de duplications segmentales aboutissant à des génomes mérodiploïdes transitoires qui subissent par la suite des délétions de gènes (Llorente *et al.*, 2000).

Malgré la distance évolutive qui sépare *Streptomyces* et *Saccharomyces*, ils partagent le caractère linéaire de leurs réplicons. Par conséquent, certains mécanismes d'évolution associés aux réplicons linéaires pourraient également être communs.

Alors que 98% des gènes présentent une organisation identique entre les génomes de *S. cerevisiae* et de *Saccharomyces bayanus* var. *uvarum*, un biais dans la localisation des ruptures de synténie a été démontré : 46% des ruptures observées concernent les régions subtélomériques qui ne représentent que 10% du génome (Fischer *et al.*, 2001). Ces régions terminales sont riches en gènes appartenant à de grandes familles. Ainsi, ce biais pourrait résulter d'une probabilité plus élevée de recombinaison à l'intérieur des régions subtélomériques qui portent un grand nombre de gènes présents en plusieurs copies (Fischer *et al.*, 2001).

Un second argument en faveur d'une compartimentation de la variabilité chez *S. cerevisiae* est la présence accrue en pseudogènes dans les régions terminales des chromosomes (Lafontaine *et al.*, 2004). Des traces d'anciennes régions codantes, appelées gènes reliques, ont été recherchées dans les séquences intergéniques. Parmi les 120 gènes reliques identifiés dans les 16 chromosomes, 61% sont localisés dans les régions terminales représentant 6% du génome (~730 kb réparties sur 16 chromosomes, 12160 kb). La formation des gènes reliques par accumulation de mutations semble faire suite à des duplications de segments d'ADN (Fischer *et al.*, 2001).

La plupart des gènes identifiés chez *S. cerevisiae* possèdent un orthologue clairement identifiable (meilleur match réciproque) chez *Saccharomyces paradoxus*, *Saccharomyces mikatae* et *S. bayanus*. Cependant, pour 211 ORF de *S. cerevisiae,* les relations d'orthologie sont plus ambiguës et 80% d'entre elles sont localisées dans les régions subtélomériques (mesurant de 7 à 52 kb) (Kellis *et al.*,

2003). Les événements de translocations réciproques sont également plus fréquemment retrouvés dans ces mêmes régions. Enfin, les 53 ORF spécifiques identifiées sont portées à 69% par ces régions terminales.
La présence de familles de gènes subtélomériques et leur caractère spécifique d'espèce a été confirmée dans 12 génomes d'hémiascomycètes (Fabre *et al.*, 2005). Les gènes de la famille *FLO* ont par ailleurs été directement impliquée dans la formation de réarrangements des séquences subtélomériques (Carro *et al.*, 2003).

Les régions subtélomériques évoluent donc différemment du reste du chromosome et leur particularité a pu être précisée en suivant le devenir des cassures double-brin en fonction de leur localisation chromosomique. Chez *S. cerevisiae*, la fréquence de survie des cellules à de telles cassures varie en fonction de leur localisation sur le chromosome (Ricchetti *et al.*, 2003). La génération artificielle de cassures le long du chromosome XI (665 kb) d'une souche de levure haploïde a permis de montrer que la fréquence des événements de réparation augmente de façon graduelle à mesure que la cassure se produit près d'une extrémité (Ricchetti *et al.*, 2003). Ces événements affectent les régions subtélomériques (environ 30 kb) et sont de diverses natures : addition de télomères, BIR ("break induced replication"), incorporation de séquences plasmidiques et conversion génique. Ils se produisent lorsque la cassure a lieu dans un locus situé entre l'extrémité du chromosome et le premier gène essentiel et s'accompagnent, dans la plupart des cas, d'une perte du fragment chromosomique distal. Les raisons d'un tel gradient de fréquence de recombinaison reste encore mal connues. Il n'est pas corrélé à la redondance plus forte de certains gènes dans ces régions. Certes, la présence de séquences répétées influe sur l'efficacité du mécanisme de BIR mais elle n'explique pas le gradient observé. En effet, une forte proportion des réarrangements terminaux observés est due à des insertions de séquences plasmidiques via la présence de microhomologies.

Chez la levure, les régions subtélomériques sont donc bien distinguables du reste du chromosome de par leurs mécanismes d'évolution. Etant donnée l'absence de gènes essentiels dans ces régions, des événements d'acquisitions, de délétions, d'échanges et de duplications peuvent être maintenus sans affecter la viabilité des cellules. Les régions terminales apparaissent donc comme le "laboratoire" de la variabilité génomique (Ricchetti *et al.*, 2003) associé à l'adaptation (Fabre *et al.*, 2005). Diverses analyses de génomes suggèrent, en effet, que les gènes de contingence responsables de l'adaptation à l'environnement sont concentrés dans les régions subtélomériques, non seulement chez les levures (Fairhead et Dujon, 2006), mais aussi chez d'autres eucaryotes (Barry *et al.*, 2003).

Des caractéristiques communes à celles décrites pour les génomes de levures ressortent de l'analyse des 14 chromosomes du parasite humain *Plasmodium falciparum* (Gardner *et al.*, 2002). En effet, les extrémités chromosomiques concentrent des familles de gènes hautement variables. Parmi celles-ci sont retrouvées les gènes *var* et *rif* codant des protéines de surface et impliquées dans la variation antigénique responsable de l'adaptation au système immunitaire. Pour chaque famille de gènes, un grand nombre de pseudogènes issus de mutations ponctuelles et/ou de troncatures est également détecté. Vingt quatre des 28 extrémités chromosomiques possèdent un gène *var* comme première ORF et des événements de recombinaison permettant des échanges d'extrémités sont fortement soupçonnés.

A travers l'exploration de ces différentes caractéristiques et notamment des régions répétées, les structures subtélomériques représenteraient environ 120 kb (pour chaque extrémité). Le génome de *P. falciparum* est un exemple marquant d'hypervariabilité au niveau des extrémités chromosomiques.
Le génome de *Tetraodon nigroviridis* présente aussi une compartimentation marquée. Les éléments transposables et certains pseudogènes (parfois plus de 200 copies) sont concentrés dans l'hétérochromatine (Dasilva *et al.*, 2002). Une telle compartimentation n'est pas détectée dans les grands génomes de mammifères mais serait plutôt une caractéristique des génomes compacts.

4. *Vitesse d'évolution des séquences et localisation chromosomique*

S'il est prouvé que la variabilité génomique ne s'installe pas de façon aléatoire le long des réplicons, la vitesse d'évolution des séquences des gènes dépend elle aussi, dans certains génomes, de leur localisation. Cette observation avait été faite avant l'apparition des génomes complets dans les banques de données (Sharp *et al.*, 1989) et a pu être précisée par la suite (Mira et Ochman, 2002). En effet, en comparant des fréquences de substitutions survenant aux positions synonymes (soumises de façon moindre à la sélection), il a été montré que plus un gène est proche du terminus de réplication, plus sa vitesse d'évolution est rapide. L'étude ainsi réalisée sur 14 paires d'espèces proches a cependant révélé deux exceptions. La première est *Chlamydia trachomatis* et *Chlamydia muridarum* pour qui aucun effet de la distance par rapport à l'origine de réplication sur le taux de mutations fixées n'est détecté. La seconde est *M. leprae* et *M. tuberculosis* pour lesquelles l'effet observé est inversé.
Une hypothèse historiquement formulée pour expliquer cette corrélation consistait à penser que l'efficacité de la réparation par conversion génique était plus forte près de l'origine de réplication du fait de l'effet dose (Sharp *et al.*, 1989). Même si un effet peut être attribué à la réparation des mésappariements, cette théorie n'explique pas pourquoi l'effet est plus marqué sur les transversions que les transitions (Mira et Ochman, 2002). L'hypothèse selon laquelle la fréquence de mutations augmente avec la distance par rapport à l'origine de réplication a donc été favorisée. La corrélation observée ne reste toutefois qu'une tendance et des variations individuelles pour chaque gène le long du chromosome existent (Hudson *et al.*, 2002).

A l'échelle du chromosome, le pourcentage en bases G/C (et A/T) est significativement différent entre les régions entourant l'origine et le terminus de réplication (Ussery et Hallin, 2004a). Un enrichissement significatif en bases A/T de la région du terminus de réplication a été montré (les régions testées recouvrent 8% du chromosome).
La localisation chromosomique est donc un facteur influençant la composition nucléotidique des gènes et, par conséquent, leur évolution. Une étude réalisée sur 48 chromosomes procaryotes, visant à suivre l'évolution du pourcentage en bases G+C à la troisième position des codons (GC3), a montré qu'une structuration du GC3 est une caractéristique commune des génomes bactériens (Daubin et Perriere, 2003). Une richesse en AT3 caractérise la région entourant le terminus de réplication chez 42 des 48 chromosomes testés. Ces variations caractéristiques du GC3 sont souvent responsables de variations du CAI (usage des codons) des gènes en fonction de leur localisation. Les taux de mutation affectant des positions synonymes (et également les positions non synonymes dans certains cas) augmentent de façon significative avec la distance à l'origine de réplication. Une hétérogénéité des vitesses

d'évolution des gènes en fonction de leur position sur le réplicon a été constatée. Cette particularité serait le résultat d'une pression de mutations plus élevée au niveau du terminus de réplication et non celui d'une sélection plus faible sur les séquences.

Quelles hypothèses peuvent expliquer un tel profil particulier des gènes entourant le terminus de réplication ?

Tout d'abord, cette richesse en AT3 s'explique en partie par le fait que cette région est une cible privilégiée du transfert horizontal. Cependant, le biais observé persiste si l'on ne garde que les gènes conservés dans l'analyse. Un enrichissement intrinsèque de cette région serait donc responsable de l'établissement de ce biais. Pour expliquer la vitesse accrue de fixation de mutations au niveau du terminus de réplication, Sharp *et al.* avaient émis l'idée que les séquences proches du terminus sont en copie unique pendant la plupart du cycle cellulaire. Elles auraient donc moins de possibilités d'engager une réparation des mutations par conversion génique que ne l'ont les séquences proches de l'origine de réplication (Sharp *et al.*, 1989). Cependant, des travaux antérieurs ont montré que les origines de réplication ségrégent vers des pôles opposés dans la cellule en division, ne permettant pas ou peu de possibilités de recombiner (Sawitzke et Austin, 2001). Daubin et Perriere suggèrent une hypothèse alternative impliquant également des mécanismes de réparation : la présence de complexes *ter*/Tus au terminus de réplication pourrait inhiber les hélicases normalement impliquées dans la reprise de la réplication au niveau des lésions de l'ADN. Ainsi, la réparation ferait intervenir le mécanisme de translésion qui incorpore préférentiellement de l'AMP dans les sites abasiques conduisant à un enrichissement de cette région en bases A+T (Daubin et Perriere, 2003).

Cette hétérogénéité de l'usage des codons intrinsèque à tout génome est une source de surestimation du nombre de gènes acquis par transfert horizontal comme ce fut le cas chez *E. coli* (Lawrence et Ochman, 1998).

D. Plasticité génomique chez *Streptomyces*

1. *Les caractéristiques phénotypiques et génotypiques originales des Streptomyces*

Les *Streptomyces* sont des bactéries Gram positive à haut pourcentage en bases G+C appartenant à l'ordre des *Actinomycetales*. Leur habitat naturel est le sol. Parmi les bactéries cultivables, les *Streptomyces* possèdent des caractéristiques tout à fait originales. Elles se développent selon un cycle cellulaire complexe caractérisé par une différentiation morphologique et biochimique poussée (Chater, 1993) (Fig. 12). Sur milieu de culture solide, la germination d'une spore produit des filaments mycéliens à croissance apicale (par les extrémités) et capables de se ramifier. Ce mycélium végétatif se différencie ensuite en mycélium aérien lorsque le milieu devient limitant en éléments nutritifs. Enfin, les hyphes aériens forment des chaînes de spores par septation. Chaque spore ne contient, en temps normal, qu'un seul exemplaire du chromosome alors que de multiples nucléoïdes coexistent dans le mycélium.

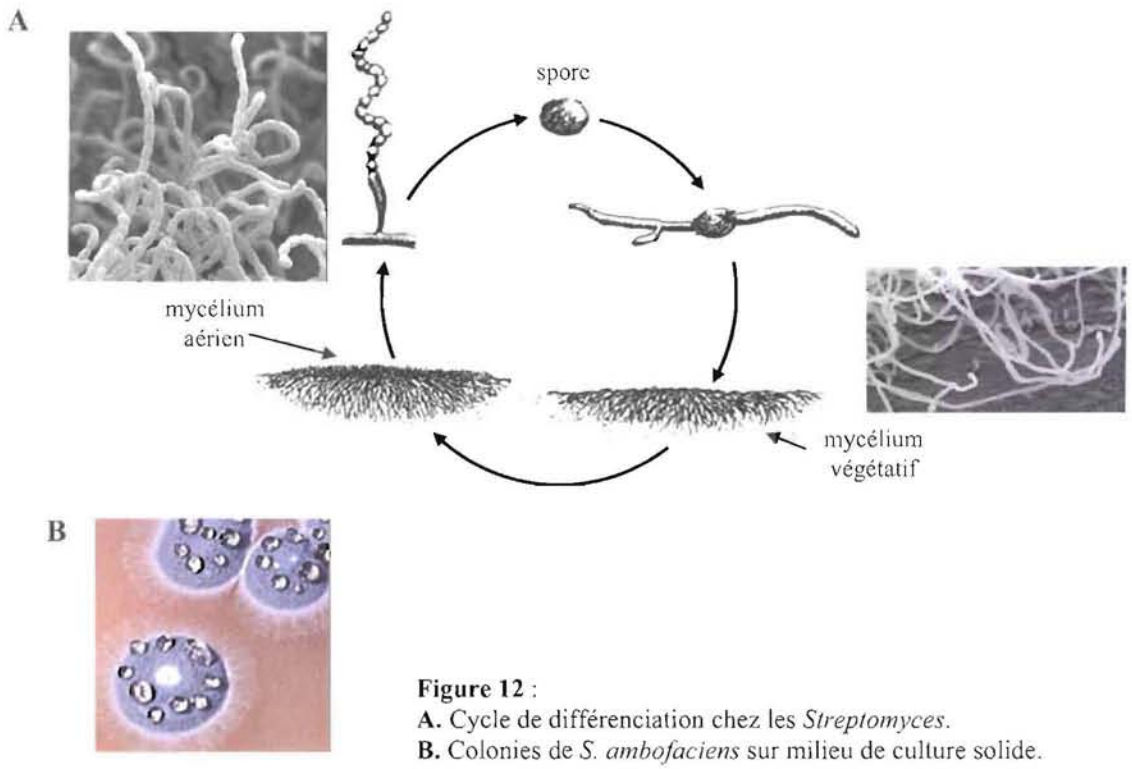


Figure 12 :
A. Cycle de différenciation chez les *Streptomyces*.
B. Colonies de *S. ambofaciens* sur milieu de culture solide.

Les *Streptomyces* sont d'une grande importance économique. Ils sont notamment producteurs de la majorité des molécules antibiotiques utilisées en thérapie humaine mais synthétisent également de nombreux métabolites d'intérêt biotechnologique. Il a été estimé que sur 16500 antibiotiques connus, 8700 (53%) sont produits par les Actinomycètes dont 6550 (40%) par des espèces de *Streptomyces* (Berdy, 2005). La différenciation morphologique s'accompagne en effet d'une différentiation métabolique. Un métabolisme secondaire se met en place dans les phases tardives du développement,

donnant lieu à la biosynthèse de composés d'une extraordinaire diversité de structures et d'activités biologiques. Le séquençage des génomes a ouvert des nouvelles perspectives dans la recherche de nouveaux métabolites puisqu'un important réservoir encore inconnu de diversité génétique et métabolique réside dans les génomes des souches isolées.

Leurs caractéristiques génomiques sont tout aussi originales dans la mesure où tous les génomes caractérisés de *Streptomyces* se composent d'un chromosome linéaire de grande taille (8,7 Mb pour *S. coelicolor*, 10,1 Mb pour *S. scabies*) avec un pourcentage très élevé en bases G+C (71% à 73%). De plus, certaines espèces peuvent posséder des plasmides linéaires et/ou circulaires. Comme de nombreux réplicons linéaires, ceux des *Streptomyces* possèdent des **répétitions terminales inversées** appelées **TIR** de composition et de taille variables. L'origine de réplication *oriC* est localisée au centre des réplicons et fonctionne de façon bidirectionnelle aussi bien pour les chromosomes (Musialowski *et al.*, 1994) que pour les plasmides (Chang et Cohen, 1994).

Les extrémités chromosomiques sont les loci où la réplication se termine. Comme l'extrémité 3' des réplicons linéaires ne peut pas être répliquée de façon continue, les 280 nucléotides terminaux restent temporairement sous forme simple-brin (Chang et Cohen, 1994). Ces séquences télomériques sont constituées de palindromes potentiellement capables de se replier en tiges boucles pour former une structure complexe (Huang *et al.*, 1998) (Fig. 13). Les structure et séquence des télomères sont très conservées à travers les espèces de *Streptomyces*. Cependant, quelques télomères atypiques ont été mis en évidence à l'extrémité de certains réplicons. C'est le cas du plasmide SCP1 de *S. coelicolor* (Kinashi *et al.*, 1991) et du chromosome de *Streptomyces griseus* (Goshi *et al.*, 2002). Ces derniers sont également constitués de palindromes mais sont totalement différents entre eux et différents des télomères classiquement retrouvés chez *Streptomyces*.

L'extrémité 5' des télomères est liée de façon covalente à une protéine terminale appelée Tpg (Terminal Protein Gene, (Bao et Cohen, 2001)). Les Tpg sont capables d'interagir entre elles, donnant aux réplicons linéaires une architecture circulaire (Wang *et al.*, 1999 ; Yang et Losick, 2001). L'analyse du voisinage immédiat du gène *tpg* a révélé la présence du gène *tap* (Terminal Associated Protein gene), formant vraisemblablement un opéron avec ce dernier (Bao et Cohen, 2003). Les protéines Tpg et Tap forment un complexe protéique terminal essentiel au maintien de l'ADN chromosomique sous forme linéaire (Bao et Cohen, 2003). Elles sont impliquées dans la réplication des télomères. Tap recruterait Tpg qui servirait d'amorce pour l'initiation de la réplication des télomères sur le brin discontinu (Bao et Cohen, 2003). Deux autres protéines, PolA (ADN polymérase) et TopA (topoisomerase I) font partie de ce complexe télomérique (Bao et Cohen, 2004).

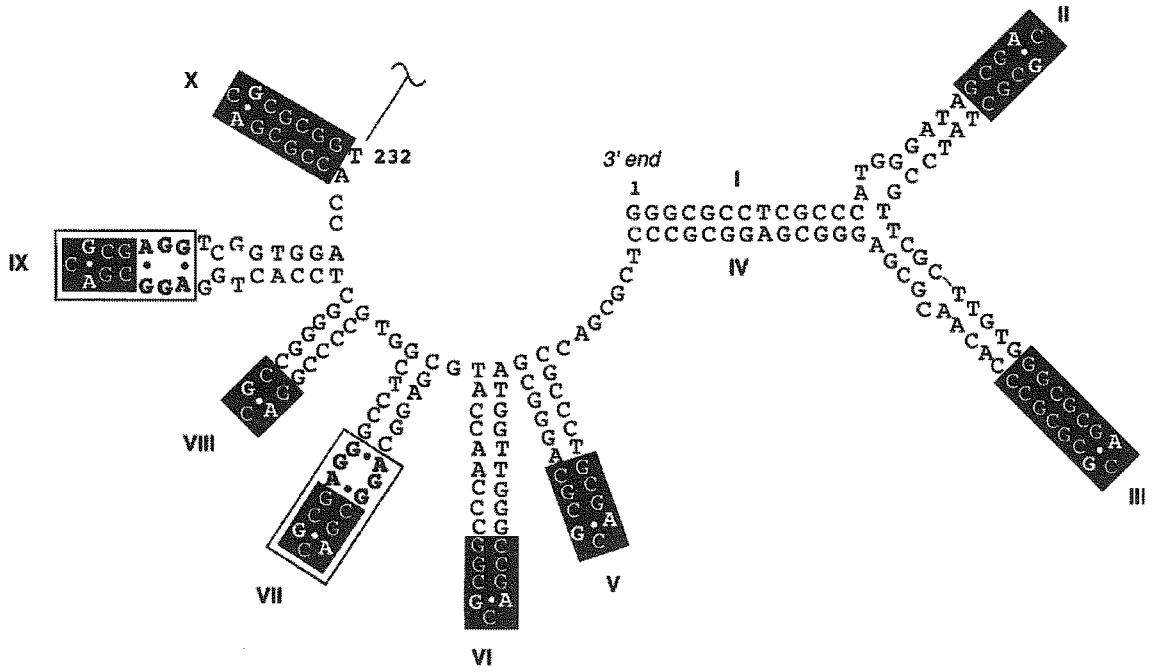


Figure 13 : Structure secondaire prédite des télomères du chromosome de *Streptomyces lividans*. L'extrémité 3' ne peut être répliquée de façon continue. Sous forme simple brin, les séquences palindromiques peuvent se replier pour former des tiges et boucles G.C.A caractéristiques. Chaque palindrome est numéroté. Les derniers nucléotides en 3' s'apparient avec le palindrome IV formant ainsi une structure en Y (structure en "foldback") potentiellement importante pour la réplication des télomères. D'après (Bey *et al.*, 2000).

2. *L'instabilité génomique chez Streptomyces*

L'apparition de mutants au sein d'une population est à la base de l'évolution des organismes vivants mais son ampleur est remarquablement élevée chez les *Streptomyces* (pour revues : (Chen *et al.*, 2002 ; Leblond et Decaris, 1994 ; Volff et Altenbuchner, 2000)). En effet, chez *S. ambofaciens*, des variants phénotypiques affectés dans la pigmentation apparaissent à haute fréquence (1%) dans la descendance de la souche sauvage (Leblond *et al.*, 1989 ; Martin *et al.*, 1998). Cette caractéristique est partagée par toutes les espèces de *Streptomyces* chez lesquelles le phénomène a été étudié. L'instabilité génétique se caractérise par la perte de marqueurs génétiques tels que le gène *cmlR*, qui confère la résistance au chloramphénicol, ou *argG*, impliqué dans la biosynthèse de l'arginine. Dans la descendance de certains mutants, l'instabilité génétique atteint un niveau extrêmement élevé (87%). Ce phénomène a été appelé "hypervariabilité" (Leblond *et al.*, 1989).

L'instabilité génétique a pu être corrélée à la formation de réarrangements de grande ampleur affectant le chromosome (Birch *et al.*, 1989 ; Leblond *et al.*, 1991). Ces réarrangements correspondent à des délétions de grands fragments d'ADN et à des amplifications de loci particuliers appelés AUD (Amplifiable Unit of DNA). Les amplifications sont le plus souvent retrouvées aux bornes des régions délétées suggérant des mécanismes liés.

Les régions affectées par ces réarrangements correspondent aux régions terminales du chromosome linéaire, aussi appelées régions instables (Leblond *et al.*, 1991 ; Leblond et Decaris, 1999 ; Redenbach *et al.*, 1993). Chez *S. lividans*, 1 Mb peut être perdu en conditions de laboratoire (Redenbach *et al.*,

1993). Chez *S. ambofaciens*, les régions dispensables représentent 2,3 Mb, soit environ un quart du chromosome englobant les extrémités chromosomiques (Fischer *et al.*, 1997a). Les délétions peuvent être internes à un bras chromosomique, c'est-à-dire qu'elles laissent les deux télomères intacts. Cependant, le maintien du chromosome sous forme circulaire est possible. Il a été observé chez certains mutants affectés par des délétions spontanées chez *S. ambofaciens* (Fischer *et al.*, 1997a). De plus, la circularisation du chromosome a pu être provoquée artificiellement (Volff *et al.*, 1997). La viabilité des mutants obtenus montre que la ségrégation et la réplication du chromosome restent fonctionnelles malgré la forme circulaire.

Les mutants possédant un chromosome circularisé présentent un niveau d'instabilité élevé, parfois plus élevé que la souche sauvage suggérant qu'il n'existe pas de lien strict entre linéarité et instabilité (Fischer *et al.*, 1997a ; Lin et Chen, 1997 ; Volff *et al.*, 1997).

E. Objectifs de la thèse

Les études portant sur l'instabilité ont permis de caractériser la structure du chromosome de différents mutants isolés de *S. ambofaciens* (Fischer *et al.*, 1997a). Des délétions de grande taille ont ainsi pu être mises en évidence chez ces mutants dérivant de la souche de collection DSM40697 (Fig. 14). La plus grande délétion a été observée chez la souche NSA857 pour laquelle 2335 kb d'ADN terminaux ont été perdus. Les régions dispensables, appelées régions instables, s'étendent sur environ 1,3 Mb sur chaque bras chromosomique.

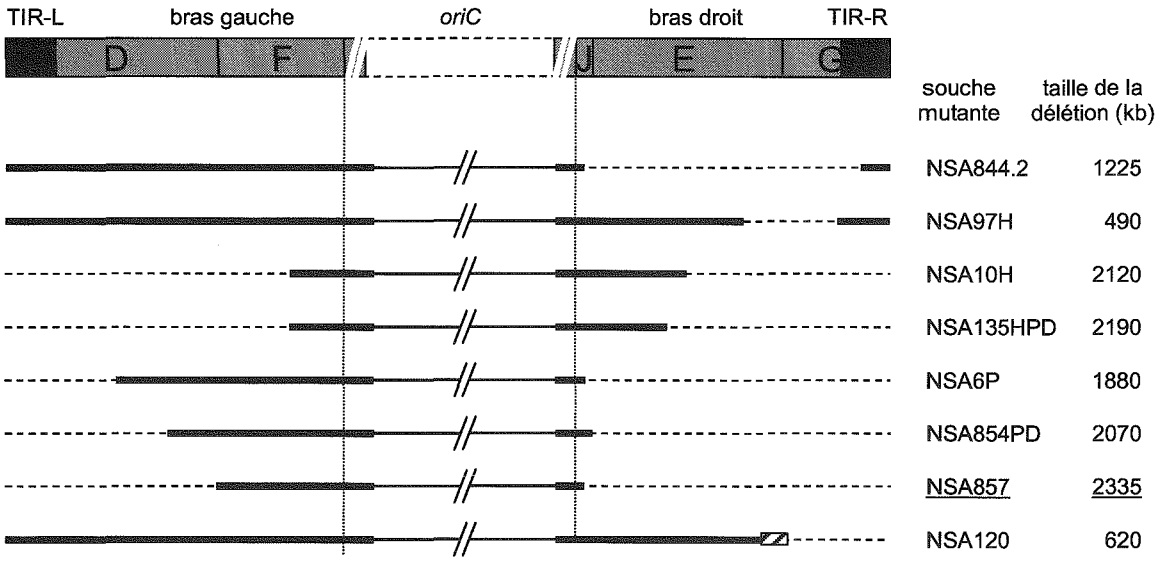


Figure 14 : Taille et localisation des régions délétées chez les différentes souches mutantes de *S. ambofaciens* (Fischer *et al.*, 1997a). La carte de restriction *Ase*I des régions terminales du chromosome de la souche parentale DSM40697 est indiquée et les traits pointillés représentent les régions délétées chez les différents mutants. Les souches pour lesquelles la délétion englobe les deux télomères, possèdent un chromosome circularisé. Le rectangle hachuré indique la présence d'une amplification bordant la région délétée. NSA857 est la souche montrant la plus grande délétion retrouvée chez *S. ambofaciens*. D'après (Fischer *et al.*, 1997a).

Le chromosome des *Streptomyces* semble donc posséder une structure compartimentée avec une région centrale qui concentre les gènes essentiels alors que les régions terminales ne seraient porteuses que de gènes non essentiels au développement végétatif. Cette particularité a été confirmée par le séquençage du génome de *S. coelicolor* publié en 2002 (Bentley *et al.*, 2002). Sur la base de l'annotation des fonctions des 7825 gènes codant des protéines chez *S. coelicolor*, Bentley *et al.* ont défini une région "core" et des régions de contingence. La première correspond à la région centrale (4,9 Mb) et regroupe les gènes essentiels. Les régions de contingence correspondent aux régions terminales et sont estimée à 1,5 Mb (bras gauche) et 2,3 Mb (bras droit) (Bentley *et al.*, 2002). Toutefois ces limites sont discutables. En effet, les 1,6 Mb terminaux du bras droit portent des gènes essentiels. Ceci fut montré par l'isolement d'une souche mutante de *S. coelicolor* contenant deux chromosomes issus de la recombinaison du chromosome sauvage et du plasmide SCP1 (Yamasaki et Kinashi, 2004).

La région "core" montre une synténie avec les chromosomes circulaires d'autres actinomycètes comme *M. tuberculosis*, *Frankia* sp. Cci3, *Nocardia farcinica*, *C. diphteriae* et *Thermobifida fusca* (Fig. 15). Elle constituerait par conséquent la partie ancestrale du chromosome.
La comparaison du chromosome de *S. coelicolor* avec celui de *S. avermitilis*, publié en 2003, a montré que les régions terminales du chromosome concentrent la majeure partie des gènes spécifiques d'espèce (Ikeda *et al.*, 2003). Au contraire, le centre du chromosome est fortement synténique entre ces deux espèces (Fig. 15C). Cette synténie ne se limite pas à la région core mais englobe une partie des régions de contingence.

Dans la poursuite des études concernant l'instabilité chez *S. ambofaciens*, le laboratoire de Génétique et Microbiologie (Nancy), en collaboration avec le Génoscope (Centre National de Séquençage à Evry) et l'Institut de Génétique et Microbiologie (J.-L. Pernodet, Orsay), a initié un programme de séquençage des régions instables chez la souche ATCC23877 de *S. ambofaciens*. De plus, le séquençage des répétitions terminales inversées de la souche *S. ambofaciens* DSM40697 a été également engagé afin d'évaluer la variabilité terminale au niveau intraspécifique. Des expériences d'hybridations avaient permis de mettre en évidence l'existence de régions spécifiques de souches aux extrémités des chromosomes de *S. ambofaciens* ATCC23877 et DSM40697. Ces régions sont inclues dans les TIR.

Le premier objectif de ma thèse était donc de réaliser l'assemblage et l'annotation des séquences produites par le Génoscope et d'identifier les gènes des régions instables chez *S. ambofaciens*.
Par la suite, des comparaisons de génomes entre les différents *Streptomyces* ont été réalisées. *S. ambofaciens* et *S. coelicolor* qui constituent un couple d'espèces très proches phylogénétiquement (1,1% de divergence entre les séquences d'ARN 16S ; Fig. 16) alors que *S. avermitilis* représente une espèce plus éloignée des deux premières (2,9% de divergence avec *S. ambofaciens*). La possibilité de comparer des génomes d'espèces proches et éloignées a permis d'obtenir une vision dynamique de l'évolution du contenu en gènes et de la structure du chromosome linéaire.

Ces analyses avaient pour but d'aborder les questions suivantes :

- Quelle est l'ampleur de la variabilité génomique, notamment dans les régions terminales ? Comment évolue cette variabilité en fonction de la distance phylogénétique des espèces considérées ?
- Quel est l'impact du transfert horizontal sur l'organisation des régions terminales chez *Streptomyces* ?
- Quels mécanismes maintiennent la compartimentation génomique ?
- Existe-il un lien entre linéarité chromosomique et instabilité ?

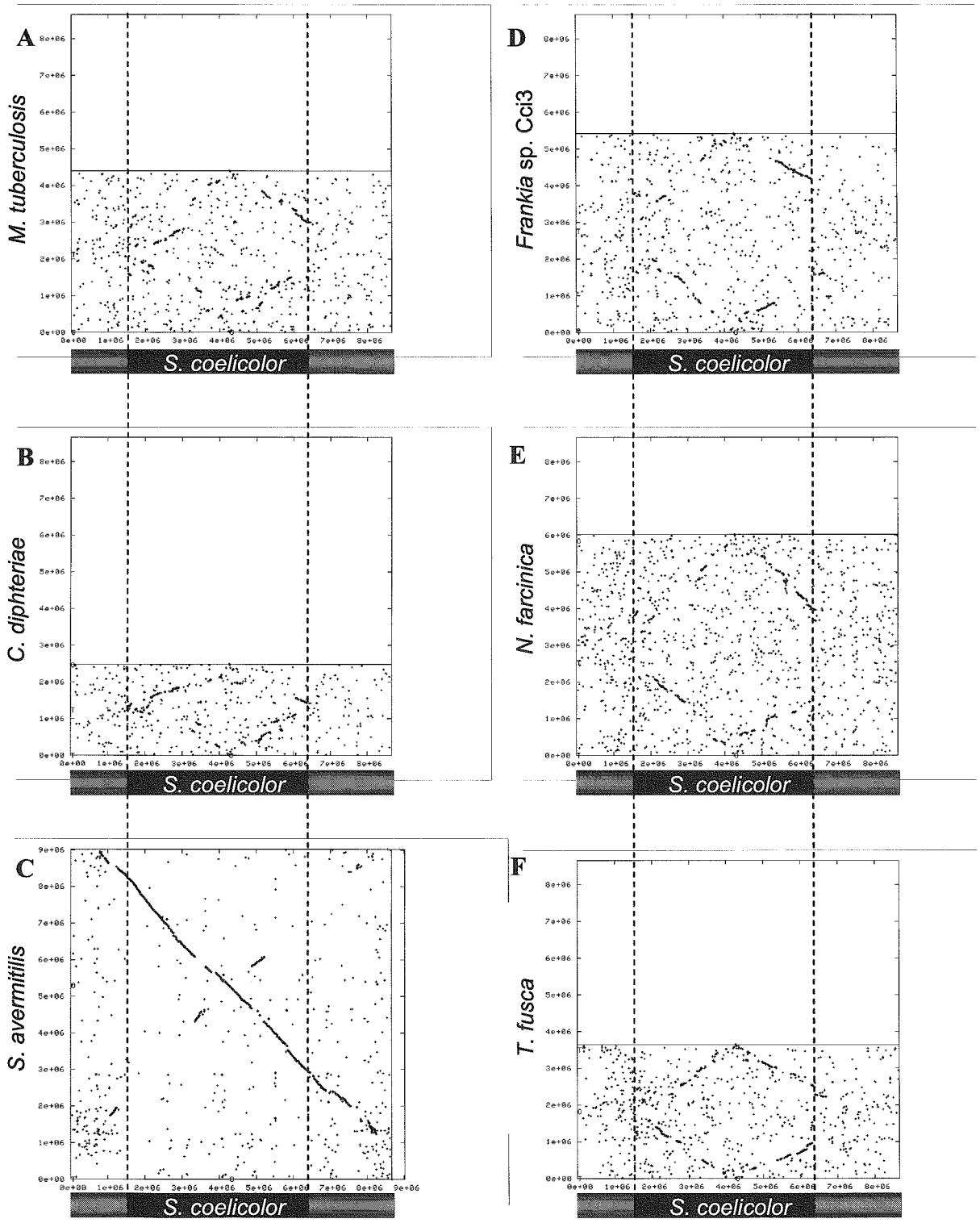


Figure 15 : Comparaison par dot plots du chromosome de *S. coelicolor* avec celui d'autres actinomycètes dont le génome est séquencé : *S. avermitilis*, *M. tuberculosis*, *C. diphteriae*, *Frankia sp. Cci3*, *Nocardia farcinica* et *Thermobifida fusca*.
Chaque point sur les graphes représente les coordonnées d'un couple de gènes codant des protéines orthologues. Les origine et terminus de réplication de chaque chromosome sont respectivement représentés par les lettres "O" et "T". Les régions "core" et de contingence (définies par Bentley *et al.*) sont représentées respectivement en noir et gris.

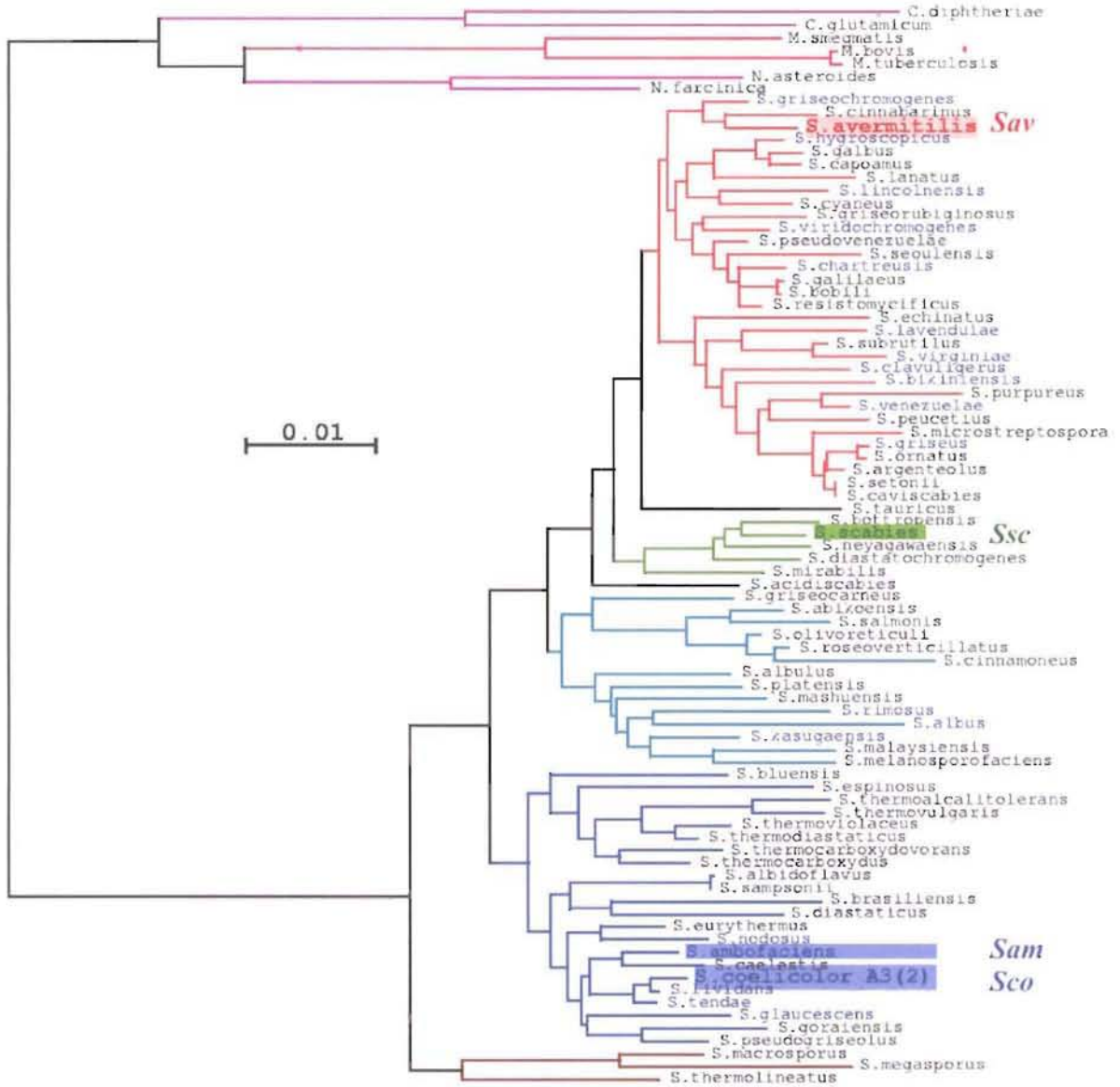


Figure 16 : Arbre phylogénétique des *Streptomyces* réalisé à partir des séquences des ADNr 16S et enraciné avec des séquences provenant d'autres actinomycètes (*Nocardia*, *Corynebacterium* et *Mycobacterium* ; cet arbre est disponible sur http://avermitilis.ls.kitasato-u.ac.jp). *Sam* : *S. ambofaciens*, *Sco* : *S. coelicolor*, *Sav* : *S. avermitilis*, *Ssc* : *S. scabies*.

RESULTATS

RESULTATS

A. Séquençage du génome de *S. ambofaciens*

Le programme de séquençage des régions terminales du chromosome de *S. ambofaciens* ATCC23877 a été initié en juillet 2000. Il est le fruit d'une collaboration entre le Centre National de Séquençage (équipe supervisée par Valérie Barbe), l'Institut de Génétique et Microbiologie d'Orsay (équipe de Jean-Luc Pernodet ; Université Paris Sud XI) et le Laboratoire de Génétique de Microbiologie de Nancy (équipe de Pierre Leblond ; Université Nancy 1).

L'assemblage et l'annotation du génome partiel de *S. ambofaciens* furent les premiers objectifs de ma thèse. Quelle stratégie a été employée pour séquencer partiellement le chromosome ? A quels problèmes a-t-il fallu faire face et quelles solutions ont permis de mener le programme de séquençage à son terme ?

La décision d'annoter au fur et à mesure de la production des données a été prise afin d'exploiter les séquences produites. Cependant, réaliser des analyses sur des séquences en cours de production nécessite l'établissement de procédures d'annotation et de comparaison de génomes les plus automatisées possibles. En effet, il est beaucoup plus simple et plus efficace de recommencer le processus d'annotation d'un génome que de faire évoluer une annotation existante sur des séquences changeant au gré des étapes d'assemblage et de correction des incertitudes.
Par conséquent, le développement de méthodes bioinformatiques adaptées à cette situation a donc été envisagé et a constitué le premier objectif de ce travail.

1. Stratégies de séquençage

a) Choix des souches

Deux souches de *S. ambofaciens* isolées indépendamment sont disponibles dans les collections de souches. La première, **ATCC23877**, isolée en 1954 du sol de Peronne en France (Pinnert-Sindico, 1954), est celle choisie pour le séquençage car elle est utilisée en industrie pharmaceutique pour la production de l'antibiotique macrolide spiramycine. La seconde étudiée au cours de ce travail est la souche **DSM40697** isolée du sol en Italie (Hütter, 1967). Les études portant sur les phénomènes d'instabilité ont été réalisées sur ces deux souches (Leblond et Decaris, 1999).

Ces dernières correspondent à deux isolats indépendants et leur appartenance à une même espèce est, à l'origine, basée sur des caractères phénétiques. En effet, elles sont toutes deux productrices des antibiotiques spiramycine, congocidine et alpomycine (Pang *et al.*, 2004 ; Pinnert-Sindico, 1954). Au contraire, un profil de production d'antibiotiques très différent a été décrit pour l'espèce phylogénétiquement proche *S. coelicolor* dont la séquence d'ADNr 16S ne diverge que de 1,1% par rapport à celle de *S. ambofaciens* ATCC23877. L'appartenance à la même espèce des deux souches de

S. ambofaciens a ensuite été confirmée par des méthodes moléculaires. En effet, les profils de restriction *Ase*I (endonucléase à sites de coupure rare chez *Streptomyces*, AT^TAAT) de leur chromosome présentent une grande similitude (Fig. 17)(Leblond *et al.*, 1996). Ils ne diffèrent que par quatre sites. Des expériences d'hybridation utilisant des clones de liaison (fragments d'ADN contenant un site *Ase*I utilisés pour la cartographie du chromosome) ont permis de mettre en évidence que la majeure partie des sites *Ase*I est partagée entre les deux souches. Une preuve supplémentaire de leur proche parenté est la présence d'une duplication de gènes codant des facteurs sigma alternatifs, les gènes *has* (Roth *et al.*, 2004). Cette duplication est partagée entre les deux souches de *S. ambofaciens* alors qu'un seul homologue aux gènes *has* est retrouvé dans le génome de *S. coelicolor*.

Enfin, la comparaison des séquences ITS (internal transcribed spacers) des six opérons ribosomiques, qui sont couramment utilisées pour préciser les relations phylogénétiques à courtes distances évolutives, a montré que les deux souches de *S. ambofaciens* partagent les mêmes séquences (100% d'identité). En revanche, un polymorphisme existe avec toutes les autres espèces de *Streptomyces* testées dont *S. coelicolor* (Wenner *et al.*, 2002). Bien que des événements de conversion génique ont été mis en évidence entre les différents loci *rrn* depuis leur divergence récente, aucun crossing-over n'a été fixé, laissant ainsi les deux chromosomes colinéaires (Wenner *et al.*, 2002).

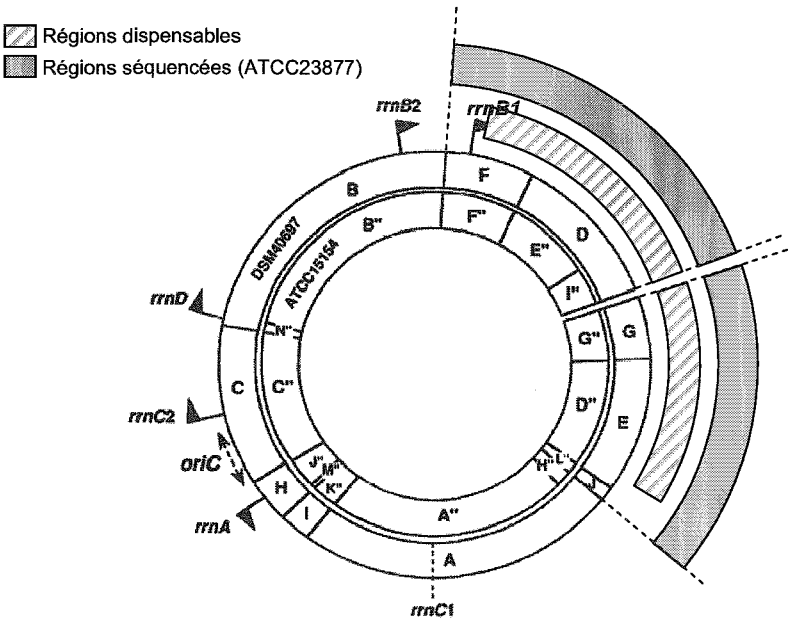


Figure 17 : Localisation des régions séquencées par rapport aux régions dispensables sur les deux souches de *S. ambofaciens*.
Cartes de restriction *Ase*I des chromosomes de *S. ambofaciens* ATCC15154 (un dérivé de la souche ATCC23877 ; cercle interne) et DSM40697 (cercle externe).
La localisation et le sens de transcription des opérons ribosomiques (*rrn*) sont indiqués par des drapeaux. Selon (Fischer *et al.*, 1998a).

b) Banques d'ADN génomique

Deux banques de BAC (Bacterial Artificial Chromosome) recombinants ont été réalisées à partir de l'ADN génomique de *S. ambofaciens* ATCC23877 digéré partiellement par l'endonucléase *Sau*3A (^GATC). Les banques ont été effectuées par F. X. Francou (IGM, Orsay).

Pour la première banque (banque A), aucune sélection de la taille des fragments clonés n'a été réalisée. Ainsi, les fragments de restriction obtenus ont été clonés directement par ligation dans le vecteur pBeloBAC11 (Wang *et al.*, 1997) à l'aide du site de clonage unique *Bam*HI (G^GATCC), les deux

enzymes de restriction générant des extrémités cohésives compatibles. Ce vecteur appartient à la seconde génération de vecteurs BAC dérivant de pBAC108L (Shizuya *et al.*, 1992) qui sont présents en très faible nombre de copies par cellule (1 à 2 copies).
En réalisant la seconde banque d'ADN génomique de *S. ambofaciens* (banque **B**), une sélection sur la taille des fragments clonés a été effectuée au moyen d'une étape d'encapsidation des BAC recombinants. Les inserts ainsi sélectionnés présentent une taille comprise entre 30 kb et 53 kb. La banque A est composée de 1737 clones tandis que 3072 clones composent la banque B. Au total, 4809 clones sont ainsi disponibles ce qui représente un taux de couverture du génome de 21x.

Par ailleurs, une banque de cosmides recombinants recouvrant l'intégralité des répétitions terminales inversées (TIR ; >200 kb) avait été réalisée et ordonnée avant mon arrivée au laboratoire (Berger *et al.*, 1996). Des digestions partielles *Bam*HI du génome de *S. ambofaciens* ATCC23877 ont été effectuées puis les fragments obtenus ont été clonés dans le cosmide Supercos1 (Stratagene). La taille des inserts ainsi clonés est en moyenne de 41 kb.

Concernant la souche DSM40697, une banque de cosmides (digestions partielles *Bam*HI ; clonage dans Supercos1) recouvrant partiellement les régions terminales (3 Mb) du chromosome a été réalisée et ordonnée.

c) Réactions de séquençage

Les réactions de séquençage du chromosome de la souche ATCC23877 et l'obtention des chromatogrammes correspondants ont été réalisées sous la responsabilité de Valérie Barbe au Centre National de Séquençage (Génoscope). Les cosmides et BAC recombinants ont été sous-clonés à l'aide du vecteur pCNS (un dérivé de pSU18 (Bartolome *et al.*, 1991)) et introduits dans *E. coli* DH10B. Pour chaque BAC, 384 réactions de séquençage ont été effectuées, générant ainsi un taux de recouvrement de **8 x** pour un insert de **38 kb** (en moyenne). Au terme de cette première étape, les régions présentant des incertitudes dues à un taux de recouvrement localement faible et les régions non recouvertes (appelés "gaps") ont été amplifiées par PCR avant d'être séquencées.

d) Séquençage des TIR des deux souches de S. ambofaciens étudiées

En premier lieu, l'obtention des séquences des TIR a nécessité le séquençage de respectivement 10 et 8 cosmides recombinants ordonnés pour les souches ATCC23877 (Fig. 18A) et DSM40697 (Fig. 18B). La taille des inserts des cosmides (respectivement 41 kb et 38 kb en moyenne) étant inférieure à celle des TIR (respectivement 200 et 210 kb), l'origine des inserts (bras chromosomique gauche ou droit) recouvrant les TIR ne peut pas être identifiée. Ainsi, une seule séquence a été obtenue pour les deux copies des TIR (du cosmide C7 à F6 pour la souche ATCC23877; du cosmide AD91 à AD68 pour la souche DSM40697). Au contraire, les cosmides F9/G6/F19/H2 (ATCC23877) et AD50/AD118 (DSM40697) portent des séquences spécifiques de chaque bras (Fig. 18).

e) Séquençage du génome de la souche ATCC23877 à faible taux de couverture

Avant d'entamer le séquençage complet des régions terminales par marche sur le chromosome, un séquençage systématique des extrémités des inserts des 4809 BAC composant les deux banques a été entrepris. Ainsi, 8457 séquences (appelées **BES** pour "BAC End Sequence") de 417+/-122 nucléotides ont été obtenues, ce qui représente 3.524.440 nucléotides. La taille du chromosome de *S. ambofaciens* étant estimée à 8,5 Mb, ce séquençage systématique a permis d'obtenir un taux de couverture d'environ 0,4 x représentant statistiquement 1 BES/kilobase.

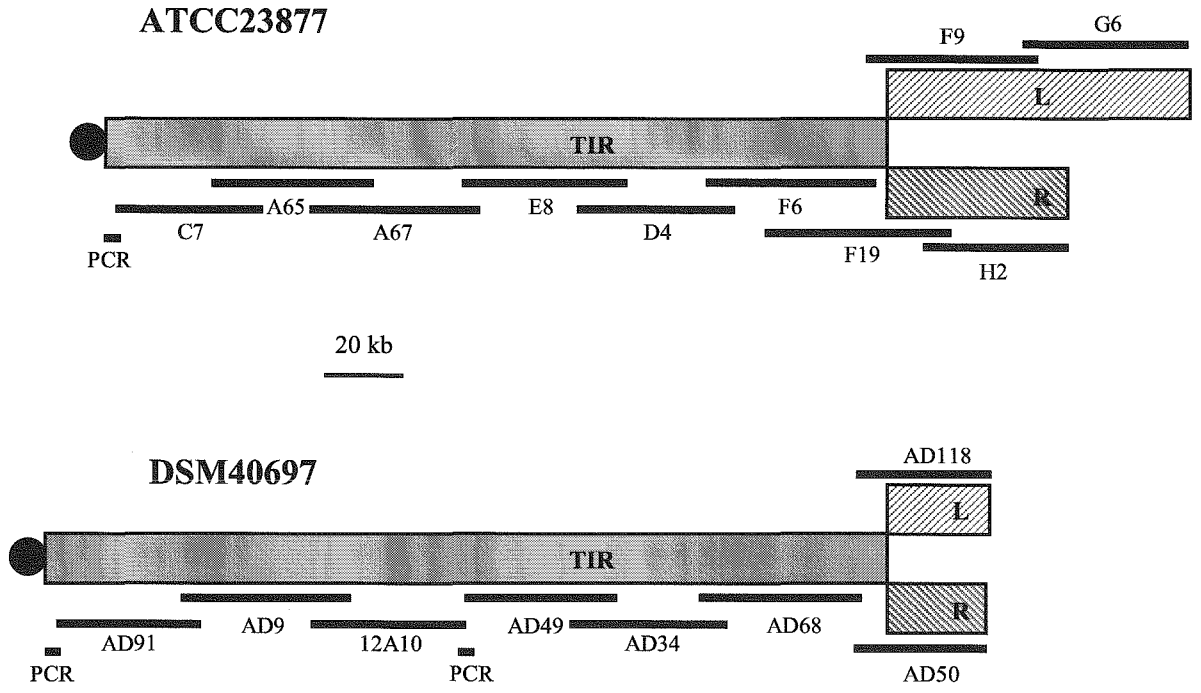


Figure 18 : Organisation des cosmides recouvrant les TIRs (et les séquences flanquantes) du chromosome des souches de *S. ambofaciens* ATCC23877 et DSM40697.
Les fragments d'ADN terminaux (non présents dans les banques génomiques) et la jonction entre les séquences portées par 12A10 et AD49 (non chevauchants) ont été obtenus par PCR. Les protéines terminales liées aux télomères sont représentées par les disques noirs.

f) Séquençage complet des régions terminales du chromosome (1544 kb et 1367 kb)

L'objectif était de séquencer complètement les régions instables du chromosome, c'est-à-dire les régions délétées chez les mutants spontanés. La plus grande délétion terminale observée est de 2335 kb chez le mutant NSA857 dérivant de la souche DSM40697 (Fischer *et al.*, 1997a). Une cartographie des chromosomes des deux souches de *S. ambofaciens* avait été réalisée par électrophorèse en champ pulsé (PFGE) des fragments de restrictions obtenus en utilisant les endonucléases *Ase*I et *Dra*I (Leblond *et al.*, 1996) (Fig. 17). Les régions délétables recouvrent l'intégralité du fragment *Ase*I D et une partie du fragment F sur le bras gauche. Sur le bras droit, elles englobent le fragment G et une partie du fragment E. Afin de recouvrir l'intégralité de ces régions chez la souche ATCC23877, les sites *Ase*I séparant les fragments de restriction F" de B" (bras gauche) et

séparant L" de H" (bras droit) ont été choisis comme limites internes au programme de séquençage (Fig. 17).

Le choix des BAC recouvrant les régions terminales du chromosome a été rendu possible grâce à la banque de cosmides ordonnés de la souche DSM40697. En effet, les hybridations de ceux-ci sur les profils de restriction *Ase*I du chromosome de la souche ATCC23877 ont permis de conclure que les chromosomes des deux souches possèdent une structure très similaire sur la majeure partie des régions terminales (Leblond *et al.*, 1996).
Au vu de la conservation des chromosomes des deux souches, certains cosmides ordonnés issus de la souche DSM40697 ont été utilisés afin d'identifier des BAC issus de la souche ATCC23877 dont les inserts recouvrent les régions terminales du chromosome (sur environ 2,9 Mb). La stratégie a consisté, tout d'abord, en un séquençage d'une extrémité des inserts clonés dans les cosmides ordonnés issus de la souche DSM40697 (Fig. 19). Par la suite, la connaissance de ces séquences a permis d'amplifier par PCR des fragments correspondants. Ces produits de PCR ont ensuite été marqués et utilisés comme sondes pour hybrider l'ADN des BAC extrait des 3072 clones d'*E. coli* composant la banque **B** puis déposé sur filtre.
Par cette approche, respectivement 5 et 10 BAC ont été sélectionnés pour la première étape de séquençage des régions terminales gauche et droite (Fig. 19). Ces 15 BAC candidats ont été appelés "points de nucléation" ("seed clones"). La stratégie mise en place visait à joindre les points de nucléation entre eux en "marchant" sur le chromosome. Connaissant les séquences des points de nucléation, il devenait ainsi possible, par comparaison avec les BES, de sélectionner des BAC flanquants les régions séquencées. Par itérations successives, la marche sur le chromosome devait permettre de relier entre eux les différents points de nucléation pour obtenir deux contigs correspondant aux deux régions terminales.

Au final, 88 BAC et 10 cosmides, dont l'ordre est représenté sur la Figure 20, ont été nécessaires pour séquencer complètement les régions instables du chromosome de *S. ambofaciens* ATCC23877. Les contigs ainsi obtenus pour les bras chromosomiques gauche et droit représentent respectivement 1.544.032 pb (numéro d'accession : AM238663) et 1.367.119 pb (numéro d'accession : AM238664).

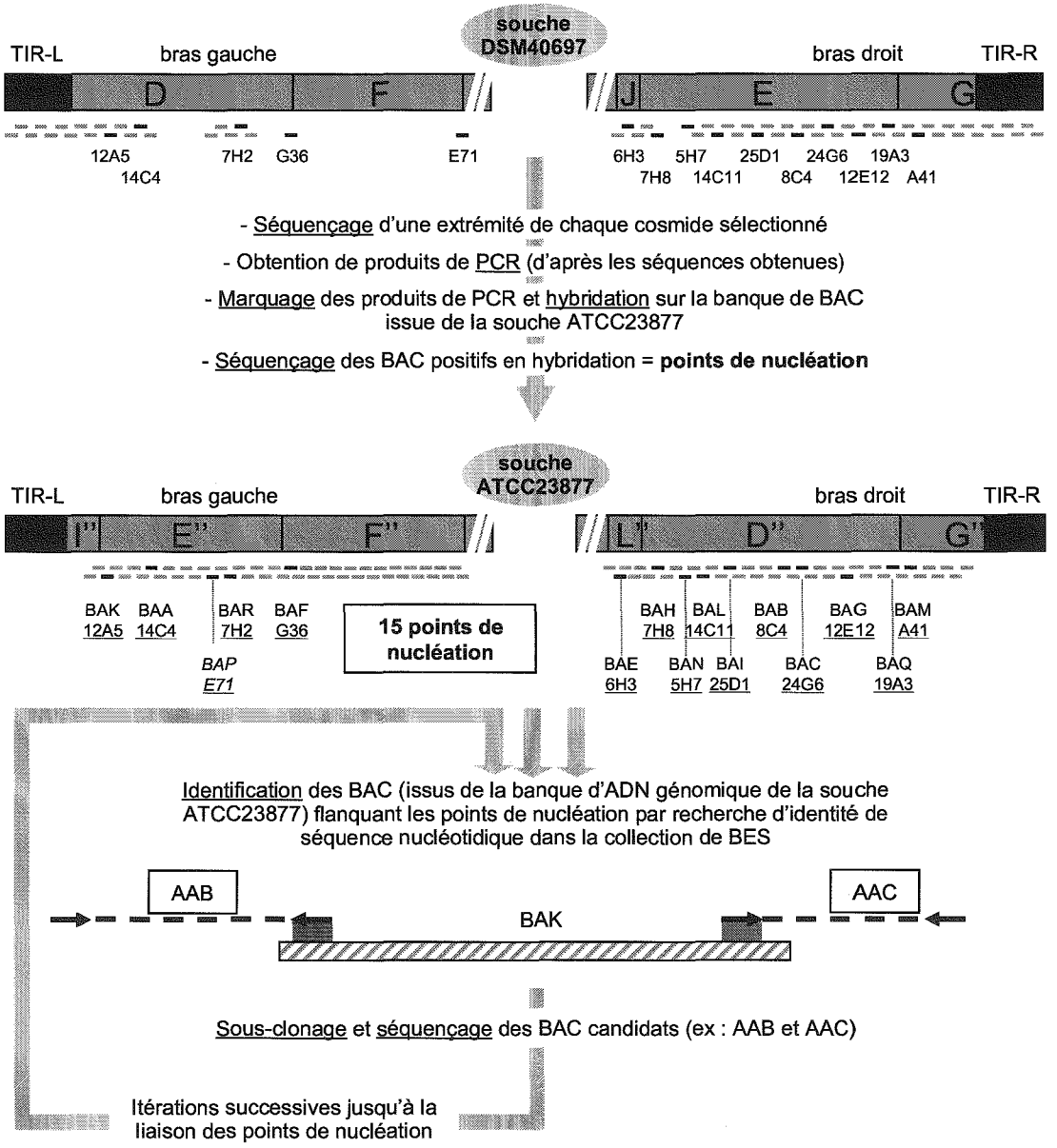


Figure 19 : Stratégie de séquençage des régions terminales du chromosome de *S. ambofaciens* ATCC23877. Quinze clones issus de la banque d'ADN génomique ordonnée de la souche DSM40697 (segments noirs) ont été utilisés afin d'identifier quinze "points de nucléation" parmi les clones de la banque d'ADN génomique de la souche ATCC23877. Leur séquence a servi de points d'ancrage pour le séquençage des régions terminales. La présence de BAC contenant des séquences chimériques issues du chromosome de la souche ATCC23877 a engendré une erreur dans le choix des points de nucléation : une extrémité de l'insert du cosmide E71 (DSM40697) hybride l'ADN de BAP (ATCC23877) dont l'insert est chimérique. Une stratégie de "marche" (étapes itératives) sur le chromosome de la souche ATCC23877 a permis de recouvrir l'intégralité des régions terminales, c'est-à-dire les fragments de restriction *Ase*I I", E", F" (bras gauche), G", D" et L" (bras droit).

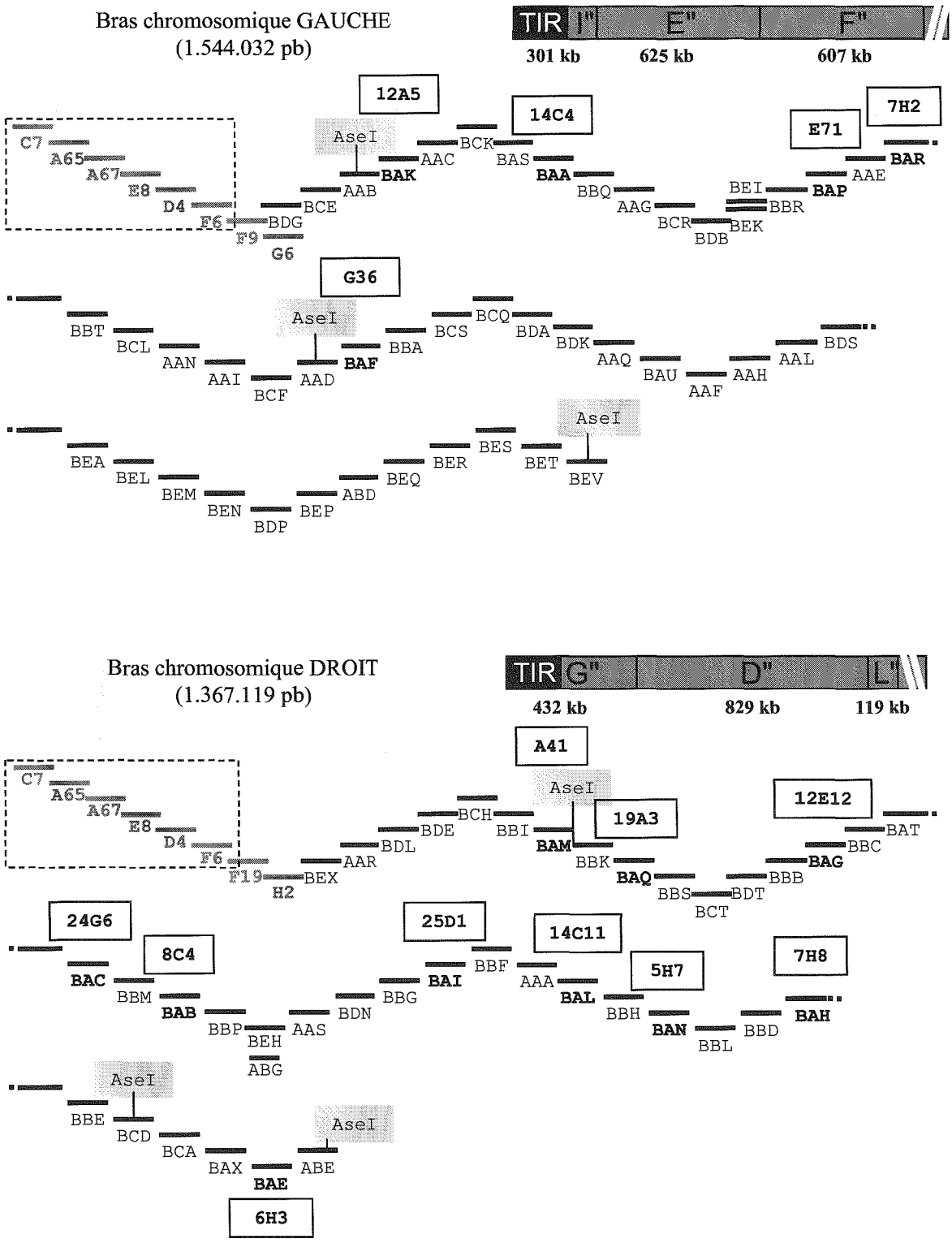


Figure 20 : Organisation des BAC et cosmides recouvrant les régions terminales du chromosome de *S. ambofaciens* ATCC23877.
Les cosmides recouvrant les TIRs (encadrées) sont représentés en gris. Les points de nucléation sont indiqués en gras et les cosmides (issus de la souche DSM40697) ayant permis leur mise en évidence sont encadrés.

2. *Problèmes rencontrés et solutions mises en œuvre*

Avec l'évolution des algorithmes d'assemblage, les génomes bactériens sont maintenant séquencés en une seule étape (stratégie appelée "shot-gun"). Séquencer un génome à partir de clones recombinants chevauchants, comme ce fut le cas pour *S. ambofaciens*, est beaucoup plus long, d'autant qu'aucune banque ordonnée n'était disponible pour *S. ambofaciens* ATCC23877. En effet, la stratégie de marche sur le chromosome est une démarche séquentielle où la connaissance de la séquence d'un BAC est un préalable au choix du BAC suivant.

a) Inserts chimériques

La présence en proportions non négligeables, dans les banques d'ADN génomique de *S. ambofaciens*, de BAC dont les inserts sont des séquences chimériques, s'est avérée un problème majeur. Une séquence d'insert chimérique résulte du clonage dans le vecteur BAC de plusieurs fragments d'ADN génomique non contigus sur le chromosome mais concaténés artificiellement au moment de l'étape de ligation. Au vu de la stratégie de marche sur le chromosome employée, chaque séquence chimérique obtenue a engendré des erreurs dans le choix des BAC flanquants. Parmi les 108 inserts séquencés au cours de ce projet, respectivement 19 et 89 proviennent des banques A et B. Les 18 inserts chimériques identifiés appartiennent tous à la banque B (20% des clones de cette banque).
Les difficultés d'assemblage des génomes peuvent être grandement minimisées lorsqu'un ou plusieurs génomes d'espèces (voire de souches) phylogénétiquement proches sont disponibles. Or, cette facilité n'était pas envisageable pour l'assemblage de *S. ambofaciens*. En outre, la justification principale du séquençage des régions terminales tient dans leur caractère spécifique.

La présence d'inserts chimériques a été systématiquement contrôlée *a posteriori*, au moment de l'assemblage des contigs, grâce à l'analyse des BES disponibles.

b) Défaut de points de nucléation pour le bras chromosomique gauche

Chez la souche DSM40697, la banque de cosmides ordonnés ne recouvre que partiellement le bras chromosomique gauche. Par conséquent, le nombre de points de nucléation disponibles pour ce bras était faible (5 clones). Par exemple, les clones de liaison G36 et E71 sont séparés par plus de 600 kb (Fig. 19). Enfin, le point de nucléation BAP, détecté par hybridation d'une extrémité d'E71, s'est révélé être un faux positif (Fig. 19). Par conséquent, aucun point de nucléation n'était disponible dans le fragment *Ase*I F".

c) Identification de BAC candidats pour le séquençage

Au printemps 2004, le séquençage du bras chromosomique droit était en cours de finition. En revanche, trois contigs étaient séquencés pour le bras chromosomique gauche de *S. ambofaciens* et les trous les séparant étaient estimés à environ 400 kb. La présence d'inserts chimériques et de séquences répétées empêchait de continuer la marche sur le chromosome.
Deux approches ont été envisagées pour combler ces espaces. La première est expérimentale et la seconde est une approche *in silico* (Fig. 21 et 22).

- Approche expérimentale

Afin d'identifier les BAC dont les inserts correspondent à des régions appartenant aux fragments de restriction *Ase*I E" et F", l'ADN de ces deux fragments isolés par électrophorèse en champs pulsés (PFGE) a été purifié puis marqué. Il a ensuite été utilisé comme sonde dirigée contre les ADN des 3072 clones composant la banque d'ADN génomique "B". Cette expérience a été réalisée par l'équipe de J. L. Pernodet (IGM Orsay).

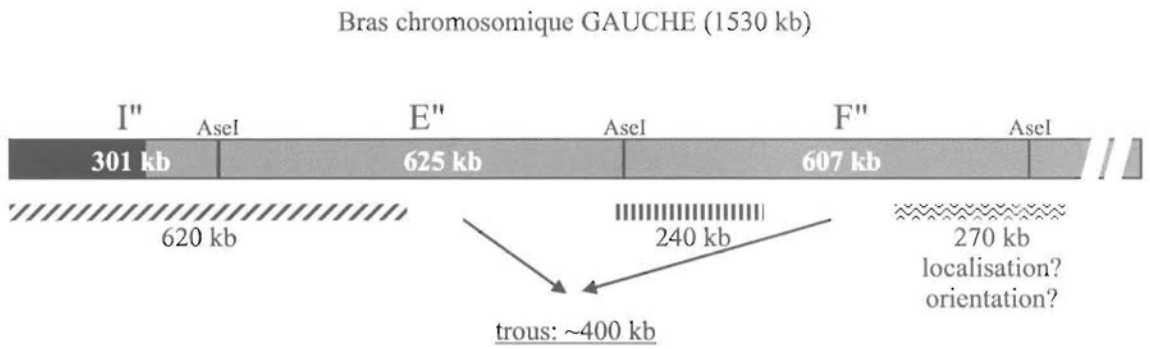
Cette sonde correspondait à 1,2 Mb d'ADN génomique, soit 15% du chromosome. Son utilisation lors d'une expérience d'hybridation s'est avérée très délicate étant donnée l'importance du bruit de fond (séquences répétées et hybridations aspécifiques) généré par une sonde de très grande taille (Fig. 21). Néanmoins, en considérant un seuil d'intensité minimale, il a été possible de sélectionner 445 clones. La proportion de clones positifs (15%) est en accord avec la taille de la sonde.

Parmi les 445 BAC positifs, la majorité possédait des extrémités correspondant à des séquences déjà disponibles et appartenant aux fragments de restriction E" et F", confirmant la validité des résultats d'hybridation. Après élimination de ces clones, 87 BAC candidats ont été retenus.

Dans l'objectif d'établir des groupes d'inserts chevauchants parmi les 87 candidats, les profils de digestion *Bam*HI ont été réalisés et comparés. De plus, des hybridations permettant de confirmer et de préciser les chevauchements prédits ont été entreprises (résultats non montrés).

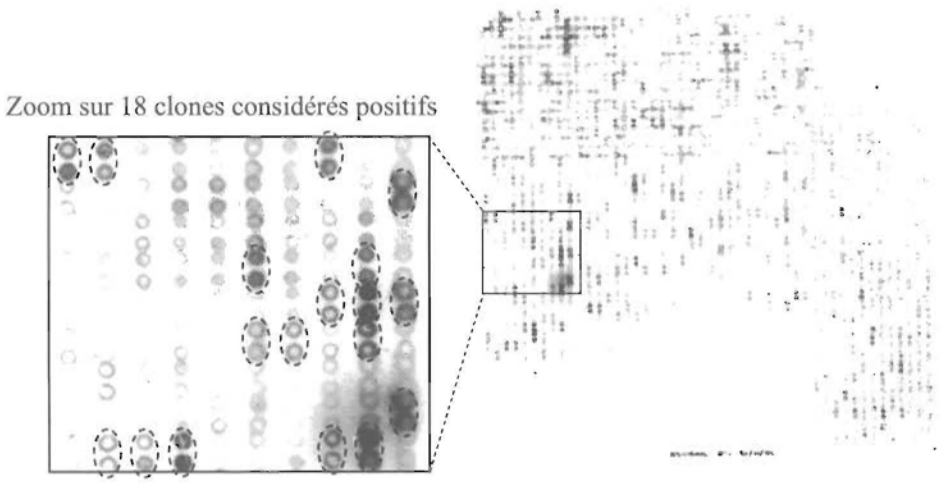
- Approche in silico

L'approche *in silico* est basée sur la comparaison des BES de *S. ambofaciens* avec le chromosome de *S. coelicolor* (voir paragraphe "Analyse des extrémités de BAC" ci-après). La région non séquencée du fragment *Ase*I F" montre, en effet, une organisation génétique conservée avec la région 1.200.000-1.600.000 du chromosome de *S. coelicolor* (Fig. 22). Cette conservation est détectable en alignant les BAC de *S. ambofaciens* sur le chromosome de *S. coelicolor* d'après leurs BES. Cette approche a permis d'identifier les onze derniers BAC permettant de terminer le programme de séquençage des régions terminales du chromosome de *S. ambofaciens*.



- Marquage des fragments AseI E" et F"

- Hybridation de la sonde E"/F" sur la banque B (3072 clones)



- Analyse visuelle des clones positifs

  445 clones montrant un signal positif en hybridation (15% des clones)

  87 clones dont les deux BES ne correspondent pas à des séquences déjà disponibles

- Profils de restriction *Bam*HI → regroupement des inserts potentiellement chevauchants

- Hybridations de BAC marqués sur les profils de digestions *Bam*HI des 87 clones candidats

Figure 21 : Détermination de BAC candidats à la complétion des trous (~400 kb) présents dans les fragments *Ase*I E" et F" par une approche expérimentale.
Les trois contigs alors totalement séquencés sont représentés par des rectangles hachurés. La localisation et l'orientation du contig de 270 kb n'étaient pas connues.

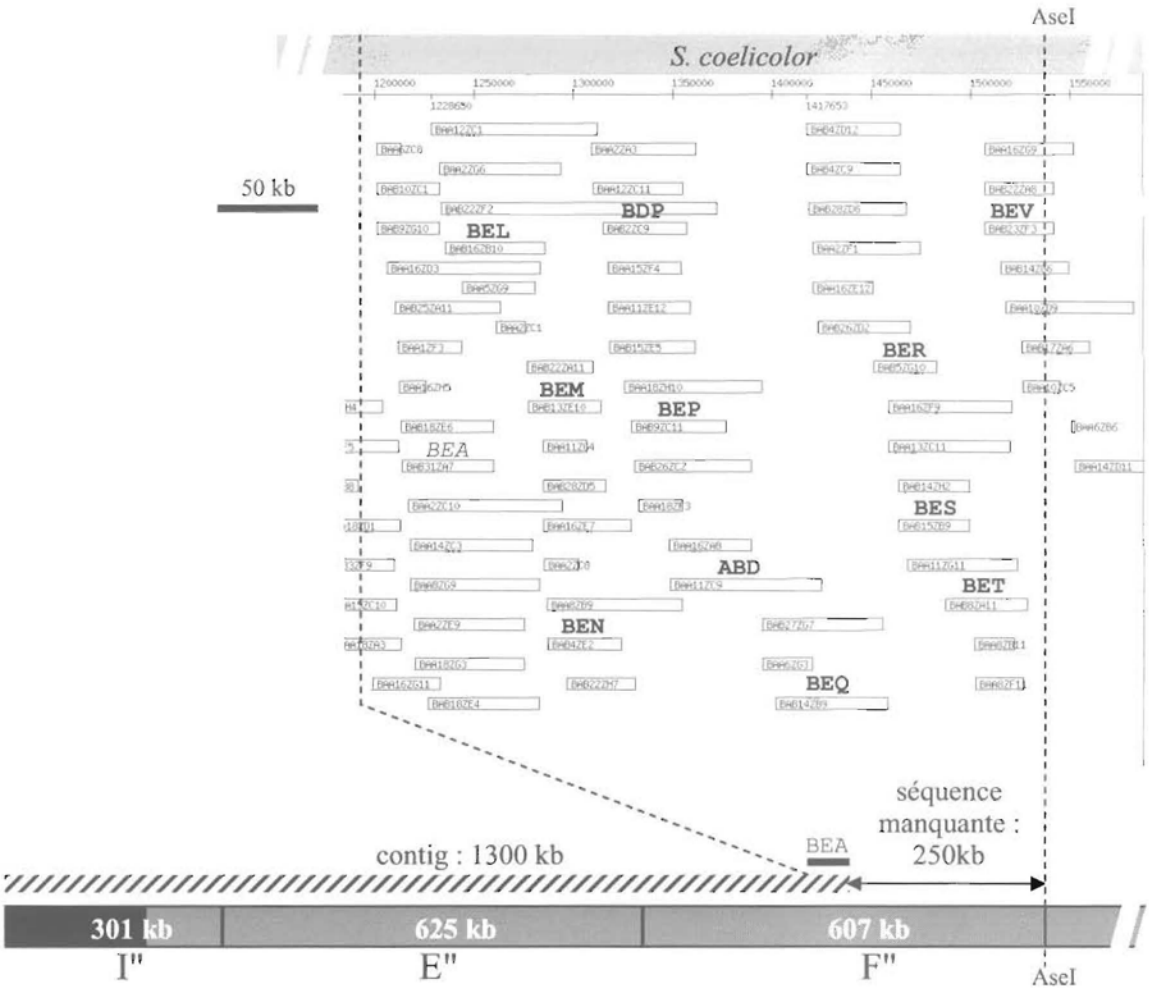


Figure 22 : Identification des BAC porteurs de régions appartenant au fragment *Ase*I F" du bras chromosomique gauche de *S. ambofaciens* par approche *in silico*.
Un contig d'environ 1300 kb se terminant par le clone BEA et recouvrant les fragments *Ase*I I", E" et une partie de F" était alors disponible. Les BES disponibles ont permis de générer un alignement des BAC de la banque génomique de *S. ambofaciens* en fonction des similarités retrouvées avec le chromosome de *S. coelicolor*. Ainsi, la région non séquencée du fragment *Ase*I F" présente une organisation génétique conservée avec la région 1.200.000-1550.000 du chromosome de *S. coelicolor*. Onze BAC, dont les noms sont indiqués sur l'alignement, ont ainsi été choisis pour être séquencés afin de recouvrir l'intégralité de la région manquante.

B. Approches bioinformatiques développées

Les programmes d'analyse se sont multipliés ces dernières années, accompagnant la génération exponentielle des données de séquences. Différentes plateformes d'annotation ont été développées pour les biologistes. Elles intègrent la plupart des outils nécessaires à l'annotation dans un logiciel qui articule les différents programmes habituels tels que BLAST (Altschul *et al.*, 1997). Cependant, ces plateformes ne permettent pas d'effectuer toutes les fonctionnalités spécifiques à chaque projet de séquençage.
Une suite de programmes adaptée à ce projet d'annotation et de génomique comparée a été développée. L'objectif était de maîtriser les paramètres d'analyse, les formats de données et, surtout, de ne pas être limité aux fonctionnalités présentes dans les plateformes existantes.

Travailler sur l'évolution des génomes implique trois étapes : annotation, comparaison et traitement des résultats. Les deux premières phases fournissent une quantité de données bien trop importante pour accéder aux informations interprétables sans traitement *a posteriori*. Une importance toute particulière a donc été accordée à la structuration des données en vue d'un traitement automatisé permettant leur visualisation (notamment la synténie) et leur interprétation en terme d'évolution.

1. Matériel et logiciels utilisés

Ce travail a été réalisé sur un serveur Compaq Proliant Intel® Xeon™ 2x2.20 GHz - mémoire : 3 Go tournant sous GNU/Linux Mandrake9.2.
Le langage Perl ("Practical Extraction and Report Language") a été choisi pour le développement des programmes gérant toutes les tâches de l'assemblage des contigs à la génomique comparative en passant par l'annotation. Il s'agit d'un logiciel libre qui présente de nombreux avantages dans le domaine de la bioinformatique et l'un des intérêts de ce langage est l'existence d'une quantité importante de modules disponibles et conçus spécifiquement pour la bioinformatique connus sous le nom de Bioperl (Stajich *et al.*, 2002). Il s'agit d'un projet communautaire en constant développement qui rassemble les méthodes utilisées fréquemment en bioinformatique et génomique dans un ensemble de modules Perl.

Dans la mesure où les analyses de séquences génèrent une quantité difficilement gérable de fichiers de résultats, les contraintes de stockage et d'accès aux données se sont posées. Dans le but de structurer les données de façon hiérarchisée, l'utilisation de bases de données relationnelles s'est imposée comme la solution la plus efficace.
PostgreSQL est un système de gestion de bases de données relationnelles libre et très performant. La version 7.3 de postgreSQL a permis la création de toutes les bases de données relatives à ce travail. Le module Perl::DBI a également été utilisé afin d'automatiser l'accès aux bases de données biologiques créées au cours de ce travail. Elles sont au nombre de trois :
- SAMDB qui contient l'intégralité des données issues de l'annotation des régions terminales de *S. ambofaciens* et celles issues de la génomique comparée des *Streptomyces*.

- EXTRBAC qui contient les données du séquençage des 8457 BES de *S. ambofaciens* et leur comparaison aux chromosomes de *Streptomyces*.
- GENOMECOMP qui contient les données provenant des analyses de comparaisons systématiques effectuées à partir des 312 génomes de bactéries et 12 génomes d'archées séquencés qui sont disponibles dans la banque EMBL (à la date du 10 mai 2006 ; Tableau annexe).

2. *Développement d'une plateforme d'annotation*

L'objectif du développement d'une plateforme d'annotation est de pouvoir automatiser au maximum la manipulation des données, depuis les chromatogrammes obtenus par électrophorèse des produits de réaction de séquençage jusqu'à l'obtention de données interprétables concernant l'annotation et la comparaison des génomes.
Les modules d'annotation (SAMANNOT) et de génomique comparée (SAMCOMP) s'articulent autour de SAMDB. Cette base de données permet le stockage et la mise en relation des données issues des analyses de séquences des *Streptomyces*. Elle permet notamment de faciliter considérablement le traitement des résultats et, par conséquent, leur interprétation en termes biologiques.
L'articulation des applications au sein des différents modules et l'architecture partielle de SAMDB sont schématisées dans les Figures 23 et 24.

a) Le module d'assemblage : SAMASSEMBLER

La suite logicielle phred/phrap/consed a été utilisée pour le traitement et l'assemblage des séquences issues des électrophorégrammes (Ewing et Green, 1998 ; Ewing *et al.*, 1998 ; Gordon *et al.*, 1998). L'assemblage de chaque insert de cosmides ou de BAC recombinant a été réalisé séparément. Les tentatives d'assemblage automatique avec Phrap (variante long_reads) de l'ensemble des différents inserts chevauchants ont échoué à cause de la présence d'inserts chimériques. SAMASSEMBLER regroupe plusieurs programmes dont *VecScreenBAC.pl* qui permet d'éliminer les traces de vecteur dans les séquences ainsi que le couple *OverlapDeterm.pl* et *Makecontig.pl* qui a été développé afin d'automatiser l'assemblage des bras chromosomiques de *S. ambofaciens*. Il utilise le programme BLAST2 (Altschul *et al.*, 1997) qui permet d'identifier les positions des régions de chevauchement entre les séquences des différents clones.
Afin d'éviter les erreurs issues de traitements manuels, toutes les manipulations des fichiers de séquences ont été automatisées à l'aide de scripts simples (*revcom.pl* : écriture du complément inverse, *subseq.pl* : extraction de sous séquences et *cat_fasta.pl* : concaténation de séquences).

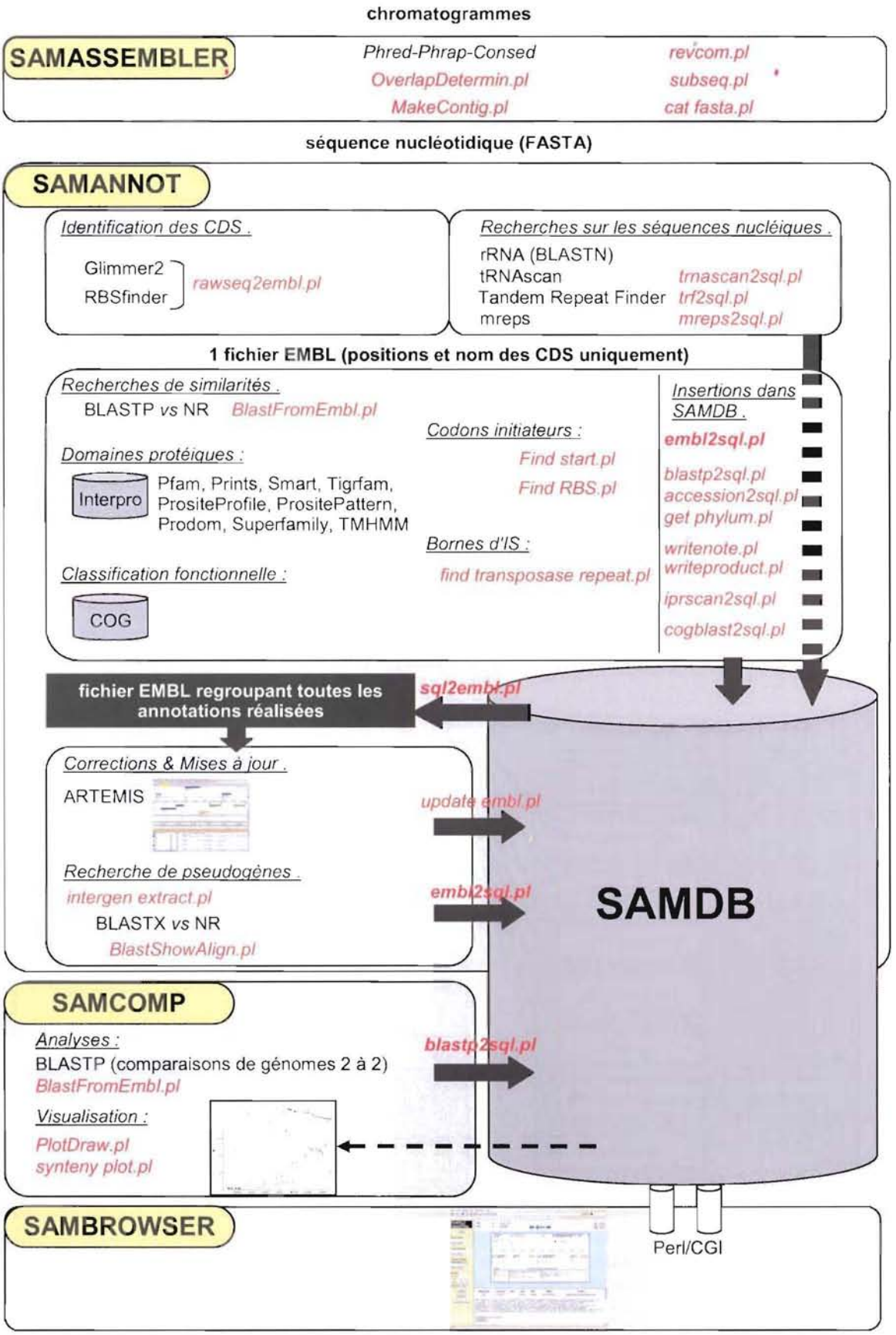


Figure 23 : Schéma de la plateforme dédiée à l'assemblage (SAMASSEMBLER), l'annotation (SAMANNOT) et la génomique comparée (SAMCOMP) du chromosome de *S. ambofaciens*. SAMBROWSER permet de parcourir l'annotation à travers une interface graphique et permet l'interrogation de SAMDB *via* Internet : **http://www.weblgm.scbiol.uhp-nancy.fr/ambofaciens/**

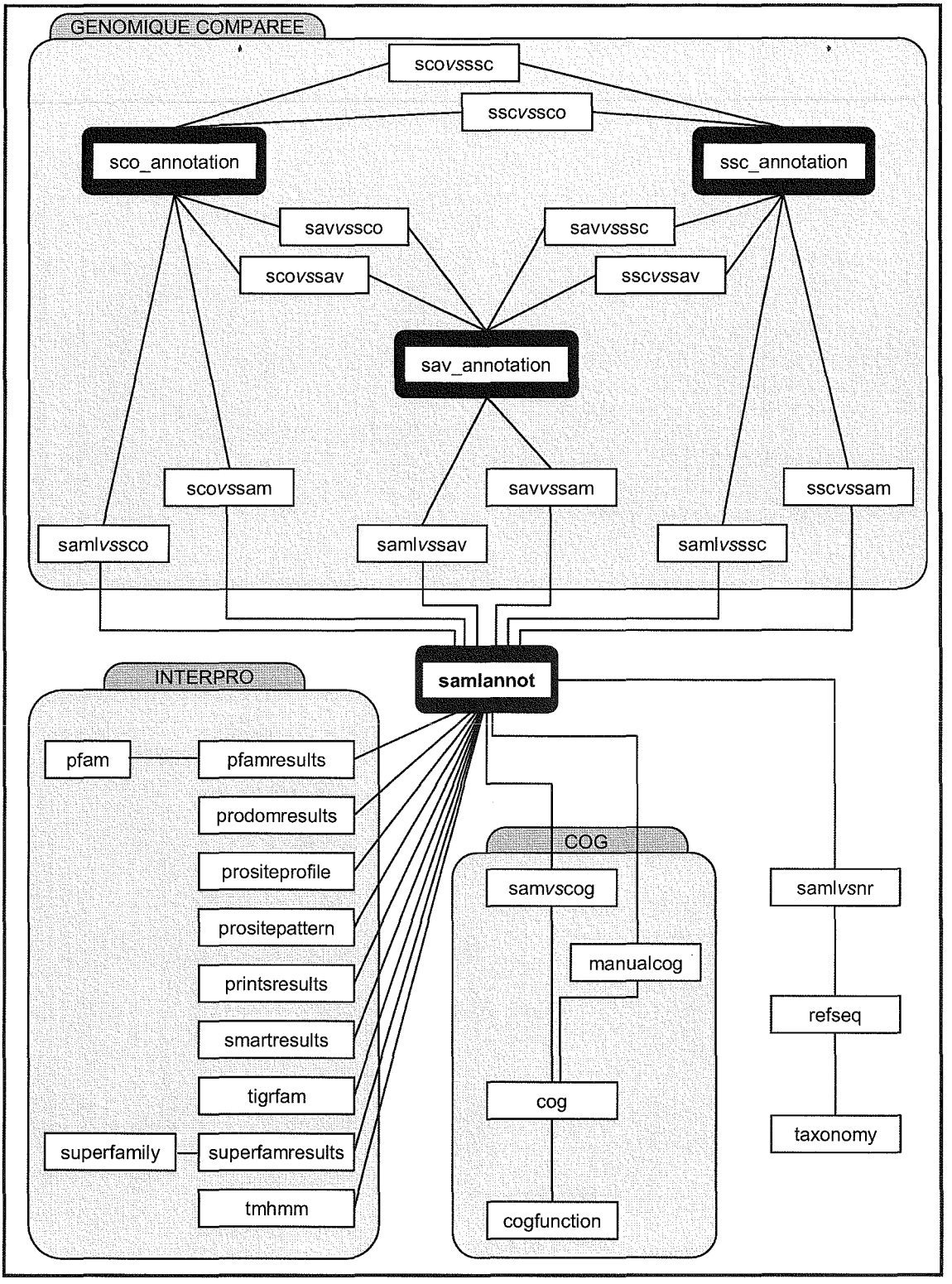


Figure 24 : Architecture partielle de la base de données SAMDB.
Deux tables contenant respectivement l'annotation des bras chromosomiques gauche (*samlannot*) et droit (*samrannot*) de *S. ambofaciens* ont été créées et présentent les mêmes relations dans la base de données, cependant, seule la première est représentée. Les tables d'annotation sont représentées en rouge. Elles sont mises en relation grâce aux tables de comparaison (x *vs* y). Par ailleurs, seules les relations avec les chromosomes des *Streptomyces* sont représentées sur cette figure. Les comparaisons réalisées avec d'autres Actinomycètes ont été stockées dans SAMDB avec le même modèle relationnel.

b) Le module d'annotation : SAMANNOT

La précision de l'annotation peut être plus ou moins poussée selon les annotateurs. Celle-ci peut se limiter à l'identification des positions de gènes potentiels sur une séquence et à la recherche d'homologies dans les bases de données. Elle peut aussi intégrer des recherches de motifs potentiels (ex : terminateurs, RBS) ou encore de séquences répétées. La fonction des gènes codant des protéines peut être prédite par recherche de similarité de séquences dans les banques (avec BLASTP). Celle-ci peut être précisée par l'identification de domaines protéiques. Les pseudogènes peuvent aussi être annotés. Enfin, la qualité et la quantité des commentaires attachés à chaque caractéristique des séquences sont également très variables selon les annotateurs.
La variabilité des annotations de génomes dans les banques publiques (EMBL, GenBank, DDBJ) est à l'image de la variabilité génomique ! Les génomes des *Streptomyces* constituent en cela un exemple frappant : l'annotation de *S. coelicolor* contient, en plus des gènes potentiels, 2576 RBS potentiels, 562 régions répétées, 195 tiges-boucles et 6671 autres caractéristiques diverses (des domaines protéiques, par exemple). En revanche, l'annotation du chromosome de *S. avermitilis* renseigne uniquement les positions des gènes potentiels.
Dans certains cas (ex : *M. tuberculosis* CDC1551 et *H. pylori* J99), ni les gènes codant les ARNr, ni ceux codant les ARNt ne sont renseignés.

Le développement du module SAMANNOT avait pour but d'automatiser la procédure d'annotation du génome de *S. ambofaciens*. Il est en fait utilisable pour tous les génomes procaryotes.

Dans le module SAMANNOT a été intégré un maximum de recherches dans le but de rendre l'annotation la plus riche possible (détaillé plus bas).
Un premier programme, *rawseq2embl.pl*, identifie les séquences codantes (CDS) en utilisant Glimmer2, un programme de prédiction de gènes basé sur des modèles de Markov (Delcher *et al.*, 1999). L'apprentissage de Glimmer2 a été réalisé avec 3000 ORF de *S. coelicolor*. Un seuil minimal classique de 120 nt a été fixé arbitrairement. Les résultats obtenus sont automatiquement redirigés vers RBSfinder dans le but de corriger les positions des codons d'initiation de la traduction en fonction de la présence ou non d'un RBS potentiel (Suzek *et al.*, 2001). Le programme *rawseq2embl.pl* formate les résultats sous forme d'un fichier EMBL contenant les positions corrigées des CDS potentielles et un nom paramétrable par l'utilisateur.

A partir du fichier EMBL, un second programme intégré dans SAMANNOT, *BlastFromEmbl.pl*, traduit les séquences nucléiques de chaque CDS en séquence protéique et recherche pour chacune d'elles des similarités de séquences à l'aide du programme d'alignement BLASTP (Altschul *et al.*, 1997). Tout d'abord, la requête est effectuée contre la banque généraliste NR (Non Redundant) du National Centre for Biotechnology Information (NCBI) qui regroupe toutes les séquences disponibles (3.658.078 séquences en juin 2006). Dans l'objectif des comparaisons par paires de génomes (module SAMCOMP), *BlastFromEmbl.pl* réalise en parallèle les recherches de similarités contre chacun des génomes de *Streptomyces* disponibles et ceux d'autres *Actinomycetes*.

- Validation des prédictions

Malgré l'efficacité des programmes Glimmer2 et RBSfinder, l'annotation automatique génère des erreurs de prédiction. La nécessité d'une étape "manuelle", c'est-à-dire d'une intervention de l'annotateur après le processus automatique, semble indispensable dans l'objectif d'obtenir une annotation la plus proche possible de la réalité biologique envisagée. La validation manuelle des CDS a été réalisée sous Artemis (Rutherford *et al.*, 2000).

Deux paramètres doivent être contrôlés : la validité de chaque CDS prédite en tant que gène potentiel et la position du codon d'initiation de la traduction. En effet, l'annotation du génome des *Streptomyces* est rendue difficile du fait de la pauvreté des séquences en codons stop (riches en A/T) dans les différentes phases de lecture. Cette caractéristique résulte du pourcentage très élevé des bases G+C. Ainsi, le nombre de phases ouvertes de lecture de taille compatible avec celle d'un gène codant une protéine est si grand qu'il n'est pas possible de s'en servir comme d'un critère de détection, contrairement aux génomes à faible pourcentage en bases G+C.

Cependant, un paramètre très utile pour l'annotation des génomes à haut pourcentage en G+C est le biais très prononcé à la troisième position des codons (GC3 ; Fig. 25) (Wright et Bibb, 1992). En effet, ce biais atteint 92% chez *S. coelicolor* (91% chez *S. avermitilis*). Pour les deux autres positions des triplets, ce pourcentage est de 73% et 52% chez *S. coelicolor* (1ère et 2ème base du triplet respectivement). Cette particularité permet souvent d'éliminer les erreurs de prédiction de CDS. Par ailleurs, le second critère utilisé afin de confirmer l'existence d'une CDS est l'existence d'homologues dans les banques de séquences.

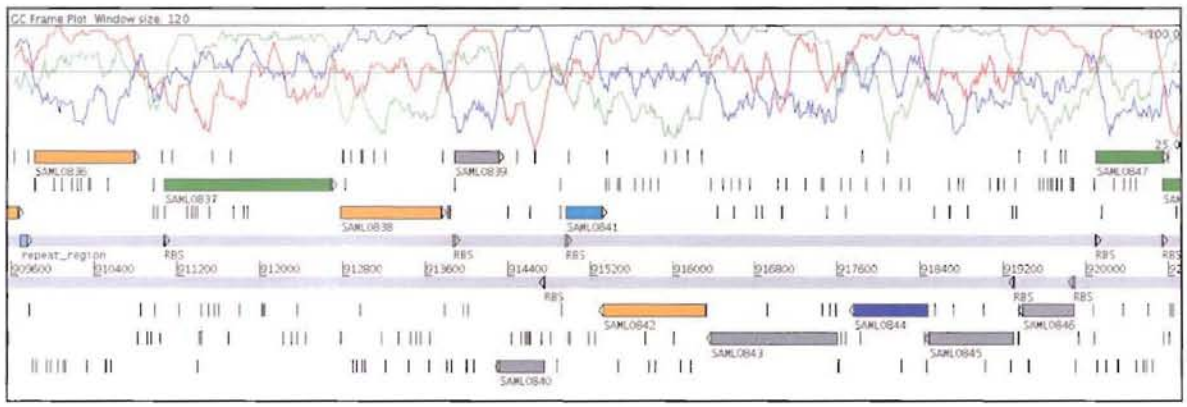


Figure 25 Profil de GC frameplot chez *S. ambofaciens* ATCC23877.
Le pourcentage en bases G et C est calculé dans les 3 phases de lecture dans une fenêtre glissante (120 nucléotides pour cet exemple), chaque phase étant représentée par une couleur distincte. Etant donné le fort biais en bases G et C (92%) à la troisième base des codons chez *S. ambofaciens*, ce signal est fréquemment utilisé pour prédire la présence et les limites des CDS dans les génomes à haut pourcentage en G+C.

- Choix du codon d'initiation de la traduction

La présence d'un RBS potentiel a été recherchée. Chez les *Streptomyces*, quatre codons d'initiation de la traduction sont connus et retrouvés à des fréquences diverses (pourcentages prédits chez *S. coelicolor*/*S. avermitilis* : ATG (62%/59%), GTG (35%/37%), TTG (3%/4%) et ATT (0%/0%)). Des études expérimentales ont permis de définir une séquence consensus pour les RBS chez *Streptomyces* : a/gGGAGG (Strohl, 1992). De plus, la séquence de l'ARN 16S de *S. ambofaciens* est

en accord avec ce consensus. Cependant, la présence d'une séquence RBS proche du consensus n'est pas systématique chez les *Streptomyces*, ce qui rend le choix du codon d'initiation de la traduction problématique. La présence d'un RBS détectable s'avère souvent trop incertaine pour en faire un critère de choix. Ainsi, les deux critères précédemment explicités (GC3 et présence de séquences homologues) ont été utilisés préférentiellement en vue de choisir le codon d'initiation de la traduction le plus probable.

SAMANNOT intègre un programme, *Find_start.pl*, permettant de suggérer la position du codon initiateur le plus probable en fonction des séquences homologues retrouvées dans la banque NR. Cette application parcourt les résultats de BLAST de chaque CDS afin de déceler des incohérences entre les prédictions et les séquences déjà annotées dans d'autres génomes. Si la partie N-terminale de l'une ou l'autre des séquences protéiques comparées par BLASTP ne s'aligne pas, le programme émet une réserve quant à la localisation du codon initiateur et indique sa position corrigée. L'utilisateur choisit ensuite en fonction de la correction proposée et en fonction du GC3 visualisé sous Artemis (profil de GC-Frameplot). Lorsque qu'aucun codon initiateur n'est privilégié, le codon le plus en amont de l'ORF a été attribué par défaut.

Après avoir effectué le choix des codons initiateurs de traduction de toutes les CDS prédites, le programme *Find_RBS.pl* permet de prédire la position d'un RBS potentiel en amont de chacune d'elles. Ce programme recherche la présence d'au moins 4 des 5 nucléotides GGAGG constituant le consensus retrouvé chez les *Streptomyces* entre 4 et 13 pb en amont du codon d'initiation de la traduction (Strohl, 1992).

- Recherche de domaines protéiques dans les séquences

De nombreuses banques de domaines protéiques sont disponibles et il n'existe en réalité aucun critère objectif pour en privilégier certaines dans la mesure où chacune d'elles possède ses propres caractéristiques. Afin de préciser au maximum la fonction des gènes prédits dans les régions terminales du chromosome de *S. ambofaciens*, toutes les banques de domaines protéiques regroupées dans le projet collaboratif INTERPRO (Apweiler *et al.*, 2001) ont été interrogées à l'aide d'InterProScan v3.3 (Zdobnov et Apweiler, 2001) :

- PROSITE patterns (Hofmann *et al.*, 1999)

  PROSITE regroupe les séquences d'un ensemble de familles de protéines et de domaines. En focalisant sur les régions conservées des familles ainsi définies, des signatures typiques de chacune d'elles ont pu être établies. ScanRegExp permet la recherche de motifs sous forme d'expressions régulières.

- PROSITE profiles (Hofmann *et al.*, 1999)

  Pour un certain nombre de familles de protéines et de domaines structuraux, les expressions régulières ne peuvent être utilisées étant donnée la divergence entre les séquences. Ainsi, pfscan permet une recherche contre PROSITE grâce à l'utilisation de matrices de poids.

- PRODOM (Corpet *et al.*, 1999)

  PRODOM regroupe des familles de domaines protéiques classifiées automatiquement par analyse des banques SWISS-PROT et TrEMBL. BlastProDom.pl permet la recherche de similarités avec ces familles en utilisant BLAST.

- PRINTS (Attwood *et al.*, 2000)

  PRINTS est une collection d'"empreintes" spécifiques de familles de protéines, chaque empreinte étant constituée de plusieurs motifs conservés. Le programme FingerPrintScan (Scordis *et al.*, 1999) a été utilisé afin de rechercher la présence de telles empreintes dans les séquences protéiques de *S. ambofaciens*.

- PFAM (Bateman *et al.*, 2000)

  PFAM est une collection d'alignements de séquences de familles de protéines et de domaines. Tous les alignements utilisent les séquences présentes dans SWISS-PROT et TrEMBL. Un modèle de Markov est généré pour chaque alignement et l'interrogation de nouvelles séquences se fait *via* HMMER2.2 (Eddy, 1998). Une partie de PFAM (PFAM_B) provient des familles de la base de données PRODOM.

- SMART (Schultz *et al.*, 2000)

  SMART est une banque dédiée aux domaines "mobiles", c'est-à-dire retrouvés en association avec d'autres dans des protéines contenant de multiples domaines. Elle fonctionne également *via* des modèles de Markov.

- TIGRFAM (Haft *et al.*, 2001)

  TIGRFAM est également une banque de données de domaines protéiques. Elle est focalisée sur une classification des protéines par groupes présentant une fonction identique (appelés "équivalogues") et non uniquement selon la parenté phylogénétique ("orthologues").

- SUPERFAMILY (Gough *et al.*, 2001)

  SUPERFAMILY est également une collection de modèles de Markov créés à partir de protéines de structures connues.

- TMHMM (Sonnhammer *et al.*, 1998)

  TMHMM est un programme de prédiction de domaines transmembranaires également basé sur des modèles de Markov.

- Recherche des gènes codant les ARNt et les ARNr

Le programme tRNAscan-SE a été utilisé afin de détecter les gènes codant les tRNA (Lowe et Eddy, 1997). Quant aux ARN ribosomiques, leurs positions ont été détectées par BLAST2 en utilisant les séquences d'ARNr déjà connues dans le genre *Streptomyces*.

- Recherche de motifs répétés dans les séquences nucléotidiques

Certains types de répétitions dans les séquences d'ADN ont également été recherchés à l'aide de mreps (Kolpakov *et al.*, 2003). Le programme *find_transposase_repeat.pl* (qui implémente BLAST2) intégré dans SAMANNOT, permet de rechercher automatiquement la présence de répétitions inversées aux bornes des séquences d'insertion (IS) potentielles.

- Ecriture d'une note et d'une fonction affectée à chaque CDS

Après la prédiction des positions des gènes, les recherches de similarités et les recherches de domaines protéiques, se pose le problème de la lisibilité de ces annotations. Deux types d'annotation sont prévus dans les fichiers EMBL. Le premier ("/product") vise à préciser la fonction potentielle du produit de gène et le second ("/note") permet d'écrire des commentaires plus détaillés concernant l'annotation de

ce gène. Deux scripts, *writeproduct.pl* et *writenote.pl*, ont été développés dans l'objectif d'automatiser l'écriture de ces annotations. Ces deux programmes fonctionnent à partir des résultats de recherche de similarités effectués par BLASTP contre la banque NR pour chaque CDS. Le script *writenote.pl* extrait les informations utiles (ex : numéro d'accession, Evalue) de chaque séquence (au maximum 5) montrant une similarité significative (Evalue inférieure à 10^{-3} pour les protéines de plus de 100 codons, 10^{-2} pour les autres) avec la protéine prédite.
Afin d'écrire la fonction potentielle de chaque CDS ("/product"), *writeproduct.pl* n'utilise pas, quant à lui, les descriptions des séquences contenues dans les résultats de BLASTP du fait de leur complexité. Il enregistre les numéros d'accession des séquences homologues en vue d'interroger la banque de données RefSeq du NCBI, *via* Internet, et de récupérer leur fonction prédite.

- Classification en catégories fonctionnelles

La base de données COG (Cluster of Orthologous Groups, (Tatusov *et al.*, 2001)) a été utilisée afin d'affecter une catégorie fonctionnelle à chaque protéine. COG définit 18 catégories regroupées en 4 grandes familles de fonctions.
Elle compte 77114 séquences protéiques regroupées en 3307 COG, eux mêmes répartis dans les 18 catégories fonctionnelles.
Ces séquences ont été formatées en base de données BLAST afin de réaliser une recherche de similarités de chaque protéine prédite chez *S. ambofaciens*. Le script *cogblast2sql.pl* permet un traitement automatique de ces résultats. Il parcourt les fichiers BLASTP et affecte à chaque protéine de *S. ambofaciens* la catégorie fonctionnelle à laquelle appartient l'homologue le plus proche (avec plus de 30% d'identité recouvrant au moins 70% de la taille de la protéine).
Un total de 1317 protéines (sur 2532 produits de gènes prédits, soit 52%) a pu être ainsi automatiquement classé dans une catégorie fonctionnelle. Pour les protéines ne présentant pas de similarité significative avec une séquence de COG mais pour lesquelles une fonction potentielle est identifiable par des recherches de similarités dans la banque NR ou par la présence de domaines protéiques, une catégorie a été assignée manuellement.

- Insertion des données dans SAMDB

SAMANNOT intègre les fonctionnalités nécessaires à l'insertion des résultats des programmes de recherche décrits précédemment dans la base de données relationnelle appelée SAMDB.

Les annotations contenues dans un fichier EMBL sont automatiquement extraites à l'aide de l'application *embl2sql.pl*. Ce programme parcourt les fichiers EMBL et écrit le code SQL nécessaire à l'insertion dans SAMDB des informations relatives aux différentes caractéristiques annotées ("Features"). En parallèle, ce programme génère le code SQL permettant la création d'une table correspondante.
Deux tables nommées *samlannot* et *samrannot* (Fig. 24) ont donc été créées pour contenir les annotations des contigs obtenus pour les régions terminales des bras chromosomiques gauche et droit de *S. ambofaciens*.

Les résultats de BLASTP contre la banque NR ont été analysés afin d'en extraire les informations pertinentes pour ensuite les insérer dans SAMDB (tables *samlvsnr* et *samrvsnr*). Le programme *blastpvsnr2sql.pl* a été développé dans cet objectif. Les valeurs correspondant au score, à l'Evalue et au pourcentage d'identité sont stockées pour le premier hit du BLASTP uniquement. De plus, la liste des numéros d'accessions des 10 séquences présentant la similarité la plus forte avec chaque protéine de *S. ambofaciens* est également créée et intégrée dans SAMDB.
A partir des alignements générés par BLASTP, une homologie a été considérée comme significative dès lors que le pourcentage d'identité entre la séquence d'entrée et une séquence de la banque est supérieur ou égal à 30% et que la longueur de l'alignement représente au minimum 80% de la taille de la protéine d'entrée.

Par ailleurs, les descriptions des séquences de la banque NR ne mentionnent pas, dans la plupart des cas, l'origine chromosomique ou plasmidique du gène séquencé. Cette information était nécessaire au vu des intégrations d'ADN plasmidique suspectées dans les extrémités chromosomiques. Le programme *accession2sql.pl* a été développé pour interroger, *via* Internet, la base de données RefSeq (NCBI Reference Sequences) et permettre ainsi d'intégrer dans SAMDB (table refseq) l'origine plasmidique ou chromosomique d'un gène (et d'autres informations complémentaires). Par ailleurs, la classification phylogénétique des organismes présentant des séquences homologues à celles de *S. ambofaciens* a été intégrée dans la table *taxonomy* de SAMDB à l'aide de l'application *get_phylum.pl*.

Les résultats issus des recherches de domaines protéiques (InterPro) et ceux correspondant à la classification fonctionnelle (COG) ont eux aussi été stockés automatiquement dans SAMDB à l'aide des programmes *iprscan2sql.pl* (tables *prositepattern*, *prositeprofile*, *prodomresults*, *printsresults*, *pfamresults*, *tigrfamresults*, *smartresults*, *superfamresults* et *tmhmm*) et *cogblast2sql.pl* (classification automatique avec COG : *samvscog*, classification manuelle : *manualcog*).
Enfin, les résultats des recherches d'ARN non codant (ARNt et ARNr) et de motifs répétés ont été stockés de façon automatique dans SAMDB.

- Recherche de gènes non prédits et de pseudogènes dans les régions intergéniques

Une procédure automatisée a été mise en place afin d'extraire les séquences intergéniques d'un génome donné et d'y rechercher des traces de gènes, correspondant parfois à des gènes fonctionnels non prédits par Glimmer2 et parfois à des pseudogènes.
L'extraction des séquences intergéniques a été réalisée avec *intergen_extract.pl* qui effectue des requêtes sur SAMDB en vue de capturer les positions de début et fin des gènes prédits. Seules les régions intergéniques de plus de 20 nt ont été analysées.
Les séquences codantes potentielles ont été recherchées par BLASTX (traduction dans les six phases de lecture pour chaque région intergénique) contre la banque NR. A partir de ces résultats, le script *BlastShowAlign.pl* a été développé dans le but de détecter automatiquement les similarités significatives (plus de 40% d'identité) et de générer un fichier EMBL contenant les gènes/pseudogènes nouvellement prédits. Le passage par un fichier EMBL permet la visualisation (avec Artemis) des

prédictions et ainsi leur validation manuelle. Les résultats retenus ont ensuite été stockés dans SAMDB (*embl2sql.pl*).

- Corrections et mises à jour et des annotations

Toutes les informations concernant l'annotation sont stockées dans SAMDB mais une représentation des données sous forme graphique est indispensable afin de détecter des erreurs éventuelles et faire évoluer les annotations. C'est dans cet objectif qu'a été développé le script *sql2embl.pl*. Il permet de générer un fichier EMBL, à partir des données stockées dans SAMDB, contenant toutes les caractéristiques annotées d'un génome ou d'une partie d'un génome (un "cluster" par exemple). Ce fichier EMBL temporaire est éditable sous Artemis.
Les changements apportés manuellement au fichier EMBL (ex : changements dans les positions de début et fin de certains gènes) peuvent ensuite être transmis à SAMDB *via* le script *update_anno.pl.* qui permet la mise à jour de SAMDB.

Par ailleurs, la banque NR intègre de nouvelles séquences chaque mois. Une procédure de téléchargement automatique de cette banque a été mise en place afin de réaliser les recherches de similarités sur les données les plus récentes. Le programme *update_blastpvsnr.pl* permet la mise à jour de SAMDB (tables *samlvsnr* et *samrvsnr*) en fonction des résultats de BLASTP contre la nouvelle banque NR.

c) Le module dédié à la génomique comparée : SAMCOMP

Les comparaisons de paires de génomes requièrent un module d'analyse spécifique. Il intègre les fonctionnalités permettant la réalisation des comparaisons de génomes deux à deux, la création des tables dédiées aux comparaisons, l'intégration et la mise en relation des résultats dans SAMDB. De plus, ce module regroupe les programmes capables d'automatiser la visualisation des comparaisons sous forme graphique.
Le script *BlastFromEmbl.pl* permet de réaliser une comparaison de chaque gène prédit dans un fichier EMBL contre chaque génome d'intérêt de façon individuelle (ceux des *Streptomyces*, d'autres Actinomycètes...). Les résultats obtenus ont été analysés par *blastp2sql.pl* qui permet la création des tables de comparaison (par exemple : samvssco pour la comparaison *S. ambofaciens*/*S. coelicolor*) et l'insertion des informations pertinentes dans ces différentes tables. Les comparaisons réciproques ont systématiquement été effectuées bien que le génome de *S. ambofaciens* ne soit que partiellement séquencé.
Les relations ainsi établies entre les tables dédiées aux annotations et celles dédiées aux comparaisons définissent le schéma relationnel de SAMDB (Fig. 24).

Deux séquences protéiques ont été considérées comme orthologues quand elles présentent réciproquement les meilleurs scores d'alignement (BLASTP) au sein des deux génomes comparés. De plus, elles doivent partager plus de 30% d'identité sur plus de 80% de leur longueur respective (alignée comme une seule HSP (High Scoring Pair)). Ainsi, en effectuant une requête sur la base de données en respectant ces critères, il est possible d'identifier les orthologues potentiels entre tous les couples de génomes comparés.

- Dot-plot

La visualisation graphique des comparaisons peut s'envisager sous plusieurs formes. L'une des méthodes les plus couramment utilisées est le "dot-plot" qui permet de représenter les positions relatives des gènes homologues entre deux réplicons sur un graphe à deux dimensions.

Le programme *PlotDraw.pl* a été développé dans cet objectif. Grâce à la structuration des données de comparaisons dans SAMDB, ce programme réalise des requêtes afin de détecter les positions relatives des orthologues sur les deux réplicons comparés. Ce script écrit les résultats correspondants sous forme d'un tableau de coordonnées. Chaque point identifie les coordonnées x (position sur le génome A) et y (position sur le génome B) d'un couple d'orthologues. Enfin, ce script écrit les commandes exécutables ensuite par le logiciel Gnuplot (http://www.gnuplot.info/) dédié au traçage du nuage de points (Fig. 26A).

- Conservation de l'Ordre des Gènes (GOC) : une mesure du niveau de synténie

La mesure du GOC a été utilisée dans l'objectif de décrire la stabilité des génomes bactériens (Rocha, 2006). Comme un dot plot, le calcul du GOC implique la comparaison de paires de génomes. Cet indice définit la fréquence avec laquelle deux gènes contigus dans un génome possèdent leur orthologue contigus dans le second génome.

Par l'utilisation d'une fenêtre glissante, la mesure du GOC permet de suivre le niveau de conservation de l'ordre et du contenu en gènes le long d'un chromosome. Le script *synteny_plot.pl* a été développé afin de suivre l'évolution du GOC à partir des données contenues dans les tables de comparaison de SAMDB. Sur le même principe que le script précédent, *synteny_plot.pl* écrit les mesures du GOC sous forme d'un tableau de valeurs et le tracé est réalisé par Gnuplot (Fig. 26B).

Deux formules de calcul du GOC ont été utilisées au cours de ce travail.

$$GOC_1 = \frac{N_{ortho\_contigus}}{N} \qquad GOC_2 = \frac{N_{ortho\_contigus}}{N_{ortho}}$$

N représente le nombre total de CDS dans la fenêtre glissante.

$N_{ortho\_contigus}$ représente le nombre de paires de gènes qui sont orthologues et contigus (adjacents) dans les deux génomes comparés (dans la fenêtre glissante).

N_{ortho} représente le nombre de gènes (dans la fenêtre glissante) ayant un orthologue dans le génome comparé (si $N_{ortho}=0 \Rightarrow GOC_2=0$).

- Visualisation de la dégénérescence de la synténie

Le logiciel ACT (Artemis Comparison Tool) permet une visualisation graphique des régions homologues entre deux séquences nucléotidiques. ACT ne réalise pas lui-même les comparaisons de séquences mais il affiche graphiquement les résultats issus d'une comparaison réalisée par BLASTN.

Cependant, la comparaison des génomes au niveau nucléotidique n'a d'intérêt que pour des organismes très proches (ex : des souches d'une même espèce). Elle s'applique mal à la recherche de synténie entre les génomes des *Streptomyces*, notamment les espèces éloignées.

Ce logiciel de visualisation a donc été détourné de son utilisation initiale pour permettre d'afficher des résultats de comparaisons des séquences protéiques (Fig. 26C). A l'aide d'une requête sur SAMDB, il est facile de connaître les positions relatives des gènes orthologues (sur la base de la comparaison des séquences protéiques) sur deux génomes comparés. Les positions et les niveaux d'identité entre orthologues ont donc été automatiquement extraits de SAMDB afin d'écrire ces résultats sous un format reconnu par le logiciel ACT.
Cette fonctionnalité du module de génomique comparée permet de préciser la nature et le nombre d'événements mutationnels (réarrangements, insertions, délétions, remplacements de gènes) qui font varier le niveau de GOC. Les régions de synténie dégénérée ont pu être mises en évidence grâce à cette représentation.
Enfin, ACT a été utilisé pour visualiser les comparaisons des séquences nucléotidiques très proches des TIR des deux souches de *S. ambofaciens* étudiées au cours de ce travail.

d) Le serveur Internet de SAMDB : SAMBROWSER

SAMDB est consultable à l'adresse suivante :

http://www.weblgm.scbiol.uhp-nancy.fr/ambofaciens/

Une interface graphique HTML fonctionnant sur un serveur HTTP (Apache) a été développée afin de rendre l'annotation des régions terminales du chromosome de *S. ambofaciens* accessible pour la communauté scientifique *via* Internet. Les Figures 27-28 montrent des captures d'écran de l'interface de SAMBROWSER. Elle se décompose en trois cadres :
- le cadre "REQUEST" permet d'interroger SAMDB par fonction ou nom de gène. De plus, une fonctionnalité permet de rechercher les homologues potentiels chez *S. ambofaciens* de gènes prédits chez les autres génomes de *Streptomyces* séquencés et annotés.

- le cadre "IMAGE" affiche les caractéristiques prédites dans les séquences sous forme d'une image interactive. Toutes ces caractéristiques, et notamment les gènes, sont cliquables afin d'obtenir les renseignements collectés, qui s'affichent alors dans le cadre "ANNO".

- le cadre "ANNO" permet l'affichage de l'annotation de chaque caractéristique : sa position sur le génome, ses séquence nucléotidique et protéique, sa description et sa fonction putative. De plus, les résultats concernant les domaines protéiques détectés et la classification par catégorie fonctionnelle sont également affichés. L'alignement réalisé par BLASTP contre la banque NR est consultable par ailleurs. Enfin, pour chaque gène prédit, les homologues dans les génomes de *S. coelicolor* et de *S. avermitilis* sont indiqués de même que les 10 séquences de Genbank présentant le plus de similarité avec la protéine d'intérêt. Afin d'améliorer l'interactivité de l'interface, des liens vers les sites hébergeant l'annotation des génomes des *Streptomyces* et vers le site du NCBI hébergeant Genbank sont créés automatiquement pour les résultats de comparaisons de génomes.

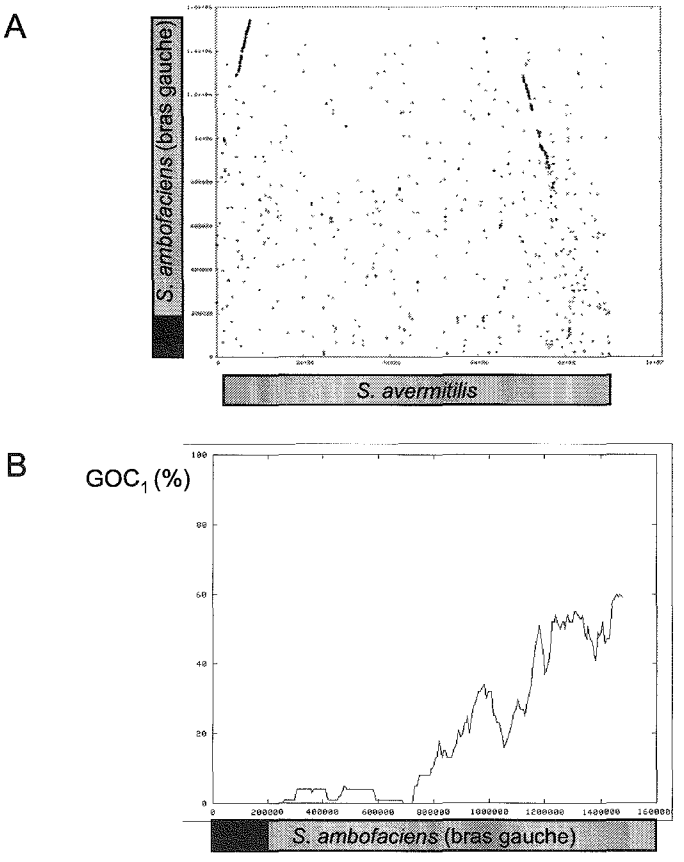


Figure 26 : Trois modes de visualisation graphique développés au cours de ce travail et dédiés aux comparaisons de génomes.

A. Comparaison du bras chromosomique gauche de *S. ambofaciens* avec le chromosome de *S. avermitilis* ; représentation sous forme de dot-plot, chaque point représentant les coordonnées d'un couple de gènes orthologues.

B. Profil de GOC_1 du bras gauche de *S. ambofaciens* comparé au chromosome de *S. avermitilis*. La TIR gauche du chromosome de *S. ambofaciens* est représentée par un rectangle noir.

C. Exemple de visualisation de la synténie entre *S. ambofaciens*, *S. coelicolor* et *S. avermitilis*. Le logiciel ACT a été détourné de son utilisation habituelle pour visualiser des comparaisons au niveau protéique.

Le programme BLAST a également été implémenté sur ce serveur pour permettre aux utilisateurs d'effectuer des recherches de similarité dans les séquences nucléotidiques et protéiques de *S. ambofaciens* (Fig. 27C). Enfin, SAMBROWSER intègre la recherche de motifs au sein des séquences nucléotidiques et protéiques de *S. ambofaciens* (Fig. 27D). Cette application utilise *scan_for_matches* développé par Ross Overbeek (Dsouza *et al.*, 1997).

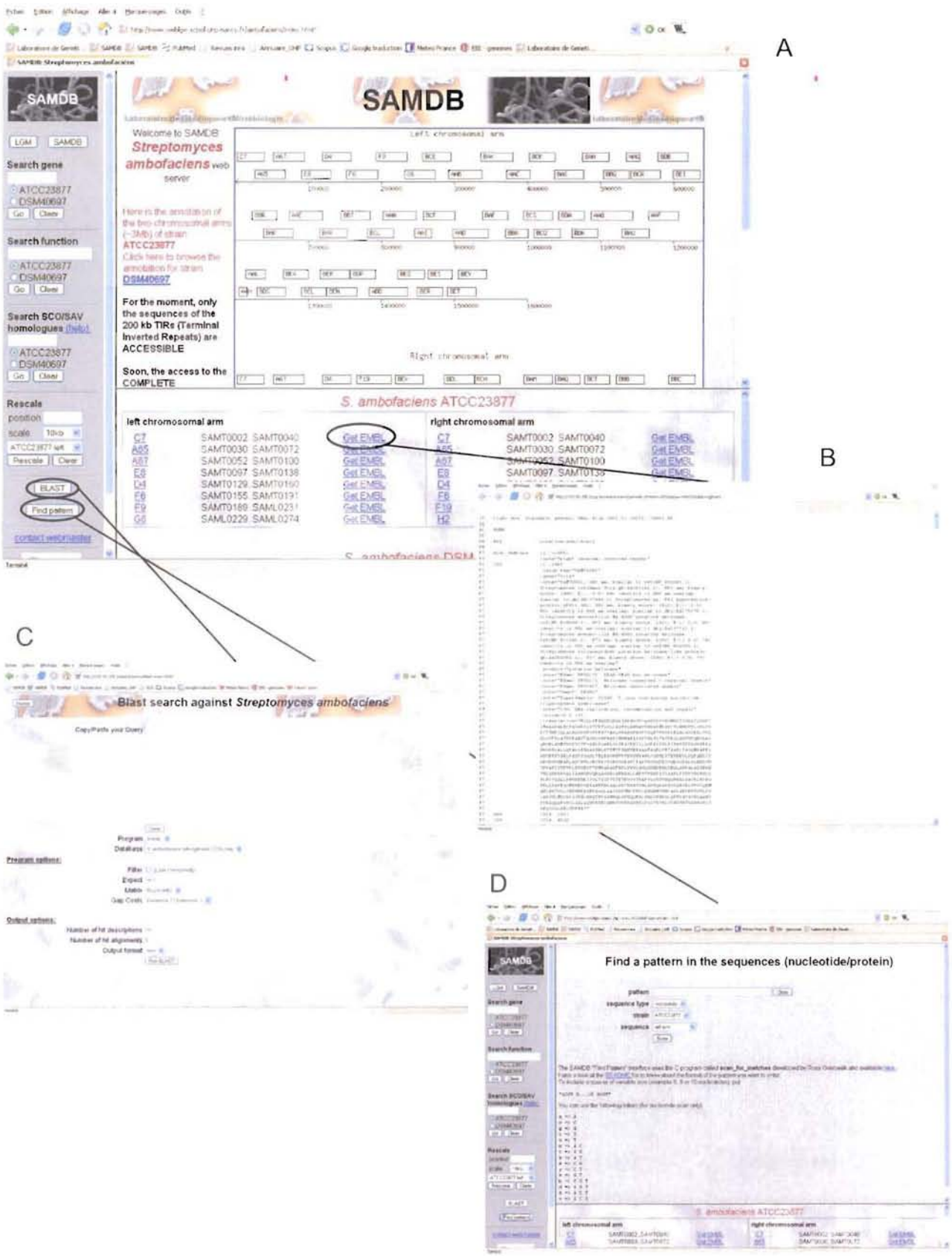


Figure 27 : Interface HTML de SAMBROWSER.
A. Page d'accueil. **B**. Extraction de l'annotation du cosmide C7 au format EMBL. **C**. Recherches de similarités avec BLAST contre les séquences de *S. ambofaciens*. **D**. Recherches de motifs avec *scan_for_matches* dans les séquences de *S. ambofaciens*.

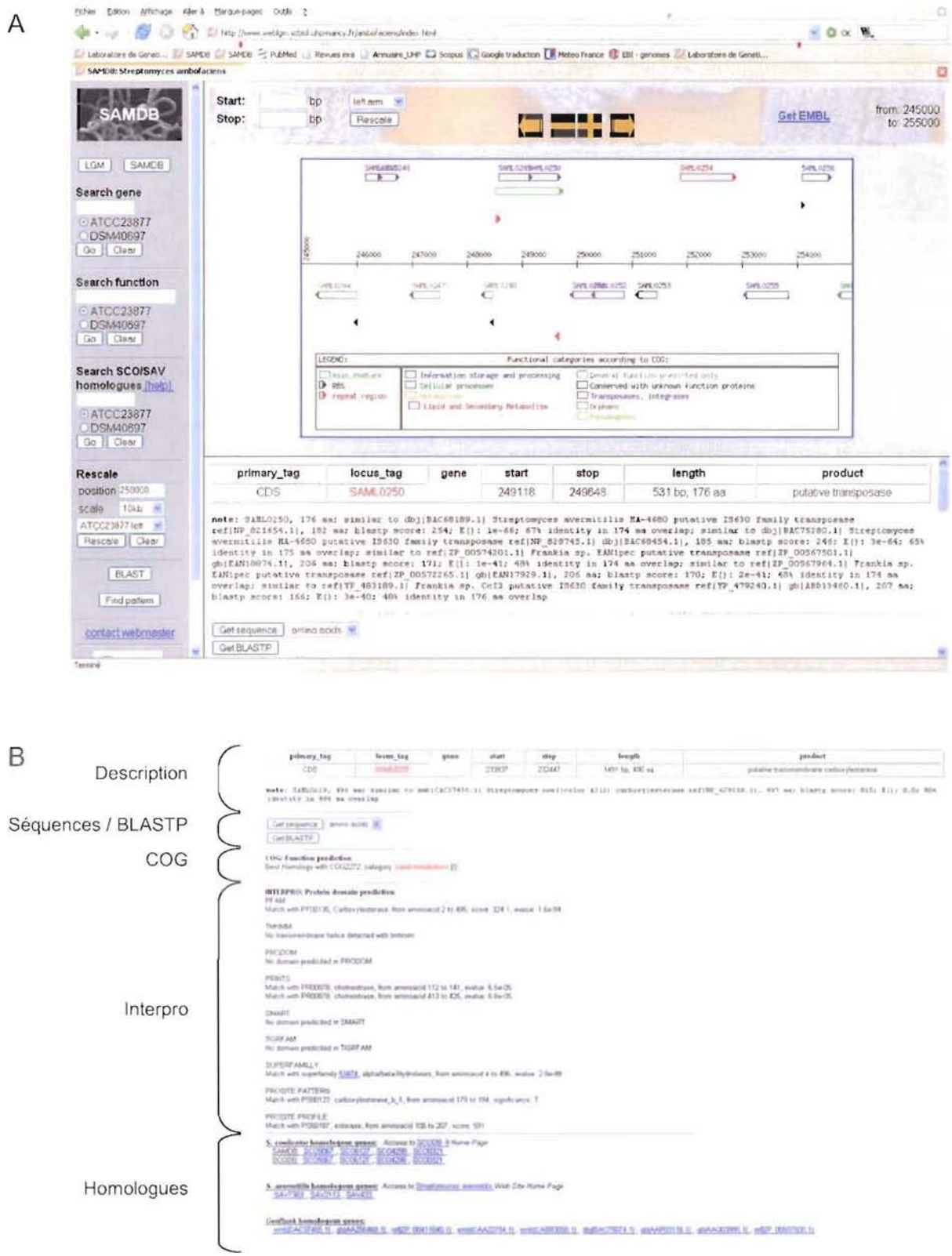


Figure 28 : Interface HTML de SAMBROWSER.
A. Visualisation graphique de l'annotation. Chaque caractéristique annotée apparaît sous forme de flèche cliquable. **B**. Affichage de l'annotation pour la CDS SAML0229.

3. *Analyses des extrémités de BAC (BES)*

L'étape préliminaire de séquençage systématique des extrémités des inserts de tous les BAC disponibles a permis d'obtenir 8457 séquences brutes (appelée BES pour BAC End Sequence) d'une longueur moyenne de 417+/-122 nt. Cette banque de séquences représente 3.524.440 nucléotides soit une couverture partielle 0,4 x du génome de *S. ambofaciens*.

Parmi les 4809 BAC disponibles, 3916 possèdent deux extrémités séquencées (soit 7832 BES) et 625 n'en possèdent qu'une seule (625 BES). Au-delà du taux de couverture qui peut sembler faible, c'est le lien physique existant entre chacun des 3916 couples de BES qui est exploitable pour comparer la structure du génome de *S. ambofaciens* aux autres génomes de *Streptomyces*.
Une analyse par BLASTN a été réalisée pour chaque BES disponible contre les séquences chromosomiques de *S. coelicolor*, *S. avermitilis* et *S. scabies*. Afin de traiter de façon automatique les 8457 résultats de BLASTN, une base de données relationnelle (EXTRBAC) a été créée. Les résultats de BLASTN contre les autres *Streptomyces* ont été intégrés (script *besmatch2sql.pl*) dans des tables dédiées aux comparaisons avec les autres *Streptomyces*.
Le programme *draw_bes_match.pl* utilise la mise en relation de ces données pour réaliser des cartes représentant l'alignement des BAC de *S. ambofaciens* sur les chromosomes des différents *Streptomyces* en fonction des résultats de BLASTN (Fig. 29).
Si les deux BES d'un BAC donné possèdent chacun des séquences homologues dans le chromosome de *S. coelicolor* par exemple, la distance séparant ces deux loci peut alors être calculée. Lorsqu'elle est compatible avec la longueur d'un insert (c'est-à-dire <80 kb), la région clonée dans ce BAC a été considérée comme conservée entre les deux espèces. De plus, chaque couple de BES correspond à deux brins d'ADN différents (réactions de séquençage effectuées sur des brins différents). Ainsi, pour détecter les BAC porteurs de régions conservées, le programme d'alignement (*draw_bes_match.pl*) vérifie non seulement la distance séparant les loci homologues mais également leur orientation sur le chromosome. Enfin, un score de BLASTN supérieur à 100 est requis pour considérer la similarité comme significative.
L'opération a donc été effectuée sur les 3916 couples de BES disponibles. Un exemple d'alignement est exposé dans la Figure 29B.

L'alignement ainsi réalisé entre les BAC de *S. ambofaciens* et les chromosomes de *Streptomyces* ont permis de localiser visuellement les régions spécifiques d'espèce et les bornes de réarrangements (notamment d'inversions) dans la région centrale du chromosome.

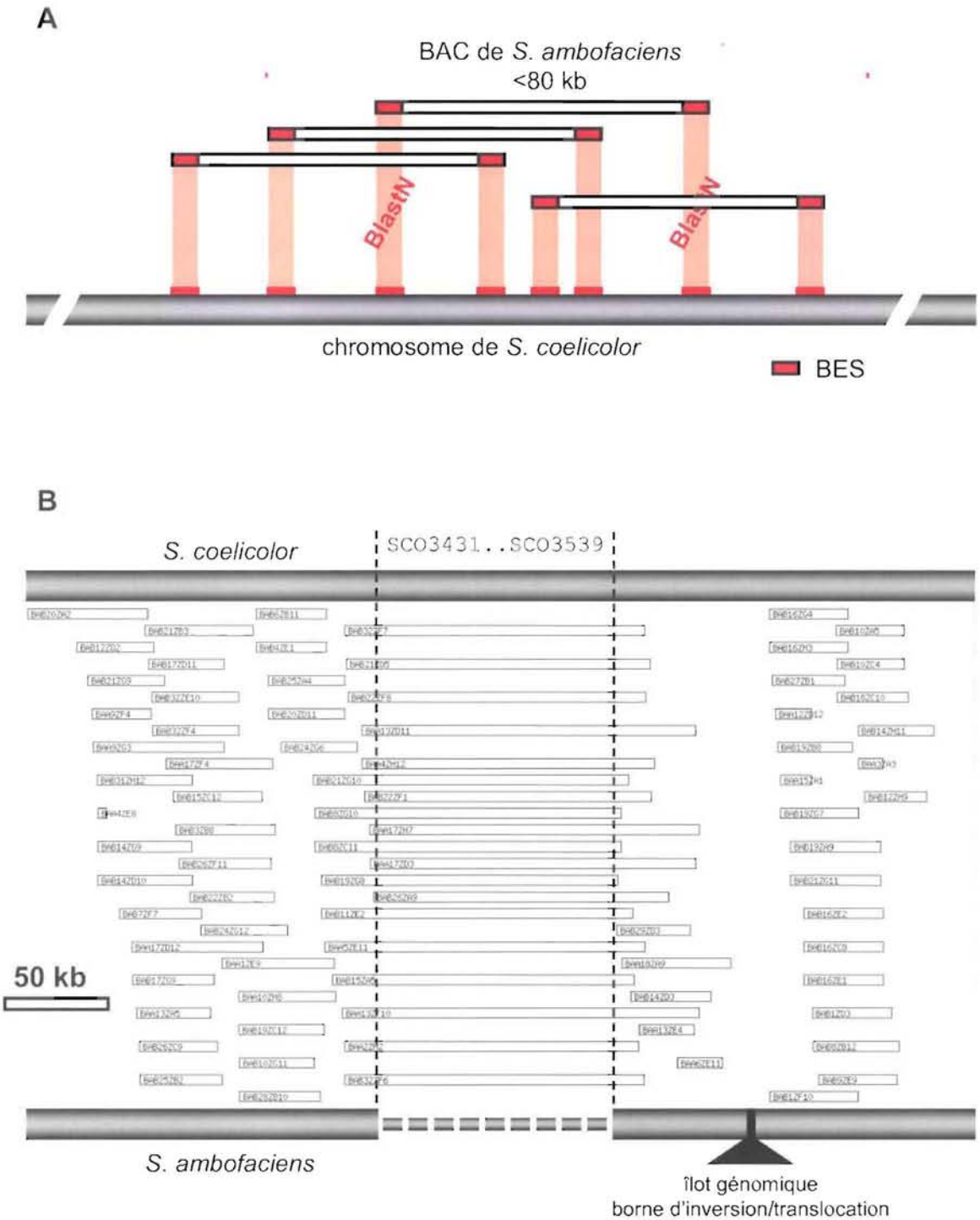


Figure 29 : Alignement des BAC de *S. ambofaciens* sur le chromosome de *S. coelicolor*.
A. Schéma du principe de l'analyse des extrémités de BAC (BES). Chaque couple de BES est comparé par BLASTN au chromosome de *S. coelicolor*, par exemple. Quand des inserts se chevauchent, la région comparée peut être considérée comme conservée entre les deux génomes. Cette analyse a été réalisée avec 3916 couples de BES.
B. Exemple de carte d'alignement des BAC de *S. ambofaciens* avec le chromosome de *S. coelicolor*. Les BAC sont représentés sous forme de rectangles blancs et localisés en fonction des similarités de séquences entre leur deux BES et le chromosome de *S coelicolor*. La présence d'une région portée par le chromosome de *S. coelicolor* absente chez *S. ambofaciens* (SCO3431..SCO3539) est repérable sur l'alignement par l'identification de BAC dont la taille n'est pas compatible avec la taille d'un insert (jusqu'à 80 kb). De plus, les îlots génomiques de *S. ambofaciens* absents chez *S. coelicolor* sont repérables par l'absence de chevauchement entre BAC, mais cela peut aussi correspondre à des bornes de réarrangements tels que les inversions.

4. *Analyse de l'ensemble des génomes procaryotes disponibles*

Après avoir étudié l'évolution de la structure du génome chez *Streptomyces*, s'est posée la question des similitudes et des différences dans l'organisation de la variabilité génomique chez les autres espèces procaryotes. Le module SAMCOMP, développé à l'origine pour la comparaison des génomes de *Streptomyces*, a été utilisé pour des analyses de génomique comparative portant sur l'ensemble des génomes procaryotes disponibles (312 génomes bactériens et 26 génomes d'archés en mai 2006, http://www.ebi.ac.uk/genomes/bacteria.html).

La base de données GENOMECOMP a été créée dans cette perspective. Elle est basée sur le même schéma relationnel que SAMDB (tables d'annotation et de comparaison).

Dans le but d'étudier la relation entre variabilité et localisation chromosomique, il était nécessaire de connaître les positions des origine et terminus de réplication de chaque chromosome. Worning *et al.* ont développé une méthode de prédiction de ces loci pour les chromosomes circulaires basée sur la recherche de biais de distribution d'octomères (Worning *et al.*, 2006). Les prédictions, disponibles dans GenomeAtlas (http://www.cbs.dtu.dk/services/GenomeAtlas/), ont été récupérées et intégrées dans GENOMECOMP.

Les données préliminaires concernant la fouille des génomes seront abordées dans la discussion.

C. Annotation des séquences des régions instables du chromosome de *S. ambofaciens*

Article en préparation :

Sequence analysis of the unstable regions of the *Streptomyces ambofaciens* linear chromosome

Choulet F., Aigle B., Gerbaud C., Decaris B., Pernodet J.-L., Leblond P.

1. Caractéristiques générales des régions séquencées

Les régions terminales du chromosome de la souche ATCC23877 ont été intégralement séquencées, conduisant ainsi à l'assemblage de deux contigs de, respectivement, 1.544.032 et 1.367.119 pb pour les bras gauche et droit. Les caractéristiques générales issues de l'annotation sont représentées dans la Figure 30 et détaillées dans le Tableau 1.

Tableau 1 : Caractéristiques générales des régions séquencées du chromosome de *S. ambofaciens*.

taille des régions séquencées	bras gauche	1.544.032 pb
	bras droit	1.367.119 pb
	total	2.911.151 pb
pourcentage en bases G+C		72,3%
taille des TIR		197.936 pb
gènes d'ARNt		3
opéron ribosomique		1
nombre de CDSs prédites (pseudogènes)		2532 (43)
CDS prédites dans les TIR (pseudogènes)		194 (3)
absence d'homologue dans la banque NR		264 (10,4%)
homologue de fonction inconnue		487 (19,2%)
homologue de fonction prédite		1781 (70,3%)
pas d'homologue dans COG		595
homologue dans COG		1186
taille moyenne des CDS (nt/aa)		986 / 327
densité en séquences codantes (%)		84,6
nombre de RBS prédits		1592
nombre de répétitions en tandem détectées (de 18 à 133 nt)		97

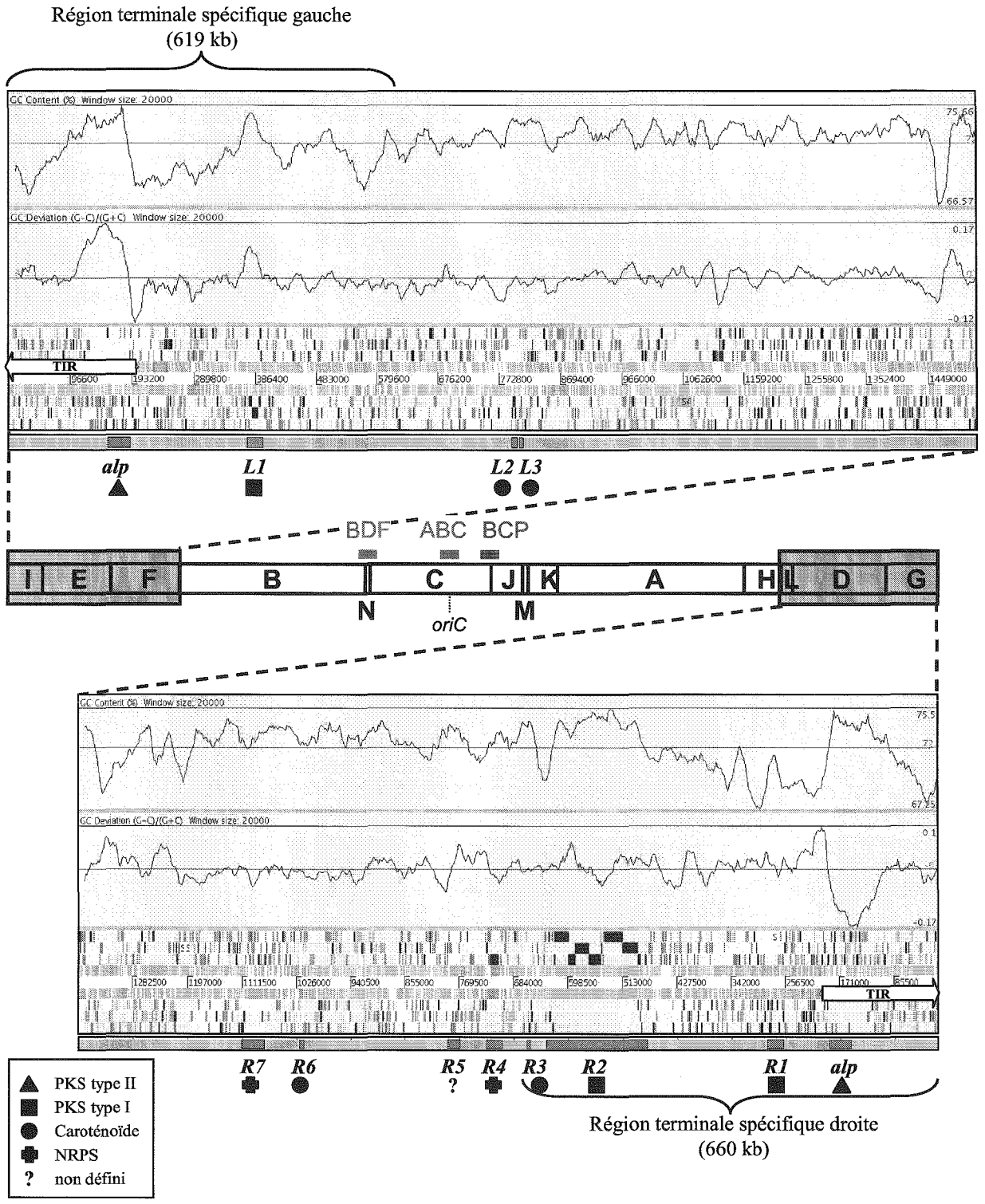


Figure 30 : Vue d'ensemble (sous Artémis) des régions séquencées du chromosome de *S. ambofaciens* ATCC23877. La carte *Ase*I du chromosome est représentée et les parties séquencées sont grisées. Trois BAC séquencés appartenant à la région centrale sont positionnés sur le schéma (BDF, ABC et BCP). ABC contient l'origine de réplication. Les CDS prédites sont représentées dans les six phases de lecture et les TIR sont indiquées par des flèches. Les profils de pourcentages en bases G+C et GC skew (G-C/G+C) sont calculés dans une fenêtre glissante de 20 kb et la valeur moyenne est représentée par une ligne horizontale sur chaque graphe. Les régions terminales spécifiques de *S. ambofaciens*, définies par comparaison avec *S. coelicolor*, sont délimitées par une accolade. Les positions des "clusters" prédits comme étant impliqués dans le métabolisme secondaire sont également indiquées (de **L1** à **L3** pour le bras gauche ; de **R1** à **R7** pour le bras droit). Le "cluster" ***alp*** (responsable de la production d'alpomycine) est dupliqué car inclus dans les TIR.
PKS : polycétide synthase, NRPS : synthase de peptide non ribosomique.

Par comparaison avec *S. coelicolor*, qui est l'espèce la plus proche dont le génome est séquencé, les extrémités du chromosome de *S. ambofaciens* ont été définies comme spécifiques d'espèce (voir *publication n°1* pour une analyse détaillée des comparaisons de génomes). Ces extrémités spécifiques ont été évaluées à **619 kb** et **660 kb** et sont indiquées sur la Figure 30. Ces régions se caractérisent par l'absence de synténie avec d'autres chromosomes séquencés. Au contraire, les régions internes des bras chromosomiques séquencés montrent une conservation, d'un degré variable, avec les génomes séquencés des autres *Streptomyces*.

Le pourcentage en bases G+C évolue de façon similaire sur les bras chromosomiques : une baisse est observée au niveau des extrémités chromosomiques sur plusieurs centaines de kilobases (Fig. 30). Seul les loci potentiellement impliqués dans le métabolisme secondaire dont notamment *alp*, impliqué dans la biosynthèse de l'antibiotique alpomycine (Pang *et al.*, 2004), présentent un contenu en G+C plus élevé que la moyenne (Fig. 30).
Afin de mieux caractériser ce phénomène, les variations des pourcentages en G+C global et G+C3 (à la troisième position des codons) des CDS ont été calculées le long du chromosome (programme *gc3cumul_plot.pl*). La somme cumulée des écarts à la moyenne (GCc et GC3c ; Fig. 31) permet de distinguer deux régions pour chaque bras chromosomique : les régions terminales enrichies en bases A+T sur **635 kb** (bras gauche) et **645 kb** (bras droit) et les régions internes plus riches en bases G+C. Cette baisse du contenu en G+C dans les régions terminales reflète donc une composition nucléotidique différente des CDS. Par ailleurs, cette diminution est majoritairement le fait de la troisième base des codons. En effet, l'amplitude de la variation du GC3c est deux fois plus forte que celle du GCc (échelle des ordonnées différentes sur graphes de la Fig. 31).

Cette richesse en bases A+T peut être le signe d'une acquisition récente par transfert horizontal. De plus, les limites des régions riches en A+T et celles des régions spécifiques d'espèces correspondent de façon surprenante, renforçant fortement l'hypothèse d'une acquisition des extrémités par transfert horizontal.

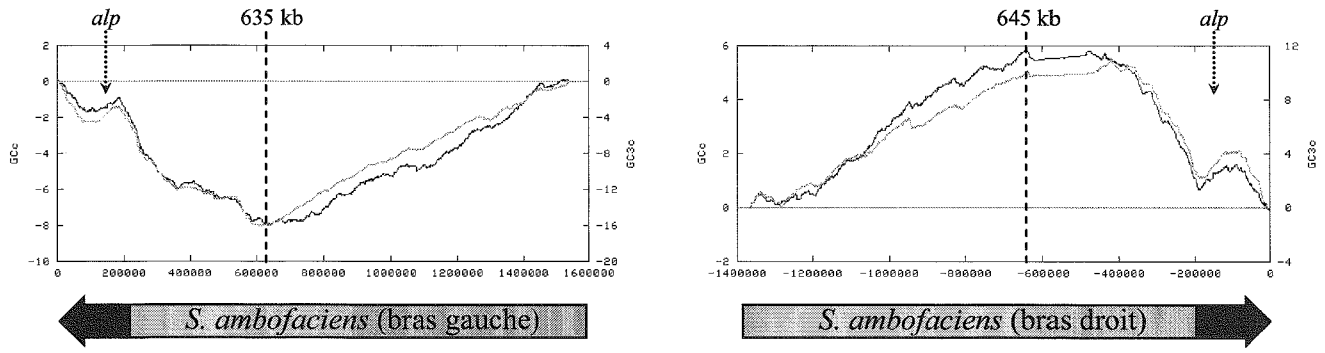


Figure 31 : Sommes cumulées des écarts à la moyenne pour le pourcentage en G+C des CDS (GCc, courbe noire) et le pourcentage en G+C à la troisième base des codons (GC3c, courbe grise) pour les bras chromosomiques gauche et droit de *S. ambofaciens*. Les TIR sont représentées par les flèches noires. Les phases décroissantes et croissantes des courbes révèlent, respectivement, une richesse en bases A+T et G+C. Chez *S. ambofaciens*, les extrémités gauche et droite présentant une richesse en A+T couvrent respectivement 635 kb et 645 kb.

Comme c'est le cas pour les autres génomes de *Streptomyces*, aucune structuration du GC skew n'est détectée (Fig. 30). Toutefois, le profil reste stable dans la majeure partie des régions séquencées alors qu'il est beaucoup plus chaotique au niveau des TIR. Il révèle notamment la présence d'une séquence de composition atypique au niveau de la limite interne des TIR (voir *publication n°2*).

2. *Métabolisme secondaire*

Le séquençage des génomes de *S. coelicolor* et *S. avermitilis* a révélé que le potentiel des *Streptomyces* à synthétiser des métabolites secondaires était beaucoup plus important que celui estimé par les méthodes de criblages classiques. En effet, *S. coelicolor* était connu pour produire 4 antibiotiques différents mais 20 voies de biosynthèse de métabolites ont été prédites dans son génome (Bentley *et al.*, 2002). Seule l'avermectine était connue chez *S. avermitilis* mais son génome contient 30 "clusters" potentiellement associés au métabolisme secondaire, représentant 6,6% des gènes codant des protéines (Ikeda *et al.*, 2003).
Un "cluster" définit un locus contenant plusieurs gènes impliqués dans une même fonction.

S. ambofaciens était connu pour synthétiser 2 antibiotiques : la spiramycine et la congocidine (Pinnert-Sindico, 1954). La recherche des tels clusters dans les régions séquencées a révélé la présence de 12 loci candidats (Tableau 2).

Tableau 2 : Clusters impliqués (ou potentiellement impliqués) dans le métabolisme secondaire et présents dans les régions terminales du chromosome de *S. ambofaciens*.

		limites des clusters	nombre de CDSs	taille (kb)	nature du cluster et/ou du métabolite
TIR	***alp***	SAMT0158..SAML0185	28	33	PKS II (alpomycine)
bras gauche	**L1**	SAML0370..SAML0383	14	24	NRPS / PKS I
	L2	SAML0729..SAML0735	7	8	phytoène synthase
	L3 (*tpc1*)	SAML0739..SAML0744	6	5	terpène synthase
bras droit	**R1**	SAMR0265..SAMR0278	14	25	PKS I
	R2	SAMR0454..SAMR0485	32	156	PKS I
	R3	SAMR0510..SAMR0513	4	5	lycopène cyclase
	R4 (*cch*)	SAMR0548..SAMR0559	12	24	NRPS
	R5	SAMR0594..SAMR0609	16	18	ACP, oxoacyl-ACP synthase
	R6 (*tpc2*)	SAMR0831..SAMR0836	6	6	terpène cyclase
	R7 (*cgc*)	SAMR0894..SAMR0921	28	34	NRPS
		total	**195***	**371***	

* *le cluster alp étant dupliqué, le nombre de CDS et sa taille ont été comptés deux fois.*

PKS I : polycétide synthase de type I, PKS II : polycétide synthase de type II., NRPS : synthase de peptide non ribosomique.

L'un d'entre eux est le cluster *alp* qui est responsable de la biosynthèse d'un pigment orangé et d'un antibiotique de la famille des angucyclines appelé alpomycine (Pang *et al.*, 2004). Il regroupe 28 CDS sur 33 kb et est dupliqué car intégralement inclus dans les TIR. Les deux copies de *alp* sont fonctionnelles (Pang *et al.*, 2004).
Les 10 autres clusters découverts grâce au séquençage ne sont pas répartis de façon équivalente sur les deux bras : 3 sont portés par la région gauche et 7 par la région droite. Leurs limites ont été définies de

façon arbitraire. Lorsqu'une synténie était observable avec d'autres génomes, ce critère a été retenu pour attribuer des bornes aux clusters. Dans le cas contraire, les fonctions putatives des gènes ont guidé ce choix.
Néanmoins, 195 gènes couvrant 371 kb seraient associés au métabolisme secondaire. Ces données indiquent que près de 15% des séquences des bras chromosomiques sont dédiées à la production de métabolites secondaires chez *S. ambofaciens*. Sur l'ensemble du chromosome de *S. avermitilis*, 30 clusters ont été décrits dont 17 sont portés par les régions terminales (Ikeda *et al.*, 2003). Un enrichissement des régions terminales en fonctions associées au métabolisme secondaire semble être une caractéristique commune des *Streptomyces*.

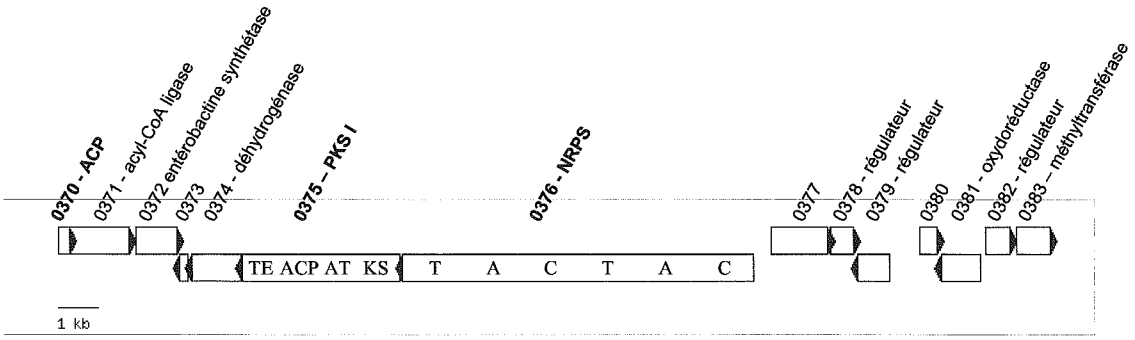
Parmi les 10 loci nouvellement détectés, 4 sont potentiellement associés à la biosynthèse de caroténoïdes : L2, L3 (*tpc1*), R3 et R6 (*tpc2*). Les caroténoïdes sont des dérivés lipidiques impliqués dans des mécanismes de protection contre les réactions d'oxydation (photoprotecteurs). Ils sont largement utilisés dans l'industrie agroalimentaire pour leurs propriétés colorantes et en pharmacologie pour leurs capacités antioxydantes.

Trois clusters codent putativement des enzymes appartenant à la famille des PKS de type I (L1, R1 et R2). Cette famille regroupe des enzymes modulaires de très grande taille (plusieurs milliers de résidus acides aminés) où chaque module possède au minimum trois domaines (KS : cétosynthase, AT : acyltransférase et ACP : acyl carrier protein) catalysant un cycle d'élongation permettant l'incorporation d'un acide carboxylique dans la chaîne polycétonique.

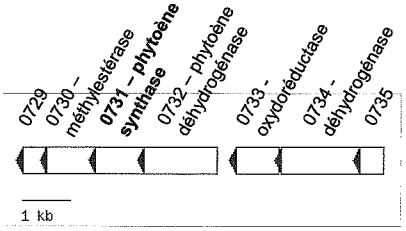
Les clusters R4 et R7 sont, quant à eux, impliqués potentiellement dans la biosynthèse non ribosomique de peptides (enzymes de type NRPS). Des études expérimentales ont permis de mettre en évidence le rôle du cluster *cch* dans la synthèse d'un sidérophore appelé coelichéline (Barona-Gomez *et al.*, 2006), décrit initialement chez *S. coelicolor* (Challis et Ravel, 2000).
Le rôle du cluster *cgc* (R7) dans la synthèse de congocidine a également été démontré (Pernodet J.-L., comm. pers.). La congocidine est un oligopeptide (structure pyrrole amide, (Bialer *et al.*, 1980)) capable de se lier à l'ADN. Elle possède des propriétés antibiotique, antitumorale et également antivirale.

Enfin, l'implication du cluster R5 dans le métabolisme secondaire est plus litigieuse. Il constitue un îlot spécifique de *S. ambofaciens* interrompant la synténie avec le chromosome de *S. coelicolor*, suggérant une acquisition récente et plusieurs fonctions codées semblent être associées au métabolisme secondaire ou au métabolisme des lipides (ex : ACP, ACP-synthase, acyltransférase, monooxygénase, épimérase, glycosyltransférase).
Une description détaillée de chaque cluster est représentée sur la Figure 32.

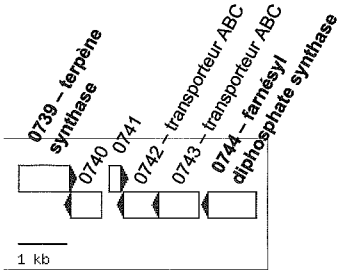
L1
PKS I/
NRPS



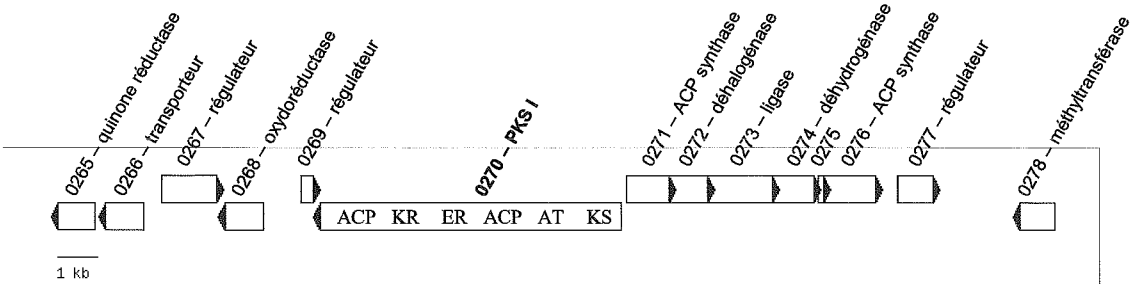
L2
caroténoïde



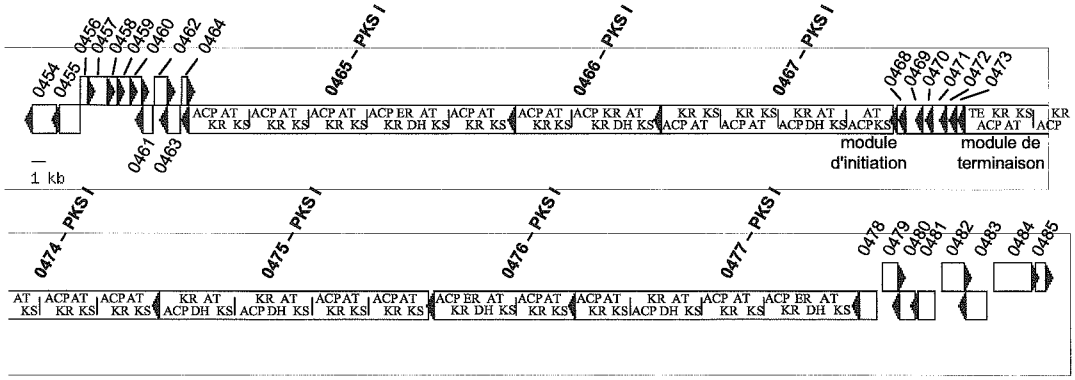
L3
caroténoïde



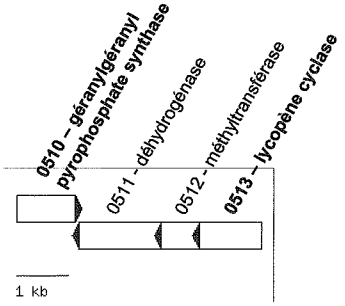
R1
PKS I



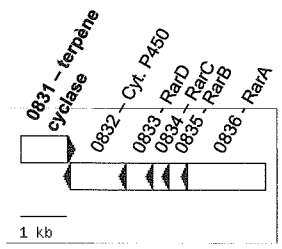
R2
PKS I



R3
caroténoïde



R6
caroténoïde



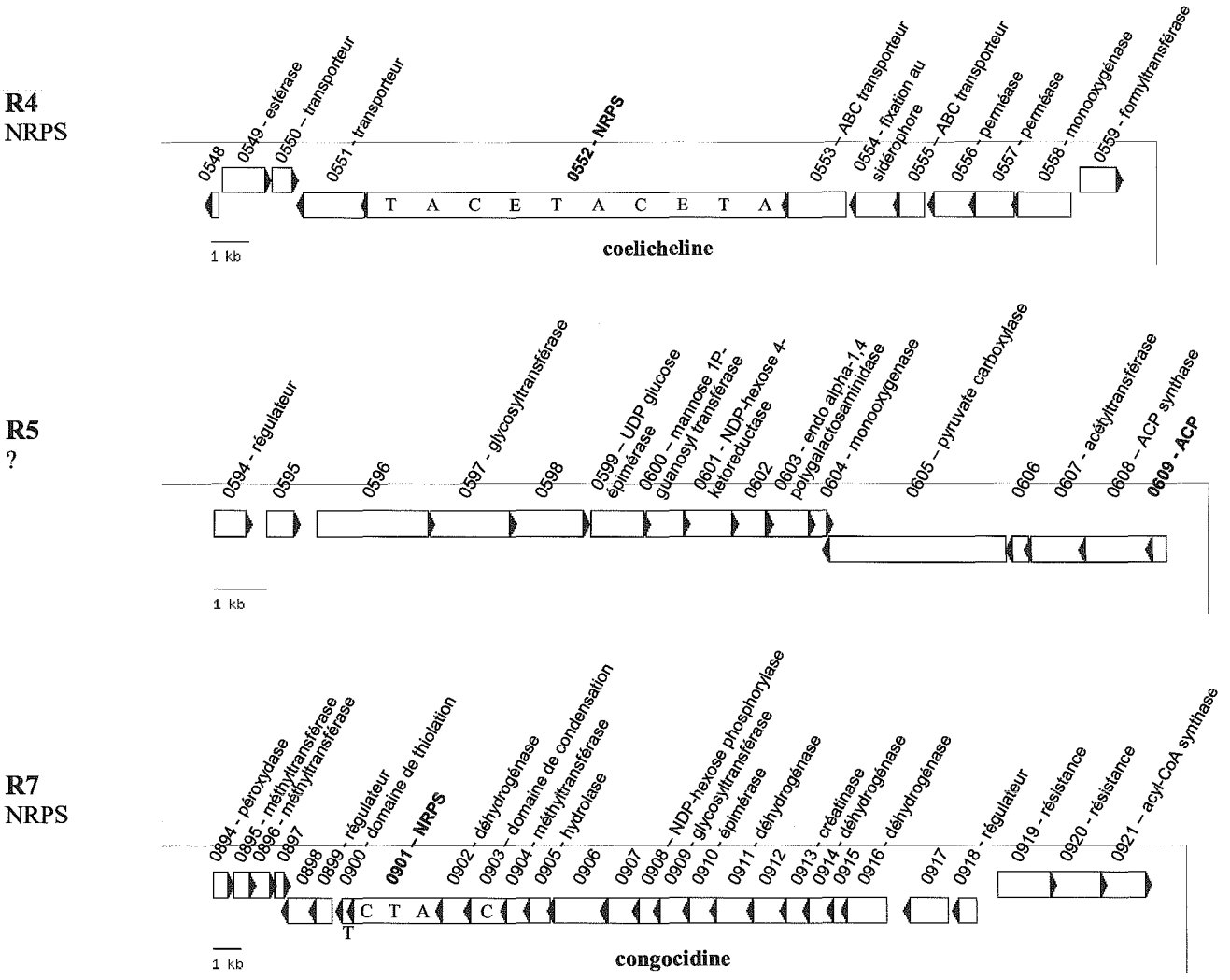


Figure 32 : Schéma des clusters potentiellement impliqués dans la biosynthèse de métabolites secondaires chez *S. ambofaciens*. Les numéros correspondent aux noms des gènes sans le préfixe "SAM[LR]".
Les domaines catalytiques des NRPS et PKS de type I sont indiqués. Pour les NRPS : A, domaine d'adénylation ; C, domaine de condensation ; T, domaine de thiolation ; E, domaine d'épimérisation. Pour les PKS : ACP, acyl carrier protein ; KS, cétosynthase ; AT, acyltransférase ; DH, déshydrogénase ; ER, énoylréductase ; KR, cétoréductase ; TE, thioestérase. PKS I/II : polycétide synthase de type I/II, NRPS : synthase de peptide non ribosomique. La recherche de ces domaines a été effectuée à l'aide du serveur http://www.nii.res.in/nrps-pks.html.

3. Spécificité et caractère chimérique des voies de biosynthèse de métabolites secondaires

La moitié des 12 clusters (10 + *alp* dupliqué) est portée par les extrémités spécifiques d'espèce comparées à *S. coelicolor* couvrant au total 1,28 Mb (619 kb+660 kb). Malgré la présence des 6 autres loci dans les régions montrant une synténie avec le chromosome de *S. coelicolor*, un seul est conservé de façon ancestrale entre ces deux espèces : R4 (*cch*) responsable de la biosynthèse de coelichéline. Les autres constituent des îlots spécifiques interrompant la synténie. Ces résultats confirment la forte spécificité d'espèce des clusters de synthèse de métabolites secondaires. Cette variabilité du potentiel de biosynthèse de métabolites chez les *Streptomyces* et autres Actinomycètes est très probablement

associée à des transferts horizontaux, comme cela a été suggéré pour les clusters de PKS (Ginolhac *et al.*, 2005 ; Metsa-Ketela *et al.*, 2002).

Les trois clusters codant des PKS de type I ne présentent pas une organisation conservée avec d'autres loci connus. De plus, les gènes codant les PKS présentent de faibles niveaux d'identité, moins de 50% en général, voire aucune similarité significative avec ceux identifiés chez d'autres *Streptomyces*. En cela, ils constituent de nouvelles pistes de recherche de métabolites d'intérêts. Par exemple, le gène SAMR0270 codant la PKS du locus R1 ne possède aucun homologue chez *S. coelicolor* et *S. avermitilis* et la séquence la plus proche similaire (37% d'identité) retrouvée dans la banque NR appartient à *Burkholderia thailandensis* (Protéobactérie beta).

Certains clusters présentent une structure chimérique et sont issus du réassortiment de différents loci. C'est par exemple le cas du cluster *alp*. Une partie est hautement conservée avec le cluster de biosynthèse de la kinamycine chez *Streptomyces murayamaensis* (numéro d'accession : AY228175) alors que la seconde montre une similarité avec un autre cluster, porté par le plasmide pSLA2-L de *Streptomyces rochei* (Mochizuki *et al.*, 2003). De plus, le module de régulation de *alp* est, quant à lui, similaire à celui de la tylosine chez *Streptomyces fradiae* (Bate *et al.*, 1999).
Les clusters L2 et R3, tous deux potentiellement associés à la biosynthèse de dérivés lipidiques, sont portés par deux bras différents chez *S. ambofaciens* et sont homologues à deux parties distinctes du même cluster *crtYTUVBIE* impliqué dans la production du caroténoïde isoréniératène chez *Streptomyces griseus* (Krugel *et al.*, 1999). Un cluster homologue possédant une organisation identique à celui de *S. griseus* a été identifié chez *S. coelicolor* (SCO0185-91). Son caractère réarrangé est illustré à la Figure 33. Chez *S. coelicolor*, la présence de pseudogènes dérivant de *crtU* et *crtV* au niveau du locus orthologue au cluster L2 de *S. ambofaciens* semble indiquer que *crt* était présent de façon ancestrale et a subi une délétion depuis la divergence des deux espèces. Ce cluster semble soumis à de fréquents réarrangements puisqu'il est organisé de trois façons différentes dans les quatre génomes étudiés (Fig. 33).

Les réarrangements auxquels sont soumises les régions terminales pourraient être le moteur du réassortiment de gènes impliquées dans des voies de biosynthèses de métabolites secondaires et donc de leur diversité.

De façon intéressante, un lien a été établi entre le phénotype de certains mutants de *Streptomyces* et des réarrangements affectant certaines voies de biosynthèse de métabolites. En effet, des mutants présentant des amplifications spontanées de fragments d'ADN (Amplified DNA sequence, ADS) correspondant à des voies du métabolisme secondaire ont été détectés. Chez *S. ambofaciens*, deux loci amplifiables correspondent à des clusters caractérisés au cours de ce programme de séquençage. Tout d'abord, un fragment de 37 kb correspondant au cluster *alp* est amplifié chez le mutant 29C1 qui génère une pigmentation verte des colonies (Catakli *et al.*, 2003). De plus, le locus amplifiable AUD90 (Aigle *et al.*, 1996) correspond au cluster R2, porteur de plusieurs gènes modulaires codant des PKS de type I. Chez ces mutants, l'amplification provoque une diminution de la synthèse de spiramycine.

Chez *Streptomyces rimosus*, le cluster (30 kb) de synthèse de l'oxytétracycline est amplifié chez certains mutants (Gravius *et al.*, 1993). De même, chez *Streptomyces kanamyceticus* le cluster de biosynthèse de la kanamycine est impliqué dans une amplification (Yanai *et al.*, 2006).

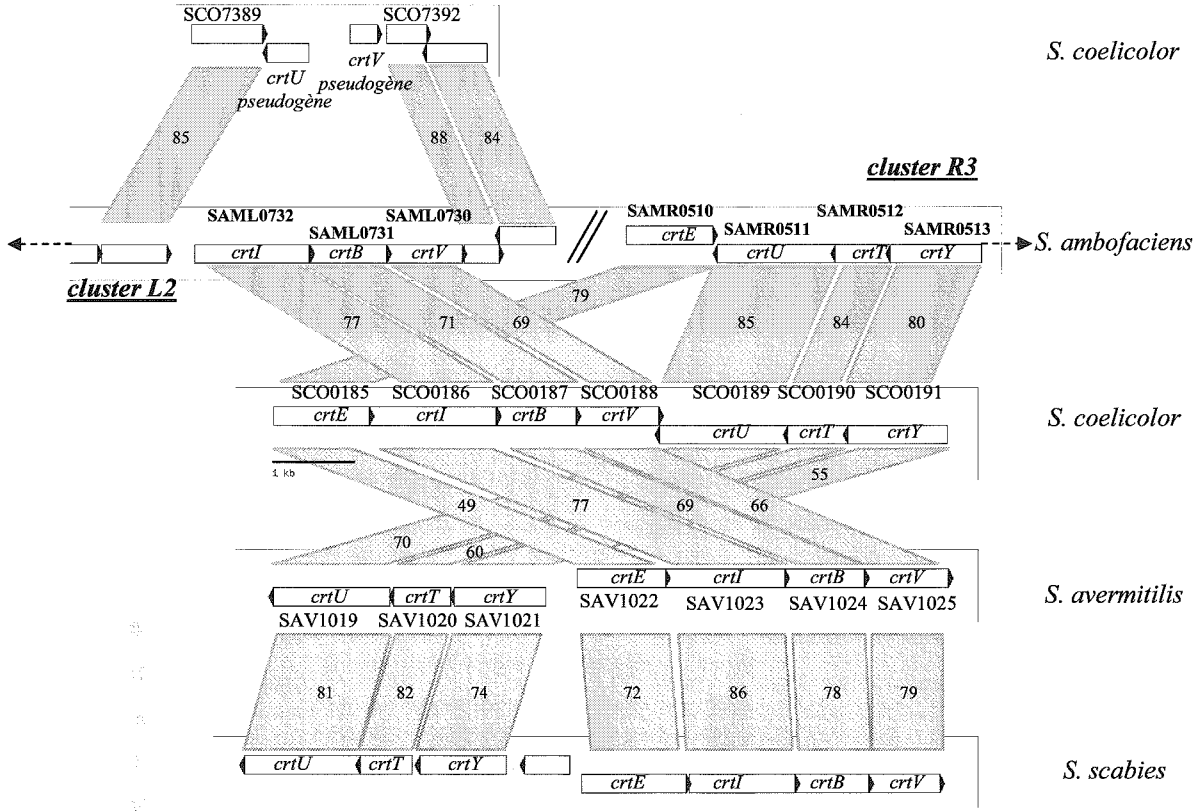


Figure 33 : Conservation et remaniements du cluster *crt* (isoréniératène) dans les génomes de *S. coelicolor*, *S. avermitilis*, *S. scabies* et *S. ambofaciens*. Les aires grisées relient des gènes homologues et les valeurs d'identité en acides aminés sont indiquées.
crtU : déhydrogénase, *crtT* : méthyltransférase, *crtY* : lycopène cyclase, *crtE* : géranylgéranyl diphosphate synthétase, *crtI* : phytoène déhydrogénase, *crtB* : phytoène synthase, *crtV* : méthylestérase.

Le mécanisme moléculaire d'amplification pourrait faire intervenir une réplication par cercle roulant comme proposé par (Young et Cullum, 1987). La présence de motifs répétés espacés de façon régulière dans le cluster de biosynthèse de la kanamycine pourrait jouer un rôle dans son amplification (Yanai *et al.*, 2006). Ainsi chez *S. ambofaciens*, la structure modulaire et répétée (Fig. 32) des gènes du cluster R2 pourrait être à l'origine de son amplification et donc être responsable d'une partie de l'instabilité génomique. En revanche, le cluster *alp* ne contient pas de gènes à structure modulaire répétée (PKS de type II). En revanche, plusieurs motifs répétés en tandem pouvant être impliqués dans son amplification ont été détectés.

4. *Fonctions prédites des gènes portés par les bras chromosomiques de S. ambofaciens*

Une classification des CDS selon leur fonction prédite a été réalisée à l'aide de la base de données COG (Tableau 3). Ce processus automatique a permis d'attribuer une catégorie fonctionnelle à 1317 CDS (52%) et les CDS restantes ont été classées manuellement parmi les 18 catégories définies par

COG. Au final, une fonction putative a pu être attribuée à 75% des CDS, laissant 25% de CDS de fonction inconnue.

Tableau 3 : Classification fonctionnelle des 2532 CDSs prédites dans les régions séquencées de *S. ambofaciens* d'après les 18 catégories de COG (les catégories les plus représentées sont indiquées en gras).

Classe	[Code] Catégorie fonctionnelle	nombre de CDS	%
Processus cellulaires	[D] Division cellulaire et partition	5	0,2
	[M] Synthèse de l'enveloppe cellulaire, membrane externe	81	3,2
	[N] Motilité des cellules et sécrétion	11	0,4
	[O] Modifications post-traductionnelles, recyclage des protéines, chaperonnes	30	1,2
	[P] Métabolisme et Transport des ions inorganiques	55	2,2
	[T] Transduction du signal	109	4,3
Stockage et traitement de l'information	[J] Traduction, structure et synthèse des ribosomes	27	1,1
	[K] **Transcription**	257	10,2
	[L] Réplication d'ADN, recombinaison et réparation	107	4,2
Métabolisme	[C] Energie, production et conversion	101	4,0
	[E] Métabolisme et transport des acides aminés	117	4,6
	[F] Métabolisme et transport des nucléotides	20	0,8
	[G] **Métabolisme et transport des sucres**	164	6,5
	[H] Métabolisme des coenzymes	50	2,0
	[I] Métabolisme des lipides	65	2,6
	[Q] **Synthèse, catabolisme et transport des métabolites secondaires**	154	6,1
Fonction peu ou pas caractérisée	[R] **Fonction peu caractérisée**	537	21,2
	[S] **Fonction inconnue**	521	20,6
	Aucune similarité ni aucun domaine identifié	121	4,8
	Fonctions prédites	1890	74,6
	Fonctions inconnues	642	25,4
	Total	2532	

- Familles de gènes paralogues

Les régions instables du chromosome de *S. ambofaciens* présentent un niveau élevé de redondance fonctionnelle. Malgré le séquençage partiel du génome, un regroupement des CDS par famille de paralogues a été entrepris avec Blastclust (Altschul *et al.*, 1997) en utilisant un seuil de 30% d'identité et un chevauchement d'au moins 80%. Cette analyse a révélé la présence 33 familles possédant plus de 5 membres paralogues dans les régions terminales (en ne tenant compte que d'une copie des TIR).

Cette redondance est attribuée à la complexité du cycle de développement, c'est-à-dire à la différenciation des *Streptomyces*. Des enzymes paralogues peuvent représenter parfois des isoenzymes (possédant la même fonction) mais dont l'activité dépend du stade de développement. Un exemple a été décrit chez *S. coelicolor* où deux copies d'un cluster de gènes dupliqués et impliqués dans le métabolisme du glycogène, agissent à des stades de développement distincts (Schneider *et al.*, 2000).

Dans les bras chromosomiques de *S. ambofaciens*, la famille la plus représentée compte 32 paralogues codant des oxydoréductases.

- Fonctions de régulation

Une quantité tout aussi remarquable de régulateurs transcriptionnels a été observée appartenant à diverses familles : LacI, TetR, AraC, LysR, MarR, DeoR, PadR, AsnC, GntR, LuxR, MerR, FadR, ROK et SARP ("*Streptomyces* antibiotic regulatory protein"). Au total, 312 gènes, soit 12,3% des

capacités de codage, sont impliqués dans divers processus de régulation. Un pourcentage identique (12,3%) avait été révélé par l'analyse du génome complet de *S. coelicolor*, les gènes de régulation étant répartis uniformément le long du chromosome (Bentley *et al.*, 2002).

Les génomes de *Streptomyces* sont connus pour coder un nombre très élevé de facteurs sigma : respectivement 65 et 63 ont été prédits chez *S. coelicolor* et *S. avermitilis*. Ceci contraste avec les 17 prédits chez *B. subtilis* et 7 chez *E. coli* K12. *S. ambofaciens* n'échappe pas à cette règle puisque 27 facteurs sigma potentiels ont été détectés dans les 2,9 Mb disponibles. Par ailleurs, 15 anti-anti-facteurs sigma (et 4 anti-facteurs sigma) ont été détectés alors que respectivement 15 et 13 (4 et 4) sont retrouvés dans les chromosomes complets de *S. coelicolor* et *S. avermitilis*. Chez ces derniers, les régions de contingence ne présentent pas d'enrichissement en fonctions régulatrices. Cette importance des fonctions régulatrices n'est pas spécifique au genre *Streptomyces* mais se retrouve chez l'ensemble des bactéries possédant un génome de grande taille (Konstantinidis et Tiedje, 2004 ; Ranea *et al.*, 2004).

- Transposons, IS et dérivés

Les régions séquencées sont riches en éléments transposables et en leurs dérivés non fonctionnels. Au total, 53 ORF sont similaires à des IS, dont au moins 30 semblent être des pseudogènes issus de troncatures. La quasi-totalité (50/53) de ces ORF est portée par les régions terminales spécifiques. Un tel biais est également observable chez *S. coelicolor* (Chen *et al.*, 2002) et *S. avermitilis* pour lesquels respectivement 45% et 79% des IS et leurs dérivés sont localisés au niveau des extrémités (représentant 25% de la taille du chromosome).

Des arguments en faveur de la mobilité de certaines de ces séquences sont apportées par le niveau de conservation de certaines copies et la présence de motifs répétés inversés flanquants. En effet, quatre copies strictement identiques d'une région de 1166 pb, flanquée par une répétition inversée de 25 nucléotides, sont dispersées dans les extrémités séquencées. Cette région coderait deux CDS de 179 et 176 résidus dans des phases de lecture décalées. Le dernier codon de la CDS amont chevauche le premier de celle située en aval. Cette situation rappelle celle décrite pour les membres de la famille IS*3* pour lesquels l'expression d'une transposase fonctionnelle est dépendante d'un événement de décalage (-1) du cadre de lecture du ribosome (Polard *et al.*, 1992 ; Sekine *et al.*, 1994).

Ces quatre couples d'ORF (SAML0249-50, SAMR0214-5, SAMR0283-4 et SAMR0289-90) sont homologues à IS*630* détectée en copies multiples et dans la même organisation chez *S. avermitilis* (61% et 67% d'identité en acides aminés) et chez *Frankia* sp. Cci3 (50% et 40% d'identité). En revanche, aucun homologue n'a été détecté chez *S. coelicolor*.

Un transposon composite de 5,1 kb constitué de 4 ORF flanquées par deux IS étroitement apparentées (99% d'identité) et flanquées par les mêmes motifs répétés inversés (29 pb) a été détecté à environ 12 kb des TIR (SAMR0213-18). Ces deux IS ne présentent pas d'homologie avec les génomes de *Streptomyces* mais avec celui de *Frankia* sp. EAN1pec. De plus, ce transposon est porteur d'une copie d'IS*630* décrit précédemment et de deux gènes orphelins. Il aurait donc été acquis récemment par transfert horizontal et pourrait être à l'origine de la prolifération d'IS*630* chez *S. ambofaciens*.

Dans la partie centrale du chromosome de *S. ambofaciens* a été détecté un élément intégratif et conjugatif appelé pSAM2 (Pernodet *et al.*, 1984). Chez *S. coelicolor*, 5 îlots dérivés de pSAM2 ont été décrits sur l'ensemble du chromosome, tous portés par la région "core" (Bentley *et al.*, 2002). Aucun dérivé de cet élément conjugatif ni aucun autre type d'éléments potentiellement impliqué dans le transfert horizontal, par conjugaison ou transduction, n'a été détecté dans les régions terminales du chromosome de *S. ambofaciens*.

5. *Présence de gènes dupliqués*

La présence de TIR chez *Streptomyces* implique la coexistence au sein du même génome de nombreux gènes de fonction identique. L'information dupliquée n'est pas limitée aux TIR. Chez *S. ambofaciens*, hormis les IS en copies multiples, 5 couples supplémentaires de fragments d'ADN (contenant entre 1 et 3 gènes) semblent être issus d'événements récents de duplication. En effet, ces gènes présentent plus de 85% d'identité nucléotidique alors que leur homologue dans les génomes de *S. coelicolor* et *S. avermitilis* sont soit, présents en copie unique soit, totalement absents. Le premier cas concerne les gènes *has* codant des facteurs sigma alternatifs (Roth *et al.*, 2004). *hasL* (SAML0459) et *hasR* (SAMR0651) partagent 98% d'identité nucléotidique et sont portés en orientation inverse par deux bras chromosomiques différents. Ils peuvent être les substrats d'événements de recombinaison homologue conduisant au remplacement d'un bras par recopiage de l'autre extrémité chromosomique selon un mécanisme de réplication induite par cassure appelé BIR (break induced replication) (Fischer *et al.*, 1998b).
Une seconde paire de gènes dupliqués codant des facteurs sigma a été détectée par cette analyse : SAML0525/SAMR0899. Ils partagent 99% d'identité nucléotidique et sont également portés par deux bras différents en orientation opposée. De plus, cette région dupliquée englobe les gènes adjacents (SAML0524/SAMR0898) de fonction inconnue et qui partagent également 99% d'identité.
Une autre paire de gènes distants d'environ 20 kb sur le même bras (SAML0574/SAML0593 ; 287 codons), codant putativement des hydrolases sécrétées, partage un niveau très élevé d'identité (99%) suggérant une duplication récente.

Certains segments dupliqués peuvent contenir jusqu'à trois gènes. C'est le cas des clusters SAMR0444/5/6 et SAMT0045/6/7 qui partagent 85% d'identité.
Enfin, un couple de gènes adjacents est présent en trois copies très similaires dans les bras chromosomiques : SAMT0034/35, SAML0518/19 et SAMR1095/96 (quatre en considérant la duplication engendrée par la présence de TIR). Ils partagent entre 89% et 92% d'identité et sont homologue à SCO0072-73 de *S. coelicolor* codant putativement une protéine sécrétée et une protéine membranaire.

Ces données pourraient refléter une grande efficacité de duplication de fragments d'ADN chez *Streptomyces*. Toutefois, la présence de gènes dupliqués peut aussi résulter de plusieurs événements indépendants d'acquisitions par transfert horizontal. Il n'existe en réalité aucun argument pour favoriser l'un ou l'autre de ces mécanismes.

6. *Spécificité et adaptation*

Au total, 987 gènes spécifiques de *S. ambofaciens* par rapport à *S. coelicolor* et *S. avermitilis* ont été identifiés, dont 570 ont une fonction prédite (et non prédits comme pseudogènes).
En plus des régions terminales de 1,28 Mb qui incluent la grande majorité des gènes spécifiques de *S. ambofaciens*, des îlots génomiques interrompant la synténie avec les autres génomes ont été identifiés (Tableau 4).

Tableau 4 : Ilots génomiques spécifiques de *S. ambofaciens* contenant au moins 5 CDS interrompant la synténie observée entre les régions séquencées gauche (A) et droite (B) et le génome de *S. coelicolor*. Les clusters impliqués potentiellement dans le métabolisme secondaire sont indiqués en gras.

	Limites	nombre de CDS	taille (kb)	Fonctions codées et commentaires
A				
	SAMT0001..SAML0572	572	619,0	Région terminale spécifique du bras gauche [***alp*, L1**]
1	SAML0586..SAML0591	6	6,3	cystéine désulfurase, acétyltransférases
2	SAML0663..SAML0668	6	6,8	nitrilotriacetate monooxygénase, transporteurs
3	SAML0730..SAML0734	5	6,5	[**L2**]
4	SAML0739..SAML0744	6	5,0	[**L3**]
5	SAML0864..SAML0874	11	11,7	chitinase, aldéhyde déhydrogénase, alkanesulfonate monooxygénase, acyl-CoA déhydrogénase, cytochrome c oxydase
6	SAML0921..SAML0925	5	6,5	inhibiteur d'initiation de la traduction, protéines membranaires
7	SAML0970..SAML0998	29	55,9	hyaluronidase, protéines de sporulation, translocase FtsK/SpoIIE, beta-glucosidase, rhamnosidase, transporteur de sucres
8	SAML1007..SAML1013	7	6,1	3-oxoacid-CoA-transférase, monooxygénase, dioxygénase, maleylpyruvate isomérase
9	SAML1209..SAML1215	7	7,4	lipase, alkylhydropéroxydase
B				
	SAMT0001..SAMR0510	510	660,0	Région terminale spécifique du bras droit [**R1, R2, R3**]
1	SAMR0592..SAMR0604	13	14,6	[**R5**]
2	SAMR0638..SAMR0644	7	6,5	phosphatase, énoyl-CoA hydratase/isomérase
3	SAMR0754..SAMR0764	11	8,2	3 acétyltransférases, glucosyltransférase
4	SAMR0831..SAMR0836	6	5,5	[**R6**]
5	SAMR0894..SAMR0921	28	33,9	[**R7**]
6	SAMR0998..SAMR1021	24	40,3	AUD6, métabolismes des sucres
7	SAMR1058..SAMR1104	47	56,8	3 pseudogènes similaires à des gènes de réplication/partition du plasmide linéaire SCP1 : polymérase III, ParA et ParB
	total :	**1300**	**1557**	

Les gènes des régions spécifiques codent par exemple des enzymes extracellulaires impliquées dans la dégradation des polymères organiques tels que la chitine, le chitosane et la cellulose. La chitine est un polymère de résidus glucoses aminés et s'avère être l'un des constituants essentiels de l'exosquelette des insectes et des parois protectrices des cellules fongiques. Le chitosane est un dérivé de cette dernière. La cellulose est, quant à elle, l'un des principaux composants de la paroi des cellules végétales. Ainsi, quatre loci codant putativement des chitinases (et protéines associées), un gène codant une chitosanase et un autre codant une cellulase ont été détectés dans les séquences de *S. ambofaciens*. Ces capacités métaboliques sont bien connues chez les différentes espèces de *Streptomyces*, qui jouent un rôle dans la biodégradation de la matière organique. Les gènes codant des

chitinases sont largement distribués à travers les espèces *Streptomyces* (Kawase *et al.*, 2004). La forte conservation de leur séquence avec celles retrouvées chez les Actinobactéries suggère que cette fonction aurait été acquise par un ancêtre des *Streptomyces* et dispersée par transfert horizontal aux Actinobactéries (Kawase *et al.*, 2004).

Le cluster SAML0279-80 code putativement une cyclo-inulinase (cycloinulino-oligosaccharide fructofuranosidase [CFTase]) et une endo-inulinase. L'inuline est un polymère constitué de résidus fructose d'origine végétale (Vandamme et Derycke, 1983). *S. ambofaciens* possède donc les capacités métaboliques pour dégrader l'inuline. Cette molécule, et notamment ses dérivés, possèdent des propriétés d'intérêt dans les domaines de la pharmacologie et de l'agroalimentaire. La première enzyme détectée est responsable de la cyclisation de l'inuline tandis que la seconde est impliquée dans sa dégradation. Toutes deux sont capables de produire des inulo-oligosaccharides possédant de nombreux effets bénéfiques pour la santé. Ces deux gènes retrouvés chez *S. ambofaciens* ne possèdent pas d'homologues dans les génomes de *Streptomyces*, cependant, une activité inulinase a déjà été décrite chez une souche de *Streptomyces* (Gill *et al.*, 2003).

Deux clusters non apparentés, tous deux impliqués dans la production de vésicules gazeuses, ont été détectés. Le premier (SAMR0685-94 ; comprenant 8 gènes *gvp*) est conservé avec les clusters *gvp* retrouvés chez de nombreux autres *Streptomyces*. En revanche, le second est spécifique d'espèce. Il comprend 4 gènes (SAML0452-5) ayant une organisation identique et une meilleure similarité (de 29% à 56% d'identité) avec les clusters *gvp* identifiés chez les archées *Methanosarcina barkeri* (CP000099) et *Halobacterium salinarium* (Englert *et al.*, 1992).

Deux loci putativement impliqués dans la dégradation de l'acide phénylacétique (*paa*) sont présents dans les bras chromosomiques de *S. ambofaciens*. L'acide phénylacétique est un composé central dans le métabolisme des composés polluants comme le styrène.
Ces deux loci constituent un nouvel exemple de clusters impliqués dans des fonctions identiques mais possédant une origine différente. Le premier (SAML0646-52) est l'orthologue (90% d'identité en moyenne et synténie conservée) du cluster *paaA-K* (SCO7469-75) de *S. coelicolor* tandis que le second (SAML0265-9), localisé dans les régions terminales spécifiques, est plus divergent (moins de 50% d'identité). Le locus *paa* est largement distribué chez les bactéries du sol (Abe-Yoshizumi *et al.*, 2004) renforçant ainsi la probabilité d'acquisition par transfert horizontal de ce second cluster chez *S. ambofaciens*.

Les souches de *Streptomyces* ne sont pas seulement un réservoir de molécules antibiotiques, ils sont également un réservoir de gènes de résistance. Ainsi, 28 gènes associés à des résistances vis-à-vis de divers composés (ex : antibiotiques, métaux lourds) ont été mis en évidence dans les bras chromosomiques chez *S. ambofaciens*. Certains d'entre eux sont spécifiques d'espèces comme SAML0988 potentiellement impliqué dans la résistance à la vancomycine.
S. coelicolor et *S. avermitilis* sont tous deux résistants au chloramphénicol mais grâce à des mécanismes différents. Le premier code une protéine responsable de l'efflux de cet antibiotique alors que *S. avermitilis* code une phosphotransférase responsable de sa modification (inactivation).

S. ambofaciens possède les deux systèmes. Les régions terminales portent trois gènes de résistance à cet antibiotique. SAML0610 et SAML0533 sont homologues à la protéine CmlR de *S. coelicolor* (SCO7526) impliquée dans le mécanisme d'efflux, le premier étant probablement l'orthologue (92% d'identité dans une région synténique) alors que le second est un paralogue plus divergent (41% d'identité). Enfin, SAMR0442 code une phosphotransférase homologue (87% d'identité) à SAV877 identifiée chez *S. avermitilis* (Mosher *et al.*, 1995).

Deux gènes spécifiques, présents dans les régions dispensables, codent putativement une phénylalanyl-tRNA synthétase (PheRS ; SAML0296) et une thréonyl-tRNA synthétase (ThrRS ; SAMR0317). La première montre la plus forte similarité avec une tRNA synthétase prédite chez *Anaeromyxobacter dehalogenans* (Protéobactérie delta du sol). La seconde est similaire à une tRNA-synthétase retrouvée chez *Frankia* (Actinomycète du sol). Les gènes codant les PheRS et ThrRS sont en copie unique, donc essentiels, dans les génomes de *S. coelicolor* et *S. avermitilis*. L'analyse des BES chez *S. ambofaciens* a montré que les orthologues correspondants sont présents dans la région centrale du chromosome. Ceci explique la présence de ces copies supplémentaires dans des régions dispensables. Elles sont peu apparentées aux synthétases endogènes et proviennent probablement de transferts horizontaux. Des études ont par ailleurs montré l'existence de transferts horizontaux de gènes de tRNA synthétases (Diaz-Lazcoz *et al.*, 1998).
Des antibiotiques peuvent inhiber spécifiquement certaines tRNA-synthétases. En effet, une classe d'antibiotiques, (sulfonamides phényl-thiazolylurée) possède une activité inhibitrice spécifique de la PheRS (Beyer *et al.*, 2004). La borrelidine est, quant à elle, connue pour inactiver sélectivement la ThrRS (Freist et Gauss, 1995). La capacité de coder deux enzymes peu apparentées, comme chez *S. ambofaciens*, pourrait donc conférer une résistance et donc un avantage adaptatif. Cette hypothèse a par ailleurs été proposée (Ikeda *et al.*, 2003) pour expliquer le maintien de deux tryptophanyl-tRNA synthétases à la fois chez *S. coelicolor* et chez *S. avermitilis* qui sont naturellement résistants à l'indolmycine, un inhibiteur de cette enzyme.

En conclusion, 653 CDS identifiées dans les régions séquencées de *S. ambofaciens* restent de fonction inconnue, voire orpheline. Elles représentent 26% du contenu en gènes des régions instables.

D. Evolution du génome des *Streptomyces*

Publication n°1 :

Evolution of the terminal regions of the *Streptomyces* linear chromosome

Choulet F., Aigle B., Gallois A., Mangenot S., Gerbaud C., Truong C., Francou F.-X., Fourrier C., Guérineau M., Decaris B., Barbe V., Pernodet J.-L., Leblond P.

Molecular Biology and Evolution, 2006, Vol. 23, No. 11, sous presse.

S. ambofaciens et *S. coelicolor* constituent un couple d'espèces très proches en termes phylogénétiques au sein du genre des *Streptomyces*. Au contraire, *S. avermitilis* est une espèce nettement plus éloignée et son ancêtre commun avec le couple *S. ambofaciens/S. coelicolor* est en fait l'ancêtre commun de la quasi totalité des espèces de *Streptomyces* identifiées (Fig. 16).

Un quatrième chromosome de *Streptomyces*, celui de l'espèce phytopathogène *S. scabies*, est disponible mais non annoté (http://www.sanger.ac.uk/Projects/S_scabies/). Les relations phylogénétiques établies sur la base des séquences des ARNr 16S montrent que *S. scabies* possède une position intermédiaire sur l'arbre des *Streptomyces*. Cette espèce est éloignée des trois autres mais semble toutefois plus proche de *S. avermitilis* que du couple *S. ambofaciens/S. coelicolor*.
Etant donnée l'absence d'une annotation disponible, cette publication ne mentionne pas les comparaisons avec *S. scabies*. Cependant, une annotation automatique de son génome a été réalisée à l'aide de la plateforme développée au cours de ce travail (voir chapitre "Approches bioinformatiques"). L'identification des CDS a permis de réaliser des comparaisons de génomes à partir du contenu en gènes et pas uniquement à partir de la séquence nucléotidique. Les résultats correspondants sont exposés dans ce chapitre.

1. Conservation de la région centrale du chromosome

Les séquences des extrémités de BAC (BES pour BAC End Sequence) de la banque d'ADN génomique de *S. ambofaciens* ont été alignées par BLASTN avec la séquence du chromosome de l'espèce proche *S. coelicolor* (voir "*Approches bioinformatiques*"). Cette approche a permis de mettre en évidence un niveau très élevé de conservation entre les régions centrales de *S. ambofaciens* et *S. coelicolor* (Fig. 34A). Par ailleurs, deux inversions péricentriques ont été fixées depuis leur divergence. En combinant cette analyse avec les comparaisons par dot-plot des autres génomes de *Streptomyces*, il a été possible de placer ces inversions sur l'arbre phylogénétique (Fig. 34C). Ainsi, les chromosomes de *S. ambofaciens* et de *S. scabies* sont colinéaires, tandis que ceux de *S. coelicolor* et de *S. avermitilis* diffèrent par quatre événements d'inversions. La plupart des bornes d'inversion (6 sur 8) ne sont pas clairement identifiables car elles correspondent à des régions variables du génome. Par exemple, l'une de ces bornes est un dérivé de l'élément mobile pSAM2. Il semblerait donc que les inversions seraient préférentiellement fixées lorsqu'elles se produisent dans des régions tolérantes aux réarrangements.

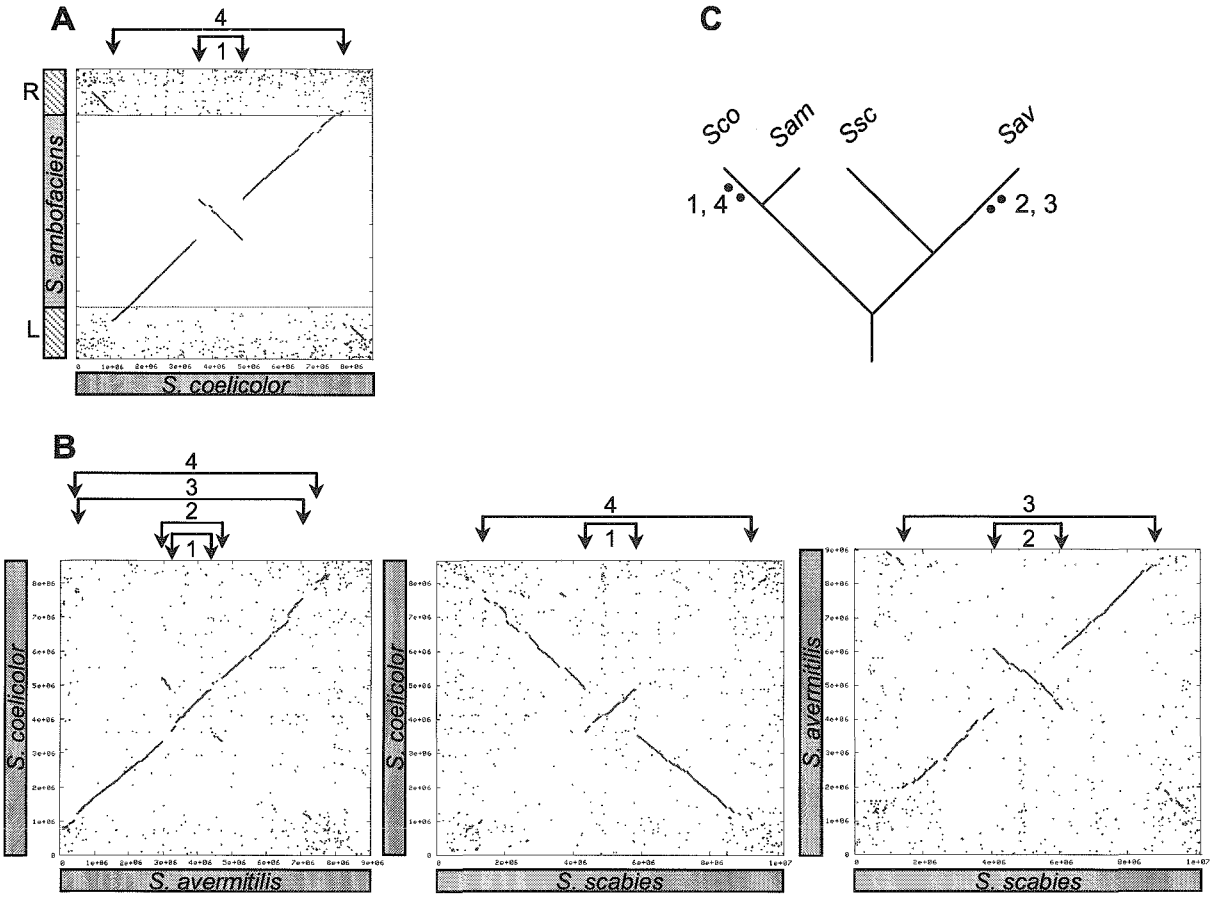


Figure 34 : Comparaisons par dot-plot des chromosomes de *Streptomyces*.
A. Comparaison des chromosomes de *S. ambofaciens* (*Sam*) et *S. coelicolor* (*Sco*). Pour *S. ambofaciens*, les rectangles hachurés représentent les bras chromosomiques totalement séquencés (L : bras gauche, R : bras droit). Chaque point représente la position d'une CDS de *S. ambofaciens* et de son homologue (meilleur score de BLASTP) chez *S. coelicolor*. L'alignement de la partie centrale du chromosome a été dessiné à partir de la comparaison des BES par BLASTN. Pour représenter l'alignement en deux dimensions, les critères suivants ont été utilisés : les régions identifiées comme conservées à partir des analyses de BES ont été considérées de même taille chez les deux espèces ; comme la taille des régions spécifiques du chromosome de *S. ambofaciens* ne peut pas être déterminée par cette approche, une taille fixe de 80 kb (longueur maximale d'un insert) leur a été attribuée pour l'alignement. **B.** Comparaison des chromosomes de *S. coelicolor*, *S. avermitilis* (*Sav*) et *S. scabies* (*Ssc*). Les inversions centrées sur l'origine de réplication sont indiquées par des flèches, numérotées et positionnées (**C.**) sur l'arbre phylogénétique.

Afin de préciser le niveau de conservation des gènes portés par la région centrale du chromosome de *S. ambofaciens*, trois BAC, dont les inserts correspondent à des loci portés par cette région, ont été séquencés. Ces trois loci recouvrent 117 kb et contiennent 91 gènes prédits. Tous ces gènes présentent une organisation conservée et un haut niveau d'identité avec les autres génomes de *Streptomyces*. En moyenne, les gènes de *S. ambofaciens* portés par la région centrale partagent respectivement 89%, 82% et 80% d'identité en acides aminés avec *S. coelicolor*, *S. avermitilis* et *S. scabies* (respectivement 90%, 84% et 82% d'identité nucléotidique). Une seule rupture de synténie est observée parmi ces 117 kb comparés au chromosome de *S. coelicolor*.

Malgré les inversions péricentriques, la région centrale du chromosome montre un niveau élevé de synténie entre les quatre espèces de *Streptomyces*. Néanmoins, l'analyse des BES a révélé 12 clusters, de 26 à 149 kb, présents chez *S. coelicolor* et absents chez *S. ambofaciens* (tableau supplementary material publication n°1). Parmi ceux-ci, sont retrouvés les clusters impliqués dans la synthèse d'antibiotiques : actinorhodine (*act*), undécylprodigiosine (*red*) et *cda* (calcium dependent antibiotic). Réciproquement, 4 régions spécifiques du chromosome de *S. ambofaciens* ont ainsi pu être localisées grâce à l'analyse des BES. L'une d'elles correspond au cluster de biosynthèse de l'antibiotique spiramycine. Cependant, la taille de ces régions spécifiques ne peut pas être identifiée par cette méthode.

Cette méthode de comparaison de génomes utilisant les BES permet d'obtenir, à partir d'un taux de recouvrement faible d'une partie du génome, une idée précise quant à son niveau de conservation. Cependant, cette technique est limitée par le niveau d'identité des séquences homologues et par le niveau de conservation du contenu en gènes.

2. *La taille des extrémités chromosomiques spécifiques d'espèce augmente avec la distance phylogénétique*

La comparaison des génomes de *Streptomyces* permet de mettre en évidence le confinement de la variabilité génomique au niveau des régions terminales. Cette caractéristique se vérifie non seulement entre espèces éloignées mais également entre espèces proches comme le sont *S. ambofaciens* et *S. coelicolor* (Fig. 34).
De façon surprenante, les bornes des régions spécifiques d'espèce changent en fonction du couple d'espèces comparées. De plus, la limite entre les régions terminales spécifiques et la région centrale conservée est floue. Du fait de la dégénérescence très prononcée de la synténie aux bornes des régions terminales spécifiques, il n'est en effet pas possible de déterminer la position du dernier gène conservé sur chaque bras chromosomique. Ainsi, le niveau de GOC (Gene Order Conservation) a été utilisé comme indice de mesure du niveau de synténie le long du chromosome (Fig. 35). Un seuil de 20% de GOC_1 a été choisi afin de fixer une limite, certes arbitraire mais définie d'un point de vue statistique, permettant de distinguer les régions terminales spécifiques de la région centrale conservée.
Chez *S. ambofaciens*, la taille des régions terminales spécifiques passe de 1279 kb (619+660 kb pour les régions terminales gauche et droite respectivement) quand la comparaison est réalisée avec *S. coelicolor*, à 1878 kb (889+989 kb) quand celle-ci est réalisée avec *S. avermitilis* et enfin à 1883 kb (889+994 kb) par rapport à *S. scabies*.
Lorsque ces analyses sont appliquées aux chromosomes de *S. coelicolor*, *S. avermitilis* et *S. scabies*, la taille minimale des régions spécifiques est respectivement de 753 kb, 1393 kb et 1199 kb.

Réciproquement, la taille de la région centrale conservée, donc héritée de l'ancêtre commun, diminue avec la distance phylogénétique qui sépare les espèces comparées. Ces données suggèrent donc que l'acquisition de gènes par transfert horizontal s'effectue préférentiellement au niveau des extrémités chromosomiques chez *Streptomyces* et que les flux de gènes faisant disparaître la synténie seraient relativement constants depuis l'ancêtre des *Streptomyces*.

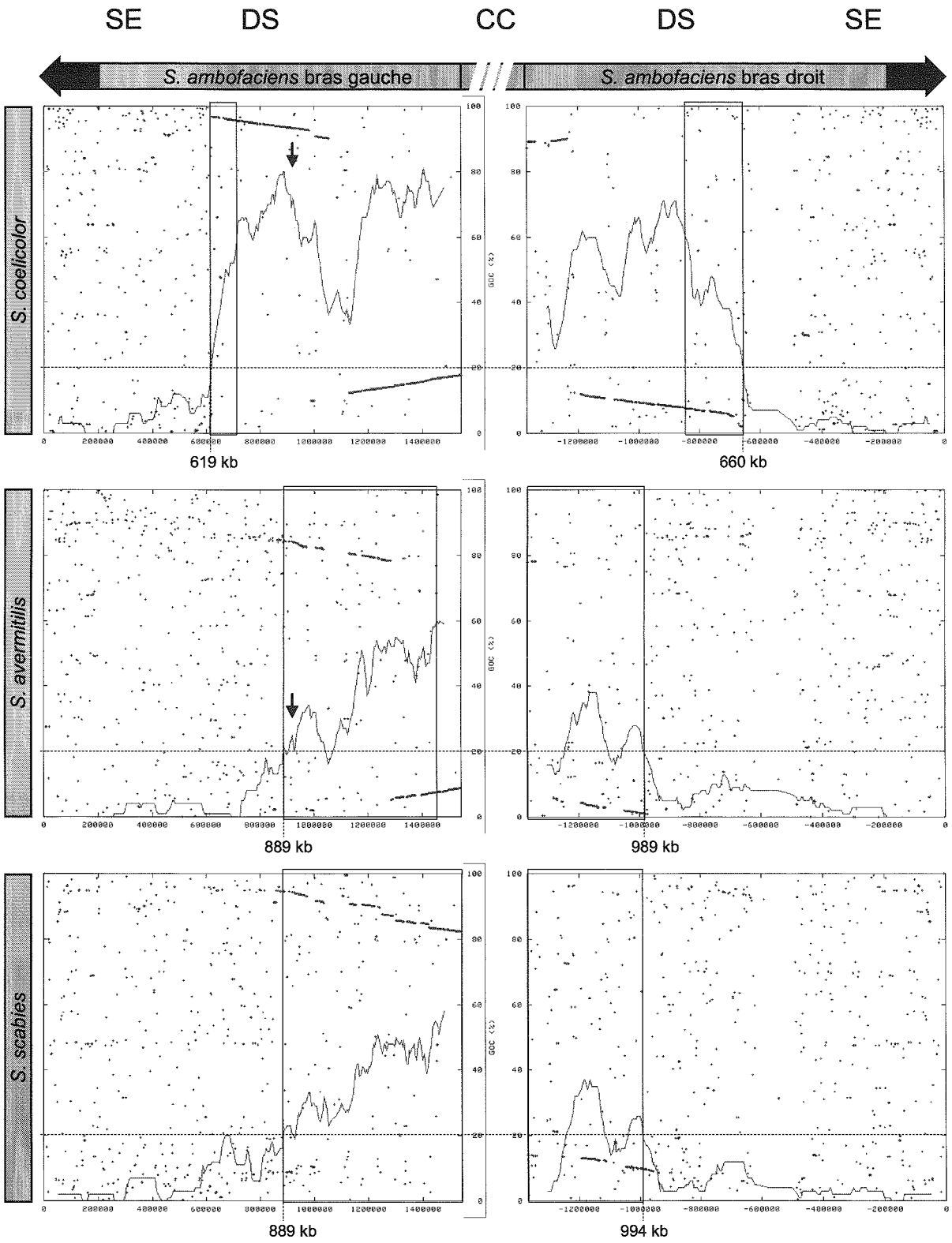


Figure 35 : Profil de GOC_1 le long des bras chromosomiques gauche et droit de *S. ambofaciens* comparés aux génomes de *S. coelicolor*, *S. avermitilis* et *S. scabies*.
Sur chaque graphe, la courbe du GOC_1 est superposée au dot-plot. Un seuil minimal de GOC_1 de 20% a été choisi pour délimiter les régions terminales spécifiques d'espèces. SE : extrémités spécifiques ; DS : régions de synténie dégénérée ; CC : région centrale conservée. La flèche indique les régions comparées sur la Figure 36. Les régions de synténie dégénérée (entre 20 et 60% de GOC) sont encadrées. Calculs réalisés avec une fenêtre glissante de 100 gènes et un pas de 5 gènes.

3. *Niveau de spécificité des régions terminales*

Tous les gènes portés par les régions terminales ne sont pas spécifiques d'espèce. Par ailleurs, les notions de spécificité et de similarité significative entre deux séquences restent arbitraires. Deux séquences peuvent être apparentées et ne présenter aucune similarité significative. A l'inverse, homologie ne signifie pas orthologie. En d'autres termes, une similarité entre deux séquences ne signifie pas forcément transmission héréditaire de cette séquence depuis l'ancêtre commun.
Le génome des *Streptomyces* comporte, certes, un nombre élevé de gènes mais beaucoup sont redondants et appartiennent à de grandes familles de séquences apparentées. Par exemple, chez *S. avermitilis*, 721 familles de gènes paralogues comportant de 2 à 91 membres par famille ont été identifiées. Cette redondance concerne 35% des CDS prédites chez cette espèce.

Les extrémités chromosomiques du génome de *S. ambofaciens* ne sont synténiques avec aucun génome de *Streptomyces* séquencé. Bien que des clusters peuvent présenter une organisation conservée avec d'autres génomes, leur nombre est limité et leur taille reste petite (moins de 8 gènes). Au total, 1082 CDS sont prédites dans les 1279 kb identifiés comme spécifiques aux extrémités du chromosome de *S. ambofaciens*. Parmi celles-ci, seulement 9% possèdent une CDS similaire partageant au moins 60% d'identité dans les génomes de *S. coelicolor* et de *S. avermitilis* (7% si *S. scabies* est inclus).
Comparée au génome du plus proche voisin, *S. coelicolor*, 63% des CDS possèdent un homologue. Cependant, seulement 13% des 1082 CDS portées par ces régions terminales possèdent un homologue fortement conservé (plus de 80% d'identité en acides aminés), c'est-à-dire à un niveau équivalent à celui des orthologues de la région centrale du chromosome. En effet, ces valeurs contrastent avec celles décrites pour la région centrale où 100% des CDS prédites dans les 117 kb séquencés présentent plus de 80% d'identité avec *S. coelicolor*.
Comparée à *S. avermitilis*, le niveau de spécificité est encore plus élevé puisque 56% des 1082 CDS possèdent un homologue chez cette espèce et seulement 4% de ces dernières présentent un niveau élevé de similarité.

Ce faible niveau de conservation, tant en terme d'identité de séquences qu'en terme de synténie, suggère fortement que la majorité des homologies détectées ne reflète pas la présence de gènes orthologues mais plutôt l'introduction massive d'allèles étrangers par transfert horizontal (xénologues). Le niveau de spécificité des extrémités chromosomiques est donc très élevé non seulement entre espèces éloignées mais aussi entre espèces très proches.

4. *Régions de synténie dégénérée*

Entre la région centrale conservée, montrant un niveau élevé de GOC_1, et les régions terminales spécifiques, des régions de synténie dégénérée ont été caractérisées. Ces régions présentent un niveau intermédiaire de GOC_1 (entre 20% et 60%). Pour chaque paire de génomes comparés, la synténie observée dans la région centrale dégénère progressivement sur plusieurs centaines de kilobases avant d'atteindre les extrémités spécifiques d'espèce (Fig. 36). La dégénérescence se caractérise par une multitude d'événements d'insertion/délétions (indels) fixés au cours de la divergence des espèces. Cette

baisse du GOC_1 ne reflète donc pas une augmentation du taux de réarrangements de l'information endogène mais plutôt une augmentation des flux de gènes (acquisition/perte). En effet, lorsque le nombre de gènes orthologues devient trop faible pour détecter une synténie, le GOC_1 devient inférieur à 20%. Les Figures 36 et 37 montrent des exemples de loci affectés par cette dégénérescence. Pour une région donnée, l'intensité de la dégénérescence suit les liens phylogénétiques. Le nombre de ruptures de synténie (insertion, délétion, remplacement de gènes) a été évalué sur 100 kb (99 CDS) du bras gauche de *S. ambofaciens* (de SAML0798 à SAML0896) et, comparé au chromosome de *S. coelicolor* (de SCO7238 à SCO7327), 10 événements se sont produits depuis leur divergence. En revanche, 29 et 25 indels sont observables lorsque la comparaison est réalisée avec *S. avermitilis* (de SAV1085 à SAV1268) et *S. scabies* (du nucléotide 9.442.000 à 9.726.000) respectivement. Cependant, le nombre de réarrangements est probablement sous-estimé lorsqu'il dépasse un seuil (saturation d'événements). La proportion de gènes orthologues diminue avec la chute du niveau de GOC_1. Pour cette dernière région, respectivement 83, 52 et 49 des 99 CDS prédites chez *S. ambofaciens* possèdent un orthologue chez *S. coelicolor*, *S. avermitilis* et *S. scabies*.

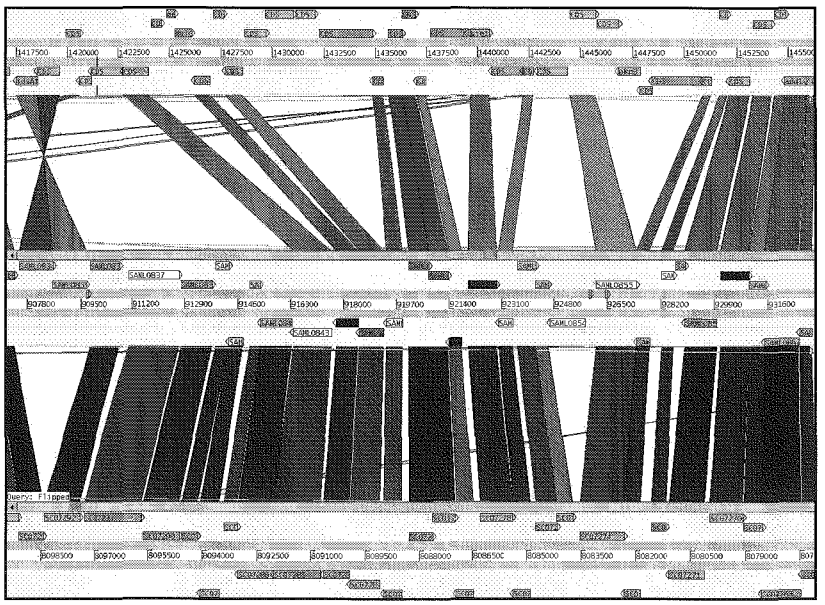
Comme le montrent les profils de GOC_1, les régions affectées par la dégénérescence ne sont pas les mêmes selon la paire de génomes comparés. En effet, de même que la taille de la région centrale ancestrale conservée diminue avec la distance phylogénétique, la dégénérescence de la synténie affecte des régions de plus en plus proches de la région "core".

A.

Sav (38 CDSs; 39 kb)
SAV1129..SAV1166

Sam (31 CDSs; 26 kb)
SAML0833..SAML0863

Sco (27 CDSs; 22 kb)
SCO7267..SCO7293



B.

Sav

Sam

Sco

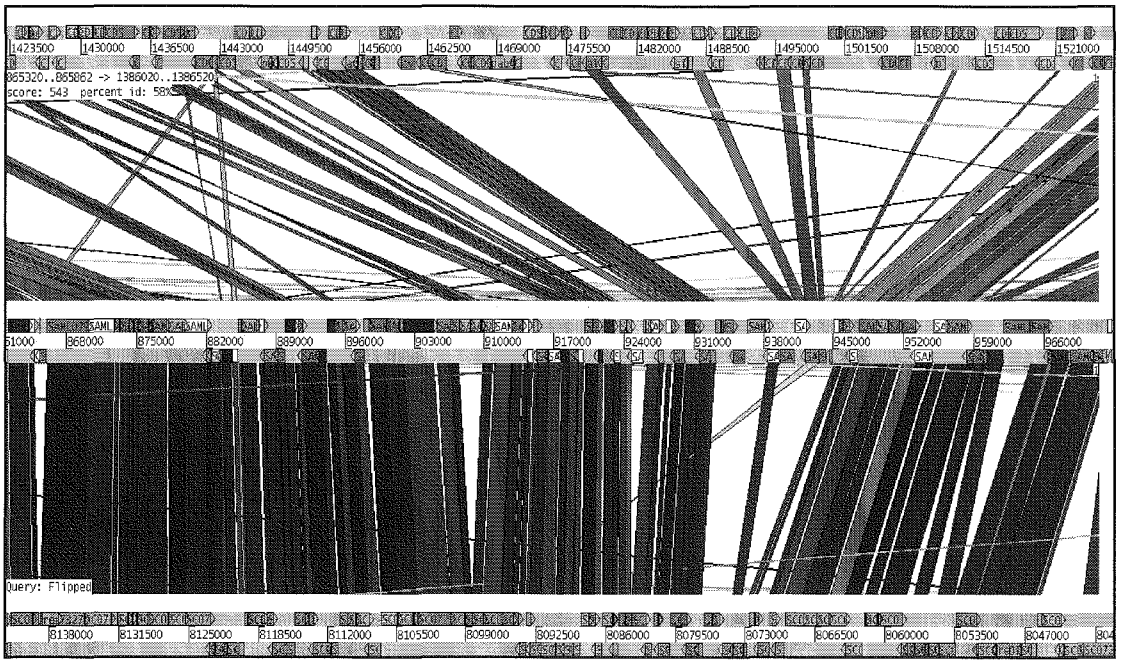


Figure 36 : Régions de synténie dégénérée entre *S. avermitilis*, *S. ambofaciens* et *S. coelicolor*.
Comparaisons des séquences protéiques des gènes visualisées à l'aide du logiciel Artemis Comparison Tools. Les gènes sont représentés par des flèches dans les six phases de lecture. Les gènes orthologues prédits entre deux génomes sont reliés par les zones grisées. **A.** Exemple d'une région de 26 kb du génome de *S. ambofaciens* incluse dans les régions de synténie dégénérée. **B.** Exemple à plus grande échelle (sur environ 100 kb incluant la région de 26 kb). *Sam* : *S. ambofaciens* ; *Sco* : *S. coelicolor* ; *Sav* : *S. avermitilis*.

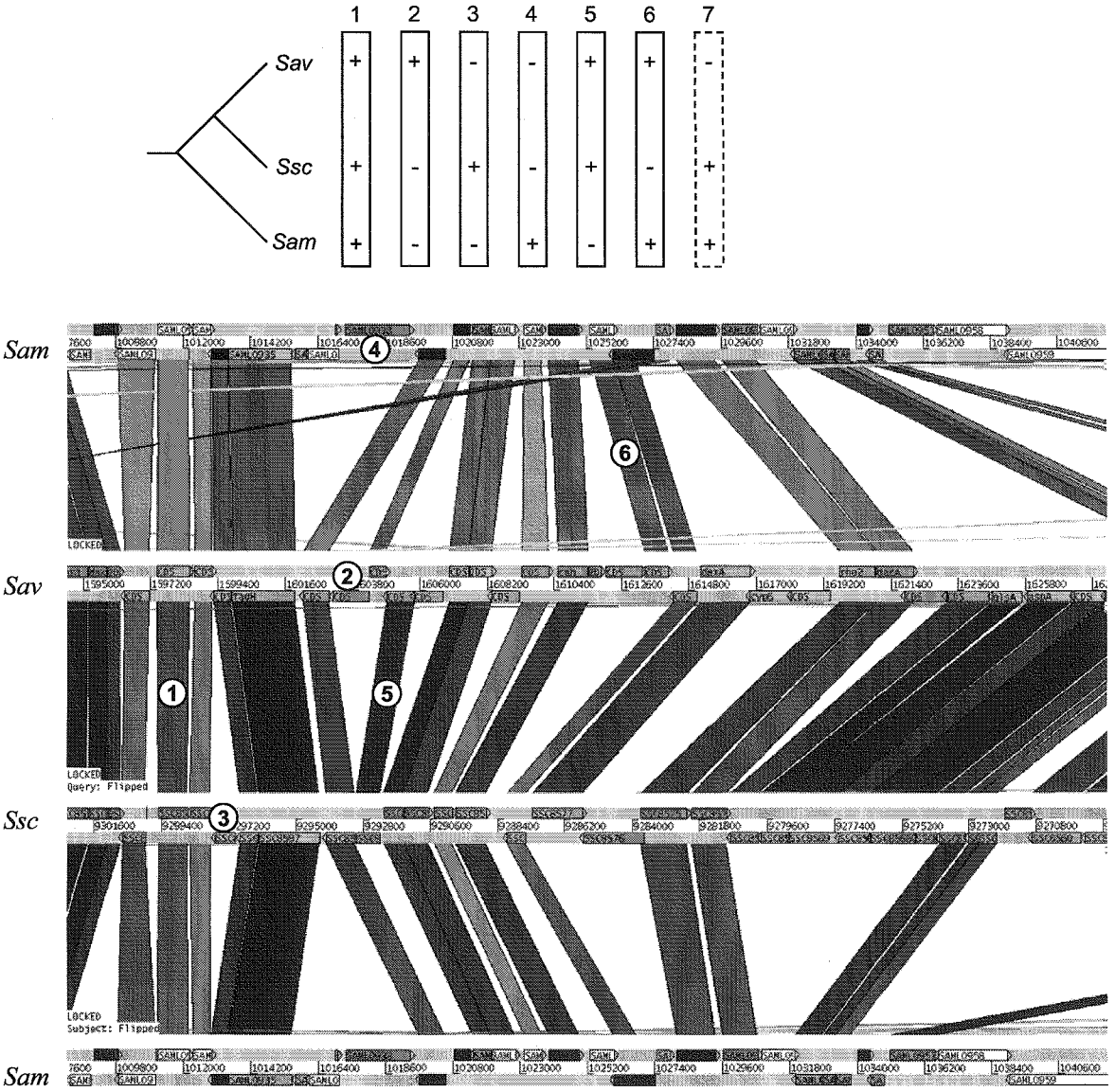


Figure 37 : Exemple de synténie dégénérée sur une région d'environ 30 kb entre *Sam* : *S. ambofaciens* ; *Ssc* : *S. scabies* et *Sav* : *S. avermitilis*.
La région de *S. ambofaciens* a été représentée en haut et en bas sur cette figure afin de pouvoir visualiser la synténie avec les deux autres chromosomes. Toutes les situations possibles de partage de gènes entre les 3 espèces sont représentées sur l'arbre (+ : gène présent, - : gène absent) et numérotées : 1. gène partagé entre les trois génomes, 2. gène spécifique de *Sav* par rapport aux deux autres, 3. gène spécifique de *Ssc*, 4. gène spécifique de *Sam*, 5. gène partagé entre *Sav* et *Ssc* mais absent chez *Sam*, 6. gène partagé entre *Sav* et *Sam* mais absent chez *Ssc*, 7. gène partagé entre *Ssc* et *Sam* mais absent chez *Sav*. Elles sont presque toutes retrouvées sur ces 30 kb. Seule la dernière n'est pas représentée sur cette région mais est retrouvée ailleurs.

Afin d'obtenir une vision sur l'ensemble du chromosome des *Streptomyces*, les comparaisons par paires des chromosomes complets de *S. coelicolor*, *S. avermitilis* et *S. scabies* ont été réalisées et l'évolution du GOC_1 a été calculée. La présence de grands îlots génomiques (plus de 30 kb) spécifiques d'espèce se traduit par une chute brutale et locale du GOC_1 (nombre d'orthologues très faible localement) qui ne reflète pas un nombre élevé d'événements (indels). Afin d'éviter ces artéfacts, une formule plus appropriée du GOC a été utilisée (GOC_2, Fig. 38). Elle calcule le nombre de paires de gènes orthologues contigus dans les deux génomes comparés divisé par le nombre de gènes

orthologues. Les courbes obtenues montrent que le niveau de synténie de la région centrale du chromosome est élevé (70% en moyenne) et stable. En revanche, le niveau de conservation diminue de façon graduelle vers les extrémités chromosomiques (de 60 à ~0%).

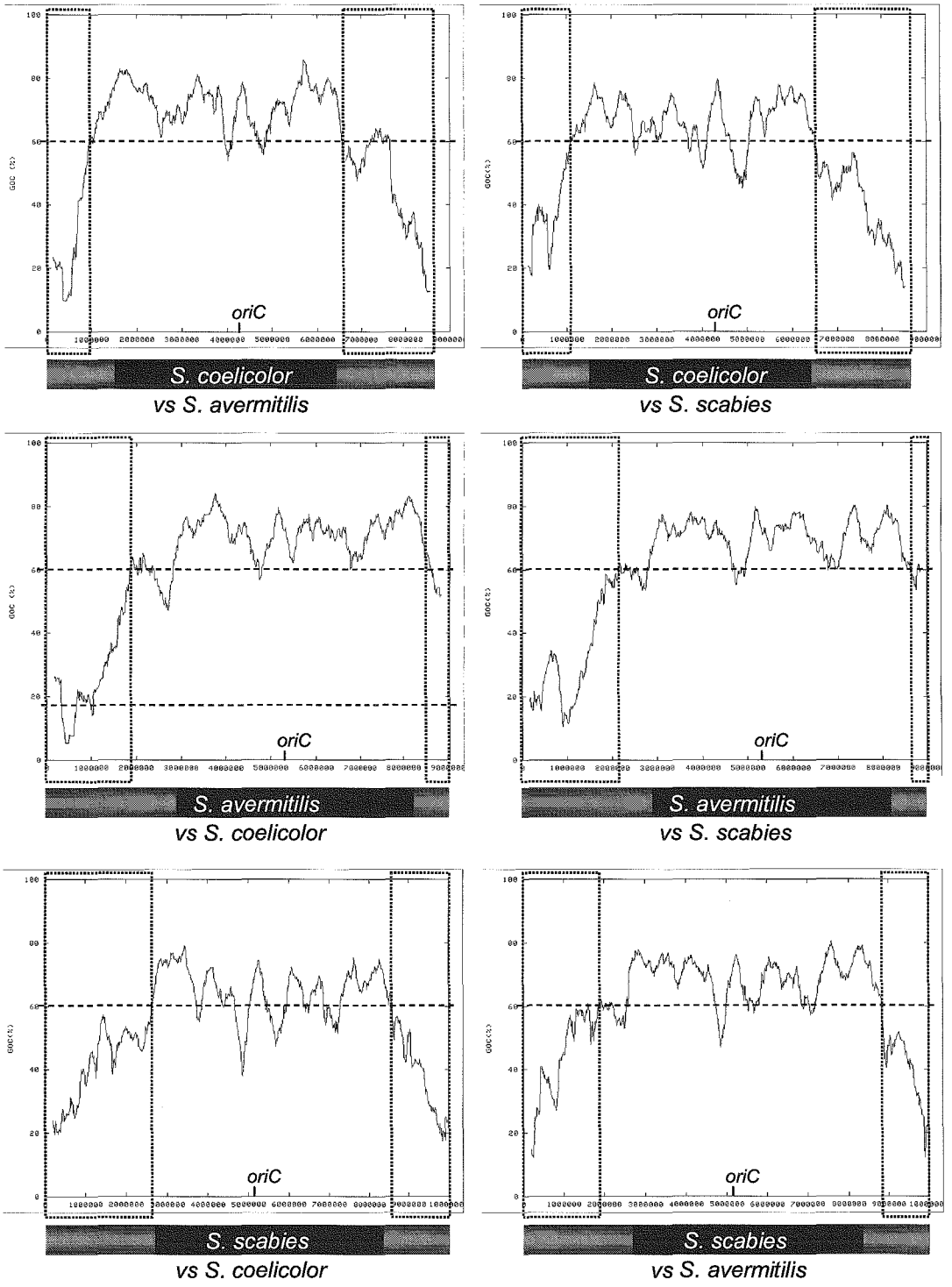


Figure 38 : Mesure du GOC_2 entre les paires de chromosomes de *Streptomyces*.
Les régions grisées représentent les régions du chromosome où la synténie diminue de façon graduelle jusqu'aux extrémités. Une valeur maximale de 60% GOC_2 a été choisie comme limite des régions de synténie dégénérée. La région "core" et les extrémités (régions de contingence) définies par Bentley *et al.* 2002 sont représentées respectivement en noir et gris. Calculs réalisés avec une fenêtre glissante de 300 gènes et un pas de 10 gènes.

Cette dégénérescence semble, en réalité, affecter l'intégralité des régions de contingence. L'augmentation du GOC_2 visible au niveau des extrémités (sur quelques centaines de kb) est un artéfact dû à un nombre de gènes orthologues quasi nul dans ces régions.

Ces résultats démontrent une dynamique particulière de l'évolution des régions terminales du chromosome chez *Streptomyces*. Deux faits majeurs permettent d'interpréter les comparaisons de génomes en terme de dynamique de l'évolution :

1. La partie centrale du chromosome conservée et partagée de façon ancestrale est d'autant plus réduite que les espèces comparées sont éloignées. Réciproquement, pour une espèce de *Streptomyces* donnée, la taille des régions terminales spécifiques est d'autant plus grande que l'espèce comparée est éloignée phylogénétiquement. Ces régions terminales spécifiques d'espèce ne sont donc pas définissables de façon statique (entre deux bornes) mais n'existent que d'un point de vue dynamique. Par conséquent, les flux de gènes n'affectent pas une région chromosomique précise mais s'étend sur l'ensemble des régions de contingence.

2. Pour une paire de génomes donnée, la dégénérescence de la synténie est graduelle, du centre vers les extrémités. Plus l'on s'approche des extrémités, plus les flux de gènes sont importants. Par saturation d'événements, les flux de gènes rendent la synténie indétectable et conduisent à l'apparition d'extrémités spécifiques d'espèce.

Pour un locus donné appartenant aux régions de synténie dégénérée, le niveau de dégénérescence est d'autant plus élevé que les espèces comparées sont éloignées phylogénétiquement (corollaire des deux premiers résultats).

Evolution of the Terminal Regions of the *Streptomyces* Linear Chromosome

Frédéric Choulet, Bertrand Aigle,* Alexandre Gallois,* Sophie Mangenot,† Claude Gerbaud,‡ Chantal Truong,† François-Xavier Francou,‡ Céline Fourrier,* Michel Guérineau,‡ Bernard Decaris,* Valérie Barbe,† Jean-Luc Pernodet,‡ and Pierre Leblond**

*Laboratoire de Génétique et Microbiologie, UMR INRA 1128, IFR 110, Université Henri Poincaré Nancy 1, Faculté des Sciences et Techniques, Vandoeuvre-lès-Nancy, France; †Génoscope, Centre National de Séquençage, Evry, France; and ‡Institut de Génétique et Microbiologie, UMR CNRS 8621, Université Paris-Sud 11, Orsay, France

Comparative analysis of the *Streptomyces* chromosome sequences, between *Streptomyces coelicolor*, *Streptomyces avermitilis*, and *Streptomyces ambofaciens* ATCC23877 (whose partial sequence is released in this study), revealed a highly compartmentalized genetic organization of their genome. Indeed, despite the presence of specific genomic islands, the central part of the chromosome appears highly syntenic. In contrast, the chromosome of each species exhibits large species-specific terminal regions (from 753 to 1,393 kb), even when considering closely related species (*S. ambofaciens* and *S. coelicolor*). Interestingly, the size of the central conserved region between species decreases as the phylogenetic distance between them increases, whereas the specific terminal fraction reciprocally increases in size. Between highly syntenic central regions and species-specific chromosomal parts, there is a notable degeneration of synteny due to frequent insertions/deletions. This reveals a massive and constant genomic flux (from lateral gene transfer and DNA rearrangements) affecting the terminal contingency regions. We speculate that a gradient of recombination rate (i.e., insertion/deletion events) toward the extremities is the force driving the exclusion of essential genes from the terminal regions (i.e., chromosome compartmentalization) and generating a fast gene turnover for strong adaptation capabilities.

Introduction

Comparisons of complete genome sequences have revealed that the level of variability in bacteria is variable. This can be related to the life style of the organism. For bacteria living in a stable environmental niche, like the intracellular pathogens, the rate of genomic variation is low compared with free-living bacteria (Mira et al. 2002). For example, no rearrangement or gene acquisition has occurred in the genome of the obligate host-associated *Buchnera aphidicola* in the past 50–70 Myr (Tamas et al. 2002). In contrast, only 39.2% of the set of proteins from 3 *Escherichia coli* strains are common to them all (Welch et al. 2002). Genome reduction by gene loss is frequent for adaptation to a stable environment (Gil et al. 2002), whereas acquisition of useful functions by lateral gene transfer (LGT) plays an important role in the evolution of free-living bacteria (Ochman et al. 2000).

Although gene content can be highly different between related organisms, the structure of the chromosome is under strong selection and is highly organized. Beneath an apparent disorder, selective pressures to maintain information sets and valuable aspects of chromosome structure restrict the rate at which diversity can be added to a genome (Lawrence and Hendrickson 2005). Many organizational features, such as gene distribution and nucleotide composition (Lobry and Louarn 2003), are related to the replication process (Rocha 2004; Boccard et al. 2005). For example, gene dosage effects would constrain the position of genes along the genome (Couturier and Rocha 2006). The terminus of replication appears as a privileged target for DNA rearrangements (Suyama and Bork 2001). Thus, the control of the level of variability is dependent on location within the genome. In addition, genome size is constrained and implies a competition between genes for their maintenance. This competition is at the origin of genomic flux (Lawrence and Roth 1999).

Streptomyces are soil bacteria belonging to the Actinomycetales order. They present a complex cell cycle characterized by both morphological and biochemical differentiation processes (Chater 1993). They exhibit a remarkable phenotypic diversity typified by the diversity of the secondary metabolites produced and are consequently of great economic interest for applications in medicine, agriculture, and biotechnology. Their chromosome is linear with a central replication origin and among the largest in bacteria, ranging from 8.7 Mb in *Streptomyces coelicolor* (Bentley et al. 2002) to 10.1 Mb in *Streptomyces scabies* (http://www.sanger.ac.uk/Projects/S_scabies/). All *Streptomyces* species studied so far have been found to be subject to a high degree of genetic instability, correlated with the formation of large rearrangements (large-scale deletions and amplifications) occurring in the terminal chromosomal regions (Leblond and Decaris 1999). The frequent loss of the terminal regions (up to 2.3 Mb in *Streptomyces ambofaciens* [Fischer et al. 1997]) in laboratory growth conditions indicates that they do not contain genes essential for vegetative growth. This organization was corroborated by the analysis of the *S. coelicolor* chromosome in which all known essential genes are located in a central "core" region (4.9 Mb), whereas the chromosomal "arms" were defined as contingency (i.e., nonessential) regions (Bentley et al. 2002). The core corresponds to the region common to both genomes of *S. coelicolor* and of the actinomycete *Mycobacterium tuberculosis* (Bentley et al. 2002).

In this article, we report the analysis of the partial genome sequence of *S. ambofaciens* focusing on comparative genomic analysis with the other available *Streptomyces* genomes: *S. coelicolor* A3(2) (Bentley et al. 2002) and *Streptomyces avermitilis* (Ikeda et al. 2003). *Streptomyces ambofaciens* and *S. coelicolor* are extremely close phylogenetically (16S rRNA divergence: 1.1%), whereas *S. avermitilis* is a more distantly related species (16S rRNA divergence from *S. ambofaciens*: 2.9%). The pairwise

Key words: comparative genomics, linear bacterial chromosome, lateral gene transfer, contingency regions, *Streptomyces*.

E-mail: leblond@nancy.inra.fr.

Mol. Biol. Evol. 23(12):2361–2369. 2006
doi:10.1093/molbev/msl108
Advance Access publication September 6, 2006

comparison of species phylogenetically closely or distantly related provides insights into the evolutionary mechanisms that shape the chromosome of *Streptomyces*.

Materials and Methods

Sequencing

Terminal inverted repeats (TIRs) were sequenced using 10 ordered cosmids constructed from partially *Bam*HI-digested *S. ambofaciens* ATCC23877 genomic DNA cloned into the Supercos1 (Stratagene, La Jolla, CA) vector. As the size of the *S. ambofaciens* TIRs (~198 kb) greatly exceeds that of a fragment readily clonable into a cosmid vector, each copy of the TIRs cannot be isolated as a single recombinant molecule. DNA libraries were constructed from partially *Sau*3A1-digested *S. ambofaciens* ATCC23877 genomic DNA cloned into the pBeloBAC11 (a derivate of pBAC108L [Shizuya et al. 1992]). A total of 4,809 recombinant bacterial artificial chromosomes (BACs) were isolated (insert average size: 37.9 ± 9.8 kb) representing a 21× covering rate. A systematic sequencing of the extremities of each of the BAC inserts was performed leading to 0.4× coverage of the complete chromosome. Finally, 88 ordered BACs and 10 cosmids were sequenced to cover both terminal regions. For sequencing, BACs and cosmids were mechanically fragmented and cloned with a *Bst*XI adaptor into either pcDNA2.1 vector (Invitrogen, Carlsbad, CA) or pCNS (Bartolome et al. 1991) and ligation products were then introduced into *E. coli* DH10B.

Annotation

Each BAC sequence was assembled using the Phred–Phrap–Consed suite (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998), and bl2seq (Altschul et al. 1997) was finally used for the overall assembly. The gene finder Glimmer2.10 (Delcher et al. 1999), trained with 3,000 coding DNA sequences (CDSs) of *S. coelicolor*, were used for CDS prediction. Results were then refined using the RBSfinder tool (Suzek et al. 2001). A potential ribosome binding site (RBS) was considered when 4 of the consensus bases 5′-GGAGG-3′ were detected upstream from the start codon (Strohl 1992). Blast 2.2.6 was used to find similarities (Altschul et al. 1997). The Interpro package was also used to describe protein domains (Zdobnov and Apweiler 2001). CDSs were assigned a functional category where their best homolog in the clusters of orthologous groups of proteins is classified (Tatusov et al. 2001). The choice of the start codon was guided by the presence of a RBS, Blast results, and the G + C frameplot pattern (Ishikawa and Hotta 1999). The most upstream start codon, minimizing overlap with the previous gene, was chosen when the situation was not clear. An annotation platform was developed using Perl scripts and the Bioperl library (Stajich et al. 2002) to deal with program outputs and sequence manipulation. A manual validation of each CDS was performed using Artemis (Rutherford et al. 2000). BlastP alignments were performed against the Non Redundant database and also against each individual *Streptomyces* proteome for comparative genomics. While comparing the predicted protein sequences with BlastP, those sharing more than 30% identity over at least 80% of the protein length were considered homologues; 2 proteins were regarded as orthologues if they are reciprocal best hits according to these criteria. Pseudogenes were identified by comparisons with their functional counterparts. A sequence was considered as a pseudogene when a coding DNA sequence has been inactivated by nonsense mutations, frameshifts, truncations, or a combination of these mechanisms. A relational database, SAMDB, is integrated into the platform to organize annotation and comparative genomics data. To visualize the degenerated synteny, protein sequence comparisons were extracted from SAMDB to be readable under ACT (Rutherford et al. 2000). Duplicated genes in the TIRs were named SAM**T**nnnn, whereas those specific only either to the left or right arms were annotated as SAM**L**nnnn and SAM**R**nnnn.

Genome Comparison using BAC End Sequences

A total of 8,457 BESs (average size: 417 ± 122 nt) were obtained from systematic sequencing of each BAC insert extremity. It resulted in 0.4× covering rate of the chromosome (~8.5 Mb), which represents approximately 1 BAC end sequence (BES) every kilobase. BlastN of each BES against the *S. coelicolor* chromosome was performed when both ends of a BAC were available and a relational database was developed in order to store the resulting data and to be able to align the *S. ambofaciens* BACs on the *S. coelicolor* central region (see fig. 1). This led us to localize the species-specific regions within the chromosome as explained in figure 1 legend.

Gene Order Conservation

For each pairwise comparison, the level of gene order conservation (GOC) was estimated along a chromosome, using a sliding window (100 CDSs with 5 CDS steps), by calculating the number of pairs of orthologues that are contiguous in the 2 compared chromosomes divided by the number of genes in the window (fig. 2). The GOC profile of the whole chromosome of *S. coelicolor* and *S. avermitilis* compared with each other (Supplementary Material online) was estimated by calculating the number of pairs of orthologues that are contiguous in the 2 chromosomes divided by the total number of orthologues in the window as defined by Rocha (2006).

The *S. ambofaciens* chromosomal arm annotation is available through the SAMDB web server on http://www.weblgm.scbiol.uhp-nancy.fr/ambofaciens/. The sequences were deposited in European Molecular Biology Laboratory under the AM238663 (left arm) and AM238664 (right arm) accession numbers.

Results

Conservation of the Central Chromosomal Region among the *Streptomyces* Genus

The terminal regions of the *S. ambofaciens* chromosome were completely sequenced over 1,544 kb and

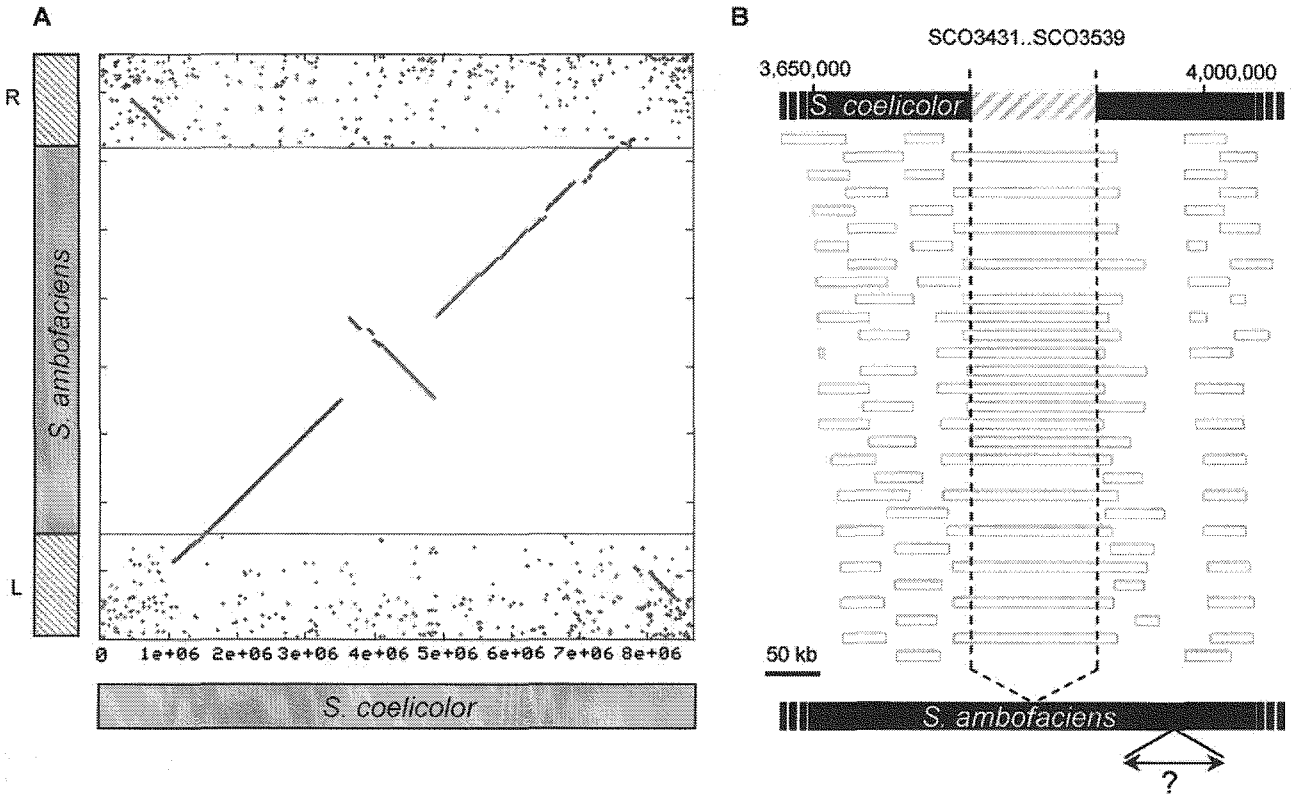


FIG. 1.—Pairwise comparison of *Streptomyces ambofaciens* and *Streptomyces coelicolor* chromosome. (*A*) Dot plot comparison of *S. ambofaciens* and *S. coelicolor*. For *S. ambofaciens*, hatched rectangles represent the completely sequenced extremities (labeled L and R). Each dot included in these extremities represents the best BlastP hit of each *S. ambofaciens* protein compared with *S. coelicolor*. The central synteny was designed using BlastN comparison of BES with *S. coelicolor* resulting in an alignment of the central chromosomal part as illustrated in *B*. For a 2-dimensional representation of this alignment, the following criteria were used: conserved regions detected with BAC alignments are considered to have the same size in both species; because the exact size of the *S. ambofaciens* central–specific regions cannot be assessed, gaps of 80 kb (i.e., the maximum insert size) were created in the alignment. (*B*) Alignment of *S. ambofaciens* BACs on the *S. coelicolor* chromosome. Each rectangle represents 1 BAC insert according to BlastN results of each BES against the *S. coelicolor* chromosome. When the distance between 2 BES matches is compatible with the length of a BAC insert (between 25 and 80 kb), the corresponding region is considered to be conserved between both species. In contrast, the set of BACs represented with a larger size reveals the presence of a *S. coelicolor*–specific region (in that case: SCO3431–SCO3539) absent in *S. ambofaciens*. Finally, the *S. ambofaciens*–specific regions can be localized by gaps in the alignment (e.g., region marked with a question mark); however, its size cannot be assessed precisely. Such gaps can also correspond to rearrangement loci, for example, inversion.

1,367 kb for the left and right extremities, respectively, and contain 2,532 CDSs, including 43 pseudogenes.

The central region was sequenced at a 0.4× covering rate using a BAC end sequencing approach (see Materials and Methods). For the comparison of the central region of the *S. ambofaciens* chromosome with *S. coelicolor*, the BESs were aligned on the *S. coelicolor* chromosome using BlastN analyses (fig. 1). This analysis reveals a high level of synteny between the central parts of the chromosome of these 2 species. Chromosome structure comparisons reveal the occurrence of 2 inversion events centered on the origin of replication, detectable as broken X patterns in the data (fig. 1*A*). They correspond to 2 out of the 4 inversions previously reported in the *S. coelicolor*/*S. avermitilis* genome comparison (Ikeda et al. 2003). Most (6/8) of these inversion points cannot be precisely localized because they are located in species-specific regions. For example, a break point falls near the pSAM2-like element (Sezonov et al. 1998) adjacent to the calcium-dependent antibiotic (*cda*) cluster in *S. coelicolor*. It thus seems that the inversion events are more likely to be fixed within regions where rearrangements are counterselected less efficiently. Further, in order to assess the level of identity between central homologous genes, 3 independent BACs from the *S. ambofaciens* central chromosomal region were arbitrarily chosen and sequenced in their entirety (91 genes; 117 kb). All predicted genes share a conserved organization and a high level of identity with the other *Streptomyces* genomes following the phylogenetic relationships (89% and 82% of amino acid identity with *S. coelicolor* and *S. avermitilis*, respectively; 90% and 84%, respectively, at the nucleotide level; data not shown) inferred by rDNA sequence analysis (http://avermitilis.ls.kitasato-u.ac.jp/tree2.html). A single synteny break is observed among the 117 kb between *S. ambofaciens* and *S. coelicolor*. Dot plots comparing homologous proteins of the 2 other *Streptomyces* reveal that the central region is highly conserved even at long evolutionary distances as shown in Ikeda et al. (2003). The synteny includes the central part, whereas the replicon extremities appear species specific.

Although the level of synteny is high in the central regions, the insertion of several specific genomic islands

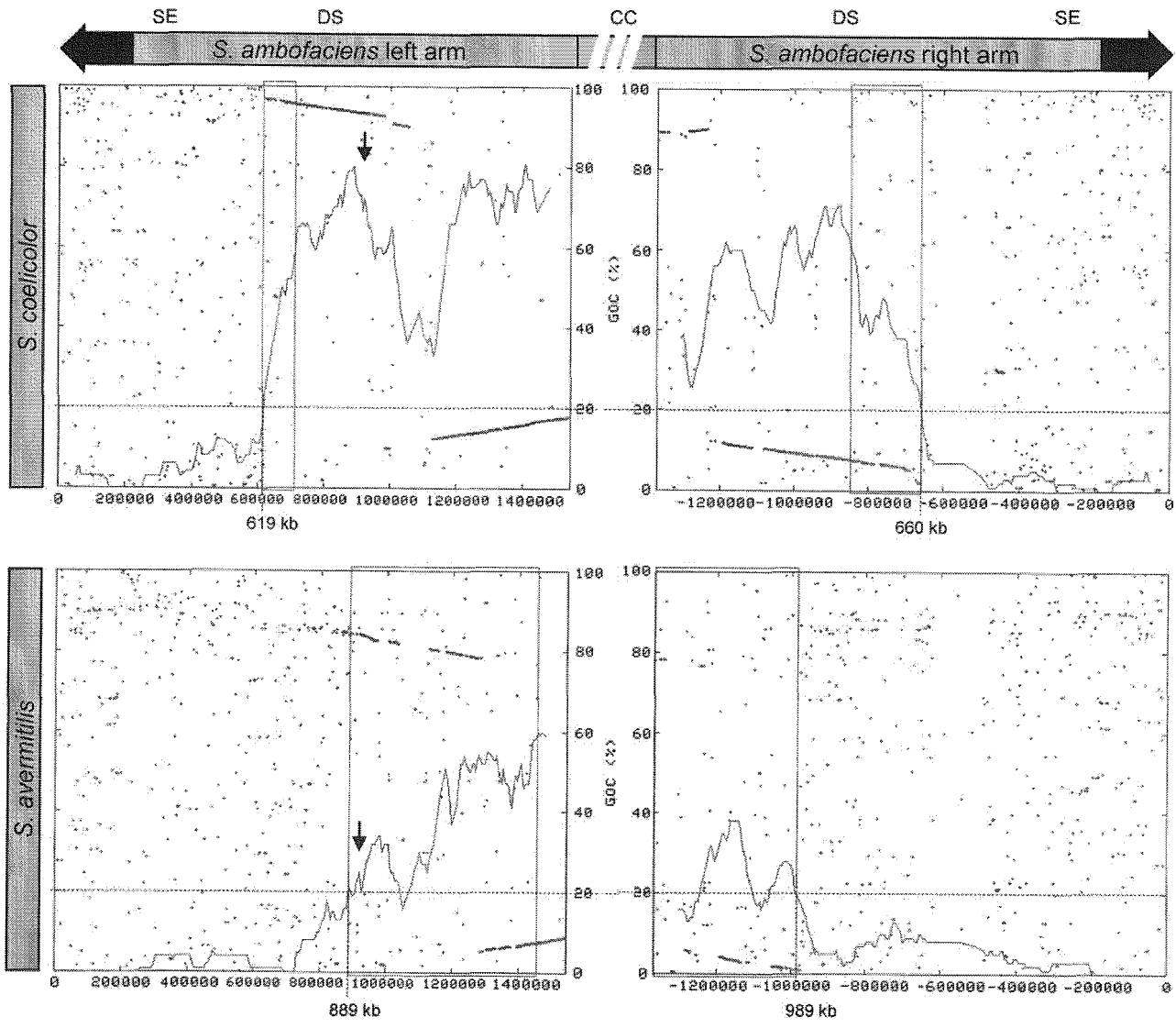


FIG. 2.—Pairwise comparison of the *Streptomyces ambofaciens* chromosomal arms with the 2 other chromosomes. Profile of GOC and protein dot plot are superimposed for a given pairwise comparison. GOC values correspond to the percent of pairs of orthologues that are contiguous in the *S. ambofaciens* genome and in the compared genome (see Materials and Methods). GOC value (*y* coordinate) was calculated using a sliding window (100 CDSs with 5 CDS steps); *x* coordinate corresponds to the position of the 50th CDS in the window. A threshold of 20% GOC (dotted horizontal lines) was chosen to delimit the end of specific extremities for each pairwise comparison, the size of which is indicated under the *x* axis. Framed areas represent the degenerated syntenic regions of *S. ambofaciens*, the location of which varies according to species compared (a threshold of 60% GOC was chosen for their internal limit). Small arrows indicate the region of *S. ambofaciens* zoomed in figure 3. Black arrows at the end of the *S. ambofaciens* chromosome represent the TIRs. SE: specific extremity; DS: degenerated syntenic part; CC: conserved central part.

has occurred. The *S. ambofaciens*/*S. coelicolor* comparison using BAC alignments (fig. 1*B*) revealed 4 clusters (larger than the maximum length of a BAC insert, i.e., circa 80 kb) specific to *S. ambofaciens*. One of them corresponds to the spiramycin antibiotic gene cluster (Richardson et al. 1990). Reciprocally, 12 *S. coelicolor*–specific regions whose sizes vary from 26 to 149 kb are located in the essential core region (Supplementary Material online). Bentley et al. (2002) defined 14 regions potentially recently laterally acquired in the *S. coelicolor* chromosome according to the G/C and gene contents. The present analysis confirms that half of them (7/14) could have a recent origin in *S. coelicolor* because they are absent in the *S. ambofaciens* chromosome, although one cannot exclude the possibility that they were once present in the *S. ambofaciens* lineage but have since been lost. They include the actinorhodin (Malpartida and Hopwood 1986), undecylprodigiosin (Rudd and Hopwood 1980), and *cda* (Wright and Hopwood 1976) gene clusters and a 41-kb region (from SCO3677 to SCO3725 including genes relevant to resistance to heavy metals) adjacent to a tRNA gene. Although tRNA genes are known targets for integrative elements, no other specific signature could be identified within this genomic island. The 7 other regions putatively recently laterally acquired have a small size (about 10 kb), and the BES analysis reveals that at least 2 of them are present in the *S. ambofaciens* chromosome.

The Size of the Terminal Species-Specific Regions Increases as the Phylogenetic Distance between Compared Species Increases

The confinement of variability at the chromosomal extremities seems a general trait of the *Streptomyces* genome. Indeed, when a given species is compared with the other 2, the minimal size of the specific information can be deduced. Thus, the specific extremities cover 1,279 kb (619 + 660 kb) in *S. ambofaciens* as illustrated in figure 2. Considering the absence of a drastic drop in synteny, the limits of the specific terminal regions were defined using a threshold of 20% of GOC for each pairwise comparison (see Materials and Methods for GOC calculation). Interestingly, the size of the species-specific regions increases with the phylogenetic distance. This is shown in figure 2, where the size of the *S. ambofaciens*–specific extremities increases from 1,279 kb (619 + 660 kb) compared with *S. coelicolor* to 1,878 kb (889 + 989 kb) compared with *S. avermitilis*. Reciprocally, the size of the central conserved part decreases with the distance. These data suggest that the specific information preferentially accumulates in the terminal regions along the evolutionary time.

When all pairwise comparisons of the terminal regions of the 2 other *Streptomyces* were carried out using the same approach (GOC calculation with the same parameters), the minimal size of the species-specific extremities was estimated to be 753 kb in *S. coelicolor* and 1,393 kb in *S. avermitilis*. Enrichment of the terminal ends in specific information is highlighted by the fact that, for example, the 753 kb specific to *S. coelicolor* represents only 9% of the whole chromosome but corresponds to 41% of the total of the regions estimated to be absent in *S. ambofaciens*.

Level of Variability of the Terminal Species-Specific Regions

Figure 2 shows the location of the ends of the synteny between *S. ambofaciens* chromosomal arms and the other genomes. The *S. ambofaciens*–specific chromosomal extremities (619 + 660 kb) are characterized by a very low level of GOC. Although conserved clusters could be observed, they are limited to very small regions (generally less than 8 genes), whatever the phylogenetic relationship considered. These terminal species-specific regions include 1,082 CDSs in *S. ambofaciens* of which only 9% share more than 60% end-to-end identity in the 2 other *Streptomyces*. When compared with its close relative *S. coelicolor*, 37% of the 1,082 CDSs are absent. Only 13% are highly conserved (>80% of identity). This contrasts with the conservation estimated in the central part, where 100% of the CDSs identified (over 117 kb) share more than 80% of identity (average: 89%). The level of variability is higher with *S. avermitilis* with 44% of the 1,082 CDSs that do not share any similarity and only 4% that are highly conserved.

The low levels of conservation (synteny and identity) between homologues strongly suggest that the majority of them are not orthologues but rather result from the massive introduction of foreign alleles (xenologues) by LGT. Thus, the specificity of the extremities probably originates from LGT. Our data reveal that the level of terminal variability is extremely high even between closely related species extending the preliminary conclusions resulting from the comparison of the *S. coelicolor* and *S. avermitilis* genome sequences (Ikeda et al. 2003).

The origins of the terminal specificity could also result from the presence of mobile genetic elements. Indeed, the terminal regions are enriched in mobile elements (insertion sequence–like elements) as previously reported (Chen et al. 2002) and proposed to mediate DNA rearrangements and integration of horizontally transferred DNA sequences. In the *S. ambofaciens* chromosomal arms, 53 transposase-encoding genes (including pseudogenes) were predicted with a strong bias in the terminal species-specific regions (50 of 53). Furthermore, some terminal CDSs show highest similarity to plasmid-associated genes (Choulet et al. 2006). In *S. ambofaciens*, 4 homologues to plasmid-associated genes lie in the terminal 50 kb, for example, the helicase-encoding gene *ttrA*, which is conserved in the 2 other species but which shares best similarity with its homologue of the plasmid SLP2 of *Streptomyces lividans* (Huang et al. 2003). Three CDSs similar to genes encoding plasmid transfer functions (*spdB*, *traB2*, and *traA2*) are also detected in the *S. avermitilis* chromosomal extremities. In *S. ambofaciens*, this observation is correlated with an average GC content (68.8% in the 50 terminal kb) that is slightly lower than that of the rest of the genome (72.3%). A lower GC content is typical from *Streptomyces* linear plasmids (Spatz et al. 2002; Huang et al. 2003; Ikeda et al. 2003; Bentley et al. 2004) and more generally from mobile elements.

Degenerated Syntenic Regions and Massive Gene Flux

Synteny analysis of the 117 kb sequenced in the center of the *S. ambofaciens* chromosome revealed a high GOC with *S. coelicolor*, and a similar level of synteny is described when compared with *S. avermitilis*. Interestingly, the synteny observed between central regions degenerates progressively over several hundreds of kilobases before reaching the terminal species-specific regions (framed areas in fig. 2). This is true whatever the pairs of species considered. In these regions, the synteny appears as gradually parceled out by multiple insertions/deletions (indels) of genes. Degeneration of synteny not only reflects an increase in the level of rearrangement of endogenous information but also reflects a rapid evolution of the gene content by gene flux, that is, by accumulation of new genes and loss of ancestral information. When the GOC level falls under 20%, the synteny becomes undetectable and the regions were consequently considered as specific. Figure 3 illustrates this degeneration phenomenon using a protein-to-protein comparison between *S. ambofaciens*, *S. coelicolor*, and *S. avermitilis* and shows that degeneration follows the phylogenetic distances.

More significantly, the number of synteny breaks (expressing the level of degeneration) was estimated over 100 kb (99 CDSs) including the locus detailed in figure 3, of the *S. ambofaciens* left arm (from SAML0798 to SAML0896). Compared with *S. coelicolor*, 10 rearrangements (from SCO7238 to SCO7327) are observed, whereas at least 29 events have led to the current genome divergence

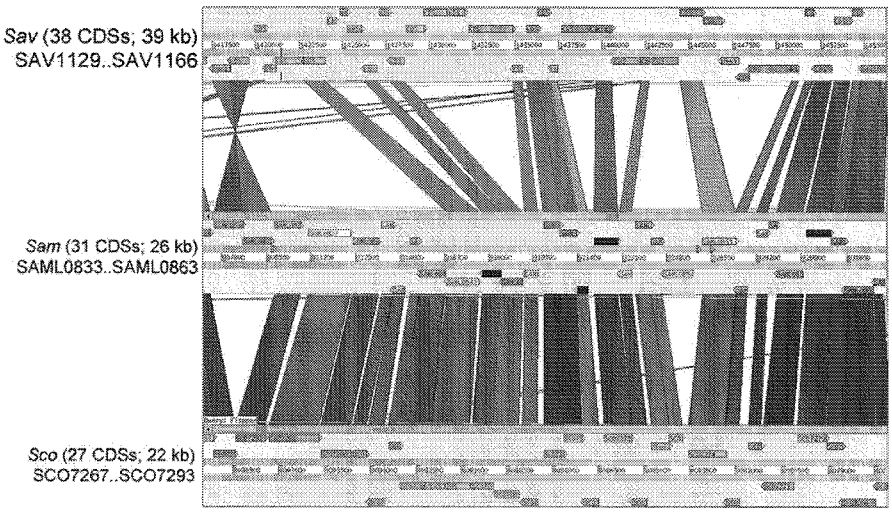


FIG. 3.—Protein-to-protein comparisons displayed with Artemis Comparison Tool (Rutherford et al. 2000). CDSs are represented by arrows in the 6 frames, and pairs of homologues are linked by a gray area. Example of a 26-kb region of the *Streptomyces ambofaciens* (Sam) left arm (from position 907,000–933,000) included into the degenerated syntenic regions compared with *Streptomyces coelicolor* (Sco) and *Streptomyces avermitilis* (Sav) (see arrows in fig. 2). As the sizes of the compared regions are not identical in the 3 species, the scales are different to fit in window.

with *S. avermitilis* (from SAV1085 to SAV1268) (in contrast, 1 and 4 synteny breaks were observed along the 117 kb of central sequences, respectively). Correlated to the decreasing GOC, the proportion of remaining orthologous genes also varies according to the phylogenetic distance (83 and 52 of 99 *S. ambofaciens* CDSs are orthologues with *S. coelicolor* and *S. avermitilis*, respectively), highlighting the importance of gene flux in these regions. However, at long phylogenetic distances, the number of indels probably becomes underestimated. Indeed, assuming an indel size of 1–10 genes (estimated with the *S. ambofaciens*/*S. coelicolor* comparison), a synteny break probably masks multiple indel events. Obviously, some indel events can involve more than 10 genes (e.g., operonic structure or gene clusters such as those involved in secondary metabolite biosynthesis), the insertion of the whole set of genes being necessary for the achievement of a selectable function. Conversely, the loss of one of these genes would favor the loss of the complete gene set. This situation was indeed observed in these degenerated synteny regions. For example, in *S. ambofaciens*, a 34.5-kb cluster (28 genes) implied in secondary metabolism is species specific. In *S. coelicolor*, this region is replaced by 30.8-kb–specific cluster (31 genes, SCO0850–SCO0880). In *S. avermitilis*, the same region (SAV7356–SAV7422) corresponds to a secondary metabolite gene cluster (*pks1*, [Ikeda et al. 2003]) different from the 2 other species. Such large genomic islands result in a drop of the GOC profile, for example, from position 1,050,000–1,125,000 of the *S. ambofaciens* left arm (fig. 2). The *S. ambofaciens* right arm exhibits higher variability than the left arm. Large specific islands inserted in the terminal regions result in falls of the GOC profile.

As shown by GOC variation (fig. 2), the regions affected by the degeneration are more internal when distant *Streptomyces* are compared. Altogether, these data suggest that the terminal regions are prone to a massive and constant gene flux. Strikingly, the profile of GOC increases in a gradual way from the terminal specific regions to the internal conserved part. The gradient of degeneration toward the extremities extends over 600 kb of the left arm of *S. ambofaciens* when it is compared with *S. avermitilis* (fig. 2). The slope of GOC variation is stronger when compared with *S. coelicolor*, but the gradient of degeneration remains detectable using a smaller sliding window as well as by dot plot comparison (data not shown).

When the whole chromosome of *S. coelicolor* and *S. avermitilis* are compared with each other, GOC profile reveals that the degenerated syntenic regions extend over the entire contingency regions as defined by Bentley et al. (2002) (Supplementary Material online). Although the GOC is high and constant in the central region, it gradually decreases toward the extremities.

Discussion

Comparative genomics within the *Streptomyces* genus has revealed a highly compartmentalized structure where variability is mostly confined to the extremities of the linear chromosome. Indeed, the core region defined by the presence of essential genes (Bentley et al. 2002) appears to be highly syntenic throughout the genus, whereas the chromosomal arms, which contain contingency genes, contain highly variable and species-specific genes. In addition, a high level of chromosome instability has long been known through genomic characterization of mutants showing terminal rearrangements (Fischer et al. 1998; Wenner et al. 2003).

Therefore, the high level of genomic diversity would occur in a highly organized structure where gene maintenance might be dependent on chromosomal location. Thus, a part of variability probably results from exchange of extremities between linear replicons (plasmids or chromosomes). Such events have already been observed in *Streptomyces* (Pandza et al. 1998; Yamasaki and Kinashi 2004) as well as in *Borrelia*

(Casjens et al. 1997, 2000). This hypothesis is supported by the identification of genes similar to plasmid-associated genes proximal to the chromosomal ends. Numerous linear conjugative plasmids are known in *Streptomyces* (Hopwood 1999), which could self-mobilize and/or mobilize chromosomal regions. The "end first" model of mobilization proposed by Chen (1996), that is, the transfer of linear replicon from one end, could explain why terminal regions might be favored during conjugational events. The transfer of terminal information could lead to replacement of the whole chromosomal end as suggested by the analysis of strain-specific regions at the end of the *S. ambofaciens* TIRs (Choulet et al. 2006). This type of fast evolution mechanism could be an advantage conferred by chromosomal linearity.

Comparative genomics revealed that the frontier between the specific and the conserved regions corresponds to a region of degenerated synteny (figs. 2 and 3). This phenomenon is observable for each pairwise comparison, and the level of degeneration is also correlated to the phylogenetic distance. This type of genome divergence cannot result from exchange of replicon extremities. These data rather reveal that *Streptomyces* terminal regions are subject to a massive and constant gene flux occurring as a result of insertions, deletions, and replacements. These events accumulate with time and gradually erase the GOC. Because the frontier between the specific and the conserved regions is not the same when considering different pairs of species, this degeneration does not seem to be locus specific but would rather affect the whole contingency regions. The earlier 2 *Streptomyces* species diverge, the shorter the conserved region. Reciprocally, the size of the species-specific regions increases with the phylogenetic distance.

Lawrence and Roth suggested that the acquisition of foreign genes contributing to cell fitness results in the loss of resident functions of lower selective value, that is, contributing weakly or not at all to the fitness (Lawrence and Roth 1999). This model applies because multiple constraints limit genome-size expansion. A minimal selective advantage, the s value corresponding to the maintenance threshold, is required for maintenance of a gene in a genome. Indeed, genes with an adaptative value under the threshold s can be either fixed or lost in a context of low or high mutation rate, respectively. In the context of a high rate of rearrangements (e.g., deletion), a stronger selective coefficient is required for a gene to be maintained. Conversely, genes located in regions of high recombination rate must have a higher adaptative value to be maintained.

Interestingly, analysis of the *Streptomyces* genomes supports a region-dependent selection pressure. In other words, we proposed that the minimal s value required for maintenance would be dependent on the chromosomal location. As the rearrangement rate (i.e., deletion rate) gradually increases while approaching the chromosomal ends, the threshold for gene maintenance and, consequently, the frequency of gene loss also gradually increases.

An alternative hypothesis would be that rearrangements preferentially occur in the nonessential regions because they are tolerated better (Chen et al. 2002). In other words, because terminal regions lack essential genes, the terminal variability would reflect a higher frequency of fixation of mutations instead of a higher frequency of rearrangements. The tendency of essential genes to be located in the central region would be the result of gene dosage effect, that is, highly expressed genes would tend to be close to the origin of replication in fast-growing bacteria as shown by Couturier and Rocha (2006). However, *Streptomyces* are slow-growing bacteria, and highly expressed genes are also present in the terminal regions, although enrichment can be noticed in the central part (Wu et al. 2005). According to this hypothesis, the gradual degeneration of the synteny would reveal that the genes are distributed according to their fitness contribution.

The increasing level of rearrangements toward the extremities may be the force driving the compartmentalization that excludes essential genes from the extremities and generates a high rate of gene flux for adaptation capabilities. Thus, the fact that the terminal regions are more tolerant to rearrangements would be true, but it would only be a consequence of the particular organization driven by a variable level of instability along the chromosome. These higher recombination frequencies could result from formation of double strand breaks (DSBs) generated by arrests of the replication fork. DSBs initiated during termination of replication were demonstrated to stimulate genetic instability in *E. coli* (Michel et al. 1997). It can also be speculated that conjugational mechanisms favor the introduction of replicon extremities into the recipient and/or the formation of DSBs in the chromosome of the donor and may thus stimulate DNA recombination in the terminal regions. The presence of multiple mobile genetic elements could also account for a high frequency of DNA breaks in the terminal regions (Gunes et al. 1999; Chen et al. 2002). These phenomena might constitute an advantage conferred by chromosomal linearity to acquired and rearranged DNA in these regions.

The core region of the *Streptomyces* genome is syntenic with the whole chromosome of *M. tuberculosis*. *Mycobacterium tuberculosis* belongs to the actinomycetes and possesses a circular chromosome. A parsimonious hypothesis about the evolution of the *Streptomyces* chromosome would be that a single event resulted in the acquisition of the contingency regions and chromosomal extremities by the ancestral chromosome (integration of a linear replicon within the ancestral chromosome) (Volff and Altenbuchner 2000). This hypothesis does not explain how the current terminal regions diverged from their original ancestral version. Our analysis has shed light on these mechanisms, proposing a gradient of chromosome instability (i.e., indel frequency) toward the chromosomal extremities that generate a gradient of selection pressure that eliminates the genes contributing weakly to the cell fitness. Hence, the closer a gene is to the end of the linear chromosome, the lower is its probability of being maintained.

Supplementary Material

1) A table outlining the locations and features of *S. coelicolor* regions absent in the *S. ambofaciens* genome and 2) GOC profiles of the *S. coelicolor* and *S. avermitilis* chromosomes compared with each other are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

F.C. and A.G. were recipients of a grant from the "Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche" (M.E.N.E.S.R.). This research was supported by a Programme d'Actions Intégrées (ALLIANCE), the "ACI Microbiologie 2003" program (funded by M.E.N.E.S.R.), and the VIth PCRDT ("ActinoGen"). Many thanks are due to K. Chater, G. Chandra, and T. Kieser (John Innes Centre, Norwich, United Kingdom) for their warm welcome and their help in the development of the computational methods. We are grateful to A. Hesketh (John Innes Centre, Norwich, United Kingdom) for critical reading of the manuscript.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Bartolome B, Jubete Y, Martinez E, de la Cruz F. 1991. Construction and properties of a family of pACYC184-derived cloning vectors compatible with pBR322 and its derivatives. Gene. 102:75–78.

Bentley SD, Brown S, Murphy LD, et al. (18 co-authors). 2004. SCP1, a 356,023 bp linear plasmid adapted to the ecology and developmental biology of its host, *Streptomyces coelicolor* A3(2). Mol Microbiol. 51:1615–1628.

Bentley SD, Chater KF, Cerdeno-Tarraga AM, et al. (43 co-authors). 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature. 417:141–147.

Boccard F, Esnault E, Valens M. 2005. Spatial arrangement and macrodomain organization of bacterial chromosomes. Mol Microbiol. 57:9–16.

Casjens S, Murphy M, DeLange M, Sampson L, van Vugt R, Huang WM. 1997. Telomeres of the linear chromosomes of Lyme disease spirochaetes: nucleotide sequence and possible exchange with linear plasmid telomeres. Mol Microbiol. 26:581–596.

Casjens S, Palmer N, van Vugt R, et al. (15 co-authors). 2000. A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. Mol Microbiol. 35: 490–516.

Chater KF. 1993. Genetics of differentiation in *Streptomyces*. Annu Rev Microbiol. 47:685–713.

Chen CW. 1996. Complications and implications of linear bacterial chromosomes. Trends Genet. 12:192–196.

Chen CW, Huang CH, Lee HH, Tsai HH, Kirby R. 2002. Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. Trends Genet. 18:522–529.

Choulet F, Gallois A, Aigle B, et al. (14 co-authors). 2006. Intraspecific variability of the terminal inverted repeats of the linear chromosome of *Streptomyces ambofaciens*. J Bacteriol. 188:6599–6610.

Couturier E, Rocha EP. 2006. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol. 59: 1506–1518.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27:4636–4641.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186–194.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8:175–185.

Fischer G, Decaris B, Leblond P. 1997. Occurrence of deletions, associated with genetic instability in *Streptomyces ambofaciens*, is independent of the linearity of the chromosomal DNA. J Bacteriol. 179:4553–4558.

Fischer G, Wenner T, Decaris B, Leblond P. 1998. Chromosomal arm replacement generates a high level of intraspecific polymorphism in the terminal inverted repeats of the linear chromosomal DNA of *Streptomyces ambofaciens*. Proc Natl Acad Sci USA. 95:14296–14301.

Gil R, Sabater-Munoz B, Latorre A, Silva FJ, Moya A. 2002. Extreme genome reduction in *Buchnera* spp.: toward the minimal genome needed for symbiotic life. Proc Natl Acad Sci USA. 99:4454–4458.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8:195–202.

Gunes G, Smith B, Dyson P. 1999. Genetic instability associated with insertion of IS6100 into one end of the *Streptomyces lividans* chromosome. Microbiology. 145:2203–2208.

Hopwood DA. 1999. Forty years of genetics with *Streptomyces*: from in vivo through in vitro to in silico. Microbiology. 145: 2183–2202.

Huang CH, Chen CY, Tsai HH, Chen C, Lin YS, Chen CW. 2003. Linear plasmid SLP2 of *Streptomyces lividans* is a composite replicon. Mol Microbiol. 47:1563–1576.

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

Ishikawa J, Hotta K. 1999. FramePlot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial DNA with a high G + C content. FEMS Microbiol Lett. 174:251–253.

Lawrence JG, Hendrickson H. 2005. Genome evolution in bacteria: order beneath chaos. Curr Opin Microbiol. 8:572–578.

Lawrence JG, Roth JR. 1999. Genomic flux: genome evolution by gene loss and acquisition in bacterial genomes. In: Charlebois RL, editor. Organization of the prokaryotic genome. Washington (DC): American Society for Microbiology. p. 263–289.

Leblond P, Decaris B. 1999. 'Unstable' linear chromosomes: the case of *Streptomyces*. In: Charlebois RL, editor. Organization of the prokaryotic genome. Washington (DC): American Society for Microbiology. p. 235–261.

Lobry JR, Louarn JM. 2003. Polarisation of prokaryotic chromosomes. Curr Opin Microbiol. 6:101–108.

Malpartida F, Hopwood DA. 1986. Physical and genetic characterisation of the gene cluster for the antibiotic actinorhodin in *Streptomyces coelicolor* A3(2). Mol Gen Genet. 205:66–73.

Michel B, Ehrlich SD, Uzest M. 1997. DNA double-strand breaks caused by replication arrest. Embo J. 16:430–438.

Mira A, Klasson L, Andersson SG. 2002. Microbial genome evolution: sources of variability. Curr Opin Microbiol. 5:506–512.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405: 299–304.

Pandza S, Biukovic G, Paravic A, Dadbin A, Cullum J, Hranueli D. 1998. Recombination between the linear plasmid pPZG101 and the linear chromosome of *Streptomyces rimosus* can lead to exchange of ends. Mol Microbiol. 28:1165–1176.

Richardson MA, Kuhstoss S, Huber ML, Ford L, Godfrey O, Turner JR, Rao RN. 1990. Cloning of spiramycin biosynthetic

genes and their use in constructing *Streptomyces ambofaciens* mutants defective in spiramycin biosynthesis. J Bacteriol. 172: 3790–3798.
Rocha EP. 2004. The replication-related organization of bacterial genomes. Microbiology. 150:1609–1627.
Rocha EP. 2006. Inference and analysis of the relative stability of bacterial chromosomes. Mol Biol Evol. 23:513–522.
Rudd BA, Hopwood DA. 1980. A pigmented mycelial antibiotic in *Streptomyces coelicolor*: control by a chromosomal gene cluster. J Gen Microbiol. 119:333–340.
Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. Bioinformatics. 16:944–945.
Sezonov G, Duchene AM, Friedmann A, Guerineau M, Pernodet JL. 1998. Replicase, excisionase, and integrase genes of the *Streptomyces* element pSAM2 constitute an operon positively regulated by the pra gene. J Bacteriol. 180:3056–3061.
Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci USA. 89: 8794–8797.
Spatz K, Kohn H, Redenbach M. 2002. Characterization of the *Streptomyces violaceoruber* SANK95570 plasmids pSV1 and pSV2. FEMS Microbiol Lett. 213:87–92.
Stajich JE, Block D, Boulez K, et al. (21 co-authors). 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12:1611–1618.
Strohl WR. 1992. Compilation and analysis of DNA sequences associated with apparent streptomycete promoters. Nucleic Acids Res. 20:961–974.
Suyama M, Bork P. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. 17:10–13.
Suzek BE, Ermolaeva MD, Schreiber M, Salzberg SL. 2001. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics. 17:1123–1130.
Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science 296:2376–2379.
Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29:22–28.
Volff JN, Altenbuchner J. 2000. A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. FEMS Microbiol Lett. 186:143–150.
Welch RA, Burland V, Plunkett G 3rd, et al. (19 co-authors). 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci USA. 99:17020–17024.
Wenner T, Roth V, Fischer G, Fourrier C, Aigle B, Decaris B, Leblond P. 2003. End-to-end fusion of linear deleted chromosomes initiates a cycle of genome instability in *Streptomyces ambofaciens*. Mol Microbiol. 50:411–425.
Wright LF, Hopwood DA. 1976. Actinorhodin is a chromosomally-determined antibiotic in *Streptomyces coelicolor* A3(2). J Gen Microbiol. 96:289–297.
Wu G, Culley DE, Zhang W. 2005. Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. Microbiology. 151:2175–2187.
Yamasaki M, Kinashi H. 2004. Two chimeric chromosomes of *Streptomyces coelicolor* A3(2) generated by single crossover of the wild-type chromosome and linear plasmid scp1. J Bacteriol. 186:6553–6559.
Zdobnov EM, Apweiler R. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 17:847–848.

Martin Embley, Associate Editor

Accepted September 1, 2006

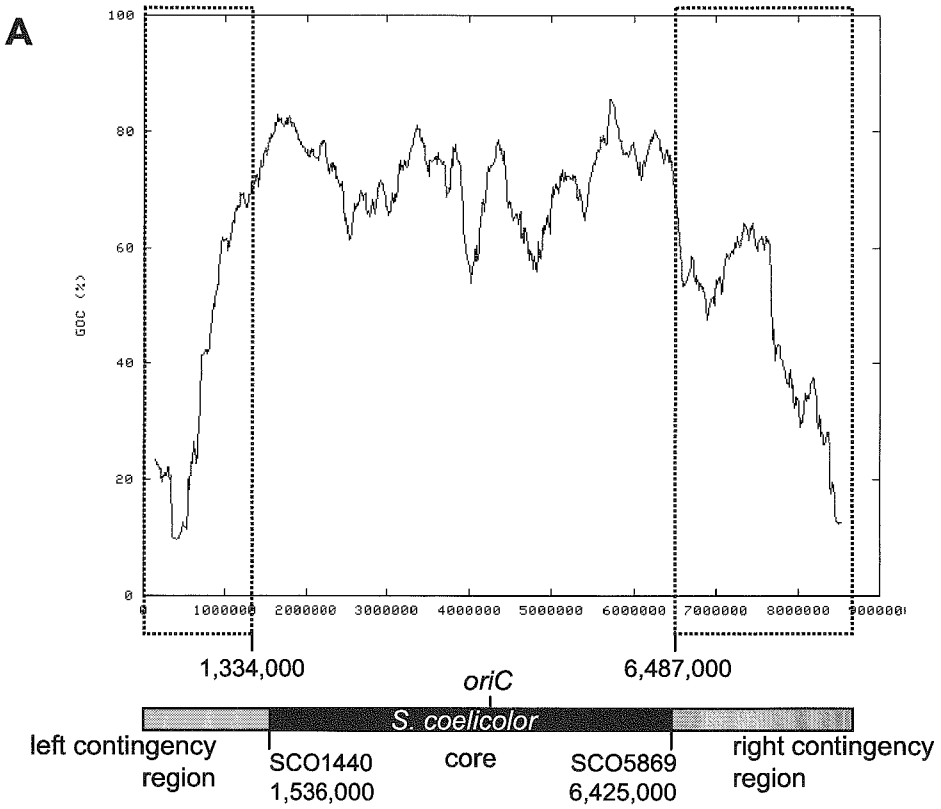
Supplementary material for Choulet *et al.*, Evolution of the terminal regions of the *Streptomyces* linear chromosome.

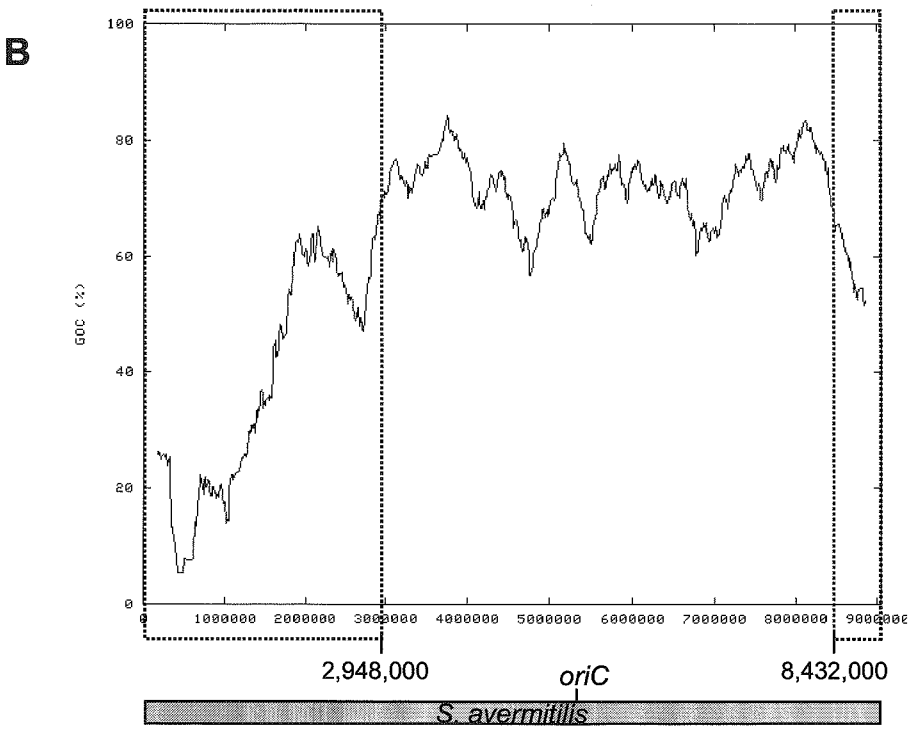
Table outlining the locations and features of *S. coelicolor* regions absent in the *S. ambofaciens* genome.

		first gene	last gene	number of CDSs	length (kb)	features
left contingency region	end	SCO0001	SCO0447	435	467	left terminal specific region
	1	SCO0850	SCO0880	31	31	cluster of membrane proteins
	2	SCO0979	SCO1009	31	33	integrase, IS*110B*
core region	3	SCO2407	SCO2429	23	28	secreted and membrane transport proteins
	4	SCO2855	SCO2892	37	31	rifampicin ADP-ribosyl transferase
	5	SCO3203	SCO3249	47	92	calcium dependent antibiotic cluster, pSAM2-like element
	6	SCO3431	SCO3539	109	117	GC: 69%, next to tRNA, transposase, IS*469*, IS*468*, IS*466*, IS*470*, carbohydrate metabolism
	7	SCO3677	SCO3725	50	41	next to tRNA, heavy metal resistance, transposase
	8	SCO4615	SCO4640	26	26	GC: 69%, next to tRNA-tyr, pSAM2-like element
	9	SCO5061	SCO5096	36	37	actinorhodin cluster
	10	SCO5312	SCO5357	46	44	GC: 69%, pSAM2-like element, IS1651, integrase next to tRNA-arg
	11	SCO5718	SCO5743	26	30	GC: 68,6%
	12	SCO5877	SCO5901	25	36	undecylprodigiosin cluster
	13	SCO6276	SCO6406	129	149	type I polyketide synthases, IS
	14	SCO6610	SCO6643	34	50	GC:69,2%, *phic31* resistance
right contingency region	15	SCO6805	SCO6953	148	155	GC: 69,3%, next to tRNA-pro, integrase, IS*468B*, polyketide synthases, arsenic resistance
	16	SCO7074	SCO7225	104	169	transposases, polyketide synthases, chitinase
	end	SCO7559	SCO7846	288	286	right terminal specific region
			contingency regions:	**1037**	**1141**	
			core region:	**588**	**681**	
			total:	**1625**	**1822**	

Supplementary material for Choulet et al., Evolution of the terminal regions of the *Streptomyces* linear chromosome.

GOC profiles (y axis) of the *S. coelicolor* (A.) and *S. avermitilis* (B.) chromosomes compared to each other (sliding window: 300 CDSs; step 10 CDSs). The GOC was estimated by calculating the number of pairs of orthologues that are contiguous in the two chromosomes divided by the number of orthologues in the window as reported in (Rocha 2006). This formula avoids the fall of GOC resulting from the presence of a single large genomic island. Thus, falls of GOC only reflect the presence of regions where multiple events breaking the synteny have occurred (i.e. the degenerated syntenic regions).
The framed areas represent the regions showing a gradually degenerated synteny toward the extremities.





E. Variabilité intraspéficique des répétitions terminales inversées du chromosome de *S. ambofaciens*

Publication n°2 :

Choulet F., Gallois A., Aigle B., Mangenot S., Gerbaud C., Truong C., Francou F.-X., Borges F, Fourrier C., Guérineau M., Decaris B., Barbe V., Pernodet J.-L., Leblond P.

Intraspecific variability of the terminal inverted repeats of the linear chromosome of *Streptomyces ambofaciens*.

Journal of Bacteriology, Sept. 2006, p. 6599-6610 Vol. 188, No. 18.

En complément de l'étude concernant la génomique comparée des génomes de *Streptomyces*, nous nous sommes intéressés à l'établissement d'une variabilité terminale à courte distance évolutive, c'est-à-dire au niveau intraspécifique.

Deux souches de *S. ambofaciens* sont disponibles dans les collections : la souche ATCC23877, dont le génome est partiellement séquencé (publication n°1), et la souche DSM40697. Il avait été montré, par des expériences de cross hybridation, que ces deux souches pouvaient être distinguées par la nature de leurs extrémités chromosomiques (Fischer *et al.*, 1997b). Ces souches sont des isolats indépendants extrêmement proches phylogénétiquement (voir "Séquençage"). Cependant un polymorphisme de longueur et de séquences des TIR a été détecté. Pour préciser la nature de ce polymorphisme et pour mieux comprendre les mécanismes évolutifs à la base de la diversification des extrémités chromosomiques chez *Streptomyces*, les TIR de la seconde souche (DSM40697) ont été totalement séquencées en collaboration avec Alexandre Gallois, doctorant dans l'équipe de Pierre Leblond. Ainsi, deux contigs de 238.517 pb (bras gauche) et 237.382 pb (bras droit) ont ainsi été obtenus (Fig. 18B).

Les TIR de la souche ATCC23877 et DSM40697 mesurent respectivement 197.936 pb et 212.655 pb. La limite interne des TIR est localisée au même locus dans les deux souches, indiquant qu'elles dérivent d'un événement de formation ancestral. Ces résultats indiquent que toutes les régions variables mises en évidence sont issues d'événements postérieurs à la formation des TIR chez *S. ambofaciens*. Pourtant, les TIR sont inclues dans les régions terminales du chromosome de *S. ambofaciens* qui ne présentent aucune synténie avec le chromosome de l'espèce proche *S. coelicolor*.

La différence de taille de TIR s'explique par la présence de régions spécifiques de souche qui représentent entre 25% et 30% du contenu en gènes des TIR. Chaque souche présente plusieurs régions variables mais l'essentiel des gènes spécifiques est localisé dans une unique région de 60 kb pour la souche DSM40697 et dans deux régions de 48 kb pour la souche ATCC23877 qui constituent les extrémités du chromosome. Au contraire, la partie interne des TIR est quasi identique sur environ 150 kb entre les deux souches. Les séquences communes partagent, en effet, 99% d'identité nucléotidique en moyenne (Art. 2, Fig. 1).

La variabilité terminale détectée ici est donc issue d'événements récents et confirme le mode d'évolution discuté dans la première partie des résultats qui décrit l'existence de flux de gènes façonnant les extrémités chromosomiques.

Les régions spécifiques présentent des similarités avec des clusters associés à des plasmides chez *Streptomyces*, notamment les plasmides linéaires SCP1 de *S. coelicolor* et SAP1 de *S. avermitilis*. De plus, les régions terminales spécifiques de souches possèdent un contenu en bases G+C plus faible que la signature du génome (environ 68%) typique de celui des plasmides retrouvés chez *Streptomyces*. Ces résultats suggèrent que des échanges d'extrémités de réplicons linéaires seraient à l'origine de la diversification des extrémités chromosomiques au niveau intraspécifique.

En raison de leur taille trop importante pour être clonée en une seule molécule, il n'a pas été possible de séquencer les deux copies des TIR de chaque souche. Néanmoins, plusieurs données suggèrent un niveau de conservation très élevé, voire identique, entre les deux copies des répétions terminales pour un même chromosome. Toutes les régions spécifiques de souche détectées dans les TIR sont donc dupliquées sur chaque bras chromosomique. Ces résultats indiquent, par conséquent, qu'il existe un mécanisme d'homogénéisation des TIR qui pourrait être responsable de la duplication des régions nouvellement acquises ou délétées dans les TIR.

JOURNAL OF BACTERIOLOGY, Sept. 2006, p. 6599–6610 Vol. 188, No. 18
0021-9193/06/$08.00+0 doi:10.1128/JB.00734-06

Intraspecific Variability of the Terminal Inverted Repeats of the Linear Chromosome of *Streptomyces ambofaciens*

Frédéric Choulet,[1]# Alexandre Gallois,[1]# Bertrand Aigle,[1] Sophie Mangenot,[2] Claude Gerbaud,[3] Chantal Truong,[2] François-Xavier Francou,[3] Frédéric Borges,[1] Céline Fourrier,[1] Michel Guérineau,[3] Bernard Decaris,[1] Valérie Barbe,[2] Jean-Luc Pernodet,[3] and Pierre Leblond[1]*

Laboratoire de Génétique et Microbiologie, UMR INRA 1128, IFR 110, Université Henri Poincaré Nancy 1, Faculté des Sciences et Techniques, BP239, 54506 Vandœuvre-lès-Nancy, France[1]; Génoscope, Centre National de Séquençage, 2 rue Gaston Crémieux CP5706 91057 Evry cedex, France[2]; and Institut de Génétique et Microbiologie, UMR CNRS 8621, Université Paris-Sud 11, Bâtiment 400, 91405 Orsay cedex, France[3]

Received 22 May 2006/Accepted 7 July 2006

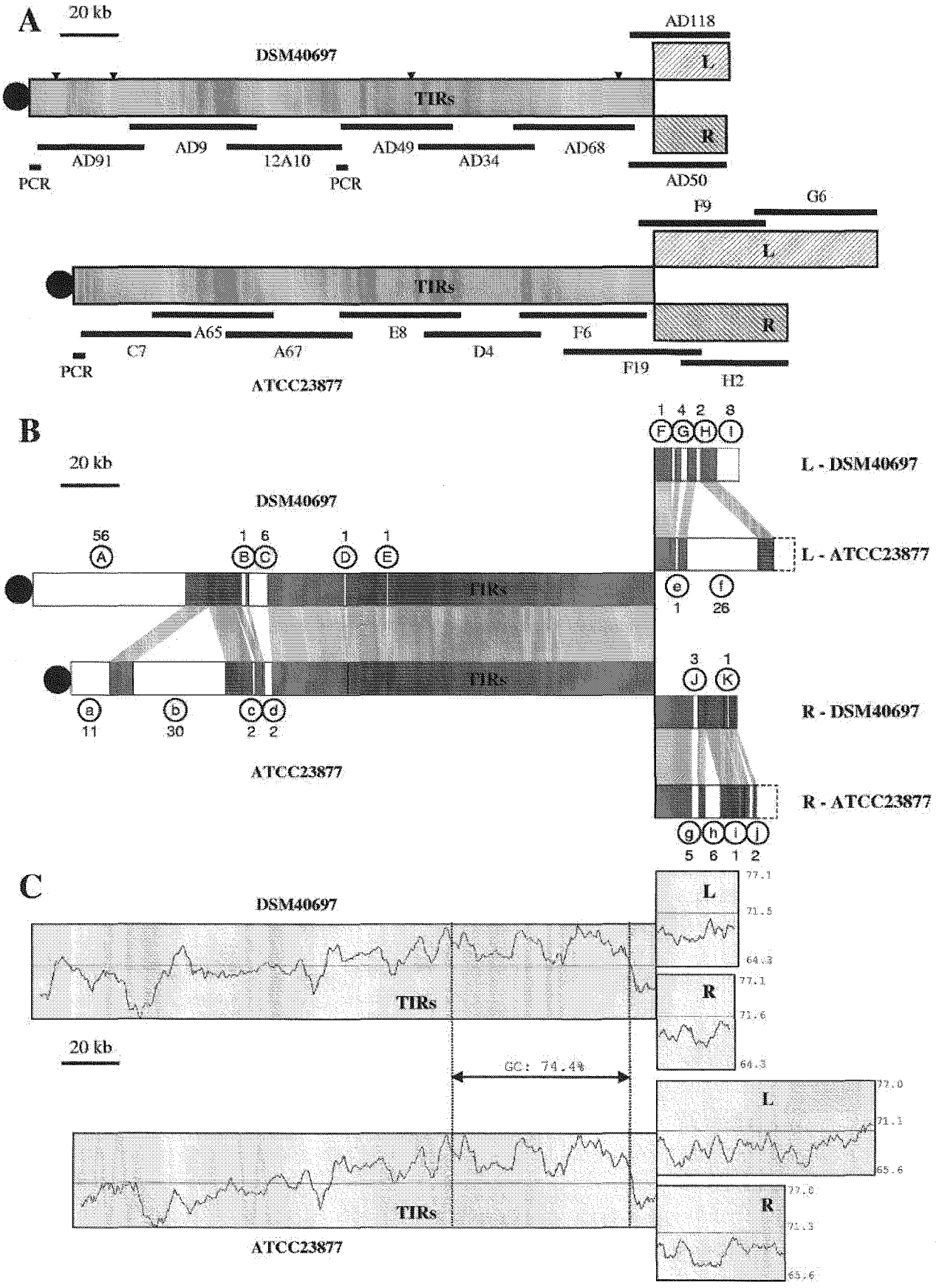
The sequences of the terminal inverted repeats (TIRs) ending the linear chromosomal DNA of two *Streptomyces ambofaciens* strains, ATCC23877 and DSM40697 (198 kb and 213 kb, respectively), were determined from two sets of recombinant cosmids. Among the 215 coding DNA sequences (CDSs) predicted in the TIRs of strain DSM40697, 65 are absent in the TIRs of strain ATCC23877. Reciprocally, 45 of the 194 predicted CDSs are specific to the ATCC23877 strain. The strain-specific CDSs are located mainly at the terminal end of the TIRs. Indeed, although TIRs appear almost identical over 150 kb (99% nucleotide identity), large regions of DNA of 60 kb (DSM40697) and 48 kb (ATCC23877), mostly spanning the ends of the chromosome, are strain specific. These regions are rich in plasmid-associated genes, including genes encoding putative conjugal transfer functions. The strain-specific regions also share a G+C content (68%) lower than that of the rest of the genome (from 71% to 73%), a percentage that is more typical of *Streptomyces* plasmids and mobile elements. These data suggest that exchanges of replicon extremities have occurred, thereby contributing to the terminal variability observed at the intraspecific level. In addition, the terminal regions include many mobile genetic element-related genes, pseudogenes, and genes related to adaptation. The results give insight into the mechanisms of evolution of the TIRs: integration of new information and/or loss of DNA fragments and subsequent homogenization of the two chromosomal extremities.

Streptomyces chromosomal DNA is linear and is among the largest described for bacteria, typically 8 to 10 Mb (6, 19). *Streptomyces* linear replicons (chromosomes and plasmids) share an invertronic structure including the presence of terminal inverted repeat sequences (TIRs) ended by bacterial telomeres covalently linked to terminal proteins (35). The lengths and sequences of the TIRs are extremely variable, and their sizes are not correlated to that of the replicon. The terminal duplications can be as large as several hundreds of kilobases, e.g., in the chromosome of *Streptomyces coelicolor* M600 (~1 Mb) (43), or can be restricted to the telomeric palindromes, e.g., in the chromosome of *Streptomyces avermitilis* (167 bp) (19). Two *Streptomyces* annotated genomes have been released so far: that for *S. coelicolor* A3(2), with 7,825 predicted coding DNA sequences (CDSs) (6), and that for *Streptomyces avermitilis* MA-4680, with 7,577 putative CDSs (19). Comparison of the two genomes revealed a common general organization with a conserved central region of about 5 Mb and terminal regions (or "arms") carrying mainly nonessential variable genes (6, 19). The terminal regions appear poorly conserved at the level of gene content and organization, which contrasts with the strong synteny observed for the central region. The fact that the terminal regions are dispensable for vegetative growth in laboratory conditions was revealed earlier by the characterization of instability phenomena (25). In *S. ambofaciens*, up to 2.5 Mb located at the ends of the chromosome can be lost and is not essential for vegetative growth in laboratory growth conditions. In addition, large DNA rearrangements, such as duplications, deletions, and amplifications, frequently occur in the subtelomeric regions. One of the most spectacular rearrangements affects the size of the TIRs, which can vary from 5 kb to 1.4 Mb in spontaneous mutant strains (46), while the wild-type strain DSM40697 harbors 210-kb TIRs (26). This variation implies nonreciprocal translocations of chromosomal extremities that in some cases result from homologous recombination between duplicated genes (14). Another phenomenon triggering TIR variation implies exchanges of replicon extremities between plasmids and chromosomes. This was demonstrated for a strain of *S. coelicolor* in which chimeric chromosomes can be generated by crossover of the wild-type chromosome and the linear plasmid SCP1 (47). This phenomenon was also shown for *Streptomyces rimosus* by interaction between plasmid pZG101 and the chromosome (30) and strongly suggested by the analysis of the terminal structure of the *Streptomyces lividans* chromosome, which could result from partial integration of the linear plasmid SLP2 (17).

In order to gain further insight into the mechanisms of chromosomal end diversification in *Streptomyces*, we analyzed

* Corresponding author. Mailing address: Laboratoire de Génétique et Microbiologie, UMR INRA 1128, IFR 110, Faculté des Sciences et Techniques, Université Henri Poincaré - Nancy 1, Boulevard des Aiguillettes, BP239, 54506 Vandœuvre-lès-Nancy, France. Phone: 33(0)3 83 68 42 07. Fax: 33(0)3 83 68 44 99. E-mail: leblond@nancy.inra.fr.

# F.C. and A.G. contributed equally to this report.


A
20 kb
DSM40697
AD118
L
TIRs
R
AD91
AD9
12A10
AD49
AD34
AD68
AD50
PCR
PCR
F9
G6
L
TIRs
R
A65
E8
F6
C7
A67
D4
F19
H2
PCR
ATCC23877
B
20 kb
DSM40697
1 4 2 8
F G H I
L - DSM40697
56
A
1 6
B C
1
D
1
E
TIRs
L - ATCC23877
e f
1 26
3 1
J K
TIRs
a b c d
11 30 2 2
R - DSM40697
R - ATCC23877
g h i j
5 6 1 2
ATCC23877
C
DSM40697
L
77.1
71.5
64.3
TIRs
R
77.1
71.6
64.3
20 kb
GC: 74.4%
L
72.0
71.1
65.6
TIRs
R
72.0
71.3
65.6
ATCC23877

the variability of the TIRs of two isolates belonging to the *S. ambofaciens* species. It was previously shown by cross-hybridization experiments that they could be distinguished by the terminal regions (13). These two independent soil isolates, ATCC23877 (32) and DSM40697 (18), were assigned to the same species according to classical morphological and physiological traits, e.g., antibiotic production. These two isolates indeed share the same antibiotic synthesis profile, both producing the three known antibiotic compounds spiramycin, congocidin, and alpomycin (31, 32). In contrast, the closely related species *S. coelicolor* (1.1% divergence of 16S rDNA sequence from that of *S. ambofaciens* ATCC23877) shows a completely different secondary metabolite profile.

This phenetic classification was supported by more-recent molecular analyses. First, at the whole-genome scale, pulsed-field gel electrophoresis analysis, which is a highly sensitive approach to distinguishing and identifying bacterial isolates, showed that the two strains differ only slightly, whereas the *S. coelicolor* chromosome diverges much more (26). Further, mapping experiments with large chromosomal DNA fragments and linking clones showed that most of them were common to both strains, leading to the conclusion that the two chromosomes show a colinear organization (26). An additional piece of evidence is the presence of a recent gene duplication affecting a sigma factor-encoding gene (*has* gene). This duplication is common to the two *S. ambofaciens* isolates, while a single copy of the *has* gene is present in the two complete genome sequences of *S. coelicolor* and *S. avermitilis* (34).

Furthermore, the most convincing evidence for their close relationships comes from the analysis of their 16S-23S internal transcribed spacer sequences, whose evolution is the most effective (among the *rrn* operon) to infer close phylogenetic relationships. The six internal transcribed spacer regions were isolated from the two strains, and their sequences were compared to those of *S. coelicolor* (45). While the two isolates share identical sequences, these sequences differ from that of *S. coelicolor*, thereby showing the divergence of the strains. Although gene conversion between the *rrn* loci could be identified, no crossover has occurred between them, maintaining the colinearity of the two chromosomes (45).

Altogether, these data show that the two isolates are closely related strains and that their assignment to the same species is supported by a multicriterion analysis. In this work, we undertook a detailed sequence analysis of the TIRs of these two *S. ambofaciens* strains to determine the strain-specific gene content and to gain a deeper insight into the mechanisms important for the evolution of chromosomal extremities.

MATERIALS AND METHODS

Sequencing. For each strain, ATCC23877 and DSM40697, a cosmid library was constructed from partially BamHI-digested *S. ambofaciens* genomic DNA cloned into the Supercos1 (Stratagene) vector. As the size of the *S. ambofaciens* TIRs greatly exceeds that of a fragment readily clonable into a cosmid vector, each copy of the TIRs cannot be isolated as a single recombinant molecule. Consequently, the sequences were obtained from a set of ordered recombinant cosmids (Fig. 1A). Note that for cosmids with insert sequences entirely corresponding to the TIRs, e.g., from cosmid C7 to F6 for strain ATCC23877, the chromosomal origin (i.e., from the right or left arm) of the recombinant cosmids cannot be deduced. Therefore, the TIR sequence is likely to consist of a chimera between the left and right repeats.

For sequencing, cosmids were mechanically fragmented and cloned with a BstXI adaptor into either pcDNA2.1 vector (Invitrogen) or pCNS (a derivate of pSU18 [4]). Ligation products were then introduced into *Escherichia coli* DH10B. It was not possible to obtain a cosmid clone containing the terminal fragment that includes the telomeres, and a PCR strategy was therefore used to amplify and sequence this region in both strains. Primers were designed from the cosmid C7 sequence ($5'$CACCCAGCGAGCCCAGCA$3'$) for strain ATCC23877 and from the cosmid AD91 sequence ($5'$AGCTGCAACGGTGCGTTCTATTGGG$3'$) for strain DSM40697, and a third primer was designed according to a consensus of the telomere sequences derived from several *Streptomyces* species ($5'$CGGAGCGGGTACCACATCGCTG$3'$). PCR was performed using 50 ng of DNA, with 800 μM deoxynucleoside triphosphates, 2 units of LA *Taq* polymerase (Takara), 2.5% dimethyl sulfoxide, and 20 pmol of each primer in a 50-μl final volume. After a denaturation step (95°C, 5 min), 30 cycles of denaturation (95°C, 30 s), annealing (58°C, 30 s), and polymerization (68°C, 3 min) were used to amplify the terminal fragment.

The 10 cosmids of strain ATCC23877 were completely sequenced (mean coverage, 10×). For strain DSM40697 (mean coverage, 10×), 17 gaps were remaining after the shotgun sequencing of the eight cosmids. PCR products were obtained for each gap and were cloned into pGEM-T Easy vector (Promega) for production of single-stranded DNA. For each gap, sequencing reactions were performed on all available templates, i.e., PCR product, double-stranded recombinant pGEM-T Easy vector, and the single-stranded DNA. In addition, sequencing reactions were carried out using two different reagents, CEQ Dye terminator (Beckman) and BigDye Terminator (Applied Biosystems). Four gaps remained, all localized in intergenic regions, suggesting they probably resulted from the formation of intrastrand secondary structures (e.g., terminators). Gap 1 (between DSMT0010 and DSMT0011) and gap 2 (between DSMT0032 and DSMT0033) are included in a strain-specific region (cosmid AD91) (Fig. 1A), and they are estimated to be less than 50 bp in length. Gap 3 (between DSMT0146 and DSMT00147) and gap 4 (between DSMT0204 and DSMT0205) belong to regions highly conserved between the two strains, and their sizes can be estimated as less than 10 bp and 200 bp, respectively, by comparison to the sequence of strain ATCC23877.

Annotation. The gene finder Glimmer2.10 (11) was used for CDS prediction, with a minimum size of 40 codons arbitrarily chosen as the threshold. Results were then refined by RBSfinder (39). The Basic Local Alignment Search Tool (BLAST 2.2.6) was used to find similarities (1), and the Interpro package was used to describe protein domains (48). CDSs were assigned a functional category where their best cluster of orthologous groups (COG) homologue is classified (40). Then, BLASTX translations were realized for each intergenic region in order to detect initially unpredicted CDSs and pseudogenes. While comparing the predicted protein sequences with BLASTP, proteins sharing more than 30% identity over at least 80% of the length of the query sequence were considered to be homologous.

FIG. 1. (A) Schematic representation of the *S. ambofaciens* ATCC23877 and DSM40697 TIRs and adjacent regions (L, left arm; R, right arm) and of the cosmids and PCR products used for the sequencing. Gap positions are represented by triangles (see Materials and Methods). (B) Comparison of the TIR sequences and flanking regions of *S. ambofaciens* ATCC23877 and DSM40697. Strain-specific regions are represented by white rectangles, whereas conserved regions are represented in gray. The strain-specific regions are named as follows: from A to K for strain DSM40697 and from a to j for strain ATCC23877, and the number of genes (including pseudogenes) carried by each specific region is indicated. Since the regions sequenced outside of the TIRs are larger for strain ATCC23877, sequences which cannot be compared because they were not sequenced in the second strain are represented by dashed rectangles. The terminal protein, covalently bound to the telomere, is represented by a black circle. (C) G+C content of TIRs and adjacent regions for the two strains (displayed with a 5,000-bp window size). Minimum, maximum, and mean (represented by a straight line) values of G+C percentages are indicated at the right side. These values are calculated for the whole contigs including the TIRs and the left or right adjacent region.

TABLE 1. General features of the TIR sequences of the *S. ambofaciens* strains ATCC23877 and DSM40697

Strain	Size of TIR (bp)	G+C content (%)	No. of predicted CDSs (no. of pseudogenes)	No. of proteins with assigned function (%)	No. conserved with proteins of unknown function (%)	No. of orphans (%)
ATCC23877	197,936	71.8	194 (3)	125 (65)	49 (25)	20 (10)
DSM40697	212,655	71.9	215 (10)	135 (63)	54 (25)	26 (12)

For strain ATCC23877, duplicated genes in the TIRs were named SAMT*nnnn*, whereas those specific only to either the left or right arm were annotated as SAML*nnnn* or SAMR*nnnn*, respectively (underlining indicates the type of specificity). A similar principle was adopted for strain DSM40697, with the prefix "DSM" used instead of "SAM." Sequences of left and right contigs for the two strains and their corresponding annotations are available through the SAMDB web server at http://www.weblgm.scbiol.ambofaciens.uhp-nancy.fr/.

Nucleotide sequence accession numbers. The sequences were deposited in EMBL under the following accession numbers: AJ937740 (ATCC23877 left TIR), AJ937741 (ATCC23877 right TIR), AM279694 (DSM40697 left TIR), and AM279695 (DSM40697 right TIR).

RESULTS

Intraspecific variability at the chromosomal ends. The large TIR sequences of two *S. ambofaciens* strains were determined using sets of recombinant cosmids spanning the terminal regions of the chromosome (Fig. 1A). Chimeric sequences of 197,936 bp and 212,655 bp (see Materials and Methods) were produced for strains ATCC23877 and DSM40697, respectively (see Table 1 for general features).

Considering a single genome, the 100% nucleotide identity of the two TIR copies was supported by two lines of evidence. First, hybridization of DNA probes corresponding to the TIRs onto genomic DNA did not reveal any polymorphism (26). Second, no mismatch was found within the overlaps of sequenced cosmids, despite the fact that they can originate from either of the chromosomal arms.

The ends of the TIRs will be defined as the first nucleotide of divergence (noted as "internal boundary of the TIRs" in Fig. 2A). Consequently, the terminal duplication does not end at exactly the same nucleotide position in the two strains. However, this situation results from point mutations and small insertions/deletions causing minor differences between the arms in both strains in the flanking small region (over about 5 kb). These regions show near identity and contain two CDSs, a probable cytochrome P450 (DSMT0215, SAML/R0195) and a homologue of *afsA* (DSML/R0216, SAML/R0196), separated by a large intergenic region that includes stretches of short repeated C/A-rich motifs (Fig. 2A). Another stretch of C/A-rich motifs is located in the 3′ part of *afsA*, and the variation in the number of motifs results in differences between the four *afsA* copies (two copies in each strain). Consequently, the *afsA* coding sequences have different sizes (three different alleles for four copies). In addition, these open reading frames (ORFs) might be pseudogenes, since only the 5′ end (about 200 codons) shows similarity (between 38% and 46%) with the *afsA*-like gene of *Streptomyces rochei* (28).

The two arm sequences diverge totally at the same nucleotide proximal to *afsA* in the two strains. Thus, the two strains share the same ancestral boundaries of TIRs (Fig. 2A).

Thus, 194 CDSs (from SAMT0001 to SAMT0194) belong to the TIRs of strain ATCC23877. This last ORF is similar to that encoding a truncated transposase. In strain DSM40697, the end of the sequences assumed to be identical occurs within the long intergenic region separating DSMT0215 (putative cytochrome P450) and DSML/R0216 (*afsA* homologues) (Fig. 2A).

When the TIRs of the two strains are compared, two regions can be distinguished: a terminal strain-specific region and a conserved region with a more internal location. In total, the syntenic regions extend over about 150 kb and include 149 genes sharing an average 99% nucleotide identity (Fig. 1B).

Within the whole TIRs, seven strain-specific segments resulting from DNA rearrangements such as insertions, deletions, and/or gene replacements involving from 1 to 56 CDSs can be delimited (Fig. 1B). For strain DSM40697, five specific regions (A to E) represent 62 kb and contain 65 CDSs (including probable pseudogenes). Region A on its own includes 56 of the 65 strain-specific CDSs and represents a quarter of the TIR size. For strain ATCC23877, the specific loci are scattered into four regions (a to d) spanning 49 kb and including 45 strain-specific CDSs. Again, 41 of these 45 specific genes are clustered into two loci, one of which constitutes the chromosomal end (region a, 14 kb).

Although the chromosomal ends are strain specific, highly similar telomere sequences were found in the two strains (96.1% nucleotide identity over 180 bp) (Fig. 3). The seven typical palindromes predicted to form the secondary structures found in other known *Streptomyces* telomeres are present in both strains (7, 17) (Fig. 3).

Outside the TIRs, intraspecific variability can also be noticed. Within the areas sequenced, at least four and five DNA rearrangements have occurred in the left and right arms, respectively (Fig. 1B; also see below).

Altogether, these results suggest that, despite their variability in gene content, the current TIRs of the two *S. ambofaciens* strains appear to derive from a single ancestral event of terminal duplication, since they share the same ancestral internal boundary.

Low G+C content. Despite the differences in gene content observed between the two strains, profiles of G+C percentages are quite similar (Fig. 1C). For each strain, a decrease in G+C content characterized the strain-specific extremities as well as the sequences outside of the TIRs. Thus, the terminal 50 kb of the ATCC23877 and DSM40697 chromosomes show G+C contents of 68.8% and 69.2%, respectively. Similar values are observed for the regions sequenced outside of the TIRs (69.1% in strain ATCC23877 and 68.9% in strain DSM40697). In contrast, G+C contents of *Streptomyces* chromosomes are 72.1% for *S. coelicolor* and 70.7% for *S. avermitilis*. In fact, the TIRs of *S. ambofaciens* seem to be present in regions of even lower G+C content, but the presence of a 61-kb cluster (which includes the alpomycin cluster [31]) showing a G+C content of

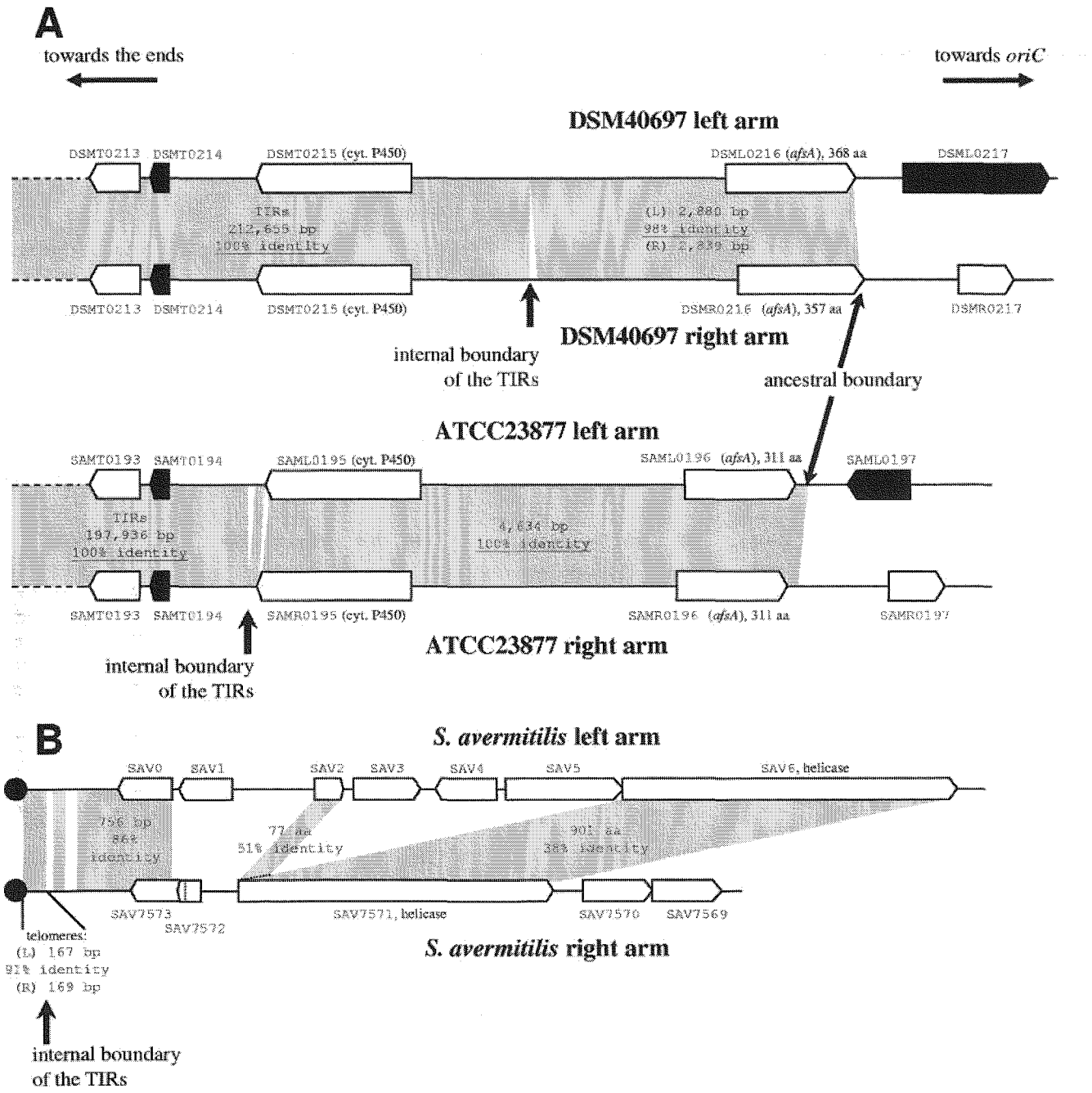


FIG. 2. Intrachromosomal comparisons of the left and right arm regions surrounding the ends of the TIRs of *S. ambofaciens* strains ATCC23877 and DSM40697 (A) and *S. avermitilis* (B). Duplicated regions between both arms are shaded in gray, and their sizes and percentages of identity are indicated. Black arrows represent ORFs similar to genes encoding transposases (or truncated transposases), and the terminal proteins are represented by the black circles. The SAV0 putative CDS was not predicted in reference 19 and was added in this work. cyt., cytochrome. aa, amino acid.

74.4% skews the data by causing a local increase in G+C content (Fig. 1C).

The low G+C content of the chromosomal extremities is reminiscent of that of *Streptomyces* plasmids (SLP2, 68.4% [17]; pSV2, 69.7% [38]; SCP1, 69.0% [5]; and SAP1, 69.2% [19] [see "Plasmid-associated genes" below]). The lower G+C content observed in the regions outside of the TIRs is related to a remarkable abundance of insertion sequences (ISs) and related genes in the strain-specific regions (found in regions F, G, H, and K in strain DSM40697 and in regions f, h, and j in strain ATCC23877) (Fig. 1B; also see "Mobile genetic element-related genes" below). The lower G+C content of mobile genetic elements in bacteria has been discussed previously (33), and their presence, together with an observed low G+C content, suggests acquisition by horizontal gene transfer.

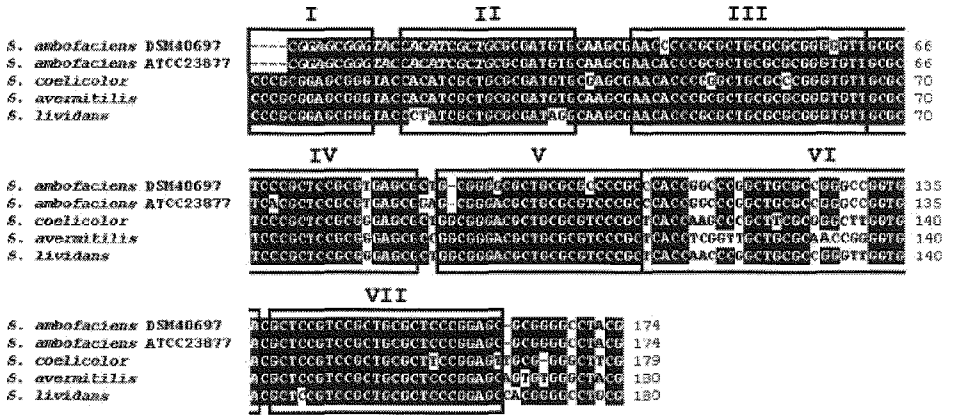


FIG. 3. Sequence alignment of the telomeres of the two *S. ambofaciens* strains with the chromosomal telomeres of *S. coelicolor* A3(2), *S. avermitilis* MA-4680, and *S. lividans* ZX7. The palindromes are labeled, and the primer sequence used for the amplification of the telomeres (see Materials and Methods) is represented in italic. This PCR strategy explains the missing nucleotides at the end of the *S. ambofaciens* telomeres. The boxes labeled I to VII indicate the positions of palindromes starting from the end of the chromosomal DNA. Numbers to the right of the sequences indicate the cumulative length for each of the aligned sequences.

Plasmid-associated genes. A large proportion of the strain-specific CDSs included in the TIRs may have a plasmid origin (see Table 2). This situation is particularly striking for strain DSM40697, for which 16 of the 56 CDSs located in the strain-specific region A show best similarity with plasmid-associated genes. For half of them, the level of identity is particularly high, i.e., more than 80% amino acid identity. Five gene products share best similarity with proteins encoded by linear plasmid SAP1 from *S. avermitilis* (19) and eight with linear plasmid SCP1 from *S. coelicolor* (5). Interestingly, one gene fragment (DSMT0048) is similar to *tpg* from *S. lividans* linear plasmid SLP2, which encodes the terminal protein involved in the replication of the telomeres (17). Some of the best hits were also with plasmids from other *Actinomycetales* spp., i.e., pNF1 (circular) from *Nocardia farcinica* (20) and pREL1 (linear) from *Rhodococcus erythropolis* (37). Furthermore, syntenic clusters between these strain-specific regions and linear plasmids (Table 2) were found, which strongly supports the hypothesis of integration of plasmid DNA into the terminal regions (9, 42). For example, the DSMT0042-45 cluster is conserved with the cluster SCP1.199-202 from linear plasmid SCP1.

The same conclusion can be inferred from the analysis of region a from strain ATCC23877. First, the probable DNA helicase TtrA (SAMT0002) shows the best BLASTP hit (84% amino acid identity) with that of plasmid SLP2 of *S. lividans*. This helicase is implicated in the conjugal transfer of the SLP2 plasmid (8). The *ttrA* gene is present at the extremities of the chromosome of each *S. ambofaciens* strain, although it is located in the strain-specific regions. It should be pointed out that a *ttrA* homologue is present close to the telomeres in almost all *Streptomyces* replicons. In fact, TtrA proteins of the two *S. ambofaciens* strains, which share 45% identity, may have a different origin: TtrA from strain ATCC23877 shares 84% identity with TtrA from plasmid SLP2, whereas that of strain DSM40697, which is truncated, shares 83% identity with TtrA from the chromosome of *S. avermitilis*. In addition, the SAMT0010 protein is similar to the KilB protein from different *Streptomyces* plasmids, pRL2 (linear) and pIJ101 (circular), in which it is implicated in conjugal transfer and intramycelial spread (36).

Mobile genetic element-related genes. Genomic islands often carry genes implicated in mobility, such as those encoding integrases, recombinases, and transposases (15). Thus, the presence of such genes in variable regions supports the idea of acquisition by horizontal gene transfer. In the conserved part of the TIRs, 149 pairs of orthologues between the two *S. ambofaciens* strains were predicted. Among them, three gene products showing similarity with transposases (DSMT0057/SAMT0012, DSMT0058/SAMT0013, and DSMT0214/SAMT0194) and one with an integrase/recombinase (DSMT0060/SAMT0015) were annotated. For three of them, the best similarity is found with transposase (or integrase) from *Frankia* species (*Actinomycetales*). Two transposase ORFs are located in a conserved region of the TIRs, just at the borders of the strain-specific regions A and a. The third one is located at the internal boundary of the TIRs as described above. However, none of the transposase-encoding genes seem to constitute a functional IS, the transposase being either inactivated by frameshift mutations or truncated, with no detectable flanking inverted repeats. The terminal strain-specific region A (DSM40697) contains a truncated IS (DSMT0046) similar to an IS from *Frankia* sp. strain Cci3 and a phage integrase (DSMT0003), close to the telomeres, sharing no homology to *Streptomyces* but with homology to *Nocardioides* species (see Table 2 for more details). Reciprocally, in the strain ATCC23877, one truncated IS (SAMT0032) is present in strain-specific region b.

Interestingly, many transposase-encoding genes are found close to the ancestral boundary of the TIRs (four in strain DSM40697 and three in strain ATCC23877), and this is a common feature of *Streptomyces* replicons. In *S. coelicolor* A3(2), an IS constitutes the ends of the chromosomal TIRs (6), while in plasmid SCP1, Tn*5714* is located 3 kb outside the left TIR and IS*466* is located at the end of the right one (5).

Given the close relationship between *S. ambofaciens* strains, it is even possible to spot recent IS- or transposon-mediated rearrangements. Indeed, outside of the TIRs, a putative com-

TABLE 2. Strain-specific genes predicted in the TIRs of *S. ambofaciens* strains DSM40697 and ATCC23877

Strain	Specific region	Gene	Product[a]	Identity (%)[b]	Overlap (%)[c]	Gene name	Plasmid	Organism
DSM40697	A	DSMT0001	Unknown					
		DSMT0002	Putative transcriptional regulator					
		DSMT0003	Putative phage integrase	35	93	NocaDRAFT_4522		*Nocardioides* sp. strain JS614
		DSMT0004	Unknown					
		DSMT0005	Putative serine/threonine protein kinase					
		DSMT0006	Unknown					
		DSMT0007	Unknown					
		DSMT0008	Putative helicase	83	88	SAV7571		*Streptomyces avermitilis*
		DSMT0009	Putative ATP-dependent DNA ligase	64	98	SAP1_90 (*lig*)	SAP1	*Streptomyces avermitilis*
		DSMT0010	Putative integral membrane transport protein	72	100	SCO6809		*Streptomyces coelicolor*
		DSMT0011	Putative secreted protein	75	100	SCO6811		*Streptomyces coelicolor*
		DSMT0012	Putative FAD-dependent oxidoreductase	58	98	pREL1_0108	pREL1	*Rhodococcus erythropolis*
		DSMT0013	Putative transcriptional regulator	95	96	pFQ25.11		*Streptomyces* sp. strain F2
		DSMT0014	Putative phosphinothricin *N*-acetyltransferase	77	99	pFQ25.10		*Streptomyces* sp. strain F2
		DSMT0015	Unknown					
		DSMT0016	Conserved hypothetical protein	85	99	pnf1840	pNF1	*Nocardia farcinica*
		DSMT0017	Putative membrane protein	76	100	SCO3280		*Streptomyces coelicolor*
		DSMT0018	Putative glycosyl hydrolase, BNR repeat	36	98	ArthDRAFT_2101		*Arthrobacter* sp. strain FB24
		DSMT0019	Putative lipoprotein	72	100	SCO4458		*Streptomyces coelicolor*
		DSMT0020	Putative membrane protein	60	96	SCO4459		*Streptomyces coelicolor*
		DSMT0021	Putative monooxygenase	78	94	SCO6838		*Streptomyces coelicolor*
		DSMT0022	Putative arsenic resistance membrane transport protein	84	99	SCO6837		*Streptomyces coelicolor*
		DSMT0023	Putative transcriptional regulator	82	100	SCO3699		*Streptomyces coelicolor*
		DSMT0024	Putative arsenate reductase	87	95	SCO6835		*Streptomyces coelicolor*
		DSMT0025	Putative thioredoxin reductase	79	99	SCO6834		*Streptomyces coelicolor*
		DSMT0026	Conserved hypothetical protein	25	75	Tfu_2935		*Thermobifida fusca*
		DSMT0027	Unknown					
		DSMT0028	Conserved hypothetical protein	87	100	SCP1.257	SCP1	*Streptomyces coelicolor*
		DSMT0029	Conserved hypothetical protein	36	83	SAP1_88	SAP1	*Streptomyces avermitilis*
		DSMT0030	Putative RNA polymerase sigma factor	37	97	SAP1_87 (*sig*)	SAP1	*Streptomyces avermitilis*
		DSMT0031	Conserved hypothetical protein	35	93	SAP1_86	SAP1	*Streptomyces avermitilis*
		DSMT0032	Putative secreted protein	84	100	SCP1.323c	SCP1	*Streptomyces coelicolor*
		DSMT0033	Putative secreted protein	89	100	SCP1.261c	SCP1	*Streptomyces coelicolor*
		DSMT0034	Conserved hypothetical protein	84	94	SCP1.262	SCP1	*Streptomyces coelicolor*
		DSMT0035	Ttra helicase fragment (pseudogene)	80	95	SAV7571		*Streptomyces avermitilis*
		DSMT0036	Conserved hypothetical protein DUF1099	63	95	ShewDRAFT_1466		*Shewanella* sp. strain PV-4
		DSMT0037	Unknown					
		DSMT0038	Conserved hypothetical protein (pseudogene)	70	99	SCO0085		*Streptomyces coelicolor*
		DSMT0039	Putative ATP-dependent DNA ligase	59	100	SAP1_90 (*lig*)	SAP1	*Streptomyces avermitilis*
		DSMT0040	Conserved hypothetical protein	71	100	SCO0048		*Streptomyces coelicolor*
		DSMT0041	Unknown					
		DSMT0042	Conserved hypothetical protein	29	95	SCP1.202	SCP1	*Streptomyces coelicolor*
		DSMT0043	Conserved hypothetical protein	84	100	SCP1.201	SCP1	*Streptomyces coelicolor*
		DSMT0044	Putative secreted protein	94	100	SCP1.200c	SCP1	*Streptomyces coelicolor*
		DSMT0045	Putative secreted esterase	93	100	SCP1.199c	SCP1	*Streptomyces coelicolor*
		DSMT0046	Putative transposase	55	71	Francci3_3385		*Frankia* sp. strain CcI3
		DSMT0047	Putative transporter	55	93	nfa29650		*Nocardia farcinica*
		DSMT0048	Tpg protein fragment (pseudogene)	75	97	*tpgSLP2*	SLP2	*Streptomyces lividans*
		DSMT0049	Putative AraC-family transcriptional regulator (pseudogene)	82	97	SCO3804		*Streptomyces coelicolor*
		DSMT0050	Conserved hypothetical protein	73	48	SCO3803		*Streptomyces coelicolor*
		DSMT0051	Putative nourseothricin acetyltransferase	69	100	*nat1*		*Streptomyces noursei*
		DSMT0052	Conserved hypothetical protein	30	93	Francci3_1866		*Frankia* sp. strain CcI3
		DSMT0053	Conserved hypothetical protein	47	76	Francci3_1863		*Frankia* sp. strain CcI3
		DSMT0054	Unknown					
		DSMT0055	Putative peptidase	38	98	SCO3610		*Streptomyces coelicolor*
		DSMT0056	Putative major facilitator superfamily	29	93	NocaDRAFT_1725		*Nocardioides* sp. strain JS614
	B	DSMT0083	Putative 3-oxoacyl-ACP synthase III	40	100	ArthDRAFT_2448		*Arthrobacter* sp. strain FB24
	C	DSMT0085	Putative ABC transport system ATP-binding protein	64	93	SCO0121		*Streptomyces coelicolor*
		DSMT0086	Putative integral membrane protein	44	93	SCO0120		*Streptomyces coelicolor*
		DSMT0087	Conserved hypothetical protein	46	76	Tfu_1509		*Thermobifida fusca*
		DSMT0088	Conserved hypothetical protein	31	98	SCO0123		*Streptomyces coelicolor*
		DSMT0089	Putative polyprenyl synthetase	61	98	SCO0568		*Streptomyces coelicolor*

Continued on following page

TABLE 2—*Continued*

Strain	Specific region	Gene	Product[a]	Identity (%)[b]	Overlap (%)[c]	Gene name	Plasmid	Organism
		DSMT0090	Putative geranylgeranyl diphosphate synthase	48	100	*ggdps*		*Streptomyces* sp. strain KO-3988
	D	DSMT0118	Conserved hypothetical protein	84	79	SCP1.218c	SCP1	*Streptomyces coelicolor*
	E	DSMT0134	Unknown					
ATCC23877	a	SAMT0001	Conserved hypothetical protein	62	92	SAV7573		*Streptomyces avermitilis*
		SAMT0002	Putative helicase	84	100	*ttrA*	SLP2	*Streptomyces lividans*
		SAMT0003	Unknown					
		SAMT0004	Conserved hypothetical protein	42	100	pFRL1.57	pFRL1	*Streptomyces* sp. strain FR1
		SAMT0005	Unknown	64	59	pFRL1.57	pFRL1	*Streptomyces* sp. strain FR1
		SAMT0006	Putative NTP pyrophosphohydrolase	39	67	PFL_4894		*Pseudomonas fluorescens*
		SAMT0007	Conserved hypothetical protein	52	95	nfa38470		*Nocardia farcinica*
		SAMT0008	Putative glyoxalase	64	96	nfa38460		*Nocardia farcinica*
		SAMT0009	Putative ferredoxin NADPH reductase	55	97	nfa38450		*Nocardia farcinica*
		SAMT0010	KilB-like protein, role in intramycelial spread	48	95	pRL2.23	pRL2	*Streptomyces* sp. strain 44414
		SAMT0011	Putative acetyltransferase	50	100	SAV2967		*Streptomyces avermitilis*
	b	SAMT0023	Unknown					
		SAMT0024	Putative hydrolase	53	90	Adeh_0548		*Anaeromyxobacter dehalogenans*
		SAMT0025	Conserved hypothetical protein	38	100	SCO7248		*Streptomyces coelicolor*
		SAMT0026	Putative phosphotransferase	54	91	*ard*		*Streptomyces avermitilis*
		SAMT0027	Unknown					
		SAMT0028	Putative secreted protein	47	80	SCO0072		*Streptomyces coelicolor*
		SAMT0029	Putative secreted protein	61	89	SCO0072		*Streptomyces coelicolor*
		SAMT0030	Unknown	30	67	Francci3_0808		*Frankia* sp. strain CcI3
		SAMT0031	Unknown					
		SAMT0032	Putative truncated transposase	71	80	*orf118*	pSLA2-L	*Streptomyces rochei*
		SAMT0033	Putative secreted protein	58	97	SCO0072		*Streptomyces coelicolor*
		SAMT0034	Putative secreted protein	86	100	SCO0072		*Streptomyces coelicolor*
		SAMT0035	Conserved hypothetical protein	60	100	SCO0073		*Streptomyces coelicolor*
		SAMT0036	Putative SAM-dependent methyltransferase	67	98	SCO2653		*Streptomyces coelicolor*
		SAMT0037	Conserved hypothetical protein	82	100	SCO0031		*Streptomyces coelicolor*
		SAMT0038	Putative transmembrane restriction endonuclease	87	100	SCO7763		*Streptomyces coelicolor*
		SAMT0039	Unknown					
		SAMT0040	Putative stress response protein	77	100	SCO3763		*Streptomyces coelicolor*
		SAMT0041	Unknown	54	48	Francci3_1126		*Frankia* sp. strain CcI3
		SAMT0042	Putative membrane protein	37	100	SAV3190		*Streptomyces avermitilis*
		SAMT0043	Unknown					
		SAMT0044	Putative regulator	39	81	SAV1103		*Streptomyces avermitilis*
		SAMT0045	Putative anti-sigma factor antagonist (pseudogene)	42	81	SCO3692		*Streptomyces coelicolor*
		SAMT0046	Putative regulator	71	96	*prpC3*		*Streptomyces avermitilis*
		SAMT0047	Putative hydrolase	79	100	SAV923		*Streptomyces avermitilis*
		SAMT0048	Putative urease beta/gamma subunit	50	94	DR_A0319		*Deinococcus radiodurans*
		SAMT0049	Putative urease alpha subunit	65	98	SCO1234		*Streptomyces coelicolor*
		SAMT0050	Putative *ureF*-like urease accessory protein	76	100	*ureF*		*Streptomyces avermitilis*
		SAMT0051	Putative *ureG*-like urease accessory protein	75	98	*ureG*		*Streptomyces avermitilis*
		SAMT0052	Putative *ureD*-like urease accessory protein	54	81	SCO1231		*Streptomyces coelicolor*
	c	SAMT0065	Putative truncated transposase	41	100	SAV18		*Streptomyces avermitilis*
		SAMT0066	Putative haloacid dehalogenase	89	99	SAV737		*Streptomyces avermitilis*
	d	SAMT0071	Putative transcriptional regulator	41	96			*Saccharopolyspora erythraea*
		SAMT0072	Putative esterase	38	100	SCO4392		*Streptomyces coelicolor*

[a] FAD, flavin adenine dinucleotide; BNR, bacterial neuraminidase repeat; ACP, acyl carrier protein; NTP, nucleoside triphosphate.
[b] Results for the best BLASTP hit are summarized.
[c] "Overlap" corresponds to the ratio between the length of the BLASTP alignment and the length of the query protein (for cases in which this ratio was >100%, 100% overlap was indicated).

posite transposon that is absent from the DSM40697 strain was detected in strain ATCC23877 (SAMR0213 to SAMR0218 [complete region h in Fig. 1B]). This transposon (5.1 kb) consists of two almost identical IS elements (99% nucleotide identity) flanking four CDSs, the products of two of which are also related to IS transposases, plus two orphans. Interestingly, the flanking ISs do not share any homology with *Streptomyces* sequences but do share homology with a transposase from *Frankia* sp. strain EAN1pec (67% amino acid identity).

Coding density and pseudogenes. Acquisition of DNA by horizontal gene transfer and loss of useless genes are the main causes of intraspecific variability, and an equilibrium between these two phenomena leads to genomic flux that shapes bacterial genomes (24). The first step of gene loss is the creation of pseudogenes either by point mutations or by truncations.

In *S. ambofaciens*, the TIRs are characterized by a low coding density. When considering pseudogenes as noncoding DNA, the coding densities of the sequenced regions are

77.1% and 75.7% for strains DSM40697 and ATCC23877, respectively. These values are not biased significantly by the presence of pseudogenes, since the coding densities are 80.1% (DSM40697) and 77.1% (ATCC23877) when pseudogenes are included as CDSs. These values contrast with values of 88.9% and 86.2% predicted for the complete chromosomes of *S. coelicolor* (6) and *S. avermitilis* (19), respectively. This is again a common trait of *Streptomyces* plasmids, such as SAP1 from *S. avermitilis*, for which the density falls to 79% (19).

Another striking feature of the terminal regions is the strong presence of pseudogenes, notably in the TIRs of strain DSM40697 (10, including two truncated transposases, representing 5% of the gene content). In contrast, pseudogenes represent less than 1% of the CDSs predicted in the whole genome of *S. coelicolor* (6). Eight out of the 10 pseudogenes are carried by the strain-specific region A. In addition to the truncation of the *ttrA* gene (DSMT0008; 219 amino acid residues out of 835 in *S. avermitilis*), two additional TtrA fragments are encoded in strain-specific regions A (DSMT0035; 44 residues [Table 2]) and G (DSML0224; 31 residues), outside of the TIRs. The first two fragments are homologous to different parts of the same gene (*ttrA* from *S. avermitilis*). However, the latter pseudogene is not related to them, suggesting acquisitions of extra copies by different integrations of parts of linear replicons. This hypothesis is further supported by the finding, in the same region A, of a gene fragment (DSMT0048; 99 bp) showing best identity with *tpgC* encoding the terminal protein of the linear plasmid SLP2 and of three gene fragments (DSMT0009, DSMT0038, and DSMT0039) similar to the *lig* gene encoding a ligase from the *S. avermitilis* linear plasmid SAP1 (SAP1_90).

In the regions sequenced outside of the TIRs, 10 additional probable pseudogenes were predicted for strain DSM40697, of which 6 belong to the strain-specific regions.

The number of pseudogenes is probably underestimated, and the low coding density could be a consequence of a high mutation rate. Altogether, these data support the hypothesis that the TIRs and the terminal regions of the genome constitute a hot spot for horizontal gene transfer events mediated by linear plasmids.

Horizontal transfer of accessory genes. Horizontal transfer mostly involves accessory genes that are able to confer a selective advantage to the recipient cell. Housekeeping genes are more recalcitrant to transfer (23). In *S. ambofaciens*, functions of many genes predicted for the strain-specific regions, such as resistance to toxic compounds, are related to adaptation to the environment. They are more particularly abundant in the terminal specific region A of strain DSM40697 (Fig. 1B). A good example is the five-gene cluster DSMT0021-25, which is highly similar (from 78% to 87% amino acid identity) to the *S. coelicolor* cluster SCO6838-34, which is implicated in the transport of and resistance to arsenate.

Antibiotic resistance genes are also present. For example, the DSMT0051 gene product shows best similarity (69%) with the *Streptomyces noursei* nourseothricin acetyltransferase, Nat1 (22). In the same way, a homologue of the Ard2-encoding gene (SAMT0026; 54% identity) from *Saccharothrix mutabilis* subsp. *capreolus* (*Actinomycetales*), which confers resistance to A201A antibiotic (3), is present in the ATCC23877 strain-specific chromosomal end.

The DSM40697 strain-specific regions B and C also carry functions related to adaptation. Indeed, the seven specific genes identified here (DSMT0083, DSMT0085-90) all have associations with secondary metabolism: DSMT0083 is a probable 3-oxoacyl-ACP synthase III-encoding gene having homologues in different *Actinomycetales* spp.; DSMT0085/86/88 are conserved, with three membrane and transport protein-encoding genes of *S. coelicolor* adjacent to the eicosapentaenoic acid cluster (SCO0124-0129); DSMT0087 shares similarity with a CDS of unknown function located in the fredericamycin biosynthesis gene cluster from *Streptomyces griseus* (44); and DSMT0089/90 are similar to genes encoding a putative polyprenyl synthase and a geranylgeranyl diphosphate synthase, respectively, from *Streptomyces* sp. strain KO-3988 (21). Interestingly, although these genes all seem to be implicated in secondary metabolism, they do not correspond to a whole conserved cluster but rather show similarity to genes from many different clusters among *Actinomycetales*. This could therefore be an example of a locus created by the association of genes derived from different pathways. Examples of transfer of secondary metabolism clusters have already been discussed in reference 27. The *alp* polyketide synthase cluster present in the TIRs of the two *S. ambofaciens* strains could also constitute an example of a chimeric cluster (31). While its left part is mostly similar to the kinamycin-biosynthetic cluster of *S. murayamaensis* (accession number AY228175), most of the right part shows high similarity to and the same genetic organization as a locus identified in the *S. rochei* linear plasmid pSLA2-L (28).

In addition, the variable region C is replaced by region d in strain ATCC23877 (Fig. 1B), in which one of the two specific CDSs identified encodes a probable transcriptional regulator sharing best identity with the OrfD regulator belonging to a cluster involved in the biosynthesis of a pigment in *Saccharopolyspora erythraea* (10).

However, a significant part of the strain-specific genes have no known function and are, in some cases, orphans.

Species specificity of the *S. ambofaciens* TIR conserved genes. The part of the TIRs conserved at the intraspecific level is highly variable at the interspecific level. No synteny can be observed the other *Streptomyces* with genomes. The regions conserved with *S. coelicolor* are limited to eight small syntenic clusters comprising two to seven CDSs (e.g., a urea degradation cluster). In contrast to what is found for *S. avermitilis*, only one cluster of four CDSs is syntenic.

Seventeen percent of the 149 pairs of orthologues between the two strains do not have any homologue in the NR database. In addition, among the 37 proteins showing best similarity with those from organisms other than *Streptomyces*, the majority (19/37) show highest similarity with those from other *Actinomycetales* genera (especially *Frankia*, *Arthrobacter*, *Nocardia*, *Nocardioides*, and *Kineococcus*). Many of these organisms are fellow soil-dwelling bacteria, which plausibly suggests lateral gene transfer events.

DISCUSSION

Acquisition of new functions by exchange of extremities between linear replicons. The *S. ambofaciens* TIRs contain a large proportion of strain- and species-specific genes as well as

sequences potentially involved in genome plasticity. Genetic organization of the TIRs is consistent with the idea that DNA rearrangements of endogenous sets of genes (duplication and translocation), integration of exogenous information, and/or deletions have been fixed during evolution from their common ancestor.

Several lines of evidence corroborate the acquisition of the terminal strain-specific regions by exchange of replicon extremities with linear plasmids: the presence of gene clusters homologous to plasmid-associated ones; the presence of genes implicated in conjugal transfer (*kilB* and *ttrA*); and their low G+C content (68.8% and 69.2%), which is characteristic of many *Streptomyces* plasmids. In addition, many similarities are found with genes carried by the extremities of the linear plasmids (*ttrA*, *tpgC*, *lig*, and *kilB*). Since the last telomere-proximal ORF conserved in the two strains is a truncated IS, it is plausible that exchange of extremities could have happened by homologous recombination involving two IS copies.

This comparison made by use of *S. ambofaciens* describes for the first time such exchanges at the origin of intraspecific variability in natural isolates. Insertion of DNA extremities of linear replicons may be a favored mechanism for gene acquisition in *Streptomyces*. Indeed, exchange of the terminal parts requires a single crossover event, the success of which would be guaranteed by the presence of the telomeres in the two replicons. In addition, the presence of a helicase-like gene (*ttrA*) closely associated with the telomere is a trait common to almost all *Streptomyces* linear replicons. However, it is truncated in strain DSM40697, and no deleterious effect could be assigned in laboratory growth conditions to the mutation of both copies of *ttrA* in *S. lividans* (17). The strong conservation of a gene in a variable region appears paradoxical, but it suggests a role for conferring some long-term advantage, e.g., conjugal transfer, as suggested by C.W. Chen in his "end-first" model (8). This model predicts that this probable helicase would be involved in conjugal transfer by acting on the DNA terminus which would correspond to an origin of transfer (7, 8).

Many of the functions predicted for the strain-specific regions are potentially related to adaptation, which consolidates the idea that horizontal transfer is at the origin of the *S. ambofaciens* terminal variability. In general, genes successfully transferred are responsible for adaptation to the environment. The presence in the DSM40697 strain chromosomal extremities of multiple resistance genes may have conferred advantages responsible for their maintenance in the bacterial population.

In addition to exchange of extremities, multiple events of insertions/deletions have occurred recently both inside and outside of the TIRs. For example, specific region b of the strain ATCC23877 chromosome could be the result of an integration event, but the absence of a detectable target site suggests either integration by illegitimate recombination or deletion of this locus in strain DSM40697. The specific regions C (DSM40697) and d (ATCC23877), which are two different variable regions located at the same locus, could be the consequence of a DNA exchange by double crossover or by different deletions at the same locus in the two strains.

All these multiple rearrangements are posterior to the formations of the TIRs that are themselves totally different from those of the very close *S. coelicolor* genome. Thus, genome plasticity is extremely strong in the terminal regions, which could be privileged targets for environmental adaptation by loss, acquisition, and creation of new functions. They might be considered as a vector for genetic transfer and also as the "Swiss army knife" of the "boy scout" *Streptomyces* (6, 16).

Origin and evolution of the size of the TIRs. Although the presence of two telomeres is known to be essential for the maintenance of chromosome linearity (2), no role has so far been attributed to the TIRs. Indeed, some replicons, such as the *S. avermitilis* chromosome, have TIRs restricted to a part of the telomeres, while others carry very large TIRs (up to 1.4 Mb [46]). In addition, the duplication of the genes located in the TIRs does not appear to be a mechanism for gene regulation. Thus, there is no obvious difference between the transcriptional levels of the genes duplicated in the *S. coelicolor* M600 TIRs (1 Mb) and those for the same set of genes present in single copy in *S. coelicolor* M145 (22-kb TIRs [43]).

The formation of TIRs could be the consequence of terminal recombination induced either by dysfunction of a telomere or by formation of double-strand breaks (DSBs). Chromosomal rescue could then be achieved by recombination with a DNA fragment including a telomere similar to the endogenous one. This fragment may be a broken daughter chromatid (46) or a plasmidic or chromosomal extremity either endogenously present or resulting from horizontal gene transfer (e.g., conjugal transfer). DSB repair might also result from the break-induced replication as described previously for *Saccharomyces cerevisiae* (29). Finally, circularization can also trigger chromosomal rescue (reviewed in reference 25).

Analysis of the *S. ambofaciens* TIR boundary reveals a preliminary step in TIR shortening (Fig. 2A). Thus, while the strict border of the TIRs is located at different positions in the two strains, the imperfect duplication includes the same genes. Point mutations and variability in tandem repeated motifs contribute to the decrease in length of the TIRs. In *S. avermitilis*, the terminal repetitions are restricted to the telomeres (19), which suggests the loss of ancestral TIRs in this species. Indeed, detailed analysis of the subtelomeric sequences reveals putative traces of ancestral TIRs (Fig. 2B). The last predicted ORF, SAV7573, is in fact imperfectly duplicated at the other end of the chromosome (Fig. 2B), and a putative CDS, which we tentatively called SAV0, can be detected between the telomere and SAV1. SAV0 (372 bp) and SAV7573 (411 bp) share 87% nucleotide identity over most of the CDS length (271/311 nucleotides). Furthermore, although only weakly similar (38% amino acid identity), a duplicated helicase gene is also present at both ends of the chromosome (SAV6/SAV7571). SAV2 also shows similarity with the helicase-like gene (Fig. 2B) and could correspond to the duplication of the 5′ end of SAV6 (69% nucleotide identity). These data suggest that once formed, the TIRs may accumulate point or insertional mutations that result in divergence which might progressively reduce and even prevent the frequent homogenization between TIRs and consequently accelerate the rate of divergence. The TIRs would then tend to degenerate by progressive shortening. The divergence of the duplicated genes present in the TIRs may result in pseudogene formation or leave functional coding sequences evolving toward new functions (34). Alternatively, the loss of TIRs in *S. avermitilis* could result from a single recombination

event between the ancestral chromosome and an exogenous DNA molecule.

The reduction in TIR size may be balanced by recombination events leading to expansion. Enlarged TIRs have been reported for mutants of *S. griseus* and *S. ambofaciens* (12, 41). In both cases, recombination events occurring between duplicated genes (each copy located on a different arm and in a divergent orientation) led to a dramatic increase of the TIR size (from 210 kb to 480 kb and 850 kb in *S. ambofaciens* and from 24 kb to 450 kb in *S. griseus*). For *S. coelicolor*, TIR length variation has been shown to be associated with strain lineage (43). The formation of large TIRs could be ascribed to homologous recombination between transposed IS*110* copies. TIR shortening from 1.06 Mb to 22 kb was observed (43). It therefore seems possible that TIR formation corresponds to a by-product of chromosomal rescue mechanisms. However, once formed, terminal duplications would provide an appropriate substrate for homologous recombination and thus for DNA repair of DSBs occurring in the terminal parts of the *Streptomyces* linear chromosome. The fact that the *S. ambofaciens* strain-specific regions are duplicated on both arms in a given chromosome strongly suggests a homogenization mechanism. In other words, these data show that DNA rearrangements occurring within a copy of TIRs can be followed by homogenization of both arms, probably by intrachromosomal recombination between duplicated sequences.

ACKNOWLEDGMENTS

F.C. and A.G. were recipients of a grant from the "Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche" (M.E.N.E.S.R.). This research was supported by the PAI (ALLIANCE), the "ACI Microbiologie 2003" programs funded by the M.E.N.E.S.R., and the VIth PCRDT ("ActinoGen").

Many thanks are due to K. Chater, G. Chandra, and T. Kieser (John Innes Centre, Norwich, United Kingdom) for their warm welcome and their help in the development of the computational methods. Many thanks to B. Segurens (Génoscope, CNS) for her help. We are grateful to A. Hesketh (John Innes Centre, Norwich, United Kingdom) and Eriko Takano (University of Groningen, Groningen, The Netherlands) for critical reading of the manuscript.

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
2. **Bao, K., and S. N. Cohen.** 2001. Terminal proteins essential for the replication of linear plasmids and chromosomes in *Streptomyces*. Genes Dev. **15:** 1518–1527.
3. **Barrasa, M. I., J. A. Tercero, and A. Jimenez.** 1997. The aminonucleoside antibiotic A201A is inactivated by a phosphotransferase activity from *Streptomyces capreolus* NRRL 3817, the producing organism. Isolation and molecular characterization of the relevant encoding gene and its DNA flanking regions. Eur. J. Biochem. **245:**54–63.
4. **Bartolome, B., Y. Jubete, E. Martinez, and F. de la Cruz.** 1991. Construction and properties of a family of pACYC184-derived cloning vectors compatible with pBR322 and its derivatives. Gene **102:**75–78.
5. **Bentley, S. D., S. Brown, L. D. Murphy, D. E. Harris, M. A. Quail, J. Parkhill, B. G. Barrell, J. R. McCormick, R. I. Santamaria, R. Losick, M. Yamasaki, H. Kinashi, C. W. Chen, G. Chandra, D. Jakimowicz, H. M. Kieser, T. Kieser, and K. F. Chater.** 2004. SCP1, a 356,023 bp linear plasmid adapted to the ecology and developmental biology of its host, *Streptomyces coelicolor* A3(2). Mol. Microbiol. **51:**1615–1628.
6. **Bentley, S. D., K. F. Chater, A. M. Cerdeno-Tarraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C. H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M. A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, and D. A. Hopwood.** 2002. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature **417:**141–147.
7. **Bey, S. J., M. F. Tsou, C. H. Huang, C. C. Yang, and C. W. Chen.** 2000. The homologous terminal sequence of the *Streptomyces lividans* chromosome and SLP2 plasmid. Microbiology **146:**911–922.
8. **Chen, C. W.** 1996. Complications and implications of linear bacterial chromosomes. Trends Genet. **12:**192–196.
9. **Chen, C. W., C. H. Huang, H. H. Lee, H. H. Tsai, and R. Kirby.** 2002. Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. Trends Genet. **18:**522–529.
10. **Cortes, J., J. Velasco, G. Foster, A. P. Blackaby, B. A. Rudd, and B. Wilkinson.** 2002. Identification and cloning of a type III polyketide synthase required for diffusible pigment biosynthesis in *Saccharopolyspora erythraea*. Mol. Microbiol. **44:**1213–1224.
11. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641.
12. **Fischer, G., B. Decaris, and P. Leblond.** 1997. Occurrence of deletions, associated with genetic instability in *Streptomyces ambofaciens*, is independent of the linearity of the chromosomal DNA. J. Bacteriol. **179:**4553–4558.
13. **Fischer, G., A. Kyriacou, B. Decaris, and P. Leblond.** 1997. Genetic instability and its possible evolutionary implications on the chromosomal structure of *Streptomyces*. Biochimie **79:**555–558.
14. **Fischer, G., T. Wenner, B. Decaris, and P. Leblond.** 1998. Chromosomal arm replacement generates a high level of intraspecific polymorphism in the terminal inverted repeats of the linear chromosomal DNA of *Streptomyces ambofaciens*. Proc. Natl. Acad. Sci. USA **95:**14296–14301.
15. **Hacker, J., and E. Carniel.** 2001. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep. **2:**376–381.
16. **Hopwood, D. A.** 2003. The *Streptomyces* genome—be prepared! Nat. Biotechnol. **21:**505–506.
17. **Huang, C. H., C. Y. Chen, H. H. Tsai, C. Chen, Y. S. Lin, and C. W. Chen.** 2003. Linear plasmid SLP2 of *Streptomyces lividans* is a composite replicon. Mol. Microbiol. **47:**1563–1576.
18. **Hütter, R.** 1967. Systematik der *Streptomycete*. Karger Verlag, Basel, Switzerland.
19. **Ikeda, H., J. Ishikawa, A. Hanamoto, M. Shinose, H. Kikuchi, T. Shiba, Y. Sakaki, M. Hattori, and S. Omura.** 2003. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. Nat. Biotechnol. **21:**526–531.
20. **Ishikawa, J., A. Yamashita, Y. Mikami, Y. Hoshino, H. Kurita, K. Hotta, T. Shiba, and M. Hattori.** 2004. The complete genomic sequence of *Nocardia farcinica* IFM 10152. Proc. Natl. Acad. Sci. USA **101:**14925–14930.
21. **Kawasaki, T., T. Kuzuyama, Y. Kuwamori, N. Matsuura, N. Itoh, K. Furihata, H. Seto, and T. Dairi.** 2004. Presence of copalyl diphosphate synthase gene in an actinomycete possessing the mevalonate pathway. J. Antibiot. (Tokyo) **57:**739–747.
22. **Krugel, H., G. Fiedler, C. Smith, and S. Baumberg.** 1993. Sequence and transcriptional analysis of the nourseothricin acetyltransferase-encoding gene *nat1* from *Streptomyces noursei*. Gene **127:**127–131.
23. **Lawrence, J. G., and H. Hendrickson.** 2005. Genome evolution in bacteria: order beneath chaos. Curr. Opin. Microbiol. **8:**572–578.
24. **Lawrence, J. G., and J. R. Roth.** 1999. Genomic flux: genome evolution by gene loss and acquisition in bacterial genomes, p. 263–289. *In* R. L. Charlebois (ed.), Organization of the prokaryotic genome. American Society for Microbiology, Washington D.C.
25. **Leblond, P., and B. Decaris.** 1999. "Unstable" linear chromosomes: the case of *Streptomyces*, p. 235–261. *In* R. L. Charlebois (ed.), Organization of the prokaryotic genome. American Society for Microbiology, Washington, D.C.
26. **Leblond, P., G. Fischer, F. X. Francou, F. Berger, M. Guérineau, and B. Decaris.** 1996. The unstable region of *Streptomyces ambofaciens* includes 210 kb terminal inverted repeats flanking the extremities of the linear chromosomal DNA. Mol. Microbiol. **19:**261–271.
27. **Metsa-Ketela, M., L. Halo, E. Munukka, J. Hakala, P. Mantsala, and K. Ylihonko.** 2002. Molecular evolution of aromatic polyketides and comparative sequence analysis of polyketide ketosynthase and 16S ribosomal DNA genes from various *Streptomyces* species. Appl. Environ. Microbiol. **68:**4472–4479.
28. **Mochizuki, S., K. Hiratsu, M. Suwa, T. Ishii, F. Sugino, K. Yamada, and H. Kinashi.** 2003. The large linear plasmid pSLA2-L of *Streptomyces rochei* has an unusually condensed gene organization for secondary metabolism. Mol. Microbiol. **48:**1501–1510.
29. **Morrow, D. M., C. Connelly, and P. Hieter.** 1997. "Break copy" duplication: a model for chromosome fragment formation in *Saccharomyces cerevisiae*. Genetics **147:**371–382.
30. **Pandza, S., G. Biukovic, A. Paravic, A. Dadbin, J. Cullum, and D. Hranueli.** 1998. Recombination between the linear plasmid pPZG101 and the linear chromosome of *Streptomyces rimosus* can lead to exchange of ends. Mol. Microbiol. **28:**1165–1176.
31. **Pang, X., B. Aigle, J. M. Girardet, S. Mangenot, J. L. Pernodet, B. Decaris,**

and P. Leblond. 2004. Functional angucycline-like antibiotic gene cluster in the terminal inverted repeats of the *Streptomyces ambofaciens* linear chromosome. Antimicrob. Agents Chemother. **48:**575–588.

32. **Pinnert-Sindico, S.** 1954. Une nouvelle espèce de *Streptomyces* productrice d'antibiotiques: *Streptomyces ambofaciens* n. sp. caractères culturaux. Ann. Inst. Pasteur (Paris) **87:**702–707.
33. **Rocha, E. P., and A. Danchin.** 2002. Base composition bias might result from competition for metabolic resources. Trends Genet. **18:**291–294.
34. **Roth, V., B. Aigle, R. Bunet, T. Wenner, C. Fourrier, B. Decaris, and P. Leblond.** 2004. Differential and cross-transcriptional control of duplicated genes encoding alternative sigma factors in *Streptomyces ambofaciens*. J. Bacteriol. **186:**5355–5365.
35. **Sakaguchi, K.** 1990. Invertrons, a class of structurally and functionally related genetic elements that includes linear DNA plasmids, transposable elements, and genomes of adeno-type viruses. Microbiol. Rev. **54:**66–74.
36. **Schully, K. L., and G. S. Pettis.** 2003. Separate and coordinate transcriptional control mechanisms link expression of the potentially lethal KilB spread locus to the upstream transmission operon on *Streptomyces* plasmid pIJ101. J. Mol. Biol. **334:**875–884.
37. **Sekine, M., S. Tanikawa, S. Omata, M. Saito, T. Fujisawa, N. Tsukatani, T. Tajima, T. Sekigawa, H. Kosugi, Y. Matsuo, R. Nishiko, K. Imamura, M. Ito, H. Narita, S. Tago, N. Fujita, and S. Harayama.** 2006. Sequence analysis of three plasmids harboured in *Rhodococcus erythropolis* strain PR4. Environ. Microbiol. **8:**334–346.
38. **Spatz, K., H. Kohn, and M. Redenbach.** 2002. Characterization of the *Streptomyces violaceoruber* SANK95570 plasmids pSV1 and pSV2. FEMS Microbiol. Lett. **213:**87–92.
39. **Suzek, B. E., M. D. Ermolaeva, M. Schreiber, and S. L. Salzberg.** 2001. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics **17:**1123–1130.
40. **Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin.** 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29:**22–28.
41. **Uchida, T., M. Miyawaki, and H. Kinashi.** 2003. Chromosomal arm replacement in *Streptomyces griseus*. J. Bacteriol. **185:**1120–1124.
42. **Volff, J. N., and J. Altenbuchner.** 2000. A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. FEMS Microbiol. Lett. **186:**143–150.
43. **Weaver, D., N. Karoonuthaisiri, H. H. Tsai, C. H. Huang, M. L. Ho, S. Gai, K. G. Patel, J. Huang, S. N. Cohen, D. A. Hopwood, C. W. Chen, and C. M. Kao.** 2004. Genome plasticity in *Streptomyces*: identification of 1 Mb TIRs in the *S. coelicolor* A3(2) chromosome. Mol. Microbiol. **51:**1535–1550.
44. **Wendt-Pienkowski, E., Y. Huang, J. Zhang, B. Li, H. Jiang, H. Kwon, C. R. Hutchinson, and B. Shen.** 2005. Cloning, sequencing, analysis, and heterologous expression of the fredericamycin biosynthetic gene cluster from *Streptomyces griseus*. J. Am. Chem. Soc. **127:**16442–16452.
45. **Wenner, T., V. Roth, B. Decaris, and P. Leblond.** 2002. Intragenomic and intraspecific polymorphism of the 16S–23S rDNA internally transcribed sequences of *Streptomyces ambofaciens*. Microbiology **148:**633–642.
46. **Wenner, T., V. Roth, G. Fischer, C. Fourrier, B. Aigle, B. Decaris, and P. Leblond.** 2003. End-to-end fusion of linear deleted chromosomes initiates a cycle of genome instability in *Streptomyces ambofaciens*. Mol. Microbiol. **50:**411–425.
47. **Yamasaki, M., and H. Kinashi.** 2004. Two chimeric chromosomes of *Streptomyces coelicolor* A3(2) generated by single crossover of the wild-type chromosome and linear plasmid SCP1. J. Bacteriol. **186:**6553–6559.
48. **Zdobnov, E. M., and R. Apweiler.** 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:** 847–848.

DISCUSSION

DISCUSSION

L'instabilité génétique chez *Streptomyces* est étudiée depuis plusieurs décennies (pour revues (Chen *et al.*, 2002 ; Leblond et Decaris, 1999 ; Volff et Altenbuchner, 2000)). L'apparition de mutants spontanés à haute fréquence (jusqu'à 1% chez *S. ambofaciens*) a été associée à des événements moléculaires de plusieurs natures : des délétions de grande taille parfois associées à des amplifications, des remplacements de bras chromosomiques, la circularisation du chromosome et des fusions de chromosomes chez *S. ambofaciens* (Fischer *et al.*, 1997a ; Fischer *et al.*, 1998b ; Wenner *et al.*, 2003).

Qu'apporte la séquence du génome à l'étude de ce phénomène ? Comment se traduit l'instabilité observée en laboratoire en terme d'évolution des chromosomes des souches naturelles de *Streptomyces* ? L'instabilité révèle-t-elle une tolérance accrue des régions terminales pour les réarrangements ou existe-t-il des mécanismes spécifiques associés à la plasticité de ces régions ?

1. *Extrémités chromosomiques et adaptation*

Les extrémités chromosomiques sont des régions hautement variables chez *Streptomyces*, tant au niveau interspécifique qu'intraspécifique. Chaque espèce de *Streptomyces* possède donc un grand nombre de fonctions spécifiques, dont les plus étudiées pour des raisons médicales et économiques sont les voies de biosynthèse de métabolites secondaires.
Le séquençage des régions instables du chromosome de *S. ambofaciens* a permis de mettre en évidence 12 voies de biosynthèse de composés potentiellement intéressants d'un point de vue applicatif. Outre leur aspect valorisable, ces voies de biosynthèse peuvent être reliées à l'adaptation de chaque espèce vis-à-vis de leur environnement.
Le sol est un milieu complexe et changeant tant au niveau des paramètres physico-chimiques qu'au niveau des interactions avec d'autres organismes. La diversité des molécules produites, dont certaines possèdent une activité antibiotique, peut être interprétée comme la mise en place d'un arsenal de défense contre les organismes en compétition pour les ressources. De plus, la diversité et la multiplicité des antibiotiques produits par une seule souche peut refléter la diversité des organismes compétiteurs qui peuvent être présents dans la même niche écologique (ex : bactéries Gram négative et Gram positive, champignons). Cependant, il n'est pas exclu que ces molécules jouent un tout autre rôle dans le développement des colonies de *Streptomyces*.
L'adaptation à l'écosystème sol peut également être reliée aux capacités de biosynthèse de nombreuses enzymes extracellulaires responsables de la dégradation de polymères retrouvés en abondance dans l'environnement (ex : cellulose, chitine, inuline).

Les régions terminales sont les loci privilégiés d'acquisition de nouvelles fonctions par transfert horizontal. Elles seraient donc très fortement impliquées dans l'adaptation à l'environnement. Les gènes portés par les régions de contingence sont peu exprimés pendant la croissance végétative mais leur transcription s'accroît pendant la phase stationnaire et au cours de différents stress (Karoonuthaisiri *et al.*, 2005). Ainsi, à l'opposé de la région "core" qui contient les gènes communs et

nécessaires à la croissance végétative, les fonctions spécifiques d'espèce semblent s'exprimer surtout pendant les phases tardives de croissance. C'est le cas du métabolisme secondaire qui se met en place au moment de la différentiation du mycélium.
Les capacités adaptatives des *Streptomyces* semblent remarquables. Ils sont retrouvés dans des milieux très différents. La plasticité des régions terminales pourrait donc être fortement associée à l'adaptabilité des *Streptomyces*.

2. *Acquisition de nouvelles fonctions par échanges d'extrémités de réplicons linéaires*

Au sein de l'espèce *S. ambofaciens*, la variabilité concerne principalement la partie terminale des TIR soit, respectivement, 60 kb et 48 kb pour les souches DSM40697 et ATCC23877. L'analyse comparative des séquences des TIR de ces deux souches a révélé que de multiples événements d'insertions/délétions ont été fixés depuis la divergence récente des deux lignées. Les régions variables représentent entre un quart et un tiers du contenu en gènes des TIR.
Plusieurs arguments suggèrent une acquisition de nouvelles fonctions par échanges d'extrémités de réplicons linéaires : la présence de clusters similaires à des gènes associés à des plasmides linéaires de *Streptomyces* dont certains ont été impliqués dans le transfert conjugatif (ex : *kilB*, *ttrA*) et le pourcentage plus faible en bases G+C (68,8% et 69,2%) typique de celui des plasmides chez *Streptomyces*.
Des interactions entre chromosomes et plasmides linéaires ont été mises en évidence expérimentalement. Par exemple, chez *S. coelicolor*, le chromosome et le plasmide SCP1 peuvent recombiner et générer deux molécules hybrides (Yamasaki et Kinashi, 2004). Un événement de recombinaison illégitime a engendré l'échange des 1,6 Mb de l'extrémité droite du chromosome contre les 130 kb terminaux du plasmide SCP1. Les deux molécules hybrides ainsi générées ont été définies comme deux chromosomes car il n'a pas été possible, après curage par divers traitements mutagènes, d'isoler des dérivés ayant perdu l'une ou l'autre de ces molécules. Chez *S. rimosus*, un événement de recombinaison illégitime entre le plasmide linéaire pZG101 et le chromosome de la souche sauvage a généré un échange d'extrémités, provoquant le passage du cluster de biosynthèse de l'oxytétracycline du chromosome vers le plasmide (Pandza *et al.*, 1998). Enfin chez *S. lividans*, un échange d'extrémités est fortement suggéré par l'analyse des régions terminales du chromosome sauvage et du plasmide linéaire SLP2 (Huang *et al.*, 2003). L'extrémité droite (15,4 kb) du plasmide est identique à celle du chromosome. De plus, le transposon Tn*4811* est localisé à la jonction des régions potentiellement échangées, suggérant qu'un événement de recombinaison homologue ou de transposition réplicative est à l'origine de la création d'un plasmide linéaire dont une extrémité provient du chromosome.
Chacun de ces exemples décrit un échange entre réplicons linéaires présents de façon endogène (intragénomique). L'étude des extrémités des TIR des deux souches de *S. ambofaciens* suggère l'implication de ce mécanisme évolutif à l'origine de la variabilité des souches naturelles.
Ce type d'échanges ne nécessite qu'un unique événement de recombinaison illégitime, de recombinaison homologue ou encore de transposition réplicative, dont le succès évolutif serait garanti par la présence de séquences télomériques aux extrémités des séquences nouvellement acquises. Il pourrait constituer un mécanisme privilégié d'acquisition de nouvelles fonctions chez *Streptomyces*.

Cette analyse permet donc d'établir un lien entre linéarité chromosomique et variabilité. En effet, le caractère linéaire pourrait être avantageux à long terme, et donc maintenu, car il favoriserait les flux de gènes aux extrémités et augmenterait donc les capacités d'adaptation des *Streptomyces*. De nombreuses fonctions putativement codées par les régions spécifiques de souches chez *S. ambofaciens* peuvent être associées directement à l'adaptation à l'environnement, notamment à la résistance à des antibiotiques (nourséotricine (Barrasa *et al.*, 1997) et A201A (Krugel *et al.*, 1993)) et à l'arsenate. La présence de fonctions non essentielles, pouvant conférer un avantage immédiat à l'organisme, et la spécificité de souche sont deux arguments en faveur d'une acquisition des extrémités par transfert horizontal.

Chez *S. ambofaciens*, la probabilité d'un échange d'extrémités est renforcée par la présence d'une IS tronquée localisée au niveau de la borne des régions spécifiques "a" (chez la souche ATCC23877) et "A" (chez la souche DSM40697, Art. 2, Fig.1). Cette ORF constitue la dernière séquence conservée entre les TIR des deux souches et pourrait avoir servi de substrat pour la recombinaison homologue. Etant donnée leur présence en copies multiples, les IS sont des substrats privilégiés pour la recombinaison intramoléculaire (réarrangements chromosomiques) et intermoléculaire (génération de coïntégrats ou échange d'ADN).
Le respect du principe d'un maximum de parcimonie conduit à favoriser l'hypothèse selon laquelle l'une des deux souches possède des TIR ancestrales (présentes dans l'ancêtre commun) alors que la seconde aurait acquis de nouvelles extrémités par échange. La souche DSM40697 correspondrait à cette dernière. En effet, elle présente le plus grand nombre d'arguments en cette faveur, notamment la présence de plusieurs clusters fortement similaires à des loci plasmidiques. Toutefois, si l'échange d'extrémités est un mécanisme privilégié d'évolution des chromosomes linéaires, il est probable que l'ancêtre commun des deux souches de *S. ambofaciens* possédait déjà des TIR dont les extrémités dérivaient de séquences plasmidiques. Ceci expliquerait pourquoi de telles séquences sont retrouvées dans les régions variables chez les deux souches.

Cependant, un fait module l'hypothèse d'un échange d'extrémités : la présence de télomères très conservés entre les deux souches de *S. ambofaciens*. Comment expliquer la conservation des 207 pb terminales incluses dans des régions totalement spécifiques de souches ? Les séquences des télomères des deux souches de *S. ambofaciens* sont, en effet, plus proches entre elles qu'elles ne le sont avec d'autres télomères connus. Néanmoins, la petite taille des télomères (<200 pb) pose le problème de la significativité des valeurs d'identités obtenues. Les télomères des deux souches de *S. ambofaciens* ont-ils une origine ancestrale ? Dans ce cas, il faut envisager que l'échange d'extrémités soit suivi d'un mécanisme de réparation du télomère perdu (probablement à partir du télomère resté intact). Deux étapes auraient donc pu être nécessaires au maintien de l'information acquise.
Dans le cas contraire, cela pourrait signifier que les échanges d'extrémités sont limités par la présence de télomères très fortement similaires entre les deux molécules partenaires.

En plus des probables échanges des extrémités, les TIR et les régions adjacentes des souches naturelles évoluent par insertions/délétions ou remplacements de gènes (Art. 2, Fig. 1). Cependant, les régions variables localisées au même locus dans les deux souches ne sont pas forcément issues de

remplacements de fragments d'ADN et donc de transferts horizontaux. Des délétions différentielles au même locus dans les deux souches peuvent aboutir au même résultat.

3. *Dynamique de l'évolution des TIR*

a) Evolution de la taille des TIR

L'analyse des limites internes des TIR révèle une étape préliminaire dans le processus de diminution de la taille des TIR. Alors que la limite exacte des régions répétées parfaitement ne se situe pas au même locus dans les deux souches, la limite ancestrale des régions répétées de façon imparfaite est identique (Art. 2, Fig. 2A). Des mutations ponctuelles et des variations du nombre de courts motifs répétés contribuent à la diminution progressive de la taille des TIR chez *S. ambofaciens*. Chez *S. avermitilis*, les TIR sont restreintes aux séquences télomériques (169 pb) mais une analyse plus détaillée révèle des traces de l'existence de répétitions ancestrales plus longues (Art. 2, Fig. 2B). L'ORF terminale SAV7573 est en réalité dupliquée au niveau de l'autre extrémité. En effet, une CDS potentielle (appelée provisoirement SAV0), non décrite dans l'annotation initiale (Ikeda *et al.*, 2003) et présentant 87% d'identité nucléotidique avec cette dernière, est présente entre les télomères et SAV1. D'autres traces de duplication des extrémités ancestrales ont par ailleurs été identifiées. Ainsi, les gènes SAV6 et SAV7571, présents à chacune des extrémités, sont également similaires.

Ces données suggèrent qu'une fois formées, les séquences des TIR divergent de façon progressive en accumulant des mutations réduisant ainsi l'efficacité de la recombinaison homologue. Alternativement, chez *S. avermitilis*, la disparition des TIR pourrait résulter d'un unique événement de remplacement d'une extrémité.

Cette réduction de la taille des TIR pourrait être contrebalancée par des événements de recombinaison provoquant leur expansion. En effet, des mutants possédant des TIR de taille largement agrandie ont été isolés chez *S. griseus* et *S. ambofaciens* (Fischer *et al.*, 1997a ; Uchida *et al.*, 2003). Celles-ci sont passées de 210 kb à 480 kb et 850 kb chez *S. ambofaciens* et de 24 kb à 450 kb chez *S. griseus*. Chez *S. coelicolor*, une variation de la taille des TIR a été associée à l'évolution des différentes lignées. Alors que la souche séquencée possède des TIR de 22 kb, de nombreux dérivés de laboratoire possèdent des TIR de 1,06 Mb. Ces dernières correspondraient à la situation ancestrale, c'est-à-dire à celle de la souche isolée originellement du sol (Weaver *et al.*, 2004). Un événement de recombinaison homologue entre deux copies d'IS aurait abouti à la réduction des TIR de 1,06 Mb à 22 kb.

b) Homogénéisation des TIR

La limite ancestrale des TIR est la même dans les deux souches de *S. ambofaciens*. Ainsi, tous les réarrangements constatés depuis la divergence des souches sont postérieurs à la formation des TIR chez cette espèce. Par ailleurs, les TIR de *S. ambofaciens* sont différentes de celles de l'espèce proche *S. coelicolor* tant en termes de taille (22 kb) que de contenu en gènes.

Ces données impliquent donc une origine extrêmement récente (à l'échelle évolutive) des régions spécifiques de souche. Malgré leur apparition récente, ces régions spécifiques sont dupliquées sur les

deux bras chromosomiques. Ceci suggère fortement un mécanisme d'homogénéisation des deux copies de TIR chez *S. ambofaciens*.
En d'autres termes, les réarrangements d'ADN survenant à l'intérieur des TIR peuvent être suivis d'une homogénéisation des deux copies. Ce processus peut aboutir à un retour aux TIR ancestrales, c'est-à-dire à une élimination de l'ADN nouvellement acquis. Dans le cas contraire, l'homogénéisation peut donner lieu à une duplication de l'ADN acquis au sein de nouvelles TIR et donc à sa fixation à court terme.
Cette propension à l'homogénéisation des deux TIR a été vérifiée expérimentalement chez *S. ambofaciens* par la caractérisation de mutants dirigés du cluster *alp* (Aigle, comm. pers.). En effet, pour certaines souches, les mutations (délétions de gènes) sélectionnées dans l'une des deux copies des TIR ont été recopiées sur la seconde copie au cours des générations suivantes.

Deux mécanismes différents, impliquant la recombinaison homologue entre des séquences dupliquées car appartenant aux TIR, peuvent être responsables de l'homogénéisation.

La première hypothèse implique une recombinaison interchromosomique (favorisée par la présence simultanée de plusieurs copies du chromosome dans le mycélium végétatif), c'est-à-dire un échange d'extrémités par crossing-over entre deux chromosomes nouvellement répliqués.

La seconde hypothèse implique un mécanisme de réparation des cassures double-brin par recopiage d'un bras chromosomique sur l'autre. Il a été proposé que ce mécanisme de remplacement de bras chromosomiques soit impliqué dans la formation de TIR de grande taille chez *S. ambofaciens* (Fischer *et al.*, 1998b). En effet, les gènes dupliqués *hasL* et *hasR* (98% d'identité nucléotidique), localisés respectivement à 480 kb et 850 kb des extrémités chromosomiques, ont été les substrats d'une recombinaison homologue générant la délétion d'un bras chromosomique et son remplacement par recopiage de l'autre bras utilisé comme matrice.

Le mécanisme de réplication induite par les cassures double-brin (appelé BIR pour "break-induced replication" chez *S. cerevisiae* (Morrow *et al.*, 1997)) pourrait être impliqué dans la génération de TIR. La région localisée au niveau de la borne des TIR présente une composition nucléotidique spécifique et différente de la signature du génome de *S. ambofaciens*. Cette région est, par ailleurs, facilement détectable par son profil de fréquences en dinucléotides (signature de Karlin, Fig. 39A). Elle est constituée d'une grande région intergénique (>2 kb) et du pseudogène *afsA* (Art. 2, Fig. 2). Cette région est enrichie en motifs répétés (riche en bases C+A : 70%). Les pentanucléotides composés de 4 C et 1 A sont fortement surreprésentés ; par exemple, le motif ACCCC est répété, respectivement, 59 et 47 fois pour les souches DSM40697 et ATCC23877 (équivalent à une surreprésentation d'un facteur 6). Ce type de micro-répétitions peut engendrer des glissements de la fourche de réplication et des pauses de la machinerie provoquant des cassures du chromosome (Viguera *et al.*, 2001). Ainsi, cette région pourrait posséder les caractéristiques structurales d'un point privilégié d'apparition de cassures double-brin. Un argument conforte l'idée d'une instabilité locale associée à cette séquence répétée : il existe un polymorphisme marqué du nombre et de l'organisation de ces répétitions entre les deux souches (Fig. 39B) alors que ce polymorphisme est quasi inexistant entre les deux bras chromosomiques d'une même souche. Ce résultat indique que cette région est remaniée à haute

fréquence (divergence entre souches) mais qu'elle est homogénéisée à une fréquence au moins comparable voire plus élevée (identité entre bras chromosomiques).
La présence d'une région possédant des caractéristiques structurales semblables n'a été détectée sur aucun autre réplicon linéaire de *Streptomyces*. Cependant, le nombre de séquences de TIR disponibles reste faible et seules celles de *S. ambofaciens* sont de grande taille (>100 kb).

4. *La présence de TIR est-elle sélectionnée ?*

Jusqu'à maintenant, aucun rôle n'a été attribué aux TIR. Chez certaines espèces de *Streptomyces*, la duplication terminale est réduite à la séquence des télomères (voire à une partie des télomères) dont la fonction essentielle dans le maintien et la réplication du chromosome linéaire est prouvée (Qin et Cohen, 1998). Ainsi, la présence de TIR (c'est-à-dire d'une duplication terminale) n'est pas essentielle à l'hérédité de la structure chromosomique. Les TIR ont-elles néanmoins une fonction ? La présence de TIR dans le plasmide SCP1 est, par exemple, nécessaire à son intégration dans le chromosome de *S. coelicolor* (Hanafusa et Kinashi, 1992 ; Kinashi *et al.*, 1992). Certains auteurs ont mis en avant l'intérêt de posséder des répétitions terminales dans les mécanismes de restauration des télomères ayant subi des dommages (Uchida *et al.*, 2003). La présence de TIR pourrait être sélectionnée car elles potentialiseraient la réparation/restauration des télomères et donc le maintien de la linéarité.
De plus, les études réalisées sur le cluster *alp* inclus dans les TIR chez *S. ambofaciens* ont montré que les deux copies de cette voie de biosynthèse sont fonctionnelles (Pang *et al.*, 2004). Ces résultats indiquent que la duplication du cluster entraîne une augmentation du niveau de production de l'alpomycine. Cette augmentation de l'activité antibiotique aurait pu être sélectionnée.

L'hypothèse selon laquelle l'existence de TIR n'est pas sélectionnée est également envisageable. Leur présence pourrait n'être qu'une conséquence du processus de réparation des cassures double-brin et des dommages causés aux télomères.

5. *Un génome très hétérogène*

a) *Compartimentation génomique chez Streptomyces*

La génomique comparée du chromosome de *S. ambofaciens* avec ceux de *S. coelicolor*, *S. avermitilis* et *S. scabies* a révélé l'existence d'une structure extrêmement compartimentée dans laquelle la variabilité est confinée au niveau des extrémités chromosomiques. Ces régions terminales variables entre espèces représentent plusieurs centaines de kilobases, c'est-à-dire entre 10% et 20% du chromosome selon les comparaisons réalisées.
La région "core" apparaît comme présentant un niveau élevé de synténie parmi le genre *Streptomyces* alors que les régions de contingence sont enrichies en gènes spécifiques. Ces observations obtenues à partir des comparaisons de génomes rejoignent les résultats expérimentaux qui démontraient une instabilité particulièrement forte des régions terminales, notamment chez *S. ambofaciens* (Fischer *et al.*, 1997a ; Leblond *et al.*, 1989 ; Wenner *et al.*, 2003). Il apparaît que les flux de gènes entraînent une diversification de l'information génétique, non pas de façon uniforme le long du chromosome, mais de façon privilégiée dans régions terminales.

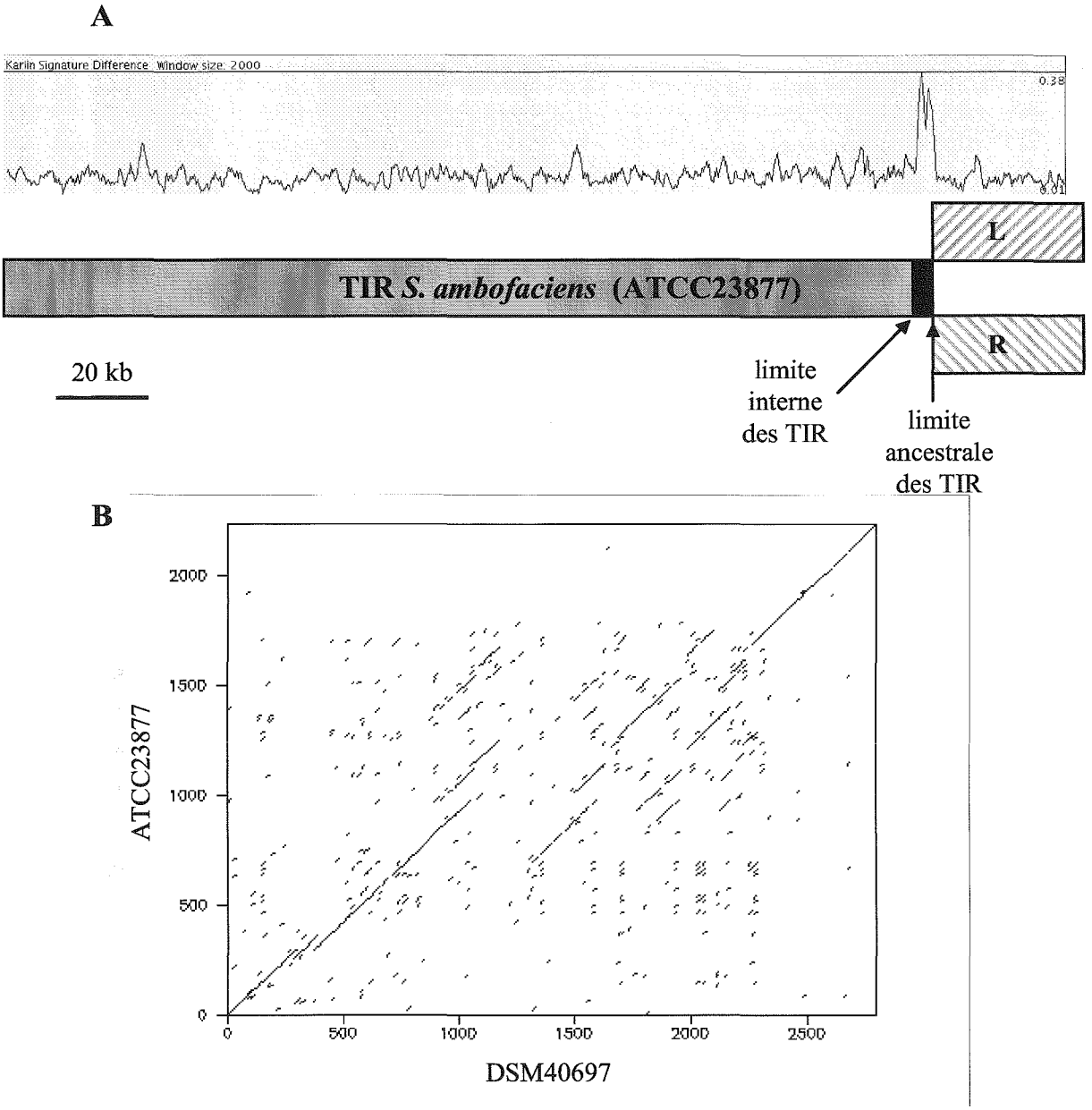


Figure 39 : Composition nucléotidique particulière de la région à la borne des TIR.
A. Profil de déviation des fréquences en dinucléotides (signature de Karlin) par rapport à la moyenne pour les TIR de *S. ambofaciens* ATCC23877 (calculée sur l'ensemble du bras chromosomique gauche de *S. ambofaciens* ATCC23877 avec une fenêtre glissante de 2000 pb et visualisée avec Artemis). Le même profil est obtenu pour les TIR de la souche DSM40697.
B. Comparaison nucléotidique par dot-plot de la région intergénique, séparant le gène codant putativement un cytochrome P450 et le pseudogène *afsA,* localisée à la borne des TIR des deux souches de *S. ambofaciens*. Chez les deux souches, cette région se compose de motifs répétés riches en bases C et A et l'absence de diagonale sur le graphe montre qu'un polymorphisme du nombre et de l'organisation de ces motifs existe entre les deux souches. Dot-plot réalisé avec une fenêtre de 10 nucléotides.

Les différences de vitesse d'évolution en fonction de la localisation chromosomique provoquent une hétérogénéité dans la composition de chacun des génomes qui associe une région ancestrale dédiée aux fonctions essentielles du développement végétatif à des régions subissant des flux responsables de l'adaptation. Si l'adaptation permet la colonisation d'un nouvel environnement et engendre un isolement génétique, elle peut être à l'initiation du processus de spéciation des *Streptomyces*.

Hormis l'enrichissement en gènes spécifiques dans les régions terminales, les signes d'une hétérogénéité sont identifiables par l'analyse de paramètres intrinsèques :

- *Contenu en bases A+T des séquences codantes*

Les chromosomes de *S. coelicolor* et *S. avermitilis* présentent un enrichissement en bases A+T au niveau des extrémités (Fig. 40) comme cela a été montré pour *S. ambofaciens* (Fig. 31). En revanche, la situation est moins claire pour le chromosome de *S. scabies*.

Cette caractéristique peut être le signe d'une acquisition récente par transfert horizontal. De plus, chez *S. ambofaciens*, les limites des régions montrant un biais en bases A+T correspondent à celles des régions spécifiques définies par rapport à *S. coelicolor* : elles mesurent 635 kb et 645 kb pour les bras chromosomiques gauche et droit (Fig. 31) alors que les régions spécifiques ont été estimées à, respectivement, 619 kb et 660 kb (Fig. 35).

Daubin et Perrière ont démontré que la région du chromosome opposée à l'origine de réplication est intrinsèquement plus riche en A+T dans la plupart des génomes étudiés, sans lien avec le transfert horizontal (Daubin et Perriere, 2003). Néanmoins, une telle correspondance entre régions spécifiques et richesse en bases A+T suggère fortement que les régions terminales du chromosome des *Streptomyces* sont majoritairement porteuses de gènes acquis récemment par transfert horizontal.

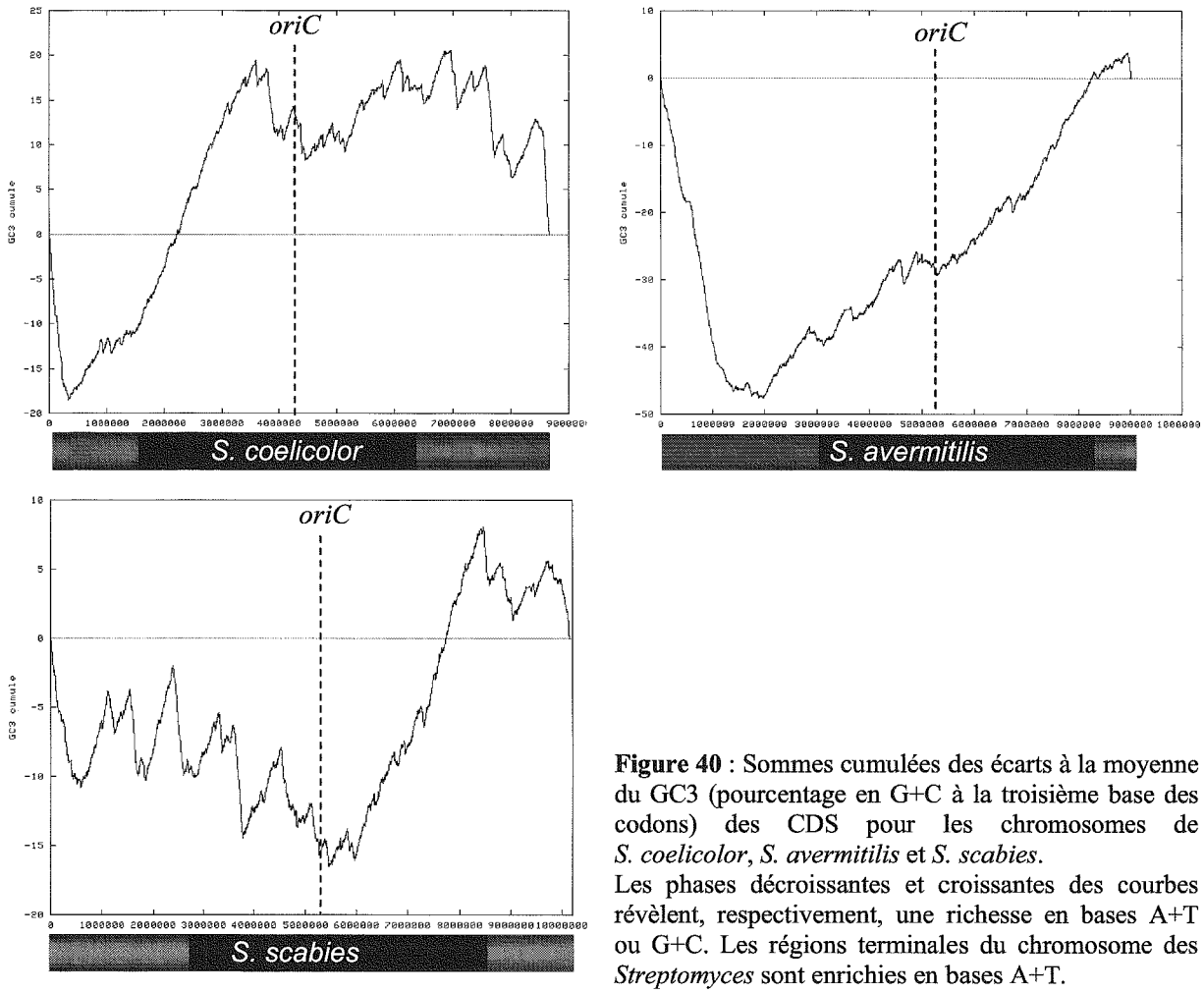


Figure 40 : Sommes cumulées des écarts à la moyenne du GC3 (pourcentage en G+C à la troisième base des codons) des CDS pour les chromosomes de *S. coelicolor*, *S. avermitilis* et *S. scabies*.
Les phases décroissantes et croissantes des courbes révèlent, respectivement, une richesse en bases A+T ou G+C. Les régions terminales du chromosome des *Streptomyces* sont enrichies en bases A+T.

- *Richesse en éléments transposables*

La variabilité et la plasticité des régions terminales sont également corrélées à une abondance en éléments transposables et en leurs dérivés (Fig. 41). Chez *S. coelicolor*, le dénombrement des ORF (et pseudogènes) codant potentiellement des transposases, intégrases et recombinases a révélé que 45% (42/94) d'entre elles sont présentes dans les extrémités, représentant 25% du chromosome (soit environ 1,1 Mb sur chaque bras). Ce biais est encore plus marqué chez *S. avermitilis* pour lequel ce pourcentage atteint 79% (99/126).

Il peut s'expliquer par deux phénomènes. Tout d'abord, une tolérance accrue des régions terminales, dépourvues de gènes essentiels, explique la contre-sélection moins efficace de la prolifération des IS. Le biais observé s'explique également par une acquisition de ces éléments et de leurs dérivés par transfert horizontal. En effet, les IS sont souvent portées par des éléments transférables. Chez *S. ambofaciens*, les régions terminales spécifiques sont porteuses d'IS (ou de leurs dérivés non fonctionnels) ne présentant aucune similarité avec les génomes de *Streptomyces* mais similaires à d'autres espèces d'Actinobactéries, *Frankia*, *Mycobacterium*, *Nocardioides* voire d'autres phyla (*Burkholderia* [*beta-Proteobacteria*] et *Nostoc* [*Cyanobacteria*] par exemple) suggérant une origine exogène.

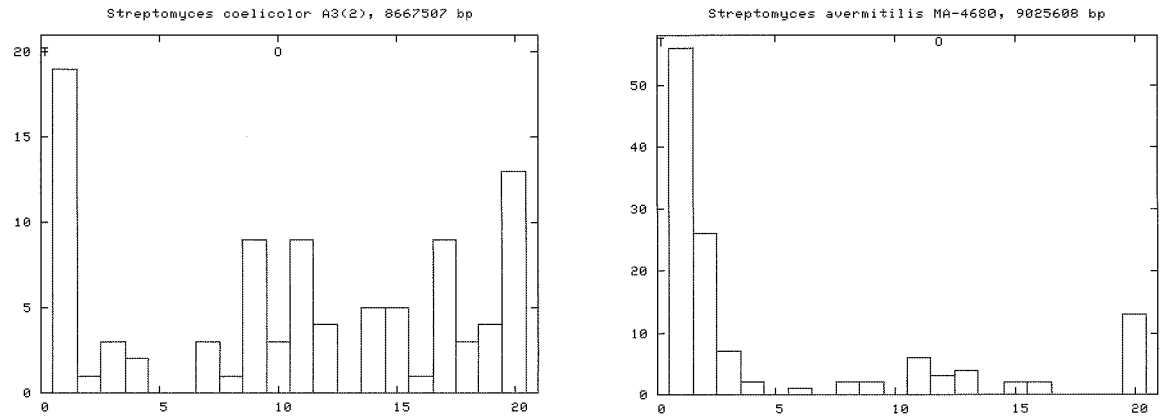


Figure 41 : Distribution des gènes associés à la plasticité génomique (transposases, intégrases, recombinases) dans les chromosomes de *S. coelicolor* (partie gauche) et *S. avermitilis* (partie droite).
Les chromosomes ont été divisés en 20 parties de taille identique et le nombre de gènes (et pseudogènes) associés à la plasticité dans chaque partie est représenté en ordonnées. "O" : origine de réplication.

- *Pseudogènes et densité en séquences codantes*

Deux autres paramètres, intimement liés, permettent de caractériser la compartimentation du génome : l'abondance en pseudogènes et la densité en séquences codantes.

Les génomes procaryotes sont en effet remarquables du point de vue de leur compacité en séquences codantes, si bien que la densité en gènes est un paramètre extrêmement peu variable chez l'ensemble des procaryotes (86,0 +/- 5,7%, calculée sur 338 génomes ; programme *gene_density.pl*). Le maintien à long terme de gènes nouvellement acquis au sein d'un génome dépend de leur fonction. L'ADN intégré au chromosome pourra être perdu à court ou moyen terme si sa fonction n'est pas sélectionnée. La création de pseudogènes est la première étape d'inactivation des gènes avant leur élimination

complète par dérive. La présence massive de gènes non fonctionnels est, en général, le reflet d'un relâchement de la sélection naturelle dans le contexte d'une adaptation à un nouvel environnement (comme c'est le cas pour les endosymbiotes ou parasites intracellulaires). Cependant, elle peut aussi être le reflet de flux de gènes de grande ampleur.

La problématique de l'identification des pseudogènes dans les génomes procaryotes est un sujet relativement récent. Les travaux réalisés par Lerat et Ochman ont permis de prendre conscience de leur abondance dans certains génomes (Lerat et Ochman, 2004, 2005). Il est donc délicat d'interpréter le nombre de pseudogènes prédits dans les annotations publiées tant il dépend du processus d'annotation lui-même.
Chez *S. ambofaciens*, 43 pseudogènes ont été initialement identifiés dans les séquences disponibles. Cependant, si les pseudogènes issus de mutations non sens et frameshifts sont aisément repérables, les séquences codantes tronquées (ou remaniées) le sont beaucoup moins. Il existe, en effet, un polymorphisme de taille des CDS homologues fonctionnelles et un seuil arbitraire doit être choisi afin de prédire l'existence d'une troncature inactivant une protéine. Chez *S. ambofaciens*, comme chez les autres *Streptomyces*, le nombre de pseudogènes prédits est donc sous-estimé (56 chez *S. coelicolor* (Bentley *et al.*, 2002) et 0 chez *S. avermitilis* (Ikeda *et al.*, 2003)). Une estimation plus précise a donc été entreprise en recherchant, parmi les CDS prédites, celles présentant une similarité d'au moins 60% avec une protéine de la banque NR mais ayant une taille inférieure à 80% de la taille de son homologue (critère choisi par (Lerat et Ochman, 2004)). D'après ces critères, 79 pseudogènes supplémentaires ont été prédits dans le génome partiel de *S. ambofaciens*. Ainsi, parmi les 2532 CDS prédites, 122 seraient non fonctionnelles, soit environ 5%. Il est à souligner que les pseudogènes inactivés par insertion d'IS n'ont pas été recherchés.
Dans les chromosomes complets de *S. coelicolor* et *S. avermitilis*, respectivement 119 et 203 séquences annotées comme CDS seraient, selon ces critères, des pseudogènes issus de troncatures. Chez *S. coelicolor*, 26% des pseudogènes prédits (175 au total) sont localisés dans les extrémités spécifiques par rapport à *S. ambofaciens* qui représentent moins de 9% du chromosome (753 kb). Chez *S. avermitilis*, 42% d'entre eux sont portés par les régions terminales spécifiques (représentant 14% du chromosome).
Chez *Streptomyces*, il existe donc un biais de distribution des pseudogènes, préférentiellement retrouvés dans les régions spécifiques.

Une preuve supplémentaire de la compartimentation du génome chez *Streptomyces* est apportée par le constat d'une hétérogénéité concernant le pourcentage en ADN codant. Chez *S. ambofaciens*, l'analyse de la proportion en séquences codantes le long des bras chromosomiques révèle que les régions terminales gauche et droite sont plus pauvres en gènes sur respectivement 600 kb et 400 kb environ (Fig. 42). A nouveau, ces régions coïncident avec la spécificité d'espèce observée et la richesse en A+T des régions terminales.
Cette analyse appliquée aux chromosomes complets de *S. coelicolor* et *S. avermitilis* révèle une situation analogue (Fig. 42). La disparité n'est pas retrouvée pour le chromosome de *S. scabies* pour qui seule une annotation automatique a été réalisée (sans correction manuelle). Cette différence pourrait s'expliquer probablement par une qualité d'annotation moindre.

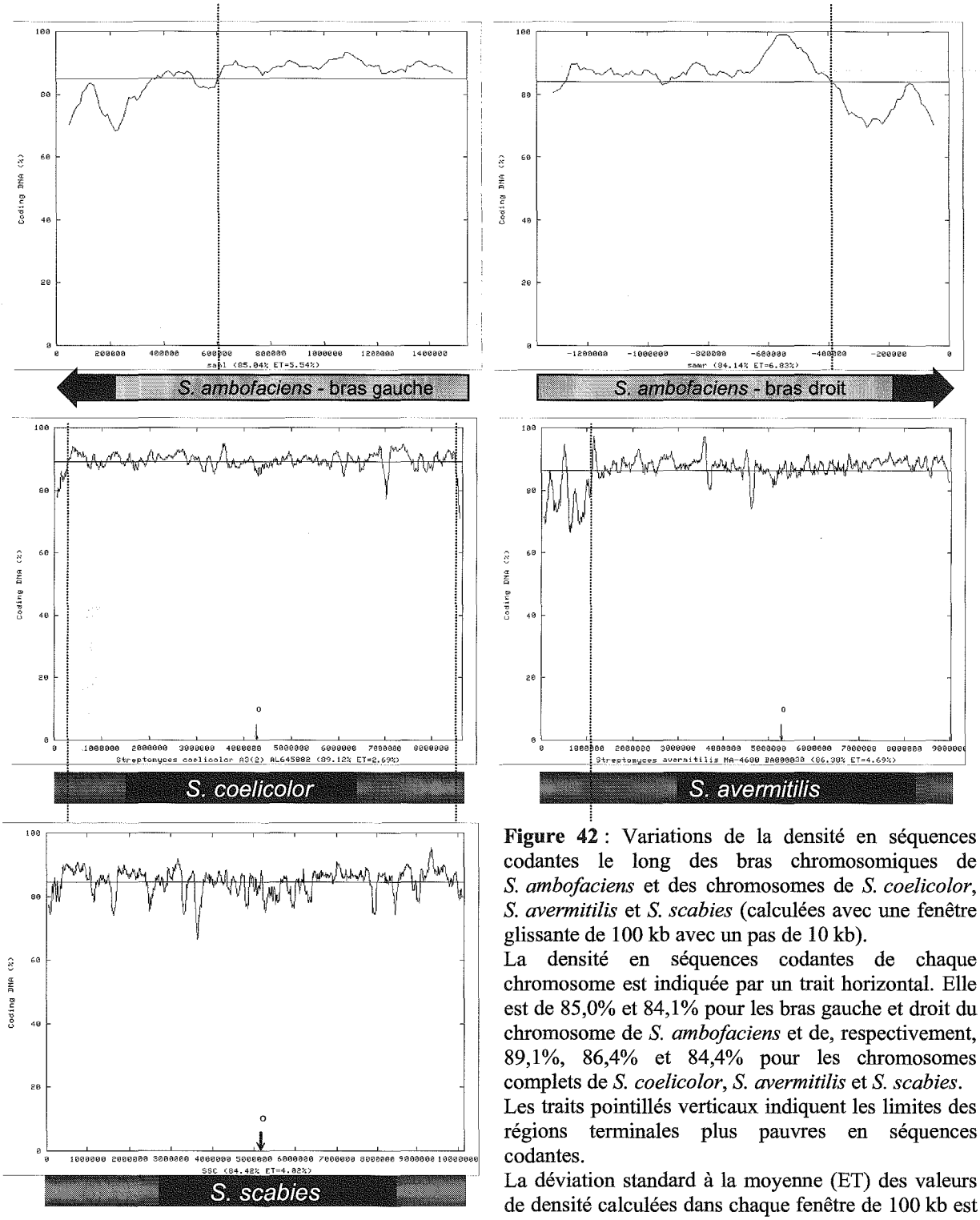


Figure 42 : Variations de la densité en séquences codantes le long des bras chromosomiques de *S. ambofaciens* et des chromosomes de *S. coelicolor*, *S. avermitilis* et *S. scabies* (calculées avec une fenêtre glissante de 100 kb avec un pas de 10 kb).
La densité en séquences codantes de chaque chromosome est indiquée par un trait horizontal. Elle est de 85,0% et 84,1% pour les bras gauche et droit du chromosome de *S. ambofaciens* et de, respectivement, 89,1%, 86,4% et 84,4% pour les chromosomes complets de *S. coelicolor*, *S. avermitilis* et *S. scabies*.
Les traits pointillés verticaux indiquent les limites des régions terminales plus pauvres en séquences codantes.
La déviation standard à la moyenne (ET) des valeurs de densité calculées dans chaque fenêtre de 100 kb est indiquée sous l'axe des abscisses. "O" : origine de réplication.

La densité plus faible en séquences codantes dans les régions terminales est intiment liée à l'abondance accrue en pseudogènes. Non seulement les gènes inactivés sont plus nombreux dans les régions terminales, mais les régions intergéniques sont plus grandes. Vu la spécificité du contenu en gènes de ces régions, il est probable que des régions intergéniques soient porteuses de pseudogènes dont aucun

homologue fonctionnel n'a encore été identifié, rendant sa détection impossible, ou de pseudogènes trop dégradés pour être identifiés par BLASTX. Par ailleurs, cette densité moindre en ADN codant peut être le reflet d'une proportion moins importante d'opérons.
Tous ces signes arguent en faveur de flux de gènes plus fréquents dans les régions terminales qu'au centre du chromosome.

b) Compartimentation génomique chez les autres espèces procaryotes

Au moment de l'écriture de cette thèse, 312 génomes bactériens et 26 génomes d'archées étaient disponibles (Tableau annexe). Afin d'élargir les conclusions tirées des comparaisons de génomes de *Streptomyces*, des analyses ont été entreprises sur l'ensemble de ces génomes.

Existe-t-il d'autres organismes présentant une compartimentation chromosomique comparable à celle des *Streptomyces* ?

L'abondance en gènes (et pseudogènes) associés à la plasticité génomique (transposases, intégrases et recombinases) est un des critères ayant permis de caractériser l'hétérogénéité du génome chez *Streptomyces*. Ainsi, la proportion de ces derniers dans la région entourant le terminus de réplication (ou dans les régions terminales pour les chromosomes linéaires) a été évaluée pour chaque génome procaryote disponible.
Considérant uniquement les 184 chromosomes pour lesquels au moins 20 de ces gènes/pseudogènes ont été prédits, 18 (10%) présentent un biais de distribution où plus de 40% des gènes associés à la plasticité sont portés par la région opposée à l'origine de réplication (25% du chromosome). Le cas de *S. avermitilis* est le plus biaisé avec 79% (Fig. 43). Parmi ces 18 organismes, sont retrouvées des espèces appartenant à différents phyla bactériens, possédant des génomes de grande taille (*Streptomyces*) et de petite taille (*Mycoplasma*, *Streptococcus*) et des modes de vie très divers.

- Hétérogénéité de la densité en ADN codant :

La densité plus faible en ADN codant dans les régions terminales du chromosome des *Streptomyces* constitue également un signe d'hétérogénéité génomique.
Les variations de ce paramètre ont donc été évaluées le long de chaque chromosome procaryote séquencé. Afin d'identifier ceux pour lesquels une hétérogénéité peut être détectée, l'écart entre les valeurs de densité minimale et maximale au sein d'un même réplicon a été retenue comme critère de sélection (calculées dans une fenêtre glissante de 100 kb). Chez *S. coelicolor* et *S. avermitilis*, cet écart est respectivement de 24% et 31%. Considérant le chromosome partiel de *S. ambofaciens*, cet écart est de 30%.

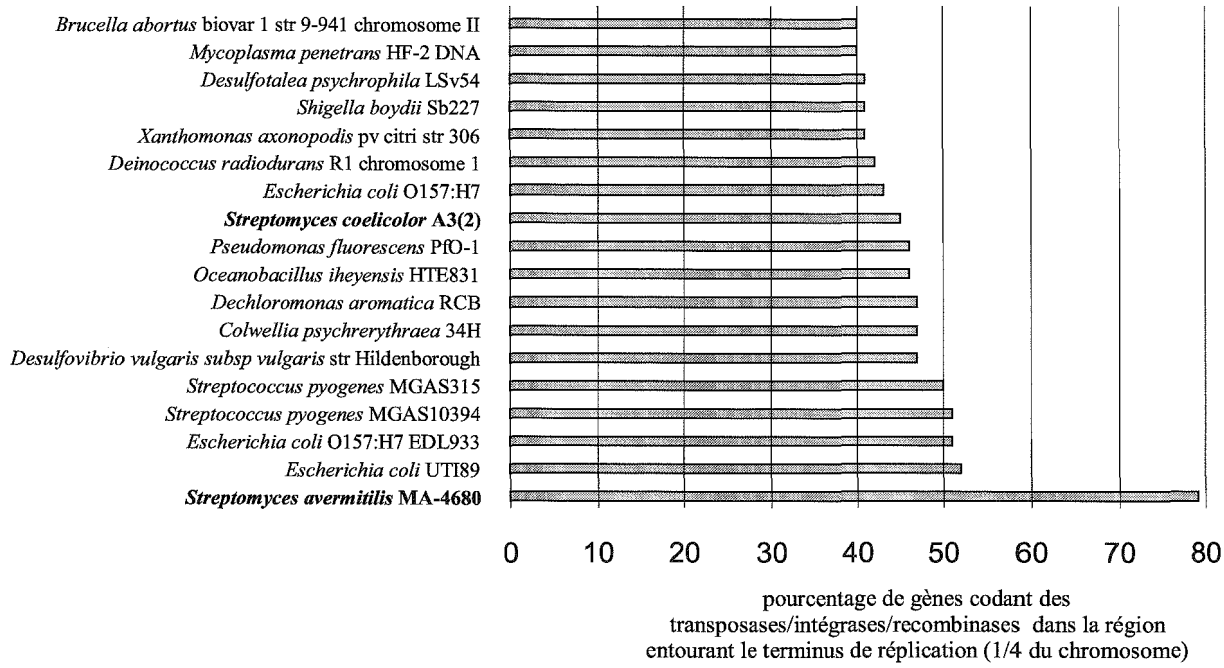


Figure 43 : Histogramme représentant le pourcentage de gènes potentiellement impliqués dans la plasticité génomique (transposases, intégrases et recombinases) portés par la région opposée à l'origine de réplication (sur 25% du chromosome). Seuls les chromosomes pour lesquels ce pourcentage est supérieur à 40% (et pour lesquels plus de 20 de ces gènes ont été prédits) ont été représentés (18 chromosomes).

Considérant les 365 réplicons procaryotes disponibles, la moyenne de cet écart est de 16 +/- 8%. Un quart d'entre eux (90 chromosomes) présente moins de 10% d'écart. Certains présentent de très faibles variations de ce paramètre, moins de 5%. C'est le cas des chromosomes de *Borrelia burgdorferi* et de plusieurs espèces de *Chlamydia* par exemple. A l'opposé, certaines espèces présentent des variations considérables. Les chromosomes de *Mycobacterium leprae*, *Prochlorococcus marinus*, *Sodalis glossinidius* et *Vibrio vulnificus* (chromosome 1) présentent un écart supérieur à 45%.

Au total, 81 réplicons (22%) présentent un écart de plus de 20% et donc une hétérogénéité comparable à celle des *Streptomyces*. Ces résultats confirment que la compartimentation observée chez *Streptomyces* n'est pas un cas isolé chez les procaryotes.
Par ailleurs, 16 chromosomes (5%) séquencés présentent une baisse sensible du pourcentage en ADN codant au niveau de la région opposée à l'origine de réplication : *Bacillus anthracis* Ames, *Bacillus anthracis* Ames ancestor, *Bacillus cereus* ATCC14579, *Bordetella parapertussis* 12822, *Burkholderia mallei* ATCC23344 chromosome 1, *Burkholderia mallei* ATCC23344 chromosome 2, *Burkholderia pseudomallei* 1710b chromosome 1, *Methanopyrus kandleri* AV19, *Prochlorococcus marinus* MIT9313, *Prochlorococcus marinus*. NATL2A (Fig. 44), *Shigella flexneri* 2a 2457T, *Staphylococcus aureus* MSSA476, *Streptomyces avermitilis* MA-4680, *Streptomyces coelicolor* A3(2), *Vibrio vulnificus* CMCP6, *Vibrio vulnificus* YJ016.
Cependant, pour d'autres espèces, la densité en ADN codant baisse au niveau de la région entourant l'origine de réplication et d'autres montrent une hétérogénéité sur l'ensemble de leur chromosome (ex : *Mycobacterium leprae*).

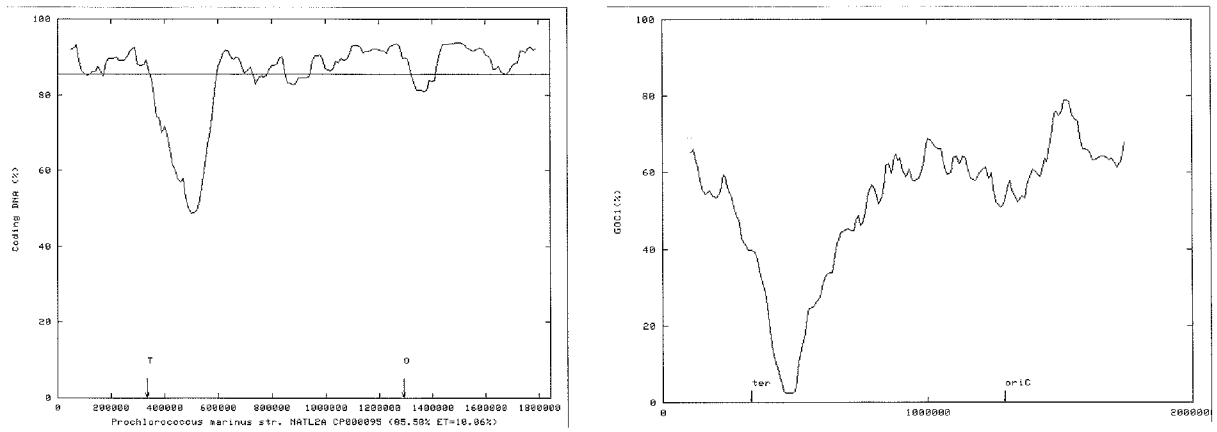


Figure 44 : Variation de la densité en ADN codant du chromosome de *Prochlorococcus marinus* NATL2A (gauche) et profil de GOC_1 (fenêtre : 200 CDS, pas : 10 CDS) de ce chromosome comparé à celui de la souche *Prochlorococcus marinus pastoris* CCMP1986 (droite). "O", "*oriC*" : origine de réplication, "T" "*ter*" : terminus de réplication.

L'exemple de *Prochlorococcus marinus* (*Cyanobacteria*, Fig. 44) montre que l'existence d'une région particulièrement pauvre en ADN codant à proximité du terminus de réplication coïncide avec la présence d'une région variable identifiable par comparaison avec la souche *Prochlorococcus marinus pastoris* CCMP1986 (Fig. 46). Ce cas rappelle les caractéristiques des chromosomes de *Streptomyces*.

- Le cas des chromosomes linéaires bactériens :

Hormis ceux de *Streptomyces*, il n'existe que 3 chromosomes bactériens linéaires séquencés : celui d'*Agrobacterium tumefaciens* (2076 kb), celui de *Borrelia burgdorferi* B31 (911 kb) et celui de *Borrelia garinii* PBi (904 kb).

Streptomyces, *Agrobacterium* et *Borrelia* appartiennent à trois phyla différents (respectivement *Actinobacteria*, *Proteobacteria* et *Spirochaetes*) et sont donc très éloignés phylogénétiquement. Par ailleurs, leur niche écologique et leur mode de vie peuvent être considérés comme divergents. Alors que les *Streptomyces* vivent de façon autonome dans l'environnement, les *Borrelia* sont des pathogènes humains (maladie de Lyme) et *Agrobacterium tumefaciens* est un pathogène de plantes.

De plus, malgré le caractère linéaire, ces réplicons sont extrêmement différents du point de vue de leur composition. Le chromosome de *Streptomyces* est de grande taille, avec un contenu en G+C très élevé et ne possède pas de biais de G ou de C (G/C skew). En revanche, les chromosomes de *Borrelia* sont de petite taille, avec un contenu en G+C très faible (28-29%) et un G/C skew marqué. Enfin, le chromosome linéaire d'*Agrobacterium tumefaciens* est de taille moyenne, avec un pourcentage en G+C moyen (59%) et une structuration du G/C skew.

Ni *Borrelia*, ni *A. tumefaciens* ne présente d'hétérogénéité de la densité en ADN codant comparable à celle observée chez *Streptomyces*.

Chez *Borrelia*, aucun gène associé à la plasticité n'a été détecté. Par conséquent, ce critère ne peut pas être utilisé pour rechercher une compartimentation génomique. Pour *Agrobacterium*, moins de 20 gènes (17 exactement) associés à la plasticité sont prédits dans le chromosome linéaire. Cependant, comme chez *Streptomyces*, un biais est observé dans les régions terminales (41% dans les 25%

terminaux). Une abondance particulière a notamment été détectée à proximité des télomères (Goodner *et al.*, 2001).

Les deux espèces de *Borrelia* possèdent un chromosome extrêmement similaire. Aucun événement d'inversion ou de translocation n'a été fixé depuis leur divergence. De plus, peu de gènes spécifiques ont été identifiés (36 chez *B. garinii* ; 55 chez *B. burgdorferi*, sans tenir compte des plasmides). Il n'est donc pas possible de comparer la compartimentation du chromosome linéaire des *Streptomyces* avec ceux des *Borrelia* qui semblent beaucoup plus stables. Néanmoins, une variabilité terminale est détectable chez ces dernières : une extrémité du chromosome de *B. burgdorferi* constitue la plus grande région spécifique du chromosome (9 kb). Ces deux espèces sont plus proches phylogénétiquement que ne le sont les espèces de *Streptomyces* étudiées ; leur séquence d'ADNr 16S ne divergent que de 0,7%.
Chez *Borrelia*, la variabilité est en fait concentrée dans les plasmides (voir *Introduction*), constituant une autre forme de compartimentation génomique. En effet, les plasmides sont principalement porteurs de gènes spécifiques, sont riches en pseudogènes et des échanges d'extrémités entre plasmides linéaires ont été mis en évidence (Casjens *et al.*, 2000). Toutes ces caractéristiques rappellent celles des régions terminales chez *Streptomyces*.
Cependant, la différence entre la stabilité du chromosome linéaire des *Borrelia* et l'instabilité de celui des *Streptomyces* tend à montrer que la linéarité n'est pas systématiquement associée à la variabilité. Il est toutefois difficile de conclure compte tenu des différences notables concernant leur mode de vie et la structure de leur génome.

6. *Les barrières au maintien de séquences nouvellement acquises et aux réarrangements*

Les contraintes structurales du génome apparaissent faibles chez les *Streptomyces*. En effet, les biais de composition en nucléotides, de distribution et d'orientation des gènes associés à la réplication (voir *Introduction*) sont de faible intensité, voire inexistants (ex : GC skew ; Fig. 5). Respectivement 55,5% et 56,2% des CDS prédites chez *S. coelicolor* et *S. avermitilis* sont portées par le brin continu. Cependant, ce biais d'orientation des gènes est, lui aussi, d'intensité variable entre les régions de contingence et la région "core". Alors que cette dernière présente un biais de respectivement 58,7% et 58,4% chez *S. coelicolor* et *S. avermitilis*, l'intensité de ce biais est quasi nulle dans les régions de contingence (respectivement 51,4% et 52,9%). Toutefois, chez l'ensemble des procaryotes, l'orientation des gènes par rapport au sens de réplication est corrélée au caractère essentiel des gènes (Rocha et Danchin, 2003a). Ainsi, l'hétérogénéité constatée chez *Streptomyces* est également associée au fait que la région "core" concentre la majeure partie, si ce n'est la totalité, des gènes essentiels.

Le pourcentage extrême en bases G+C, et l'usage du code associé, impose probablement une contrainte très forte sur l'expression des fonctions acquises par transfert horizontal et donc sur leur maintien. Parmi les 312 gènes (de taille supérieure ou égale à 300 nucléotides) de *S. ambofaciens* ne présentant aucune identité (<30%) avec les génomes de *S. coelicolor* et *S. avermitilis*, aucun ne possède une composition en nucléotide aberrante par rapport à la signature du chromosome. En effet,

le pourcentage en G+C minimum identifié pour une CDS de *S. ambofaciens* est de 60,4%. De même, toutes les CDS prédites présentent un biais marqué en G+C à la troisième position des codons. La moyenne est de 88,2% pour les CDS spécifiques (valeur minimum : 68,5%) alors qu'elle est de 92,1% sur l'ensemble du chromosome de *S. coelicolor*.

Ces données suggèrent que le pourcentage en G+C et l'usage du code associé constituent une barrière forte au maintien de gènes nouvellement acquis par les *Streptomyces*. Hormis les *Streptomyces*, peu de groupes bactériens possèdent un génome dont la composition en bases G+C est aussi biaisée. Ainsi, le succès évolutif de la majorité des transferts horizontaux prédits dans les régions terminales est probablement conditionné par l'origine du donneur. Parmi les organismes cultivables et identifiés, le donneur pourrait appartenir essentiellement aux espèces du genre *Streptomyces* et aux Actinobactéries dont le génome possède un pourcentage élevé en G+C ; par exemple, parmi les génomes séquencés : *Mycobacterium*, 65-69% ; *Leifsonia xyli*, 68% ; *Nocardia farcinica*, 71% ; *Symbiobacterium thermophilum*, 69% ; *Thermobifida fusca*, 68% ; *Frankia* sp. Cci3, 70%.

Par ailleurs, des plasmides linéaires possédant une structure invertronique, c'est-à-dire avec des TIR liées covalemment à une protéine terminale, ont été mis en évidence chez différentes espèces de *Rhodococcus* qui appartiennent aussi à l'ordre des *Actinomycetales* (Masai *et al.*, 1997 ; Sekine *et al.*, 2006 ; Shimizu *et al.*, 2001 ; Stecker *et al.*, 2003). Le transfert conjugatif de certains de ces plasmides, porteurs de gènes de résistance (mercure et arsenate) et de clusters responsables de la dégradation de composés polluants a par ailleurs été décrit (Dabrock *et al.*, 1994). L'ensemble de ces données suggère que les espèces de *Rhodococcus*, entre autres, sont des partenaires probables d'échanges et de transferts horizontaux avec *Streptomyces*.

Enfin, la plupart de ces espèces d'*Actinomycetales* sont retrouvées dans l'écosystème sol, tout comme *Streptomyces*, favorisant l'éventualité de transferts horizontaux.

7. *Influence d'un gradient de fréquence de réarrangements*

a) *Indice GOC pour identifier des régions issues de transferts horizontaux ?*

Avant de discuter des résultats de comparaisons de génomes de *Streptomyces*, il est nécessaire de préciser la raison pour laquelle l'indice GOC (Gene Order Conservation) a été utilisé.

En effet, le GOC représente le niveau de synténie entre deux séquences. Cette synténie se base donc sur les gènes partagés. Comment cet indice peut-t-il permettre de caractériser des régions variables du point de vue du contenu en gènes ?

Deux formules ont été utilisées pour évaluer le niveau de conservation le long du chromosome des *Streptomyces* (GOC_1 et GOC_2 ; voir chapitre *Approches bioinformatiques*). En réalité, seul l'indice GOC_2 permet de mesurer la conservation de l'ordre des gènes orthologues (Rocha, 2006).

L'indice GOC_1 représente le niveau de conservation du contenu en gènes. Pour cela, il calcule le pourcentage de paires de gènes orthologues contigus dans deux génomes le long du chromosome par rapport au nombre total de gènes (partagés et spécifiques). L'introduction du critère de la synténie (caractère contigu) dans ce calcul s'est avérée nécessaire afin d'éliminer de la fraction conservée tous les gènes présentant une homologie mais n'étant probablement pas des orthologues. En effet, le

génome complet de *S. ambofaciens* n'étant pas disponible, la sélection des meilleurs appariements réciproques (critère RBM : reciprocal best match; couramment utilisé dans les comparaisons des génomes) ne suffit pas pour éliminer les paralogies. Le critère de la synténie est d'autant plus important dans le cas des génomes de *Streptomyces* puisqu'ils sont caractérisés par une grande redondance de gènes paralogues, notamment dans les régions terminales. Par exemple chez *S. avermitilis*, 35% des gènes appartiennent à des familles de paralogues (Ikeda *et al.*, 2003).
La superposition des signaux de dot-plot et GOC_1 (Fig. 35) a permis de mettre en évidence la présence de régions montrant une dégénérescence de la synténie croissante du centre vers les régions terminales spécifiques. Cette dégénérescence traduit une multitude d'événements d'insertions/délétions de petites régions (en général contenant moins de 10 gènes) interrompant la synténie. Elle n'est pas causée par un bouleversement de l'ordre des gènes suite à, notamment, des inversions centrées autour de l'origine de réplication comme décrit par (Tillier et Collins, 2000a). En effet, quatre événements d'inversions (détectables par dot-plot) ont été fixés depuis la divergence entre *S. coelicolor* et *S. avermitilis*, c'est-à-dire depuis l'ancêtre commun de la plupart des *Streptomyces*.
Afin d'étendre ces résultats aux chromosomes complets (*S. coelicolor*, *S. avermitilis* et *S. scabies*), l'indice GOC_2 s'est avéré plus approprié. Le but était alors de confirmer la présence d'un gradient de fréquence d'insertions/délétions et d'identifier les limites des régions évoluant selon ce gradient. En effet, la présence d'îlots génomiques de grande taille fait chuter drastiquement la valeur du GOC_1 localement alors qu'elle ne reflète qu'un unique événement d'insertion/délétion. L'indice GOC_2 permet de s'affranchir de ce problème (Fig. 38).

b) Quelles hypothèses permettent d'expliquer la compartimentation du génome ?

La comparaison des génomes a révélé une dynamique très particulière de l'évolution des régions terminales chez *Streptomyces* :

1. Plus deux espèces de *Streptomyces* sont éloignées phylogénétiquement, plus la partie conservée du chromosome (centre) est restreinte. Réciproquement, pour le génome d'une espèce donnée, la taille des régions terminales spécifiques est d'autant plus grande que l'espèce comparée est éloignée phylogénétiquement.

2. Pour chaque paire de génomes, la dégénérescence de la synténie est graduelle, du centre vers les extrémités des deux bras chromosomiques. Plus l'on s'approche des extrémités, plus le contenu en gènes est variable, donc plus le flux de gènes (insertions/délétions) depuis la divergence des espèces considérées est important.

Le corollaire de ces deux résultats et que, pour un locus donné appartenant aux régions de synténie dégénérée, le niveau de dégénérescence est d'autant plus élevé que les espèces comparées sont éloignées phylogénétiquement.

Le premier de ces résultats révèle que, chez *Streptomyces*, les régions terminales spécifiques d'espèce ne sont définissables que d'un point de vue dynamique. Elles ne sont pas bornées et ne peuvent donc pas être considérées comme des îlots génomiques. Puisque les frontières entre régions conservées et

spécifiques ne sont pas les mêmes selon le couple d'espèces considéré, les flux de gènes affectent l'ensemble des régions de contingence (Fig. 38).

La caractérisation de la variabilité des TIR chez *S. ambofaciens* a montré que des échanges entre réplicons linéaires sont impliqués dans la spécificité des extrémités chromosomiques. Historiquement, il a été proposé que ces échanges seraient le moteur de la variabilité terminale. L'identification d'une dégénérescence graduelle de l'information génétique ancestrale indique que d'autres mécanismes interviennent dans l'établissement de cette variabilité. L'existence d'une synténie, certes dégénérée mais toujours détectable, démontre que les gènes spécifiques présents dans ces régions ne sont pas issus d'échanges entre réplicons linéaires par simples crossing-overs qui, au contraire, font varier l'ensemble du contenu en gènes depuis la borne de l'événement jusqu'au télomère.

En fait, les régions terminales évoluent principalement par fixation d'une multitude d'événements d'insertions/délétions accumulés au cours des temps évolutifs, effaçant progressivement l'information ancestrale. La dégénérescence graduelle de la synténie observée dans tous les génomes de *Streptomyces* indique que, dans les régions de contingence, la fréquence d'apparition et/ou de fixation des réarrangements est variable en fonction de la localisation chromosomique : elle serait d'autant plus élevée que l'on s'approche des télomères. Par saturation d'événements, le nombre d'insertions/délétions devient trop élevé et le contenu en gènes trop divergent pour détecter une synténie. Il est donc probable que la manière dont les flux de gènes façonnent les régions de synténie dégénérée s'applique aux extrémités spécifiques d'espèce, dans lesquelles une saturation d'événements aurait engendré un renouvellement complet de l'information génétique.
Les échanges d'extrémités de réplicons linéaires sont, certes, impliqués dans les flux de gènes au niveau des régions subtélomériques mais ils ne seraient qu'une partie minime des événements impliqués dans la diversification des régions terminales.

Deux hypothèses alternatives concernant la dynamique des génomes peuvent expliquer ces résultats :

1. les régions terminales sont variables car elles sont plus tolérantes aux réarrangements (différence de fréquences de fixation de réarrangements).
2. les régions terminales sont variables car la fréquence de recombinaison (réarrangements) est plus élevée dans ces régions (différence de fréquences d'apparition des réarrangements).

Le maintien à long terme d'une séquence dans un génome dépend de sa valeur adaptative. Plus un gène contribue de façon significative au "fitness" de l'organisme, plus sa probabilité d'être maintenu par sélection naturelle est élevée. La théorie de Lawrence et Roth implique une compétition entre les gènes au sein d'un génome (voir *Introduction*) (Lawrence et Roth, 1999). Etant données les contraintes s'appliquant sur la taille des génomes, toute acquisition de nouvelles séquences (qui apportent une fonction) entraîne la perte d'autres séquences de valeur adaptative plus faible. Il existerait donc, pour chaque génome, une valeur seuil (appelée s) de contribution au "fitness" en dessous de laquelle un gène ne peut plus être maintenu par sélection et pourrait donc être perdu par dérive. Pour préciser cette théorie, la valeur s est proportionnelle au taux de mutations et donc au taux de délétions et de réarrangements (qui génèrent des allèles non fonctionnels). En effet, plus le taux de mutations d'un

génome est élevé, plus la valeur adaptative d'un gène devra être élevée pour assurer sa maintenance dans la population.

De façon intéressante, chez *Streptomyces*, la fréquence des flux de gènes est plus élevée dans les régions terminales. Par conséquent, plus un gène est proche des extrémités, plus sa probabilité d'être maintenu à long terme est faible.

- Dans le cadre de l'hypothèse n°1, la contribution des gènes au "fitness" diminuerait du centre vers les extrémités. Ainsi, si le taux de réarrangements est constant le long du chromosome, la dégénérescence de la synténie s'expliquerait par un différentiel du taux de fixation de ces événements. Ainsi, les régions terminales seraient plus tolérantes aux réarrangements du fait de l'absence de gènes essentiels dans ces régions.
Selon cette hypothèse, la dégénérescence graduelle de la synténie vers les extrémités chromosomiques semblerait indiquer que les gènes sont organisés le long du chromosome en fonction de leur valeur adaptative. Plus un gène serait proche de l'extrémité, moins sa contribution serait importante. Dans ce cas, l'effet dose pourrait-il être la force qui organise les gènes sur le chromosome selon leur importance ?
En effet, pour certaines espèces bactériennes, une distribution préférentielle des gènes fortement exprimés à proximité de l'origine de réplication est sélectionnée (Couturier et Rocha, 2006). Cependant, cette corrélation n'est vérifiée que pour les bactéries à croissance rapide, auxquelles les *Streptomyces* n'appartiennent pas. De plus, la corrélation mise en évidence ne concerne que les gènes impliqués dans les processus cellulaires de traduction et transcription. Par ailleurs, si certains gènes essentiels sont fortement exprimés, d'autres ne le sont pas. Une étude portant sur la distribution des gènes fortement exprimés (d'après leur Index d'Adaptation du Code, CAI) chez *S. coelicolor* et *S. avermitilis* a montré que, malgré un enrichissement dans la région "core", les régions terminales sont également porteuses de gènes fortement exprimés (Fig. 45). Ces résultats sont en défaveur de l'hypothèse d'une compartimentation du génome par effet dose.

- Alternativement, dans le cadre de l'hypothèse n°2, les données indiquent que le taux de réarrangements (insertions/délétions) augmenterait de façon graduelle vers les extrémités chromosomiques. Ainsi, la valeur s ne serait pas seulement variable entre les organismes, mais le serait également entre loci d'un même génome.
L'existence d'un gradient de fréquence de recombinaison pourrait être la force qui maintient cette compartimentation et qui exclut les gènes essentiels des extrémités chromosomiques. Le passage d'un ou plusieurs gènes essentiels vers les régions terminales (par inversion interne à un réplichore par exemple) serait fortement contre-sélectionné.
L'instabilité intrinsèque des régions terminales du chromosome chez *Streptomyces* pourrait être à l'origine de la variabilité terminale et expliquerait la façon dont le génome évolue. La tolérance accrue des régions terminales aux réarrangements ne serait donc pas la raison de leur spécificité mais plutôt une conséquence de l'instabilité.

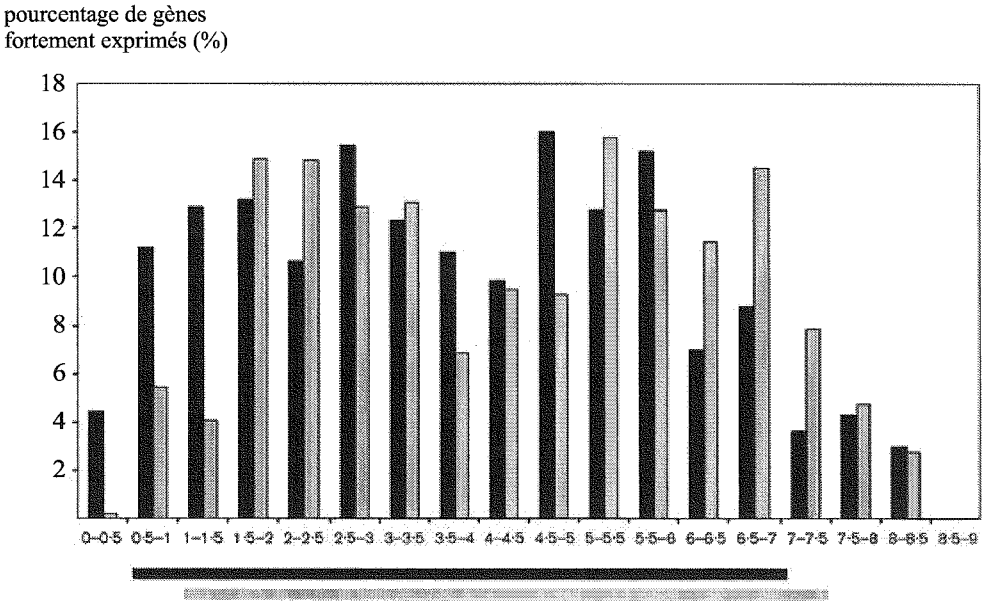


Figure 45 : Répartition des gènes fortement exprimés le long du chromosome de *S. coelicolor* (gris) et *S. avermitilis* (noir). La région core de chaque chromosome est représentée par une barre sous le graphique. Le pourcentage de gènes fortement exprimés est calculé par rapport au nombre de gènes dans chaque intervalle de 500 kb. D'après (Wu *et al.*, 2005).

8. Hypothèses concernant l'origine d'un gradient de fréquence de réarrangements

a) Les mécanismes de recombinaison et de réparation de l'ADN chez Streptomyces

Les mécanismes de recombinaison restent peu étudiés chez *Streptomyces*. Chez *E. coli* comme chez beaucoup d'autres espèces bactériennes, les cassures double-brin sont principalement prises en charge par la voie RecBCD reconnaissant des sites cibles de distribution asymétrique, appelés *chi* (Kowalczykowski, 2000 ; Singleton *et al.*, 2004). Cependant, aucun système homologue à celui-ci (ni au système analogue AddAB) n'a été trouvé chez *Streptomyces*. Seul un homologue de *recD* (nucléase) a été décrit chez *S. coelicolor* (Bentley *et al.*, 2002) et *S. avermitilis* (Ikeda *et al.*, 2003) suggérant que les mécanismes de réparation de l'ADN chez *Streptomyces* sont différents de ceux connus chez *E. coli*. La présence de *recD* lorsque *recBC* est absent se retrouve également chez de nombreuses autres espèces bactériennes (Rocha *et al.*, 2005).

La présence de sites *chi* n'a jamais été recherchée chez *Streptomyces*. Toutefois, un octomère surreprésenté sur l'un des deux brins d'ADN (une des caractéristiques des sites *chi*) a été détecté chez *S. coelicolor* par une approche bioinformatique : TGGGGGAG (Hendrickson et Lawrence, 2006).

Ce site est non seulement présent en abondance sur le brin continu mais, de plus, sa fréquence augmente vers les extrémités chromosomiques. La distribution biaisée de cet octomère a également été mise en évidence chez d'autres Actinobactéries : *M. tuberculosis* et *Nocardia farcinica* (Hendrickson et Lawrence, 2006). Les auteurs proposent que ces motifs confèrent une polarité au chromosome nécessaire au moment de la séparation des chromatides sœurs après réplication (séquences AIMS, voir *Introduction*). Ces sites seraient reconnus par des translocases FtsK capables de tracter les

chromosomes à travers le septum au moment de la division cellulaire (Fig. 10). Ces séquences pourraient-elles jouer le rôle de sites *chi* chez *Streptomyces* ?

Des gènes homologues à ceux impliqués dans le système de recombinaison RecFOR ont été identifiés chez *S. coelicolor* et *S. avermitilis*. Ce système est responsable de la réparation des "gaps" simple-brin (pour revue (Rocha *et al.*, 2005)). De plus, un homologue du système RuvABC est présent et hautement conservé à travers les génomes de *Streptomyces* étudiés. Ce système est impliqué dans la migration et le clivage des jonctions de Holliday au cours de la recombinaison homologue. L'ensemble de ces gènes est porté par la région centrale du chromosome chez *Streptomyces*.

La recombinaison homologue semble très efficace chez *Streptomyces* (Wohlleben *et al.*, 1994). Le gène central de la recombinaison homologue, *recA*, joue un rôle dans l'instabilité génomique chez *Streptomyces*. La réduction de l'activité de RecA conduit à une augmentation forte (d'un facteur 70) du niveau d'instabilité génomique (Volff et Altenbuchner, 1997). Elle joue donc un rôle important dans le maintien de l'intégrité du génome et a également été impliquée dans les événements d'amplification d'ADN (Aigle *et al.*, 1997 ; Volff et Altenbuchner, 1997).
L'obtention de mutants nuls pour *recA* a longtemps été problématique chez *Streptomyces* (Aigle *et al.*, 1997 ; Muth *et al.*, 1997). Récemment, un mutant n'exprimant plus RecA a pu être obtenu chez *S. coelicolor* démontrant que cette protéine n'est pas essentielle (Huang et Chen, 2006). Les mutants n'exprimant plus RecA ont une vitesse de croissance plus faible et produisent des spores non viables (sans ADN) à plus haute fréquence que la souche sauvage et la fréquence de recombinaison pendant la conjugaison est également affectée (Huang et Chen, 2006).

b) Apparition et réparation des cassures double-brin

Les cassures double-brin constituent l'un des dommages majeurs causés sur l'ADN. Leur apparition pourrait être associée aux réarrangements du chromosome des *Streptomyces*. En effet, les cassures génèrent des extrémités d'ADN recombinogènes et sont à l'origine de réarrangements génomiques. Plusieurs mécanismes de réparation des cassures sont connus et ont été bien étudiés notamment chez la levure. Certains font intervenir la recombinaison homologue. C'est le cas du BIR (break induced replication) et de la conversion génique. Le "raboutage" d'extrémités d'ADN double-brin non homologues appelé NHEJ (Non Homologous End-Joining) est également documenté. Au contraire de la recombinaison homologue, il répare de façon illégitime les régions subissant des cassures et peut donc être responsable de création de pseudogènes et de réarrangements chromosomiques (Daley *et al.*, 2005).

Les mécanismes de réparation des cassures double-brin restent inconnus chez *Streptomyces*. Cependant, des gènes homologues à ceux codant les protéines Ku70/Ku80 et la ligase IV (ligase ADN dépendante de l'ATP), impliquées dans la réparation par NHEJ chez les eucaryotes, ont été identifiés chez certaines espèces bactériennes dont *Streptomyces* et *Mycobacterium* (Aravind et Koonin, 2001 ; Weller *et al.*, 2002). Bien que le NHEJ requiert plusieurs complexes protéiques chez les eucaryotes, les protéines Ku et ligase IV sont suffisantes à la réparation chez *M. tuberculosis* (Pitcher *et al.*, 2005).

Des dimères de protéines Ku sont capables de reconnaître les extrémités générées par cassure. Ils recrutent la ligase IV et stimulent son activité au niveau des sites de lésion de l'ADN.
Chez *Streptomyces*, leur rôle n'a pas été étudié. Les homologues des protéines Ku70/Ku80 et de la ligase IV ont été prédits dans la région centrale du chromosome de *S. coelicolor* (SCO5308/09 (Aravind et Koonin, 2001)) et *S. avermitilis* (SAV2945/46). Chez *S. ambofaciens*, l'analyse des BES montre une situation analogue à celle de *S. coelicolor*. Le cluster *Ku-lig4* est identifiable avec une configuration différente dans le génome de *S. scabies* où les gènes *ku* et *lig4* sont séparés par *uvrA* codant une nucléase putative. De plus, bien que la protéine Ku soit similaire dans tous les génomes procaryotes étudiés, la ligase IV diffère entre les espèces. Chez *Streptomyces*, ce gène ne code qu'un domaine catalytique sur les trois habituellement rencontrés (activité polymérase putative).

Le raboutage d'extrémités d'ADN a été impliqué dans la circularisation du chromosome chez *S. griseus* (Inoue *et al.*, 2003) et dans la génération de mutants qui possèdent un chromosome issu de la fusion de deux chromosomes chez *S. ambofaciens* (Wenner *et al.*, 2003). Ces chromosomes fusionnés possèdent donc deux origines de réplication et deux loci impliqués dans la partition chromosomique. Leur niveau d'instabilité très élevé pourrait être dû à des cycles de cassure-fusion-pont comme mis en évidence chez les eucaryotes (McClintock, 1951).

Malgré la distance évolutive entre les *Streptomyces* et la levure et malgré l'apparition probablement indépendante de la linéarité chromosomique dans ces deux lignées, des similitudes dans le mode d'évolution de leurs chromosomes apparaissent. Chez la levure, les régions subtélomériques sont également des régions plus plastiques que la région centrale (Louis, 1995). De plus, elles sont riches en familles de gènes paralogues et sont dépourvues de gènes essentiels (Fabre *et al.*, 2005). Par ailleurs, chez *Kluyveromyces lactis* (hémiascomycète), 7 des 12 régions subtélomériques sont composées de séquences homologues portant des gènes dupliqués suggérant que des échanges entre extrémités chromosomiques sont responsables d'une partie de la diversification des régions terminales (Fairhead et Dujon, 2006). Ces régions portent notamment des gènes codant des protéines de localisation extracellulaire qui pourraient être impliqués dans l'adaptation à l'environnement (Fairhead et Dujon, 2006).

Toutefois, même si des analogies apparaissent entre la plasticité des régions subtélomériques chez *Streptomyces* et celles de la levure, l'ampleur de ces régions n'est pas comparable : elles représentent plusieurs centaines voire milliers de kilobases chez *Streptomyces* (un quart du génome) alors qu'elles se limitent à environ 30 kb par chromosome chez la levure.

De façon intéressante, un gradient de fréquence de survie aux cassures double-brin a été démontré chez la levure (Fig. 46, (Ricchetti *et al.*, 2003)). En effet, les cassures générées artificiellement dans les régions subtélomériques sont réparées avec une meilleure efficacité que dans la région centrale du chromosome. Cette étude a montré que la fréquence de NHEJ est constante le long du chromosome alors que les autres mécanismes de réparation des cassures voient leur efficacité augmenter de façon graduelle vers les extrémités. Ces mécanismes alternatifs impliquent en général la perte du fragment chromosomique distal. Il s'agit du BIR, de l'addition de séquences télomériques au niveau de la

cassure, de la conversion génique (à l'aide d'une séquence homologue portée par un autre chromosome et utilisée comme matrice pour la réplication de la région entourant la cassure) et de l'addition de séquences plasmidiques en tandem.

Ainsi, l'efficacité des mécanismes de réparation des cassures est variable selon la localisation chromosomique chez la levure. De plus, des insertions de séquences d'ADN mitochondrial ont été observées suite à la réparation de cassures par NHEJ (Ricchetti *et al.*, 1999). Ceci signifie que les sites de cassures double-brin sont des sites privilégiés d'intégration de fragments d'ADN. Le mécanisme NHEJ pourrait donc être impliqué dans la diversification des régions terminales chez *Streptomyces*. En effet, la multitude d'événements d'intégrations/délétions de petits fragments d'ADN (de 1 à 10 gènes en général) mis en évidence dans les régions de synténie dégénérée pourrait résulter de réactions de ligation des extrémités d'ADN recombinogènes issues de cassures avec des molécules d'ADN transitoirement présentes dans la cellule, notamment acquises par transfert horizontal.

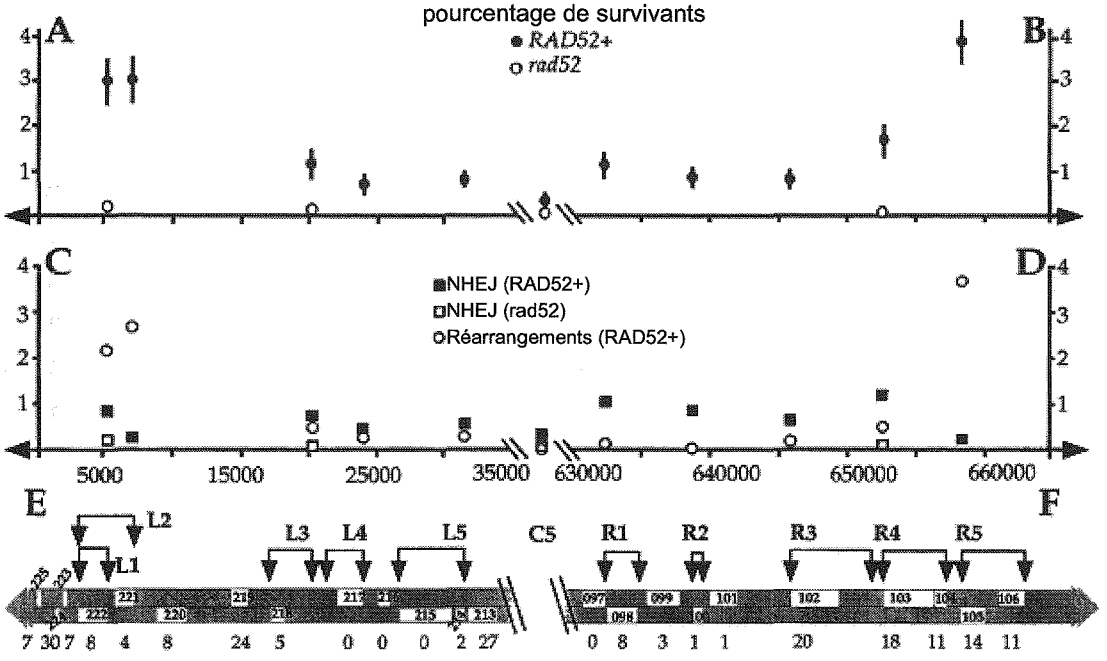


Figure 46 : Réparation des cassures double-brin le long du chromosome XI de *S. cerevisiae.*
A. et **B.** Survie aux cassures d'ADN double-brin en fonction de la localisation de la cassure dans les régions subtélomériques gauche et droite du chromosome XI. Les cercles pleins représentent les souches RAD52+ alors que les cercles vides représentent les souches mutantes rad52. RAD52 est impliquée dans la recombinaison homologue chez la levure.
C. et **D.** Efficacité des mécanismes de réparation des cassures double-brin le long du chromosome. Les valeurs indiquées en ordonnées représentent le pourcentage de chaque mécanisme ayant permis la réparation de la cassure par rapport au nombre total de cellules. Les carrés pleins et vides représentent, respectivement, le pourcentage de NHEJ chez les souches RAD52+ et rad52. Les cercles indiquent le pourcentage représenté par l'ensemble des autres mécanismes (BIR, addition de télomères, incorporation de séquences plasmidiques et conversion génique).
E. et **F.** Extrémités gauche et droite du chromosome XI. Les loci nommés L1 à L5, C5 et R1 à R5 sont les sites d'induction des cassures double brin. L'intervalle entre deux têtes de flèche représente la région remplacée par une cassette cible de coupure par l'endonucléase I-*Sce*I utilisée pour générer des cassures double-brin. Les têtes de flèches aux extrémités du chromosome représentent les télomères. Les rectangles blancs indiquent la présence d'ORF et le nombre de gènes paralogues identifiés est indiqué sous chacune d'entre elle. D'après (Ricchetti *et al.*, 2003).

Les gènes spécifiques d'espèce massivement présents dans les régions terminales ne semblent pas issus d'intégration d'éléments génétiques mobiles autonomes (ex : pSAM2). En effet, les régions de synténie dégénérée révèlent que la diversification des extrémités se fait par acquisitions/pertes de petits fragments d'ADN, ne contenant que quelques gènes (parfois un seul), et aucune signature d'éléments transférables n'est en général reconnaissable (ex : module de recombinaison). Les intégrations fixées apparaissent comme le fruit de la recombinaison illégitime qui pourrait se produire lors de la réparation des cassures double-brin par exemple.

La fréquence croissante des réarrangements (insertions/délétions) pourrait être expliquée par deux hypothèses non exclusives :
1. La fréquence de cassures double-brin croit dans les régions terminales.
2. L'efficacité des mécanismes de réparation des cassures est dépendante de la localisation chromosomique.

- La première hypothèse tente d'expliquer le gradient de fréquence d'insertions/délétions par une fréquence d'apparition des cassures double-brin variable le long du chromosome des *Streptomyces*. Deux modes de cassures double-brin peuvent être distingués de façon arbitraire : les cassures site-spécifiques et les autres. Les premières sont dues, notamment, à l'excision d'éléments mobiles comme les IS. Chez *Streptomyces*, une abondance croissante du nombre d'IS vers les extrémités chromosomiques pourrait être associée à un tel gradient d'instabilité. Par ailleurs, certains *Streptomyces* synthétisent des métabolites secondaires capables de générés des cassures site-spécifiques de l'ADN (et utilisés comme anti-tumoraux, (Xu *et al.*, 1994)).
Les cassures aléatoires peuvent, quant à elles, être provoquées par des arrêts de la fourche de réplication (Michel *et al.*, 2001). Chez *E. coli*, des cassures générées artificiellement provoquent des arrêts de la réplication et stimulent l'instabilité génomique (Michel *et al.*, 1997).

Ainsi chez *Streptomyces*, les arrêts de la fourche de réplication pourraient-ils être plus fréquents à mesure que celle-ci s'approche des extrémités ?
La fixation des protéines Tus sur les sites *ter* est responsable de l'arrêt spécifique de la fourche de réplication chez les chromosomes circulaires d'*E. coli* et *B. subtilis* et initie le processus de terminaison de la réplication (pour revue (Neylon *et al.*, 2005)). Etant donnée la linéarité chromosomique chez *Streptomyces*, l'existence d'un mécanisme dédié à la terminaison de la réplication ne semble pas nécessaire et n'a, à ce jour, pas été décrit. Cependant, la fixation spécifique ou aspécifique de protéines dans les régions terminales du chromosome pourrait engendrer des arrêts de la fourche de réplication et donc des cassures plus fréquentes.

- Selon la seconde hypothèse, les mécanismes de réparation fidèles des cassures, c'est-à-dire ne créant pas de réarrangements, seraient moins efficaces dans les régions de contingence. C'est le cas de la conversion génique. Au moment de la réplication, les régions localisées au voisinage de l'origine de réplication sont en nombre de copies plus élevé que les régions terminales (effet dose). Ainsi, la réparation d'une cassure double-brin apparue dans la région centrale a plus de chance d'être réparée par

conversion génique en utilisant une chromatide sœur comme matrice. Dans ce cas, le gradient de fréquence de réarrangements pourrait être du à une efficacité moindre de la conversion génique quand la cassure est proche des télomères (hormis dans les TIR).

c) Transfert conjugatif à partir des extrémités chromosomiques

Les *Streptomyces* ne semblent pas être naturellement transformables (Chater et Hopwood, 1984). Aucun homologue des gènes impliqués dans la compétence chez *B. subtilis* n'est prédit dans les génomes de *Streptomyces*. Par ailleurs, un seul système de transduction généralisé a été rapporté : le phage phiSV1 chez *Streptomyces venezuelae* (Chater et Hopwood, 1984).
En revanche, de nombreux éléments conjugatifs ont été identifiés depuis une cinquantaine d'années, notamment des plasmides conjugatifs et des éléments intégrés dans le chromosome (pour revues : (Grohmann *et al.*, 2003 ; Hopwood, 2006)). Les *Streptomyces* étant producteurs de molécules antibiotiques, ils sont aussi un réservoir naturel de gènes de résistance et les éléments conjugatifs présents chez les *Streptomyces* pourraient être (ou avoir été) impliqués dans la dissémination des résistances par transfert horizontal, notamment vers les bactéries pathogènes (Grohmann *et al.*, 2003).

Une grande variété de plasmides conjugatifs a été décrite chez *Streptomyces* : des petits plasmides circulaires ne codant que les fonctions de transfert et de réplication, comme pIJ101 de *S. lividans* (Kieser *et al.*, 1982), mais aussi de grands plasmides linéaires tels que SCP1 (360 kb) de *S. coelicolor* (Hopwood *et al.*, 1983). Certains des éléments transférables sont intégratifs, comme pSAM2 de *S. ambofaciens* qui est capable de se maintenir par recombinaison site-spécifique dans l'extrémité 3' d'un gène d'ARNt (Pernodet *et al.*, 1984). De plus, certaines espèces de *Streptomyces* phytopathogènes sont porteuses d'un îlot de pathogénicité de grande taille (>325 kb) mobilisable vers des espèces non pathogènes (Kers *et al.*, 2005).

La plupart des plasmides conjugatifs sont transférés avec une très grande efficacité, pouvant atteindre 100%, entre souches de *Streptomyces* (Kieser *et al.*, 1982). De plus, la mobilisation de marqueurs chromosomiques a été observée à des fréquences variant de 0,1% à 1% (Hopwood *et al.*, 1985). L'originalité du mécanisme est tout aussi remarquable que son efficacité. En effet, chez les autres espèces bactériennes, la conjugaison implique un ensemble de protéines, responsables de l'établissement d'un contact entre cellules et du passage de l'ADN transféré vers la réceptrice. L'ADN est clivé par une relaxase, transféré sous forme simple-brin et subit une ligation dans la cellule réceptrice une fois l'élément transféré.
Au contraire chez *Streptomyces*, le mécanisme nécessite l'intervention d'une seule protéine : Tra. Les études portant sur pSAM2 ont permis de démontrer que l'ADN est transféré sous forme double-brin pendant la conjugaison (Possoz *et al.*, 2001). Tra reconnaît le locus *clt* (*cis*-acting locus) nécessaire au transfert des plasmides mais pas à celui des marqueurs chromosomiques (Pettis et Cohen, 1994). *clt* ne subit pas de coupure, confirmant le transfert de l'ADN sous forme double-brin (Ducote et Pettis, 2006). Après acquisition par le mycélium récepteur, un mécanisme de dispersion intra mycélien se met en place, codé par les gènes *spd* ("spread") de l'élément (Kieser *et al.*, 1982). Il assure ainsi sa dispersion dans toutes les cellules voisines et augmente considérablement l'efficacité du transfert.

Enfin, Tra possède une activité de liaison non spécifique à l'ADN qui pourrait expliquer la mobilisation du chromosome.

Les protéines Tra des différents éléments conjugatifs découverts sont peu conservées mais présentent toutes une similarité avec les translocases de la famille FtsK/SpoIIIE impliquées dans la ségrégation chromosomique. Des analyses fonctionnelles ont d'ailleurs permis de confirmer son rôle de translocase dépendante de l'ATP (Kosono *et al.*, 1996 ; Reuther *et al.*, 2006).
Le système conjugatif chez *Streptomyces* et la machinerie de ségrégation aurait une origine commune. Ainsi, de la même façon que les chromosomes nouvellement répliqués sont tractés à travers le septum au moment de la sporulation, leur transfert conjugatif impliquerait une translocation du chromosome entre deux mycélia.

L'efficacité de la conjugaison (Troost *et al.*, 1979), son amplification par transfert intra mycélien ("spread") et la mobilisation de séquences chromosomiques suggèrent que les *Streptomyces* sont soumis à un flux de molécules d'ADN double-brin circulaires et linéaires. Le passage de plasmides complets est fréquemment observé. Etant donnée la taille du chromosome, la mobilisation de marqueurs chromosomiques réalise probablement le transfert de fragments d'ADN chromosomique sous forme linéaire, générés par cassures au moment de la rupture du contact entre les mycélia partenaires. Ainsi chez *Streptomyces*, la conjugaison pourrait aboutir à l'acquisition de molécules linéaires qui pourraient être intégrées par NHEJ au niveau des cassures chromosomiques ou par recombinaison homologue.

Un lien peut-il être établit entre conjugaison et variabilité des extrémités chromosomiques ?
Un faisceau de présomptions converge vers l'hypothèse d'un transfert conjugatif du chromosome par l'une ou l'autre (ou les deux) des extrémités. C. W. Chen a proposé un modèle de conjugaison, appelé "End-first", dans lequel les télomères joueraient le rôle d'origine de transfert (Chen, 1996). A l'origine, ce modèle se basait sur les mécanismes connus de conjugaison qui impliquaient systématiquement un passage de l'ADN sous forme simple-brin. En réalité, l'ADN est transféré sous forme double-brin mais ce modèle reste attractif. L'un des arguments en faveur d'un transfert par les extrémités chromosomique est l'existence du gène *ttrA* (pour "terminal transfer") dont la localisation télomérique est très conservée chez *Streptomyces*. Ce gène coderait une hélicase dont le rôle dans la conjugaison du plasmide SLP2 et de chromosome de *S. lividans* a été démontré (Huang *et al.*, 2003). Il est retrouvé à chacune des extrémités des chromosomes et des plasmides linéaires. Chez *S. ambofaciens*, ce gène est même présent dans les régions terminales spécifiques de souches bien qu'il soit tronqué chez la souche DSM40697.
Sa localisation conservée dans une région hypervariable suggère une forte pression de sélection sur la fonction de cette hélicase qui agirait donc en *cis*. Un argument supplémentaire en faveur d'un transfert par les extrémités des réplicons linéaires est l'absence de *ttrA* dans les plasmides circulaires chez *Streptomyces*.

Le transfert conjugatif par les extrémités pourrait être en lien avec la variabilité des régions terminales qui seraient, par conséquent, les régions les plus fréquemment transférées chez *Streptomyces*. La

compartimentation originale du chromosome des *Streptomyces* refléterait donc une organisation optimale des gènes vis-à-vis de leur dissémination par transfert horizontal. Le transfert horizontal implique majoritairement les gènes non essentiels (voir *Introduction*). Chez *Streptomyces*, une corrélation existerait donc entre la localisation des gènes fréquemment transférés et la mobilisation plus favorable des extrémités chromosomiques.
Ce concept intégrait jusqu'alors les plasmides et les éléments mobiles codant leur propre transfert. Il pourrait s'appliquer aux régions terminales du chromosome des *Streptomyces*.

Chez les souches Hfr d'*E coli*, la mobilisation du chromosome depuis une origine de transfert (*oriT*) induit un gradient de fréquence de transfert des marqueurs chromosomiques en fonction de leur distance à *oriT* (Reimmann et Haas, 1993). Par analogie avec ce système, chez *Streptomyces*, un gradient de fréquence de mobilisation pourrait exister en fonction de la distance aux télomères.

PERSPECTIVES

Ce travail a permis d'élargir notre compréhension de l'évolution du génome des *Streptomyces*. Les interprétations tirées d'analyses bioinformatiques des génomes ont amené à émettre de nouvelles hypothèses quant aux mécanismes évolutifs et ont soulevé de nouveaux questionnements. Certaines hypothèses ouvrent de nouvelles pistes de recherche.

L'hypothèse selon laquelle l'évolution du chromosome linéaire se produit selon un gradient de fréquence croissante d'insertions/délétions vers les extrémités chromosomiques pourrait être testée. Par des expériences de conjugaison entre souches de *Streptomyces*, il serait intéressant de tester l'intégration préférentielle des molécules d'ADN acquises au niveau des régions terminales.
Le mécanisme de recombinaison illégitime (NHEJ) pourrait également être étudié chez *Streptomyces*. Le rôle des protéines putatives Ku et ligase IV chez *Streptomyces* pourraient faire l'objet d'études à travers la sélection de souches mutantes chez lesquelles les ORF correspondantes auraient été inactivées par mutagenèse dirigée. Il serait intéressant de préciser l'implication potentielle de ce mécanisme dans l'établissement d'un gradient de fréquence de réarrangements dans les régions terminales. Etant donné l'environnement génétique différent des gènes potentiellement impliqués dans ce mécanisme chez *Streptomyces*, il serait intéressant de savoir s'il existe des différences notables dans l'efficacité de ces mécanismes entre les espèces.
La relation entre NHEJ et instabilité génomique pourrait également être recherchée. Quel serait le niveau d'instabilité des souches mutantes n'exprimant plus les protéines Ku et Ligase IV ?
Comme cela a été réalisé chez la levure, il serait possible de suivre le taux de survie à la formation de cassures double-brin générées artificiellement (par l'endonucléase I-*Sce*I par exemple) chez *Streptomyces*. De plus, la caractérisation des événements moléculaires associés à la réparation des cassures pourrait être envisagée. Les mécanismes de réparation diffèrent-ils en fonction des loci chromosomiques ? L'intégration de fragments d'ADN par NHEJ au niveau des sites de cassure sera-t-elle favorisée ?
Il serait également intéressant de tester l'efficacité de la recombinaison homologue en fonction de la localisation chromosomique. Par exemple, à l'aide d'un gène rapporteur localisé entre deux séquences identiques, il serait possible de mesurer la fréquence de recombinaison conduisant à la perte de ce marqueur.

L'hypothèse d'un transfert conjugatif à partir des extrémités chromosomiques pourrait également être testée, de même que le rôle de l'hélicase putative TtrA dans ce mécanisme. Chez la souche *S. ambofaciens* DSM40697, le gène *ttrA* est tronqué (~200 codons identifiables sur 886). Il serait donc intéressant de savoir si la fréquence de mobilisation du chromosome est réduite chez cette souche. A travers la sélection par différents marqueurs localisés dans les régions terminales, il serait également envisageable de vérifier l'hypothèse d'un gradient de fréquence de transfert en fonction de la distance aux télomères. De plus, l'existence de mutants possédant un chromosome circularisé pourrait permettre de réaliser des études comparées concernant les fréquences de mobilisation des chromosomes linéaires et circulaires.

La fouille des génomes ouvre des perspectives très intéressantes dans la compréhension de l'évolution. Les résultats préliminaires concernant l'analyse de l'ensemble des génomes disponibles ont permis de montrer que certains d'entre eux présentent des biais comparables à ceux observés chez *Streptomyces*. Cette analyse doit donc être poursuivie pour mieux comprendre, par exemple, les facteurs communs qui peuvent expliquer une variabilité accrue de certains loci chromosomiques. L'existence de grandes régions de synténie dégénérée chez d'autres espèces pourraient aussi être recherchée.

Par ailleurs, plusieurs projets de séquençage de génomes de *Streptomyces* sont en cours. L'apport de nouvelles données devra permettre de préciser les interprétations tirées de la comparaison des quatre premiers génomes. Quel est le niveau de corrélation entre la taille des extrémités spécifiques et la distance phylogénétique séparant deux espèces ?
Il serait également intéressant de réaliser des comparaisons de génomes complets au niveau intraspécifique. Ce travail a mis en évidence une variabilité entre souches de *S. ambofaciens* qui concerne les extrémités des TIR mais également les régions séquencées en dehors des TIR. Quelle est l'ampleur de cette variabilité à l'échelle des régions de contingence, voire à celle du chromosome complet ?

D'un point de vue appliqué, le séquençage complet de la région centrale du chromosome de *S. ambofaciens* doit être envisagé afin d'identifier les clusters potentiellement impliqués dans la biosynthèse de métabolites d'intérêt.

Enfin, grâce aux développements de nouvelles technologies telles que le séquençage 454, l'explosion des séquences de génomes va s'emballer. Elle ouvre une nouvelle ère fort excitante qui permettra, entre autre, d'améliorer notre compréhension de la dynamique des génomes.

BIBLIOGRAPHIE

RÉFÉRENCES BIBLIOGRAPHIQUES

Abe-Yoshizumi, R., Kamei, U., Yamada, A., Kimura, M., and Ichihara, S. (2004) The evolution of the phenylacetic acid degradation pathway in bacteria. *Biosci Biotechnol Biochem* **68**: 746-748.

Aigle, B., Schneider, D., Morilhat, C., Vandewiele, D., Dary, A., Holl, A.C., Simonet, J.M., and Decaris, B. (1996) An amplifiable and deletable locus of *Streptomyces ambofaciens* RP181110 contains a very large gene homologous to polyketide synthase genes. *Microbiology* **142**: 2815-2824.

Aigle, B., Holl, A.C., Angulo, J.F., Leblond, P., and Decaris, B. (1997) Characterization of two *Streptomyces ambofaciens recA* mutants: identification of the RecA protein by immunoblotting. *FEMS Microbiol Lett* **149**: 181-187.

Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M., and Aksoy, S. (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat Genet* **32**: 402-407.

Allardet-Servent, A., Michaux-Charachon, S., Jumas-Bilak, E., Karayan, L., and Ramuz, M. (1993) Presence of one linear and one circular chromosome in the *Agrobacterium tumefaciens* C58 genome. *J Bacteriol* **175**: 7869-7874.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402.

Andersson, J.O., and Andersson, S.G. (1999) Genome degradation is an ongoing process in *Rickettsia*. *Mol Biol Evol* **16**: 1178-1191.

Andersson, J.O., and Andersson, S.G. (2001) Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Mol Biol Evol* **18**: 829-839.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J., and Zdobnov, E.M. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* **29**: 37-40.

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**: 1365-1374.

Attwood, T.K., Croning, M.D., Flower, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., Selley, J.N., and Wright, W. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res* **28**: 225-227.

Bao, K., and Cohen, S.N. (2001) Terminal proteins essential for the replication of linear plasmids and chromosomes in *Streptomyces*. *Genes Dev* **15**: 1518-1527.

Bao, K., and Cohen, S.N. (2003) Recruitment of terminal protein to the ends of *Streptomyces* linear plasmids and chromosomes by a novel telomere-binding protein essential for linear DNA replication. *Genes Dev* **17**: 774-785.

Bao, K., and Cohen, S.N. (2004) Reverse transcriptase activity innate to DNA polymerase I and DNA topoisomerase I proteins of *Streptomyces* telomere complex. *Proc Natl Acad Sci U S A* **101**: 14361-14366.

Barbour, A.G., and Garon, C.F. (1987) Linear plasmids of the bacterium *Borrelia burgdorferi* have covalently closed ends. *Science* **237**: 409-411.

Barona-Gomez, F., Lautru, S., Francou, F.X., Leblond, P., Pernodet, J.L., and Challis, G.L. (2006) Multiple biosynthetic and uptake systems mediate siderophore-dependent iron acquisition in *Streptomyces coelicolor* M145 and *Streptomyces ambofaciens* ATCC23877. *Microbiology* **in press**.

Barrasa, M.I., Tercero, J.A., and Jimenez, A. (1997) The aminonucleoside antibiotic A201A is inactivated by a phosphotransferase activity from *Streptomyces capreolus* NRRL 3817, the producing organism. Isolation and molecular characterization of the relevant encoding gene and its DNA flanking regions. *Eur J Biochem* **245**: 54-63.

Barry, J.D., Ginger, M.L., Burton, P., and McCulloch, R. (2003) Why are parasite contingency genes often associated with telomeres? *Int J Parasitol* **33**: 29-45.

Bartolome, B., Jubete, Y., Martinez, E., and de la Cruz, F. (1991) Construction and properties of a family of pACYC184-derived cloning vectors compatible with pBR322 and its derivatives. *Gene* **102**: 75-78.

Bate, N., Butler, A.R., Gandecha, A.R., and Cundliffe, E. (1999) Multiple regulatory genes in the tylosin biosynthetic cluster of *Streptomyces fradiae*. *Chem Biol* **6**: 617-624.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L. (2000) The Pfam protein families database. *Nucleic Acids Res* **28**: 263-266.

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**: 141-147.

Bentley, S.D., and Parkhill, J. (2004) Comparative genomic structure of prokaryotes. *Annu Rev Genet* **38**: 771-792.

Berdy, J. (2005) Bioactive microbial metabolites. *J Antibiot (Tokyo)* **58**: 1-26.

Berger, F., Fischer, G., Kyriacou, A., Decaris, B., and Leblond, P. (1996) Mapping of the ribosomal operons on the linear chromosomal DNA of *Streptomyces ambofaciens* DSM40697. *FEMS Microbiol Lett* **143**: 167-173.

Bey, S.J., Tsou, M.F., Huang, C.H., Yang, C.C., and Chen, C.W. (2000) The homologous terminal sequence of the *Streptomyces lividans* chromosome and SLP2 plasmid. *Microbiology* **146**: 911-922.

Beyer, D., Kroll, H.P., Endermann, R., Schiffer, G., Siegel, S., Bauser, M., Pohlmann, J., Brands, M., Ziegelbauer, K., Haebich, D., Eymann, C., and Brotz-Oesterhelt, H. (2004) New class of bacterial phenylalanyl-tRNA synthetase inhibitors with high potency and broad-spectrum activity. *Antimicrob Agents Chemother* **48**: 525-532.

Bialer, M., Yagen, B., Mechoulam, R., and Becker, Y. (1980) Structure-activity relationships of pyrrole amidine antiviral antibiotics. 2. Preparation of mono- and tripyrrole derivatives of congocidine. *J Med Chem* **23**: 1144-1148.

Bigot, S., Saleh, O.A., Lesterlin, C., Pages, C., El Karoui, M., Dennis, C., Grigoriev, M., Allemand, J.F., Barre, F.X., and Cornet, F. (2005) KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase. *Embo J* **24**: 3770-3780.

Birch, A., Hausler, A., Vogtli, M., Krek, W., and Hutter, R. (1989) Extremely large chromosomal deletions are intimately involved in genetic instability and genomic rearrangements in *Streptomyces glaucescens*. *Mol Gen Genet* **217**: 447-458.

Blakely, G., May, G., McCulloch, R., Arciszewska, L.K., Burke, M., Lovett, S.T., and Sherratt, D.J. (1993) Two related recombinases are required for site-specific recombination at *dif* and cer in *E. coli* K12. *Cell* **75**: 351-361.

Boccard, F., Esnault, E., and Valens, M. (2005) Spatial arrangement and macrodomain organization of bacterial chromosomes. *Mol Microbiol* **57**: 9-16.

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**: 1554-1558.

Boucher, Y., Douady, C.J., Sharma, A.K., Kamekura, M., and Doolittle, W.F. (2004) Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. *J Bacteriol* **186**: 3980-3990.

Brewer, B.J. (1988) When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell* **53**: 679-686.

Campo, N., Dias, M.J., Daveran-Mingot, M.L., Ritzenthaler, P., and Le Bourgeois, P. (2004) Chromosomal constraints in Gram-positive bacteria revealed by artificial inversions. *Mol Microbiol* **51**: 511-522.

Carro, D., Garcia-Martinez, J., Perez-Ortin, J.E., and Pina, B. (2003) Structural characterization of chromosome I size variants from a natural yeast strain. *Yeast* **20**: 171-183.

Casjens, S., Murphy, M., DeLange, M., Sampson, L., van Vugt, R., and Huang, W.M. (1997) Telomeres of the linear chromosomes of Lyme disease spirochaetes: nucleotide sequence and possible exchange with linear plasmid telomeres. *Mol Microbiol* **26**: 581-596.

Casjens, S. (1999) Evolution of the linear DNA replicons of the *Borrelia* spirochetes. *Curr Opin Microbiol* **2**: 529-534.

Casjens, S., Palmer, N., van Vugt, R., Huang, W.M., Stevenson, B., Rosa, P., Lathigra, R., Sutton, G., Peterson, J., Dodson, R.J., Haft, D., Hickey, E., Gwinn, M., White, O., and Fraser, C.M. (2000) A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Mol Microbiol* **35**: 490-516.

Catakli, S., Andrieux, A., Leblond, P., Decaris, B., and Dary, A. (2003) Spontaneous chromosome circularization and amplification of a new amplifiable unit of DNA belonging to the terminal inverted repeats in *Streptomyces ambofaciens* ATCC 23877. *Arch Microbiol* **179**: 387-393.

Challis, G.L., and Ravel, J. (2000) Coelichelin, a new peptide siderophore encoded by the *Streptomyces coelicolor* genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol Lett* **187**: 111-114.

Chang, P.C., and Cohen, S.N. (1994) Bidirectional replication from an internal origin in a linear *Streptomyces* plasmid. *Science* **265**: 952-954.

Charlebois, R.L., and Doolittle, W.F. (2004) Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res* **14**: 2469-2477.

Chater, K.F., and Hopwood, D.A. (1984) *Streptomyces* genetics. In *The biology of the Actinomycetes*. Goodfellow, M., Mordarski, M. and Williams, S.T. (eds). London: Academic press, pp. 229-286.

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

Chen, C.W. (1996) Complications and implications of linear bacterial chromosomes. *Trends Genet* **12**: 192-196.

Chen, C.W., Huang, C.H., Lee, H.H., Tsai, H.H., and Kirby, R. (2002) Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. *Trends Genet* **18**: 522-529.

Cole, S.T., Eiglmeier, K., Parkhill, J., James, K.D., Thomson, N.R., Wheeler, P.R., Honore, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R.M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M.A., Rajandream, M.A., Rutherford, K.M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J.R., and Barrell, B.G. (2001) Massive gene decay in the leprosy bacillus. *Nature* **409**: 1007-1011.

Corpet, F., Gouzy, J., and Kahn, D. (1999) Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res* **27**: 263-267.

Corre, J., Cornet, F., Patte, J., and Louarn, J.M. (1997) Unraveling a region-specific hyper-recombination phenomenon: genetic control and modalities of terminal recombination in *Escherichia coli*. *Genetics* **147**: 979-989.

Corre, J., Patte, J., and Louarn, J.M. (2000) Prophage lambda induces terminal recombination in *Escherichia coli* by inhibiting chromosome dimer resolution. An orientation-dependent cis-effect lending support to bipolarization of the terminus. *Genetics* **154**: 39-48.

Corre, J., and Louarn, J.M. (2002) Evidence from terminal recombination gradients that FtsK uses replichore polarity to control chromosome terminus positioning at division in *Escherichia coli*. *J Bacteriol* **184**: 3801-3807.

Corre, J., and Louarn, J.M. (2005) Extent of the activity domain and possible roles of FtsK in the *Escherichia coli* chromosome terminus. *Mol Microbiol* **56**: 1539-1548.

Couturier, E., and Rocha, E.P. (2006) Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol Microbiol* **59**: 1506-1518.

Crozat, E., Philippe, N., Lenski, R.E., Geiselmann, J., and Schneider, D. (2005) Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. *Genetics* **169**: 523-532.

Dabrock, B., Kesseler, M., Averhoff, B., and Gottschalk, G. (1994) Identification and characterization of a transmissible linear plasmid from *Rhodococcus erythropolis* BD2 that encodes isopropylbenzene and trichloroethene catabolism. *Appl Environ Microbiol* **60**: 853-860.

Dai, H., Chow, T.Y., Liao, H.J., Chen, Z.Y., and Chiang, K.S. (1988) Nucleotide sequences involved in the neolysogenic insertion of filamentous phage Cf16-v1 into the *Xanthomonas campestris pv. citri* chromosome. *Virology* **167**: 613-620.

Daley, J.M., Palmbos, P.L., Wu, D., and Wilson, T.E. (2005) Nonhomologous end joining in yeast. *Annu Rev Genet* **39**: 431-451.

Danchin, A., Dondon, L., and Daniel, J. (1984) Metabolic alterations mediated by 2-ketobutyrate in *Escherichia coli* K12. *Mol Gen Genet* **193**: 473-478.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**: 324-328.

Dasilva, C., Hadji, H., Ozouf-Costaz, C., Nicaud, S., Jaillon, O., Weissenbach, J., and Roest Crollius, H. (2002) Remarkable compartmentalization of transposable elements and pseudogenes in the heterochromatin of the *Tetraodon nigroviridis* genome. *Proc Natl Acad Sci U S A* **99**: 13636-13641.

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

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636-4641.

DelVecchio, V.G., Kapatral, V., Redkar, R.J., Patra, G., Mujer, C., Los, T., Ivanova, N., Anderson, I., Bhattacharyya, A., Lykidis, A., Reznik, G., Jablonski, L., Larsen, N., D'Souza, M., Bernal, A., Mazur, M., Goltsman, E., Selkov, E., Elzer, P.H., Hagius, S., O'Callaghan, D., Letesson, J.J., Haselkorn, R., Kyrpides, N., and Overbeek, R. (2002) The genome sequence of the facultative intracellular pathogen *Brucella melitensis*. *Proc Natl Acad Sci U S A* **99**: 443-448.

Deng, S., Stein, R.A., and Higgins, N.P. (2004) Transcription-induced barriers to supercoil diffusion in the *Salmonella typhimurium* chromosome. *Proc Natl Acad Sci U S A* **101**: 3398-3403.

Dervyn, E., Suski, C., Daniel, R., Bruand, C., Chapuis, J., Errington, J., Janniere, L., and Ehrlich, S.D. (2001) Two essential DNA polymerases at the bacterial replication fork. *Science* **294**: 1716-1719.

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**: 1548-1561.

Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* **2**: 414-424.

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

Ducote, M.J., and Pettis, G.S. (2006) An in vivo assay for conjugation-mediated recombination yields novel results for *Streptomyces* plasmid pIJ101. *Plasmid* **55**: 242-248.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J.M., Beyne, E., Bleykasten, C., Boisrame, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J.M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G.F., Straub, M.L., Suleau, A., Swennen, D., Tekaia, F., Wesolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P., and Souciet, J.L. (2004) Genome evolution in yeasts. *Nature* **430**: 35-44.

Eardly, B.D., Nour, S.M., van Berkum, P., and Selander, R.K. (2005) Rhizobial 16S rRNA and dnaK genes: mosaicism and the uncertain phylogenetic placement of *Rhizobium galegae*. *Appl Environ Microbiol* **71**: 1328-1335.

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

Englert, C., Kruger, K., Offner, S., and Pfeifer, F. (1992) Three different but related gene clusters encoding gas vesicles in halophilic archaea. *J Mol Biol* **227**: 586-592.

Ewing, B., and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**: 186-194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175-185.

Fabre, E., Muller, H., Therizols, P., Lafontaine, I., Dujon, B., and Fairhead, C. (2005) Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. *Mol Biol Evol* **22**: 856-873.

Fairhead, C., and Dujon, B. (2006) Structure of *Kluyveromyces lactis* subtelomeres: duplications and gene content. *FEMS Yeast Res* **6**: 428-441.

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**: 5969-5973.

Ferdows, M.S., Serwer, P., Griess, G.A., Norris, S.J., and Barbour, A.G. (1996) Conversion of a linear to a circular plasmid in the relapsing fever agent *Borrelia hermsii*. *J Bacteriol* **178**: 793-800.

Fischer, G., Decaris, B., and Leblond, P. (1997a) Occurrence of deletions, associated with genetic instability in *Streptomyces ambofaciens*, is independent of the linearity of the chromosomal DNA. *J Bacteriol* **179**: 4553-4558.

Fischer, G., Kyriacou, A., Decaris, B., and Leblond, P. (1997b) Genetic instability and its possible evolutionary implications on the chromosomal structure of *Streptomyces*. *Biochimie* **79**: 555-558.

Fischer, G., Holl, A.C., Volff, J.N., Vandewiele, D., Decaris, B., and Leblond, P. (1998a) Replication of the linear chromosomal DNA from the centrally located oriC of *Streptomyces ambofaciens* revealed by PFGE gene dosage analysis. *Res Microbiol* **149**: 203-210.

Fischer, G., Wenner, T., Decaris, B., and Leblond, P. (1998b) Chromosomal arm replacement generates a high level of intraspecific polymorphism in the terminal inverted repeats of the linear chromosomal DNA of *Streptomyces ambofaciens*. *Proc Natl Acad Sci U S A* **95**: 14296-14301.

Fischer, G., Neuveglise, C., Durrens, P., Gaillardin, C., and Dujon, B. (2001) Evolution of gene order in the genomes of two related yeast species. *Genome Res* **11**: 2009-2019.

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

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**: 397-403.

Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K., Gwinn, M., Dougherty, B., Tomb, J.F., Fleischmann, R.D., Richardson, D., Peterson, J., Kerlavage, A.R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M.D., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fuji, C., Cotton, M.D., Horst, K., Roberts, K., Hatch, B., Smith, H.O., and Venter, J.C. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**: 580-586.

Freist, W., and Gauss, D.H. (1995) Threonyl-tRNA synthetase. *Biol Chem Hoppe Seyler* **376**: 213-224.

French, S. (1992) Consequences of replication fork movement through transcription units in vivo. *Science* **258**: 1362-1365.

Garcia-Russell, N., Harmon, T.G., Le, T.Q., Amaladas, N.H., Mathewson, R.D., and Segall, A.M. (2004) Unequal access of chromosomal regions to each other in *Salmonella*: probing chromosome structure with phage lambda integrase-mediated long-range rearrangements. *Mol Microbiol* **52**: 329-344.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan,

M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., and Barrell, B. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**: 498-511.

Gill, P.K., Sharma, A.D., Harchand, R.K., and Singh, P. (2003) Effect of media supplements and culture conditions on inulinase production by an actinomycete strain. *Bioresour Technol* **87**: 359-362.

Ginolhac, A., Jarrin, C., Robe, P., Perriere, G., Vogel, T.M., Simonet, P., and Nalin, R. (2005) Type I polyketide synthases may have evolved through horizontal gene transfer. *J Mol Evol* **60**: 716-725.

Glass, J.I., Lefkowitz, E.J., Glass, J.S., Heiner, C.R., Chen, E.Y., and Cassell, G.H. (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature* **407**: 757-762.

Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison, C.A., 3rd, Smith, H.O., and Venter, J.C. (2006) Essential genes of a minimal bacterium. *Proc Natl Acad Sci U S A* **103**: 425-430.

Glockner, F.O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., Gade, D., Beck, A., Borzym, K., Heitmann, K., Rabus, R., Schlesner, H., Amann, R., and Reinhardt, R. (2003) Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc Natl Acad Sci U S A* **100**: 8298-8303.

Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B.S., Cao, Y., Askenazi, M., Halling, C., Mullin, L., Houmiel, K., Gordon, J., Vaudin, M., Iartchouk, O., Epp, A., Liu, F., Wollam, C., Allinger, M., Doughty, D., Scott, C., Lappas, C., Markelz, B., Flanagan, C., Crowell, C., Gurson, J., Lomo, C., Sear, C., Strub, G., Cielo, C., and Slater, S. (2001) Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* **294**: 2323-2328.

Gordon, D., Abajian, C., and Green, P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res* **8**: 195-202.

Goshi, K., Uchida, T., Lezhava, A., Yamasaki, M., Hiratsu, K., Shinkawa, H., and Kinashi, H. (2002) Cloning and analysis of the telomere and terminal inverted repeat of the linear chromosome of *Streptomyces griseus*. *J Bacteriol* **184**: 3411-3415.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313**: 903-919.

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**: r43-74.

Gravius, B., Bezmalinovic, T., Hranueli, D., and Cullum, J. (1993) Genetic instability and strain degeneration in *Streptomyces rimosus*. *Appl Environ Microbiol* **59**: 2220-2228.

Grohmann, E., Muth, G., and Espinosa, M. (2003) Conjugative plasmid transfer in gram-positive bacteria. *Microbiol Mol Biol Rev* **67**: 277-301, table of contents.

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**: 41-43.

Hamilton, H.L., Dominguez, N.M., Schwartz, K.J., Hackett, K.T., and Dillard, J.P. (2005) *Neisseria gonorrhoeae* secretes chromosomal DNA via a novel type IV secretion system. *Mol Microbiol* **55**: 1704-1721.

Hanafusa, T., and Kinashi, H. (1992) The structure of an integrated copy of the giant linear plasmid SCP1 in the chromosome of *Streptomyces coelicolor* 2612. *Mol Gen Genet* **231**: 363-368.

Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Umayam, L., Gill, S.R., Nelson, K.E., Read, T.D., Tettelin, H., Richardson, D., Ermolaeva, M.D., Vamathevan, J., Bass, S., Qin, H., Dragoi, I., Sellers, P., McDonald, L., Utterback, T., Fleishmann, R.D., Nierman, W.C., White, O., Salzberg, S.L., Smith, H.O., Colwell, R.R., Mekalanos, J.J., Venter, J.C., and Fraser, C.M. (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**: 477-483.

Hendrickson, H., and Lawrence, J.G. (2006) Selection for Chromosome Architecture in Bacteria. *J Mol Evol.*

Herrmann, R., and Reiner, B. (1998) *Mycoplasma pneumoniae* and *Mycoplasma genitalium*: a comparison of two closely related bacterial species. *Curr Opin Microbiol* **1**: 572-579.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C., and Herrmann, R. (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* **24**: 4420-4449.

Hochhut, B., Jahreis, K., Lengeler, J.W., and Schmid, K. (1997) CTnscr94, a conjugative transposon found in enterobacteria. *J Bacteriol* **179**: 2097-2102.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res* **27**: 215-219.

Hopwood, D.A., Kieser, T., Wright, H.M., and Bibb, M.J. (1983) Plasmids, recombination and chromosome mapping in *Streptomyces lividans* 66. *J Gen Microbiol* **129**: 2257-2269.

Hopwood, D.A., Lydiate, D.J., Malpartida, F., and Wright, H.M. (1985) Conjugative sex plasmids of *Streptomyces*. *Basic Life Sci* **30**: 615-634.

Hopwood, D.A. (2006) Soil to Genomics: The *Streptomyces* Chromosome. *Annu Rev Genet.*

Huang, C.H., Lin, Y.S., Yang, Y.L., Huang, S.W., and Chen, C.W. (1998) The telomeres of *Streptomyces* chromosomes contain conserved palindromic sequences with potential to form complex secondary structures. *Mol Microbiol* **28**: 905-916.

Huang, C.H., Chen, C.Y., Tsai, H.H., Chen, C., Lin, Y.S., and Chen, C.W. (2003) Linear plasmid SLP2 of *Streptomyces lividans* is a composite replicon. *Mol Microbiol* **47**: 1563-1576.

Huang, T.W., and Chen, C.W. (2006) A *recA* Null Mutation May Be Generated in *Streptomyces coelicolor*. *J Bacteriol* **188**: 6771-6779.

Huang, W.M., Robertson, M., Aron, J., and Casjens, S. (2004) Telomere exchange between linear replicons of *Borrelia burgdorferi*. *J Bacteriol* **186**: 4134-4141.

Huber, K.E., and Waldor, M.K. (2002) Filamentous phage integration requires the host recombinases XerC and XerD. *Nature* **417**: 656-659.

Hudson, R.E., Bergthorsson, U., Roth, J.R., and Ochman, H. (2002) Effect of chromosome location on bacterial mutation rates. *Mol Biol Evol* **19**: 85-92.

Hutchison, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O., and Venter, J.C. (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**: 2165-2169.

Hütter, R. (1967) In *Systematik der Streptomycete*. Verlag, K. (ed). Basel.

Iida, T., Makino, K., Nasu, H., Yokoyama, K., Tagomori, K., Hattori, A., Okuno, T., Shinagawa, H., and Honda, T. (2002) Filamentous bacteriophages of vibrios are integrated into the dif-like site of the host chromosome. *J Bacteriol* **184**: 4933-4935.

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**: 526-531.

Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol* **146**: 1-21.

Ikemura, T. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J Mol Biol* **158**: 573-597.

Inoue, S., Higashiyama, K., Uchida, T., Hiratsu, K., and Kinashi, H. (2003) Chromosomal circularization in *Streptomyces griseus* by nonhomologous recombination of deletion ends. *Biosci Biotechnol Biochem* **67**: 1101-1108.

Ip, S.C., Bregu, M., Barre, F.X., and Sherratt, D.J. (2003) Decatenation of DNA circles by FtsK-dependent Xer site-specific recombination. *Embo J* **22**: 6399-6407.

Ishikawa, J., and Hotta, K. (1999) FramePlot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial DNA with a high G + C content. *FEMS Microbiol Lett* **174**: 251-253.

Itoh, T., Takemoto, K., Mori, H., and Gojobori, T. (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol* **16**: 332-346.

Jakimowicz, D., Gust, B., Zakrzewska-Czerwinska, J., and Chater, K.F. (2005) Developmental-stage-specific assembly of ParB complexes in *Streptomyces coelicolor* hyphae. *J Bacteriol* **187**: 3572-3580.

Karlin, S., and Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11**: 283-290.

Karlin, S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* **1**: 598-610.

Karlin, S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol* **9**: 335-343.

Karoonuthaisiri, N., Weaver, D., Huang, J., Cohen, S.N., and Kao, C.M. (2005) Regional organization of gene expression in *Streptomyces coelicolor*. *Gene* **353**: 53-66.

Kato-Maeda, M., Rhee, J.T., Gingeras, T.R., Salamon, H., Drenkow, J., Smittipat, N., and Small, P.M. (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res* **11**: 547-554.

Kawase, T., Saito, A., Sato, T., Kanai, R., Fujii, T., Nikaidou, N., Miyashita, K., and Watanabe, T. (2004) Distribution and phylogenetic analysis of family 19 chitinases in *Actinobacteria*. *Appl Environ Microbiol* **70**: 1135-1144.

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

Kers, J.A., Cameron, K.D., Joshi, M.V., Bukhalid, R.A., Morello, J.E., Wach, M.J., Gibson, D.M., and Loria, R. (2005) A large, mobile pathogenicity island confers plant pathogenicity on *Streptomyces* species. *Mol Microbiol* **55**: 1025-1033.

Kieser, H.M., Kieser, T., and Hopwood, D.A. (1992) A combined genetic and physical map of the *Streptomyces coelicolor* A3(2) chromosome. *J Bacteriol* **174**: 5496-5507.

Kieser, T., Hopwood, D.A., Wright, H.M., and Thompson, C.J. (1982) pIJ101, a multi-copy broad host-range *Streptomyces* plasmid: functional analysis and development of DNA cloning vectors. *Mol Gen Genet* **185**: 223-228.

Kinashi, H., Shimaji-Murayama, M., and Hanafusa, T. (1991) Nucleotide sequence analysis of the unusually long terminal inverted repeats of a giant linear plasmid, SCP1. *Plasmid* **26**: 123-130.

Kinashi, H., Shimaji-Murayama, M., and Hanafusa, T. (1992) Integration of SCP1, a giant linear plasmid, into the *Streptomyces coelicolor* chromosome. *Gene* **115**: 35-41.

Klasson, L., and Andersson, S.G. (2004) Evolution of minimal-gene-sets in host-dependent bacteria. *Trends Microbiol* **12**: 37-43.

Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., Richardson, D.L., Kerlavage, A.R., Graham, D.E., Kyrpides, N.C., Fleischmann, R.D., Quackenbush, J., Lee, N.H., Sutton, G.G., Gill, S., Kirkness, E.F., Dougherty, B.A., McKenney, K., Adams, M.D., Loftus, B., Peterson, S., Reich, C.I., McNeil, L.K., Badger, J.H., Glodek, A., Zhou, L., Overbeek, R., Gocayne, J.D., Weidman, J.F., McDonald, L., Utterback, T., Cotton, M.D., Spriggs, T., Artiach, P., Kaine, B.P., Sykes, S.M., Sadow, P.W., D'Andrea, K.P., Bowman, C., Fujii, C., Garland, S.A., Mason, T.M., Olsen, G.J., Fraser, C.M., Smith, H.O., Woese, C.R., and Venter, J.C. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**: 364-370.

Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., Boland, F., Brignell, S.C., Bron, S., Bunai, K., Chapuis, J., Christiansen, L.C., Danchin, A., Debarbouille, M., Dervyn, E., Deuerling, E., Devine, K., Devine, S.K., Dreesen, O., Errington, J., Fillinger, S., Foster, S.J., Fujita, Y., Galizzi, A., Gardan, R., Eschevins, C., Fukushima, T., Haga, K., Harwood, C.R., Hecker, M., Hosoya, D., Hullo, M.F., Kakeshita, H., Karamata, D., Kasahara, Y., Kawamura, F., Koga, K., Koski, P., Kuwana, R., Imamura, D., Ishimaru, M., Ishikawa, S., Ishio, I., Le Coq, D., Masson, A., Mauel, C., Meima, R., Mellado, R.P., Moir, A., Moriya, S., Nagakawa, E., Nanamiya, H., Nakai, S., Nygaard, P., Ogura, M., Ohanan, T., O'Reilly, M., O'Rourke, M., Pragai, Z., Pooley, H.M., Rapoport, G., Rawlins, J.P., Rivas, L.A., Rivolta, C., Sadaie, A., Sadaie, Y., Sarvas, M., Sato, T., Saxild, H.H., Scanlan, E., Schumann, W., Seegers, J.F., Sekiguchi, J., Sekowska, A., Seror, S.J., Simon, M., Stragier, P., Studer, R., Takamatsu, H., Tanaka, T., Takeuchi, M., Thomaides, H.B., Vagner, V., van Dijl, J.M., Watabe, K., Wipat, A., Yamamoto, H., Yamamoto, M., Yamamoto, Y., Yamane, K., Yata, K., Yoshida, K.,

Yoshikawa, H., Zuber, U., and Ogasawara, N. (2003) Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A* **100**: 4678-4683.

Kolpakov, R., Bana, G., and Kucherov, G. (2003) mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* **31**: 3672-3678.

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**: 3160-3165.

Koonin, E.V. (2000) How many genes can make a cell: the minimal-gene-set concept. *Annu Rev Genomics Hum Genet* **1**: 99-116.

Kosono, S., Kataoka, M., Seki, T., and Yoshida, T. (1996) The TraB protein, which mediates the intermycelial transfer of the *Streptomyces* plasmid pSN22, has functional NTP-binding motifs and is localized to the cytoplasmic membrane. *Mol Microbiol* **19**: 397-405.

Kowalczykowski, S.C. (2000) Initiation of genetic recombination and recombination-dependent replication. *Trends Biochem Sci* **25**: 156-165.

Krugel, H., Fiedler, G., Smith, C., and Baumberg, S. (1993) Sequence and transcriptional analysis of the nourseothricin acetyltransferase-encoding gene nat1 from *Streptomyces noursei*. *Gene* **127**: 127-131.

Krugel, H., Krubasik, P., Weber, K., Saluz, H.P., and Sandmann, G. (1999) Functional analysis of genes from *Streptomyces griseus* involved in the synthesis of isorenieratene, a carotenoid with aromatic end groups, revealed a novel type of carotenoid desaturase. *Biochim Biophys Acta* **1439**: 57-64.

Lafontaine, I., Fischer, G., Talla, E., and Dujon, B. (2004) Gene relics in the genome of the yeast *Saccharomyces cerevisiae*. *Gene* **335**: 1-17.

Lathe, W.C., 3rd, Snel, B., and Bork, P. (2000) Gene context conservation of a higher order than operons. *Trends Biochem Sci* **25**: 474-479.

Lau, I.F., Filipe, S.R., Soballe, B., Okstad, O.A., Barre, F.X., and Sherratt, D.J. (2003) Spatial and temporal organization of replicating *Escherichia coli* chromosomes. *Mol Microbiol* **49**: 731-743.

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

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

Lawrence, J.G., and Roth, J.R. (1999) Genomic flux: genome evolution by gene loss and acquisition in bacterial genomes. In *Organization of the Prokaryotic Genome*. Charlebois, R.L. (ed). Washington D.C.: American Society for Microbiology, pp. 263-289.

Lawrence, J.G., Hendrix, R.W., and Casjens, S. (2001) Where are the pseudogenes in bacterial genomes? *Trends Microbiol* **9**: 535-540.

Lawrence, J.G., and Hendrickson, H. (2004) Chromosome structure and constraints on lateral gene transfer. *Dynamical Genetics*: 319-336.

Lawrence, J.G., and Hendrickson, H. (2005) Genome evolution in bacteria: order beneath chaos. *Curr Opin Microbiol* **8**: 572-578.

Leblond, P., Demuyter, P., Moutier, L., Laakel, M., Decaris, B., and Simonet, J.M. (1989) Hypervariability, a new phenomenon of genetic instability, related to DNA amplification in *Streptomyces ambofaciens*. *J Bacteriol* **171**: 419-423.

Leblond, P., Demuyter, P., Simonet, J.M., and Decaris, B. (1991) Genetic instability and associated genome plasticity in *Streptomyces ambofaciens*: pulsed-field gel electrophoresis evidence for large DNA alterations in a limited genomic region. *J Bacteriol* **173**: 4229-4233.

Leblond, P., and Decaris, B. (1994) New insights into the genetic instability of *Streptomyces*. *FEMS Microbiol Lett* **123**: 225-232.

Leblond, P., Fischer, G., Francou, F.X., Berger, F., Guerineau, M., and Decaris, B. (1996) The unstable region of *Streptomyces ambofaciens* includes 210 kb terminal inverted repeats flanking the extremities of the linear chromosomal DNA. *Mol Microbiol* **19**: 261-271.

Leblond, P., and Decaris, B. (1999) 'Unstable' linear chromosomes: the case of *Streptomyces*. In *Organization of the Prokaryotic Genome*. Charlebois, R.L. (ed). Washington, D.C.: American Society for Microbiology, pp. 235-261.

Lerat, E., and Ochman, H. (2004) Psi-Phi: exploring the outer limits of bacterial pseudogenes. *Genome Res* **14**: 2273-2278.

Lerat, E., and Ochman, H. (2005) Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res* **33**: 3125-3132.

Levy, O., Ptacin, J.L., Pease, P.J., Gore, J., Eisen, M.B., Bustamante, C., and Cozzarelli, N.R. (2005) Identification of oligonucleotide sequences that direct the movement of the *Escherichia coli* FtsK translocase. *Proc Natl Acad Sci U S A* **102**: 17618-17623.

Lin, N.T., Chang, R.Y., Lee, S.J., and Tseng, Y.H. (2001) Plasmids carrying cloned fragments of RF DNA from the filamentous phage (phi)Lf can be integrated into the host chromosome via site-specific integration and homologous recombination. *Mol Genet Genomics* **266**: 425-435.

Lin, Y.S., Kieser, H.M., Hopwood, D.A., and Chen, C.W. (1993) The chromosomal DNA of *Streptomyces lividans* 66 is linear. *Mol Microbiol* **10**: 923-933.

Lin, Y.S., and Chen, C.W. (1997) Instability of artificially circularized chromosomes of *Streptomyces lividans*. *Mol Microbiol* **26**: 709-719.

Liu, B., Wong, M.L., Tinker, R.L., Geiduschek, E.P., and Alberts, B.M. (1993) The DNA replication fork can pass RNA polymerase without displacing the nascent transcript. *Nature* **366**: 33-39.

Liu, B., and Alberts, B.M. (1995) Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. *Science* **267**: 1131-1137.

Llorente, B., Malpertuy, A., Neuveglise, C., de Montigny, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., Durrens, P., Gaillardin, C., Lepingle, A., Ozier-Kalogeropoulos, O., Potier, S., Saurin, W., Tekaia, F., Toffano-Nioche, C., Wesolowski-Louvel, M., Wincker, P., Weissenbach, J., Souciet, J., and Dujon, B. (2000) Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. *FEBS Lett* **487**: 101-112.

Lobry, J.R., and Louarn, J.M. (2003) Polarisation of prokaryotic chromosomes. *Curr Opin Microbiol* **6**: 101-108.

Louarn, J., Cornet, F., Francois, V., Patte, J., and Louarn, J.M. (1994) Hyperrecombination in the terminus region of the *Escherichia coli* chromosome: possible relation to nucleoid organization. *J Bacteriol* **176**: 7524-7531.

Louarn, J.M., Louarn, J., Francois, V., and Patte, J. (1991) Analysis and possible role of hyperrecombination in the termination region of the *Escherichia coli* chromosome. *J Bacteriol* **173**: 5097-5104.

Louis, E.J. (1995) The chromosome ends of *Saccharomyces cerevisiae*. *Yeast* **11**: 1553-1573.

Lowe, T.M., and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955-964.

Manna, D., Breier, A.M., and Higgins, N.P. (2004) Microarray analysis of transposition targets in *Escherichia coli*: the impact of transcription. *Proc Natl Acad Sci U S A* **101**: 9780-9785.

Martin, P., Dary, A., and Decaris, B. (1998) Generation of a genetic polymorphism in clonal populations of the bacterium *Streptomyces ambofaciens*: characterization of different mutator states. *Mutat Res* **421**: 73-82.

Masai, E., Sugiyama, K., Iwashita, N., Shimizu, S., Hauschild, J.E., Hatta, T., Kimbara, K., Yano, K., and Fukuda, M. (1997) The bphDEF meta-cleavage pathway genes involved in biphenyl/polychlorinated biphenyl degradation are located on a linear plasmid and separated from the initial bphACB genes in *Rhodococcus* sp. strain RHA1. *Gene* **187**: 141-149.

McClintock, B. (1951) Chromosome organization and genic expression. *Cold Spring Harb Symp Quant Biol* **16**: 13-47.

McLean, M.J., Wolfe, K.H., and Devine, K.M. (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol* **47**: 691-696.

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005) The microbial pan-genome. *Curr Opin Genet Dev* **15**: 589-594.

Metsa-Ketela, M., Halo, L., Munukka, E., Hakala, J., Mantsala, P., and Ylihonko, K. (2002) Molecular evolution of aromatic polyketides and comparative sequence analysis of polyketide ketosynthase and 16S ribosomal DNA genes from various *streptomyces* species. *Appl Environ Microbiol* **68**: 4472-4479.

Michel, B., Ehrlich, S.D., and Uzest, M. (1997) DNA double-strand breaks caused by replication arrest. *Embo J* **16**: 430-438.

Michel, B., Flores, M.J., Viguera, E., Grompone, G., Seigneur, M., and Bidnenko, V. (2001) Rescue of arrested replication forks by homologous recombination. *Proc Natl Acad Sci U S A* **98**: 8181-8188.

Mira, A., Ochman, H., and Moran, N.A. (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17**: 589-596.

Mira, A., and Ochman, H. (2002) Gene location and bacterial sequence divergence. *Mol Biol Evol* **19**: 1350-1358.

Mochizuki, S., Hiratsu, K., Suwa, M., Ishii, T., Sugino, F., Yamada, K., and Kinashi, H. (2003) The large linear plasmid pSLA2-L of *Streptomyces rochei* has an unusually condensed gene organization for secondary metabolism. *Mol Microbiol* **48**: 1501-1510.

Moran, N.A. (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A* **93**: 2873-2878.

Moran, N.A., and Mira, A. (2001) The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. *Genome Biol* **2**: RESEARCH0054.

Moran, N.A. (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**: 583-586.

Morrow, D.M., Connelly, C., and Hieter, P. (1997) "Break copy" duplication: a model for chromosome fragment formation in *Saccharomyces cerevisiae*. *Genetics* **147**: 371-382.

Mosher, R.H., Camp, D.J., Yang, K., Brown, M.P., Shaw, W.V., and Vining, L.C. (1995) Inactivation of chloramphenicol by O-phosphorylation. A novel resistance mechanism in *Streptomyces venezuelae* ISP5230, a chloramphenicol producer. *J Biol Chem* **270**: 27000-27006.

Moszer, I., Rocha, E.P., and Danchin, A. (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr Opin Microbiol* **2**: 524-528.

Muller, H.J. (1964) The Relation of Recombination to Mutational Advance. *Mutat Res* **106**: 2-9.

Mushegian, A.R., and Koonin, E.V. (1996a) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A* **93**: 10268-10273.

Mushegian, A.R., and Koonin, E.V. (1996b) Gene order is not conserved in bacterial evolution. *Trends Genet* **12**: 289-290.

Musialowski, M.S., Flett, F., Scott, G.B., Hobbs, G., Smith, C.P., and Oliver, S.G. (1994) Functional evidence that the principal DNA replication origin of the *Streptomyces coelicolor* chromosome is close to the dnaA-gyrB region. *J Bacteriol* **176**: 5123-5125.

Muth, G., Frese, D., Kleber, A., and Wohlleben, W. (1997) Mutational analysis of the *Streptomyces lividans recA* gene suggests that only mutants with residual activity remain viable. *Mol Gen Genet* **255**: 420-428.

Nakamura, Y., Itoh, T., Matsuda, H., and Gojobori, T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36**: 760-766.

Neylon, C., Kralicek, A.V., Hill, T.M., and Dixon, N.E. (2005) Replication termination in *Escherichia coli*: structure and antihelicase activity of the Tus-Ter complex. *Microbiol Mol Biol Rev* **69**: 501-526.

Niki, H., Yamaichi, Y., and Hiraga, S. (2000) Dynamic organization of chromosomal DNA in *Escherichia coli*. *Genes Dev* **14**: 212-223.

Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299-304.

Olsen, G.J., Woese, C.R., and Overbeek, R. (1994) The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol* **176**: 1-6.

Orengo, C.A., Pearl, F.M., and Thornton, J.M. (2003) The CATH domain structure database. *Methods Biochem Anal* **44**: 249-271.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999a) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**: 2896-2901.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999b) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol* **1**: 93-108.

Pandza, S., Biukovic, G., Paravic, A., Dadbin, A., Cullum, J., and Hranueli, D. (1998) Recombination between the linear plasmid pPZG101 and the linear chromosome of *Streptomyces rimosus* can lead to exchange of ends. *Mol Microbiol* **28**: 1165-1176.

Pang, X., Aigle, B., Girardet, J.M., Mangenot, S., Pernodet, J.L., Decaris, B., and Leblond, P. (2004) Functional angucycline-like antibiotic gene cluster in the terminal inverted repeats of the *Streptomyces ambofaciens* linear chromosome. *Antimicrob Agents Chemother* **48**: 575-588.

Parkhill, J., Dougan, G., James, K.D., Thomson, N.R., Pickard, D., Wain, J., Churcher, C., Mungall, K.L., Bentley, S.D., Holden, M.T., Sebaihia, M., Baker, S., Basham, D., Brooks, K., Chillingworth, T., Connerton, P., Cronin, A., Davis, P., Davies, R.M., Dowd, L., White, N., Farrar, J., Feltwell, T., Hamlin, N., Haque, A., Hien, T.T., Holroyd, S., Jagels, K., Krogh, A., Larsen, T.S., Leather, S., Moule, S., O'Gaora, P., Parry, C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S., and Barrell, B.G. (2001a) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar *Typhi* CT18. *Nature* **413**: 848-852.

Parkhill, J., Wren, B.W., Thomson, N.R., Titball, R.W., Holden, M.T., Prentice, M.B., Sebaihia, M., James, K.D., Churcher, C., Mungall, K.L., Baker, S., Basham, D., Bentley, S.D., Brooks, K., Cerdeno-Tarraga, A.M., Chillingworth, T., Cronin, A., Davies, R.M., Davis, P., Dougan, G., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Karlyshev, A.V., Leather, S., Moule, S., Oyston, P.C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S., and Barrell, B.G. (2001b) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413**: 523-527.

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., Holden, M.T., Churcher, C.M., Bentley, S.D., Mungall, K.L., Cerdeno-Tarraga, A.M., Temple, L., James, K., Harris, B., Quail, M.A., Achtman, M., Atkin, R., Baker, S., Basham, D., Bason, N., Cherevach, I., Chillingworth, T., Collins, M., Cronin, A., Davis, P., Doggett, J., Feltwell, T., Goble, A., Hamlin, N., Hauser, H., Holroyd, S., Jagels, K., Leather, S., Moule, S., Norberczak, H., O'Neil, S., Ormond, D., Price, C., Rabbinowitsch, E., Rutter, S., Sanders, M., Saunders, D., Seeger, K., Sharp, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Unwin, L., Whitehead, S., Barrell, B.G., and Maskell, D.J. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nat Genet* **35**: 32-40.

Pease, P.J., Levy, O., Cost, G.J., Gore, J., Ptacin, J.L., Sherratt, D., Bustamante, C., and Cozzarelli, N.R. (2005) Sequence-directed DNA translocation by purified FtsK. *Science* **307**: 586-590.

Pernodet, J.L., Simonet, J.M., and Guerineau, M. (1984) Plasmids in different strains of *Streptomyces ambofaciens*: free and integrated form of plasmid pSAM2. *Mol Gen Genet* **198**: 35-41.

Peters, J.E., and Craig, N.L. (2000) Tn7 transposes proximal to DNA double-strand breaks and into regions where chromosomal DNA replication terminates. *Mol Cell* **6**: 573-582.

Pettis, G.S., and Cohen, S.N. (1994) Transfer of the plJ101 plasmid in *Streptomyces lividans* requires a cis-acting function dispensable for chromosomal gene transfer. *Mol Microbiol* **13**: 955-964.

Pinnert-Sindico, S. (1954) Une nouvelle espèce de *Streptomyces* productrice d'antibiotiques: *Streptomyces ambofaciens* n. sp. caractères culturaux. *Ann Inst Pasteur (Paris)* **87**: 702-707.

Pitcher, R.S., Tonkin, L.M., Green, A.J., and Doherty, A.J. (2005) Domain structure of a NHEJ DNA repair ligase from *Mycobacterium tuberculosis*. *J Mol Biol* **351**: 531-544.

Polard, P., Prere, M.F., Fayet, O., and Chandler, M. (1992) Transposase-induced excision and circularization of the bacterial insertion sequence IS*911*. *Embo J* **11**: 5079-5090.

Possoz, C., Ribard, C., Gagnat, J., Pernodet, J.L., and Guerineau, M. (2001) The integrative element pSAM2 from *Streptomyces*: kinetics and mode of conjugal transfer. *Mol Microbiol* **42**: 159-166.

Qin, Z., and Cohen, S.N. (1998) Replication at the telomeres of the *Streptomyces* linear plasmid pSLA2. *Mol Microbiol* **28**: 893-903.

Qiu, W.G., Schutzer, S.E., Bruno, J.F., Attie, O., Xu, Y., Dunn, J.J., Fraser, C.M., Casjens, S.R., and Luft, B.J. (2004) Genetic exchange and plasmid transfers in *Borrelia burgdorferi sensu stricto* revealed by three-way genome comparisons and multilocus sequence typing. *Proc Natl Acad Sci U S A* **101**: 14150-14155.

Ranea, J.A., Buchan, D.W., Thornton, J.M., and Orengo, C.A. (2004) Evolution of protein superfamilies and bacterial genome size. *J Mol Biol* **336**: 871-887.

Rebollo, J.E., Francois, V., and Louarn, J.M. (1988) Detection and possible role of two large nondivisible zones on the *Escherichia coli* chromosome. *Proc Natl Acad Sci U S A* **85**: 9391-9395.

Redenbach, M., Flett, F., Piendl, W., Glocker, I., Rauland, U., Wafzig, O., Kliem, R., Leblond, P., and Cullum, J. (1993) The *Streptomyces lividans* 66 chromosome contains a 1 MB deletogenic region flanked by two amplifiable regions. *Mol Gen Genet* **241**: 255-262.

Reimmann, C., and Haas, D. (1993) Mobilization of Chromosomes and Nonconjugative Plasmids by Conintegrative Mechanisms. In *Bacterial Conjugation*. Clewell, D.B. (ed). New York: Plenum Press.

Reuther, J., Gekeler, C., Tiffert, Y., Wohlleben, W., and Muth, G. (2006) Unique conjugation mechanism in mycelial streptomycetes: a DNA-binding ATPase translocates unprocessed plasmid DNA at the hyphal tip. *Mol Microbiol* **61**: 436-446.

Ricchetti, M., Fairhead, C., and Dujon, B. (1999) Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* **402**: 96-100.

Ricchetti, M., Dujon, B., and Fairhead, C. (2003) Distance from the chromosome end determines the efficiency of double strand break repair in subtelomeres of haploid yeast. *J Mol Biol* **328**: 847-862.

Rivera, M.C., Jain, R., Moore, J.E., and Lake, J.A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A* **95**: 6239-6244.

Rocha, E. (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol* **10**: 393-395.

Rocha, E.P., and Danchin, A. (2002) Base composition bias might result from competition for metabolic resources. *Trends Genet* **18**: 291-294.

Rocha, E.P., and Danchin, A. (2003a) Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res* **31**: 6570-6577.

Rocha, E.P., and Danchin, A. (2003b) Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet* **34**: 377-378.

Rocha, E.P. (2004a) The replication-related organization of bacterial genomes. *Microbiology* **150**: 1609-1627.

Rocha, E.P. (2004b) Order and disorder in bacterial genomes. *Curr Opin Microbiol* **7**: 519-527.

Rocha, E.P., Cornet, E., and Michel, B. (2005) Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet* **1**: e15.

Rocha, E.P. (2006) Inference and analysis of the relative stability of bacterial chromosomes. *Mol Biol Evol* **23**: 513-522.

Roth, V., Aigle, B., Bunet, R., Wenner, T., Fourrier, C., Decaris, B., and Leblond, P. (2004) Differential and cross-transcriptional control of duplicated genes encoding alternative sigma factors in *Streptomyces ambofaciens*. *J Bacteriol* **186**: 5355-5365.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16**: 944-945.

Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., Billault, A., Brottier, P., Camus, J.C., Cattolico, L., Chandler, M., Choisne, N., Claudel-Renard, C., Cunnac, S., Demange, N., Gaspin, C., Lavie, M., Moisan, A., Robert, C., Saurin, W., Schiex, T., Siguier, P., Thebault, P., Whalen, M., Wincker, P., Levy, M., Weissenbach, J., and Boucher, C.A. (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* **415**: 497-502.

Sawitzke, J., and Austin, S. (2001) An analysis of the factory model for chromosome replication and segregation in bacteria. *Mol Microbiol* **40**: 786-794.

Schmid, M.B., Kapur, N., Isaacson, D.R., Lindroos, P., and Sharpe, C. (1989) Genetic analysis of temperature-sensitive lethal mutants of *Salmonella typhimurium*. *Genetics* **123**: 625-633.

Schneider, D., Bruton, C.J., and Chater, K.F. (2000) Duplicated gene clusters suggest an interplay of glycogen and trehalose metabolism during sequential stages of aerial mycelium development in *Streptomyces coelicolor* A3(2). *Mol Gen Genet* **263**: 543-553.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., and Bork, P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res* **28**: 231-234.

Schwedock, J., McCormick, J.R., Angert, E.R., Nodwell, J.R., and Losick, R. (1997) Assembly of the cell division protein FtsZ into ladder-like structures in the aerial hyphae of *Streptomyces coelicolor*. *Mol Microbiol* **25**: 847-858.

Scordis, P., Flower, D.R., and Attwood, T.K. (1999) FingerPRINTScan: intelligent searching of the PRINTS motif database. *Bioinformatics* **15**: 799-806.

Segall, A., Mahan, M.J., and Roth, J.R. (1988) Rearrangement of the bacterial chromosome: forbidden inversions. *Science* **241**: 1314-1318.

Sekine, M., Tanikawa, S., Omata, S., Saito, M., Fujisawa, T., Tsukatani, N., Tajima, T., Sekigawa, T., Kosugi, H., Matsuo, Y., Nishiko, R., Imamura, K., Ito, M., Narita, H., Tago, S., Fujita, N., and Harayama, S. (2006) Sequence analysis of three plasmids harboured in *Rhodococcus erythropolis* strain PR4. *Environ Microbiol* **8**: 334-346.

Sekine, Y., Eisaki, N., and Ohtsubo, E. (1994) Translational control in production of transposase and in transposition of insertion sequence IS*3*. *J Mol Biol* **235**: 1406-1420.

Seshadri, R., Paulsen, I.T., Eisen, J.A., Read, T.D., Nelson, K.E., Nelson, W.C., Ward, N.L., Tettelin, H., Davidsen, T.M., Beanan, M.J., Deboy, R.T., Daugherty, S.C., Brinkac, L.M., Madupu, R., Dodson, R.J., Khouri, H.M., Lee, K.H., Carty, H.A., Scanlan, D., Heinzen, R.A., Thompson, H.A., Samuel, J.E., Fraser, C.M., and Heidelberg, J.F. (2003) Complete genome sequence of the Q-fever pathogen *Coxiella burnetii*. *Proc Natl Acad Sci U S A* **100**: 5455-5460.

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**: 1281-1295.

Sharp, P.M., Shields, D.C., Wolfe, K.H., and Li, W.H. (1989) Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* **246**: 808-810.

Shiffman, D., and Cohen, S.N. (1992) Reconstruction of a *Streptomyces* linear replicon from separately cloned DNA fragments: existence of a cryptic origin of circular replication within the linear plasmid. *Proc Natl Acad Sci U S A* **89**: 6129-6133.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera sp. APS*. *Nature* **407**: 81-86.

Shimizu, S., Kobayashi, H., Masai, E., and Fukuda, M. (2001) Characterization of the 450-kb linear plasmid in a polychlorinated biphenyl degrader, *Rhodococcus* sp. strain RHA1. *Appl Environ Microbiol* **67**: 2021-2028.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci U S A* **89**: 8794-8797.

Singleton, M.R., Dillingham, M.S., Gaudier, M., Kowalczykowski, S.C., and Wigley, D.B. (2004) Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. *Nature* **432**: 187-193.

Snel, B., Bork, P., and Huynen, M.A. (1999) Genome phylogeny based on gene content. *Nat Genet* **21**: 108-110.

Songsivilai, S., and Dharakul, T. (2000) Multiple replicons constitute the 6.5-megabase genome of *Burkholderia pseudomallei*. *Acta Trop* **74**: 169-179.

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.

Souciet, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., Casaregola, S., de Montigny, J., Dujon, B., Durrens, P., Gaillardin, C., Lepingle, A., Llorente, B., Malpertuy, A., Neuveglise, C., Ozier-Kalogeropoulos, O., Potier, S., Saurin, W., Tekaia, F., Toffano-Nioche, C., Wesolowski-Louvel, M., Wincker, P., and Weissenbach, J. (2000) Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. *FEBS Lett* **487**: 3-12.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., and Birney, E. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**: 1611-1618.

Stecker, C., Johann, A., Herzberg, C., Averhoff, B., and Gottschalk, G. (2003) Complete nucleotide sequence and genetic organization of the 210-kilobase linear plasmid of *Rhodococcus erythropolis* BD2. *J Bacteriol* **185**: 5269-5274.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E., Lory, S., and Olson, M.V. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**: 959-964.

Strohl, W.R. (1992) Compilation and analysis of DNA sequences associated with apparent streptomycete promoters. *Nucleic Acids Res* **20**: 961-974.

Suyama, M., and Bork, P. (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet* **17**: 10-13.

Suzek, B.E., Ermolaeva, M.D., Schreiber, M., and Salzberg, S.L. (2001) A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics* **17**: 1123-1130.

Tamas, I., Klasson, L., Canback, B., Naslund, A.K., Eriksson, A.S., Wernegreen, J.J., Sandstrom, J.P., Moran, N.A., and Andersson, S.G. (2002) 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**: 2376-2379.

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**: 22-28.

Teleman, A.A., Graumann, P.L., Lin, D.C., Grossman, A.D., and Losick, R. (1998) Chromosome arrangement within a bacterium. *Curr Biol* **8**: 1102-1109.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., and Fraser, C.M. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**: 13950-13955.

Tillier, E.R., and Collins, R.A. (2000a) Genome rearrangement by replication-directed translocation. *Nat Genet* **26**: 195-197.

Tillier, E.R., and Collins, R.A. (2000b) The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J Mol Evol* **50**: 249-257.

Toh, H., Weiss, B.L., Perkin, S.A., Yamashita, A., Oshima, K., Hattori, M., and Aksoy, S. (2006) Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res* **16**: 149-156.

Tourand, Y., Kobryn, K., and Chaconas, G. (2003) Sequence-specific recognition but position-dependent cleavage of two distinct telomeres by the *Borrelia burgdorferi* telomere resolvase, ResT. *Mol Microbiol* **48**: 901-911.

Troost, T.R., Danilenko, V.N., and Lomovskaya, N.D. (1979) Fertility properties and regulation of antimicrobial substance production by plasmid SCP2 of *Streptomyces coelicolor*. *J Bacteriol* **140**: 359-368.

Uchida, T., Miyawaki, M., and Kinashi, H. (2003) Chromosomal arm replacement in *Streptomyces griseus*. *J Bacteriol* **185**: 1120-1124.

Ussery, D.W., and Hallin, P.F. (2004a) Genome Update: AT content in sequenced prokaryotic genomes. *Microbiology* **150**: 749-752.

Ussery, D.W., and Hallin, P.F. (2004b) Genome update: Length distributions of sequenced prokaryotic genomes. *Microbiology* **150**: 513-516.

Valens, M., Penaud, S., Rossignol, M., Cornet, F., and Boccard, F. (2004) Macrodomain organization of the *Escherichia coli* chromosome. *Embo J* **23**: 4330-4341.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., and Medigue, C. (2006) MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res* **34**: 53-65.

Vandamme, E.J., and Derycke, D.G. (1983) Microbial inulinases: fermentation process, properties, and applications. *Adv Appl Microbiol* **29**: 139-176.

Viguera, E., Canceill, D., and Ehrlich, S.D. (2001) Replication slippage involves DNA polymerase pausing and dissociation. *Embo J* **20**: 2587-2595.

Viollier, P.H., and Shapiro, L. (2004) Spatial complexity of mechanisms controlling a bacterial cell cycle. *Curr Opin Microbiol* **7**: 572-578.

Viollier, P.H., Thanbichler, M., McGrath, P.T., West, L., Meewan, M., McAdams, H.H., and Shapiro, L. (2004) Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. *Proc Natl Acad Sci U S A* **101**: 9257-9262.

Volff, J.N., and Altenbuchner, J. (1997) Influence of disruption of the *recA* gene on genetic instability and genome rearrangement in *Streptomyces lividans*. *J Bacteriol* **179**: 2440-2445.

Volff, J.N., Viell, P., and Altenbuchner, J. (1997) Artificial circularization of the chromosome with concomitant deletion of its terminal inverted repeats enhances genetic instability and genome rearrangement in *Streptomyces lividans*. *Mol Gen Genet* **253**: 753-760.

Volff, J.N., and Altenbuchner, J. (2000) A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol Lett* **186**: 143-150.

Wang, K., Boysen, C., Shizuya, H., Simon, M.I., and Hood, L. (1997) Complete nucleotide sequence of two generations of a bacterial artificial chromosome cloning vector. *Biotechniques* **23**: 992-994.

Wang, S.J., Chang, H.M., Lin, Y.S., Huang, C.H., and Chen, C.W. (1999) *Streptomyces* genomes: circular genetic maps from the linear chromosomes. *Microbiology* **145**: 2209-2220.

Waters, E., Hohn, M.J., Ahel, I., Graham, D.E., Adams, M.D., Barnstead, M., Beeson, K.Y., Bibbs, L., Bolanos, R., Keller, M., Kretz, K., Lin, X., Mathur, E., Ni, J., Podar, M., Richardson, T., Sutton, G.G., Simon, M., Soll, D., Stetter, K.O., Short, J.M., and Noordewier, M. (2003) The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. *Proc Natl Acad Sci U S A* **100**: 12984-12988.

Weaver, D., Karoonuthaisiri, N., Tsai, H.H., Huang, C.H., Ho, M.L., Gai, S., Patel, K.G., Huang, J., Cohen, S.N., Hopwood, D.A., Chen, C.W., and Kao, C.M. (2004) Genome plasticity in *Streptomyces*: identification of 1 Mb TIRs in the *S. coelicolor* A3(2) chromosome. *Mol Microbiol* **51**: 1535-1550.

Wei, J., Goldberg, M.B., Burland, V., Venkatesan, M.M., Deng, W., Fournier, G., Mayhew, G.F., Plunkett, G., 3rd, Rose, D.J., Darling, A., Mau, B., Perna, N.T., Payne, S.M., Runyen-Janecky, L.J., Zhou, S., Schwartz, D.C., and Blattner, F.R. (2003) Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect Immun* **71**: 2775-2786.

Welch, R.A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G.F., Rose, D.J., Zhou, S., Schwartz, D.C., Perna, N.T., Mobley, H.L., Donnenberg, M.S., and Blattner, F.R. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A* **99**: 17020-17024.

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**: 1686-1689.

Wenner, T., Roth, V., Decaris, B., and Leblond, P. (2002) Intragenomic and intraspecific polymorphism of the 16S-23S rDNA internally transcribed sequences of *Streptomyces ambofaciens*. *Microbiology* **148**: 633-642.

Wenner, T., Roth, V., Fischer, G., Fourrier, C., Aigle, B., Decaris, B., and Leblond, P. (2003) End-to-end fusion of linear deleted chromosomes initiates a cycle of genome instability in *Streptomyces ambofaciens*. *Mol Microbiol* **50**: 411-425.

White, O., Eisen, J.A., Heidelberg, J.F., Hickey, E.K., Peterson, J.D., Dodson, R.J., Haft, D.H., Gwinn, M.L., Nelson, W.C., Richardson, D.L., Moffat, K.S., Qin, H., Jiang, L., Pamphile, W., Crosby, M., Shen, M., Vamathevan, J.J., Lam, P., McDonald, L., Utterback, T., Zalewski, C., Makarova, K.S., Aravind, L., Daly,

M.J., Minton, K.W., Fleischmann, R.D., Ketchum, K.A., Nelson, K.E., Salzberg, S., Smith, H.O., Venter, J.C., and Fraser, C.M. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286**: 1571-1577.

Willems, H., Jager, C., and Baljer, G. (1998) Physical and genetic map of the obligate intracellular bacterium *Coxiella burnetii*. *J Bacteriol* **180**: 3816-3822.

Wohlleben, W., Hartmann, V., Hillemann, D., Krey, K., Muth, G., Nussbaumer, B., and Pelzer, S. (1994) Transfer and establishment of DNA in *Streptomyces* (a brief review). *Acta Microbiol Immunol Hung* **41**: 381-389.

Worning, P., Jensen, L.J., Hallin, P.F., Staerfeldt, H.H., and Ussery, D.W. (2006) Origin of replication in circular prokaryotic chromosomes. *Environ Microbiol* **8**: 353-361.

Wright, F., and Bibb, M.J. (1992) Codon usage in the G+C-rich *Streptomyces* genome. *Gene* **113**: 55-65.

Wu, G., Culley, D.E., and Zhang, W. (2005) Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. *Microbiology* **151**: 2175-2187.

Xu, Y.J., Zhen, Y.S., and Goldberg, I.H. (1994) C1027 chromophore, a potent new enediyne antitumor antibiotic, induces sequence-specific double-strand DNA cleavage. *Biochemistry* **33**: 5947-5954.

Yamasaki, M., and Kinashi, H. (2004) Two chimeric chromosomes of *Streptomyces coelicolor* A3(2) generated by single crossover of the wild-type chromosome and linear plasmid scp1. *J Bacteriol* **186**: 6553-6559.

Yanai, K., Murakami, T., and Bibb, M. (2006) Amplification of the entire kanamycin biosynthetic gene cluster during empirical strain improvement of *Streptomyces kanamyceticus*. *Proc Natl Acad Sci U S A* **103**: 9661-9666.

Yang, M.C., and Losick, R. (2001) Cytological evidence for association of the ends of the linear chromosome in *Streptomyces coelicolor*. *J Bacteriol* **183**: 5180-5186.

Yap, W.H., Zhang, Z., and Wang, Y. (1999) Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J Bacteriol* **181**: 5201-5209.

Young, M., and Cullum, J. (1987) A plausible mechanism for large-scale chromosomal DNA amplification in streptomycetes. *FEBS Lett* **212**: 10-14.

Zahrt, T.C., and Maloy, S. (1997) Barriers to recombination between closely related bacteria: MutS and RecBCD inhibit recombination between *Salmonella typhimurium* and *Salmonella typhi*. *Proc Natl Acad Sci U S A* **94**: 9786-9791.

Zdobnov, E.M., and Apweiler, R. (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**: 847-848.

ANNEXE

ANNEXE

Liste des 312 génomes d'espèces bactériennes et 26 génomes d'archées totalement séquencés en mai 2006 (www.ebi.ac.uk/genomes/).

phylum	organisme (réplicon)	taille (pb)	accession
BACTERIA			
Acidobacteria	Acidobacteria bacterium Ellin345 (chr 1)	5650368	CP000360
Actinobacteria	Bifidobacterium longum NCC2705 (chr 1)	2256640	AE014295
Actinobacteria	Corynebacterium diphtheriae NCTC 13129 (chr 1)	2488635	BX248353
Actinobacteria	Corynebacterium efficiens YS-314 (chr 1)	3147090	BA000035
Actinobacteria	Corynebacterium glutamicum ATCC 13032 (chr 1)	3309401	BA000036
Actinobacteria	Corynebacterium glutamicum ATCC 13032 IS fingerprint type 4-5 (chr 1)	3282708	BX927147
Actinobacteria	Corynebacterium jeikeium K411 (chr 1)	2462499	CR931997
Actinobacteria	Corynebacterium glutamicum - (chr 1)	3309400	CT009589
Actinobacteria	Frankia sp. CcI3 CcI3 (chr 1)	5433628	CP000249
Actinobacteria	Leifsonia xyli subsp. xyli str. CTCB07 (chr 1)	2584158	AE016822
Actinobacteria	Mycobacterium avium subsp. paratuberculosis str. k10 (chr 1)	4829781	AE016958
Actinobacteria	Mycobacterium bovis subsp. bovis AF2122/97 (chr 1)	4345492	BX248333
Actinobacteria	Mycobacterium leprae TN (chr 1)	3268203	AL450380
Actinobacteria	Mycobacterium tuberculosis CDC1551 (chr 1)	4403837	AE000516
Actinobacteria	Mycobacterium tuberculosis H37Rv (chr 1)	4411532	AL123456
Actinobacteria	Nocardia farcinica IFM 10152 (chr 1)	6021225	AP006618
Actinobacteria	Propionibacterium acnes KPA171202 (chr 1)	2560265	AE017283
Actinobacteria	Streptomyces avermitilis MA-4680 (chr 1)	9025608	BA000030
Actinobacteria	Streptomyces coelicolor A3(2) (chr 1)	8667507	AL645882
Actinobacteria	Symbiobacterium thermophilum 14863 (chr 1)	3566135	AP006840
Actinobacteria	Thermobifida fusca YX (chr 1)	3642249	CP000088
Actinobacteria	Tropheryma whipplei str. Twist (chr 1)	927303	AE014184
Actinobacteria	Tropheryma whipplei TW08/27 (chr 1)	925938	BX072543
Aquificae	Aquifex aeolicus VF5 (chr 1)	1551335	AE000657
Bacteroidetes/Chlorobi	Bacteroides fragilis NCTC 9343 (chr 1)	5205140	CR626927
Bacteroidetes/Chlorobi	Bacteroides fragilis YCH46 (chr 1)	5277274	AP006841
Bacteroidetes/Chlorobi	Bacteroides thetaiotaomicron VPI-5482 (chr 1)	6260361	AE015928
Bacteroidetes/Chlorobi	Chlorobium chlorochromatii CaD3 (chr 1)	2572079	CP000108
Bacteroidetes/Chlorobi	Chlorobium tepidum TLS (chr 1)	2154946	AE006470
Bacteroidetes/Chlorobi	Pelodictyon luteolum DSM 273 (chr 1)	2364842	CP000096
Bacteroidetes/Chlorobi	Porphyromonas gingivalis W83 (chr 1)	2343476	AE015924
Bacteroidetes/Chlorobi	Salinibacter ruber DSM 13855 (chr 1)	3551823	CP000159
Chlamydiae/Verrucomicrobia	Chlamydia muridarum Nigg (chr 1)	1072950	AE002160
Chlamydiae/Verrucomicrobia	Chlamydia trachomatis A/HAR-13 (chr 1)	1044459	CP000051
Chlamydiae/Verrucomicrobia	Chlamydia trachomatis D/UW-3/CX (chr 1)	1042519	AE001273
Chlamydiae/Verrucomicrobia	Chlamydophila abortus strain S26/3 (chr 1)	1144377	CR848038
Chlamydiae/Verrucomicrobia	Chlamydophila caviae GPIC (chr 1)	1173390	AE015925
Chlamydiae/Verrucomicrobia	Chlamydophila felis Fe/C-56 (chr 1)	1166239	AP006861
Chlamydiae/Verrucomicrobia	Chlamydophila pneumoniae AR39 (chr 1)	1229853	AE002161
Chlamydiae/Verrucomicrobia	Chlamydophila pneumoniae CWL029 (chr 1)	1230230	AE001363
Chlamydiae/Verrucomicrobia	Chlamydophila pneumoniae J138 (chr 1)	1226565	BA000008
Chlamydiae/Verrucomicrobia	Chlamydophila pneumoniae TW-183 (chr 1)	1225935	AE009440
Chloroflexi	Dehalococcoides ethenogenes 195 (chr 1)	1469720	CP000027
Chloroflexi	Dehalococcoides sp. CBDB1 CBDB1 (chr 1)	1395502	AJ965256
Cyanobacteria	Anabaena variabilis ATCC 29413 (chr 1)	6365727	CP000117
Cyanobacteria	Gloeobacter violaceus PCC 7421 (chr 1)	4659019	BA000045
Cyanobacteria	Nostoc sp. PCC 7120 DNA 7120 (chr 1)	6413771	BA000019
Cyanobacteria	Prochlorococcus marinus MED4 (chr 1)	1657990	BX548174
Cyanobacteria	Prochlorococcus marinus MIT9313 (chr 1)	2410873	BX548175
Cyanobacteria	Prochlorococcus marinus str. MIT 9312 (chr 1)	1709204	CP000111
Cyanobacteria	Prochlorococcus marinus str. NATL2A (chr 1)	1842899	CP000095
Cyanobacteria	Prochlorococcus marinus subsp. marinus str. CCMP1375 (chr 1)	1751080	AE017126
Cyanobacteria	Synechococcus elongatus 6301 (chr 1)	2696255	AP008231
Cyanobacteria	Synechococcus elongatus 7942 (chr 1)	2695903	CP000100
Cyanobacteria	Synechococcus sp. CC9605 CC9605 (chr 1)	2510659	CP000110

Cyanobacteria	Synechococcus sp. CC9902 CC9902 (chr 1)	2234828	CP000097
Cyanobacteria	Synechococcus sp. JA-2-3B'a(2-13) JA-2-3B'a(2-13) (chr 1)	3046682	CP000240
Cyanobacteria	Synechococcus sp. JA-3-3Ab JA-3-3Ab (chr 1)	2932766	CP000239
Cyanobacteria	Synechococcus sp. WH8102 WH8102 (chr 1)	2434428	BX548020
Cyanobacteria	Synechocystis sp. PCC 6803 DNA PCC 6803 (chr 1)	3573470	BA000022
Cyanobacteria	Thermosynechococcus elongatus BP-1 (chr 1)	2593857	BA000039
Deinococcus-Thermus	Deinococcus geothermalis DSM 11300 (chr 1)	2467205	CP000359
Deinococcus-Thermus	Deinococcus radiodurans R1 (chr 1)	2648638	AE000513
Deinococcus-Thermus	Deinococcus radiodurans R1 (chr 2)	412348	AE001825
Deinococcus-Thermus	Thermus thermophilus HB27 (chr 1)	1894877	AE017221
Deinococcus-Thermus	Thermus thermophilus HB8 (chr 1)	1849742	AP008226
Firmicutes	Aster yellows witches'-broom phytoplasma AYWB (chr 1)	706569	CP000061
Firmicutes	Bacillus anthracis str.'Ames Ancestor' (chr 1)	5227419	AE017334
Firmicutes	Bacillus anthracis str. Ames (chr 1)	5227293	AE016879
Firmicutes	Bacillus anthracis str. Sterne (chr 1)	5228663	AE017225
Firmicutes	Bacillus cereus ATCC 10987 (chr 1)	5224283	AE017194
Firmicutes	Bacillus cereus ATCC 14579 (chr 1)	5411809	AE016877
Firmicutes	Bacillus cereus E33L (chr 1)	5300915	CP000001
Firmicutes	Bacillus clausii KSM-K16 (chr 1)	4303871	AP006627
Firmicutes	Bacillus halodurans C-125 (chr 1)	4202352	BA000004
Firmicutes	Bacillus licheniformis ATCC 14580 (chr 1)	4222334	CP000002
Firmicutes	Bacillus licheniformis DSM 13 (chr 1)	4222645	AE017333
Firmicutes	Bacillus thuringiensis serovar konkukian str. 97-27 (chr 1)	5237682	AE017355
Firmicutes	Bacillus subtilis subsp. subtilis str. 168 (chr 1)	4214630	AL009126
Firmicutes	Carboxydothermus hydrogenoformans Z-2901 (chr 1)	2401520	CP000141
Firmicutes	Clostridium acetobutylicum ATCC 824 (chr 1)	3940880	AE001437
Firmicutes	Clostridium perfringens str. 13 (chr 1)	3031430	BA000016
Firmicutes	Clostridium tetani E88 (chr 1)	2799251	AE015927
Firmicutes	Desulfitobacterium hafniense Y51 (chr 1)	5727534	AP008230
Firmicutes	Enterococcus faecalis V583 (chr 1)	3218031	AE016830
Firmicutes	Geobacillus kaustophilus HTA426 (chr 1)	3544776	BA000043
Firmicutes	Lactobacillus acidophilus NCFM (chr 1)	1993564	CP000033
Firmicutes	Lactobacillus johnsonii NCC 533 (chr 1)	1992676	AE017198
Firmicutes	Lactobacillus plantarum strain WCFS1 (chr 1)	3308274	AL935263
Firmicutes	Lactobacillus sakei strain 23K (chr 1)	1884661	CR936503
Firmicutes	Lactobacillus salivarius subsp. salivarius UCC118 (chr 1)	1827111	CP000233
Firmicutes	Lactococcus lactis subsp. lactis IL1403 (chr 1)	2365589	AE005176
Firmicutes	Listeria innocua Clip11262 (chr 1)	3011208	AL592022
Firmicutes	Listeria monocytogenes str. 4b F2365 (chr 1)	2905187	AE017262
Firmicutes	Listeria monocytogenes strain EGD (chr 1)	2944528	AL591824
Firmicutes	Mesoplasma florum L1 (chr 1)	793224	AE017263
Firmicutes	Moorella thermoacetica ATCC 39073 (chr 1)	2628784	CP000232
Firmicutes	Mycoplasma capricolum subsp. capricolum ATCC 27343 (chr 1)	1010023	CP000123
Firmicutes	Mycoplasma gallisepticum strain R (chr 1)	996422	AE015450
Firmicutes	Mycoplasma genitalium G37 (chr 1)	580076	L43967
Firmicutes	Mycoplasma hyopneumoniae 232 (chr 1)	892758	AE017332
Firmicutes	Mycoplasma hyopneumoniae 7448 (chr 1)	920079	AE017244
Firmicutes	Mycoplasma hyopneumoniae J (chr 1)	897405	AE017243
Firmicutes	Mycoplasma mobile 163K (chr 1)	777079	AE017308
Firmicutes	Mycoplasma mycoides subsp. mycoides SC str. PG1 (chr 1)	1211703	BX293980
Firmicutes	Mycoplasma penetrans HF-2 (chr 1)	1358633	BA000026
Firmicutes	Mycoplasma pneumoniae M129 (chr 1)	816394	U00089
Firmicutes	Mycoplasma pulmonis UAB CTIP (chr 1)	963879	AL445566
Firmicutes	Mycoplasma synoviae 53 (chr 1)	799476	AE017245
Firmicutes	Oceanobacillus iheyensis HTE831 (chr 1)	3630528	BA000028
Firmicutes	Onion yellows phytoplasma OY-M (chr 1)	860631	AP006628
Firmicutes	Staphylococcus aureus RF122 (chr 1)	2742531	AJ938182
Firmicutes	Staphylococcus aureus strain MSSA476 (chr 1)	2799802	BX571857
Firmicutes	Staphylococcus aureus subsp. aureus COL (chr 1)	2809422	CP000046
Firmicutes	Staphylococcus aureus subsp. aureus Mu50 (chr 1)	2878529	BA000017
Firmicutes	Staphylococcus aureus subsp. aureus MW2 strain:MW2 (chr 1)	2820462	BA000033
Firmicutes	Staphylococcus aureus subsp. aureus N315 (chr 1)	2814816	BA000018
Firmicutes	Staphylococcus aureus subsp. aureus NCTC 8325 (chr 1)	2821361	CP000253
Firmicutes	Staphylococcus aureus subsp. aureus strain MRSA252 (chr 1)	2902619	BX571856
Firmicutes	Staphylococcus aureus subsp. aureus USA300 (chr 1)	2872769	CP000255

Firmicutes	Staphylococcus epidermidis ATCC 12228 (chr 1)	2499279	AE015929
Firmicutes	Staphylococcus epidermidis RP62A (chr 1)	2616530	CP000029
Firmicutes	Staphylococcus haemolyticus JCSC1435 (chr 1)	2685015	AP006716
Firmicutes	Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305 (chr 1)	2516575	AP008934
Firmicutes	Streptococcus agalactiae 2603V/R (chr 1)	2160267	AE009948
Firmicutes	Streptococcus agalactiae A909 (chr 1)	2127839	CP000114
Firmicutes	Streptococcus agalactiae NEM316 (chr 1)	2211485	AL732656
Firmicutes	Streptococcus mutans UA159 (chr 1)	2030921	AE014133
Firmicutes	Streptococcus pneumoniae R6 (chr 1)	2038615	AE007317
Firmicutes	Streptococcus pneumoniae TIGR4 (chr 1)	2160842	AE005672
Firmicutes	Streptococcus pyogenes M1 GAS (chr 1)	1852441	AE004092
Firmicutes	Streptococcus pyogenes MGAS10270 (chr 1)	1928252	CP000260
Firmicutes	Streptococcus pyogenes MGAS10394 (chr 1)	1899877	CP000003
Firmicutes	Streptococcus pyogenes MGAS10750 (chr 1)	1937111	CP000262
Firmicutes	Streptococcus pyogenes MGAS2096 (chr 1)	1860355	CP000261
Firmicutes	Streptococcus pyogenes MGAS315 (chr 1)	1900521	AE014074
Firmicutes	Streptococcus pyogenes MGAS5005 (chr 1)	1838554	CP000017
Firmicutes	Streptococcus pyogenes MGAS6180 (chr 1)	1897573	CP000056
Firmicutes	Streptococcus pyogenes MGAS9429 (chr 1)	1836467	CP000259
Firmicutes	Streptococcus pyogenes SSI-1 (chr 1)	1894275	BA000034
Firmicutes	Streptococcus pyogenes strain MGAS8232 (chr 1)	1895017	AE009949
Firmicutes	Streptococcus thermophilus CNRZ1066 (chr 1)	1796226	CP000024
Firmicutes	Streptococcus thermophilus LMG 18311 (chr 1)	1796846	CP000023
Firmicutes	Thermoanaerobacter tengcongensis strain MB4 (chr 1)	2689445	AE008691
Firmicutes	Ureaplasma parvum serovar 3 str. ATCC 700970 (chr 1)	751719	AF222894
Fusobacteria	Fusobacterium nucleatum subsp. nucleatum ATCC 25586 (chr 1)	2174500	AE009951
Planctomycetes	Pirellula sp. strain 1 strain 1 (chr 1)	7145576	BX119912
Proteobacteria	Acinetobacter sp. ADP1 ADP1 (chr 1)	3598621	CR543861
Proteobacteria	Agrobacterium tumefaciens str. C58 (chr 1)	2841490	AE008688
Proteobacteria	Agrobacterium tumefaciens str. C58 (chr 1)	2841581	AE007869
Proteobacteria	Agrobacterium tumefaciens str. C58 (chr 2)	2074782	AE007870
Proteobacteria	Agrobacterium tumefaciens str. C58 (chr 2)	2075560	AE008689
Proteobacteria	Anaeromyxobacter dehalogenans 2CP-C (chr 1)	5013479	CP000251
Proteobacteria	Anaplasma marginale str. St. Maries (chr 1)	1197687	CP000030
Proteobacteria	Anaplasma phagocytophilum HZ (chr 1)	1471282	CP000235
Proteobacteria	Azoarcus sp. EbN1 EbN1 (chr 1)	4296230	CR555306
Proteobacteria	Bartonella henselae strain Houston-1 (chr 1)	1931047	BX897699
Proteobacteria	Bartonella quintana str. Toulouse (chr 1)	1581384	BX897700
Proteobacteria	Baumannia cicadellinicola str. Hc (Homalodisca coagulata) (chr 1)	686194	CP000238
Proteobacteria	Bdellovibrio bacteriovorus strain HD100 (chr 1)	3782950	BX842601
Proteobacteria	Bordetella avium 197N (chr 1)	3732255	AM167904
Proteobacteria	Bordetella bronchiseptica strain RB50 (chr 1)	5339179	BX470250
Proteobacteria	Bordetella parapertussis strain 12822 (chr 1)	4773551	BX470249
Proteobacteria	Bordetella pertussis strain Tohama (chr 1)	4086189	BX470248
Proteobacteria	Bradyrhizobium japonicum USDA 110 (chr 1)	9105828	BA000040
Proteobacteria	Brucella abortus biovar 1 str. 9-941 (chr 1)	2124241	AE017223
Proteobacteria	Brucella abortus biovar 1 str. 9-941 (chr 2)	1162204	AE017224
Proteobacteria	Brucella melitensis 16M (chr 1)	2117144	AE008917
Proteobacteria	Brucella melitensis 16M (chr 2)	1177787	AE008918
Proteobacteria	Brucella melitensis biovar Abortus strain 2308 (chr 1)	2121359	AM040264
Proteobacteria	Brucella melitensis biovar Abortus strain 2308 (chr 2)	1156948	AM040265
Proteobacteria	Brucella suis 1330 (chr 1)	2107794	AE014291
Proteobacteria	Brucella suis 1330 (chr 2)	1207381	AE014292
Proteobacteria	Buchnera aphidicola str. APS (Acyrthosiphon pisum) (chr 1)	640681	BA000003
Proteobacteria	Buchnera aphidicola str. Bp (Baizongia pistaciae) (chr 1)	615980	AE016826
Proteobacteria	Buchnera aphidicola str. Sg (Schizaphis graminum) (chr 1)	641454	AE013218
Proteobacteria	Burkholderia mallei ATCC 23344 (chr 1)	3510148	CP000010
Proteobacteria	Burkholderia mallei ATCC 23344 (chr 2)	2325379	CP000011
Proteobacteria	Burkholderia pseudomallei 1710b (chr 1)	4126292	CP000124
Proteobacteria	Burkholderia pseudomallei 1710b (chr 2)	3181762	CP000125
Proteobacteria	Burkholderia pseudomallei strain K96243 (chr 1)	4074542	BX571965
Proteobacteria	Burkholderia pseudomallei strain K96243 (chr 2)	3173005	BX571966
Proteobacteria	Burkholderia thailandensis E264 (chr 1)	3809201	CP000086
Proteobacteria	Burkholderia thailandensis E264 (chr 2)	2914771	CP000085
Proteobacteria	Burkholderia sp. 383 383 (chr 1)	3694126	CP000151

Proteobacteria	Burkholderia sp. 383 383 (chr 2)	3587082	CP000152
Proteobacteria	Burkholderia sp. 383 383 (chr 3)	1395069	CP000150
Proteobacteria	Campylobacter jejuni RM1221 (chr 1)	1777831	CP000025
Proteobacteria	Campylobacter jejuni subsp. jejuni NCTC 11168 (chr 1)	1641481	AL111168
Proteobacteria	Candidatus Blochmannia str. BPEN (chr 1)	791654	CP000016
Proteobacteria	Candidatus Pelagibacter HTCC1062 (chr 1)	1308759	CP000084
Proteobacteria	Caulobacter crescentus CB15 (chr 1)	4016947	AE005673
Proteobacteria	Chromobacterium violaceum ATCC 12472 (chr 1)	4751080	AE016825
Proteobacteria	Chromohalobacter salexigens DSM 3043 (chr 1)	3696649	CP000285
Proteobacteria	Colwellia psychrerythraea 34H (chr 1)	5373180	CP000083
Proteobacteria	Coxiella burnetii RSA 493 (chr 1)	1995281	AE016828
Proteobacteria	Dechloromonas aromatica RCB (chr 1)	4501104	CP000089
Proteobacteria	Desulfotalea psychrophila LSv54 (chr 1)	3523383	CR522870
Proteobacteria	Desulfovibrio desulfuricans G20 (chr 1)	3730232	CP000112
Proteobacteria	Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough (chr 1)	3570858	AE017285
Proteobacteria	Ehrlichia canis str. Jake (chr 1)	1315030	CP000107
Proteobacteria	Ehrlichia chaffeensis str. Arkansas (chr 1)	1176248	CP000236
Proteobacteria	Ehrlichia ruminantium strain Welgevonden (chr 1)	1516355	CR767821
Proteobacteria	Ehrlichia ruminantium str. Gardel (chr 1)	1499920	CR925677
Proteobacteria	Ehrlichia ruminantium str. Welgevonden (chr 1)	1512977	CR925678
Proteobacteria	Erwinia carotovora subsp. atroseptica SCRI1043 (chr 1)	5064019	BX950851
Proteobacteria	Erythrobacter litoralis HTCC2594 (chr 1)	3052398	CP000157
Proteobacteria	Escherichia coli CFT073 (chr 1)	5231428	AE014075
Proteobacteria	Escherichia coli K-12 MG1655 (chr 1)	4639675	U00096
Proteobacteria	Escherichia coli O157:H7 (chr 1)	5498450	BA000007
Proteobacteria	Escherichia coli O157:H7 EDL933 (chr 1)	5528445	AE005174
Proteobacteria	Escherichia coli UTI89 (chr 1)	5065741	CP000243
Proteobacteria	Escherichia coli W3110 (chr 1)	4646332	AP009048
Proteobacteria	Francisella tularensis subsp. holarctica strain LVS (chr 1)	1895994	AM233362
Proteobacteria	Francisella tularensis subsp. tularensis SCHU S4 (chr 1)	1892819	AJ749949
Proteobacteria	Geobacter metallireducens GS-15 (chr 1)	3997420	CP000148
Proteobacteria	Geobacter sulfurreducens PCA (chr 1)	3814139	AE017180
Proteobacteria	Gluconobacter oxydans 621H (chr 1)	2702173	CP000009
Proteobacteria	Haemophilus ducreyi strain 35000HP (chr 1)	1698955	AE017143
Proteobacteria	Haemophilus influenzae Rd KW20 (chr 1)	1830138	L42023
Proteobacteria	Haemophilus influenzae strain 86-028NP (chr 1)	1913428	CP000057
Proteobacteria	Hahella chejuensis KCTC 2396 (chr 1)	7215267	CP000155
Proteobacteria	Helicobacter hepaticus ATCC 51449 (chr 1)	1799146	AE017125
Proteobacteria	Helicobacter pylori 26695 (chr 1)	1667867	AE000511
Proteobacteria	Helicobacter pylori J99 (chr 1)	1643831	AE001439
Proteobacteria	Idiomarina loihiensis L2TR (chr 1)	2839318	AE017340
Proteobacteria	Jannaschia sp. CCS1 CCS1 (chr 1)	4317977	CP000264
Proteobacteria	Lawsonia intracellularis PHE/MN1-00 (chr 1)	1457619	AM180252
Proteobacteria	Legionella pneumophila str. Lens (chr 1)	3345687	CR628337
Proteobacteria	Legionella pneumophila str. Paris (chr 1)	3503610	CR628336
Proteobacteria	Legionella pneumophila subsp. pneumophila str. Philadelphia 1 (chr 1)	3397754	AE017354
Proteobacteria	Magnetospirillum magneticum AMB-1 (chr 1)	4967148	AP007255
Proteobacteria	Mannheimia succiniciproducens MBEL55E (chr 1)	2314078	AE016827
Proteobacteria	Mesorhizobium loti MAFF303099 (chr 1)	7036071	BA000012
Proteobacteria	Methylococcus capsulatus str. Bath (chr 1)	3304561	AE017282
Proteobacteria	Neisseria gonorrhoeae FA 1090 (chr 1)	2153922	AE004969
Proteobacteria	Neisseria meningitidis MC58 (chr 1)	2272360	AE002098
Proteobacteria	Neisseria meningitidis serogroup A strain Z2491 (chr 1)	2184406	AL157959
Proteobacteria	Neorickettsia sennetsu strain Miyayama (chr 1)	859006	CP000237
Proteobacteria	Nitrobacter hamburgensis X14 (chr 1)	4406967	CP000319
Proteobacteria	Nitrobacter winogradskyi Nb-255 (chr 1)	3402093	CP000115
Proteobacteria	Nitrosococcus oceani ATCC 19707 (chr 1)	3481691	CP000127
Proteobacteria	Nitrosomonas europaea ATCC 19718 (chr 1)	2812094	AL954747
Proteobacteria	Nitrosospira multiformis ATCC 25196 (chr 1)	3184243	CP000103
Proteobacteria	Novosphingobium aromaticivorans DSM 12444 (chr 1)	3561584	CP000248
Proteobacteria	Pasteurella multocida subsp. multocida str. Pm70 (chr 1)	2257487	AE004439
Proteobacteria	Pelobacter carbinolicus DSM 2380 (chr 1)	3665893	CP000142
Proteobacteria	Photobacterium profundum SS9 (chr 1)	4085304	CR354531
Proteobacteria	Photobacterium profundum SS9 (chr 2)	2237943	CR354532
Proteobacteria	Photorhabdus luminescens subsp. laumondii TTO1 (chr 1)	5688987	BX470251

Proteobacteria	Pseudoalteromonas haloplanktis str. TAC125 (chr 1)	3214944	CR954246
Proteobacteria	Pseudoalteromonas haloplanktis str. TAC125 (chr 2)	635328	CR954247
Proteobacteria	Pseudomonas aeruginosa PAO1 (chr 1)	6264403	AE004091
Proteobacteria	Pseudomonas fluorescens Pf-5 (chr 1)	7074893	CP000076
Proteobacteria	Pseudomonas fluorescens PfO-1 (chr 1)	6438405	CP000094
Proteobacteria	Pseudomonas putida KT2440 (chr 1)	6181863	AE015451
Proteobacteria	Pseudomonas syringae pv. phaseolicola 1448A (chr 1)	5928787	CP000058
Proteobacteria	Pseudomonas syringae pv. syringae B728a (chr 1)	6093698	CP000075
Proteobacteria	Pseudomonas syringae pv. tomato str. DC3000 (chr 1)	6397126	AE016853
Proteobacteria	Psychrobacter arcticus 273-4 (chr 1)	2650701	CP000082
Proteobacteria	Psychrobacter cryohalolentis K5 (chr 1)	3059876	CP000323
Proteobacteria	Ralstonia eutropha JMP134 (chr 1)	3806533	CP000090
Proteobacteria	Ralstonia eutropha JMP134 (chr 2)	2726152	CP000091
Proteobacteria	Ralstonia metallidurans CH34 (chr 1)	3928089	CP000352
Proteobacteria	Ralstonia solanacearum GMI1000 (chr 1)	3716413	AL646052
Proteobacteria	Rhizobium etli CFN 42 (chr 1)	4381608	CP000133
Proteobacteria	Rhizobium leguminosarum bv. viciae 3841 strain 3841 (chr 1)	5057142	AM236080
Proteobacteria	Rhodobacter sphaeroides 2.4.1 (chr 1)	3188609	CP000143
Proteobacteria	Rhodobacter sphaeroides 2.4.1 (chr 2)	943016	CP000144
Proteobacteria	Rhodoferax ferrireducens DSM 15236 (chr 1)	4712337	CP000267
Proteobacteria	Rhodopseudomonas palustris BisB18 (chr 1)	5513844	CP000301
Proteobacteria	Rhodopseudomonas palustris CGA009 (chr 1)	5459213	BX571963
Proteobacteria	Rhodopseudomonas palustris HaA2 (chr 1)	5331656	CP000250
Proteobacteria	Rhodospirillum rubrum ATCC 11170 (chr 1)	4352825	CP000230
Proteobacteria	Rickettsia bellii RML369-C (chr 1)	1522076	CP000087
Proteobacteria	Rickettsia conorii str. Malish 7 (chr 1)	1268755	AE006914
Proteobacteria	Rickettsia felis URRWXCal2 (chr 1)	1485148	CP000053
Proteobacteria	Rickettsia prowazekii str. Madrid E (chr 1)	1111523	AJ235269
Proteobacteria	Rickettsia typhi str. Wilmington (chr 1)	1111496	AE017197
Proteobacteria	Saccharophagus degradans 2-40 (chr 1)	5057531	CP000282
Proteobacteria	Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67 (chr 1)	4755700	AE017220
Proteobacteria	Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150 (chr 1)	4585229	CP000026
Proteobacteria	Salmonella enterica subsp. enterica serovar Typhi str. CT18 (chr 1)	4809037	AL513382
Proteobacteria	Salmonella enterica subsp. enterica serovar Typhi Ty2 (chr 1)	4791961	AE014613
Proteobacteria	Salmonella typhimurium LT2 (chr 1)	4857432	AE006468
Proteobacteria	Shewanella oneidensis MR-1 (chr 1)	4969803	AE014299
Proteobacteria	Shigella boydii Sb227 (chr 1)	4519823	CP000036
Proteobacteria	Shigella dysenteriae Sd197 (chr 1)	4369232	CP000034
Proteobacteria	Shigella flexneri 2a str. 2457T (chr 1)	4599354	AE014073
Proteobacteria	Shigella flexneri 2a str. 301 (chr 1)	4607203	AE005674
Proteobacteria	Shigella sonnei Ss046 (chr 1)	4825265	CP000038
Proteobacteria	Silicibacter pomeroyi DSS-3 (chr 1)	4109442	CP000031
Proteobacteria	Sinorhizobium meliloti 1021 (chr 1)	3654135	AL591688
Proteobacteria	Sodalis glossinidius str 'morsitans' (chr 1)	4171146	AP008232
Proteobacteria	Syntrophus aciditrophicus SB (chr 1)	3179300	CP000252
Proteobacteria	Thiobacillus denitrificans ATCC 25259 (chr 1)	2909809	CP000116
Proteobacteria	Thiomicrospira crunogena XCL-2 (chr 1)	2427734	CP000109
Proteobacteria	Thiomicrospira denitrificans ATCC 33889 (chr 1)	2201561	CP000153
Proteobacteria	Vibrio cholerae O1 biovar eltor str. N16961 (chr 1)	2961149	AE003852
Proteobacteria	Vibrio cholerae O1 biovar eltor str. N16961 (chr 2)	1072315	AE003853
Proteobacteria	Vibrio fischeri ES114 (chr 1)	2906179	CP000020
Proteobacteria	Vibrio fischeri ES114 (chr 2)	1332022	CP000021
Proteobacteria	Vibrio parahaemolyticus RIMD 2210633 (chr 1)	3288558	BA000031
Proteobacteria	Vibrio parahaemolyticus RIMD 2210633 (chr 2)	1877212	BA000032
Proteobacteria	Vibrio vulnificus CMCP6 I (chr 1)	3281944	AE016795
Proteobacteria	Vibrio vulnificus CMCP6 (chr 2)	1844853	AE016796
Proteobacteria	Vibrio vulnificus YJ016 (chr 1)	3354505	BA000037
Proteobacteria	Vibrio vulnificus YJ016 (chr 2)	1857073	BA000038
Proteobacteria	Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis (chr 1)	697724	BA000021
Proteobacteria	Wolbachia endosymbiont endosymbiont of Drosophila melanogaster (chr 1)	1267782	AE017196
Proteobacteria	Wolbachia endosymbiont endosymbiont strain TRS of Brugia malayi (chr 1)	1080084	AE017321
Proteobacteria	Wolinella succinogenes DSM 1740 (chr 1)	2110355	BX571656
Proteobacteria	Xanthomonas axonopodis pv. citri str. 306 (chr 1)	5175554	AE008923
Proteobacteria	Xanthomonas campestris pv. campestris str. 8004 (chr 1)	5148708	CP000050
Proteobacteria	Xanthomonas campestris pv. campestris str. ATCC 33913 (chr 1)	5076188	AE008922

Proteobacteria	Xanthomonas campestris pv. vesicatoria (chr 1)	5178466	AM039952
Proteobacteria	Xanthomonas oryzae pv. oryzae KACC10331 (chr 1)	4941439	AE013598
Proteobacteria	Xanthomonas oryzae pv. oryzae MAFF 311018 (chr 1)	4940217	AP008229
Proteobacteria	Xylella fastidiosa 9a5c (chr 1)	2679306	AE003849
Proteobacteria	Xylella fastidiosa Temecula1 (chr 1)	2519802	AE009442
Proteobacteria	Yersinia pestis biovar Medievalis str. 91001 (chr 1)	4595065	AE017042
Proteobacteria	Yersinia pestis KIM (chr 1)	4600755	AE009952
Proteobacteria	Yersinia pestis strain CO92 (chr 1)	4653728	AL590842
Proteobacteria	Yersinia pseudotuberculosis IP32953 genome (chr 1)	4744671	BX936398
Proteobacteria	Zymomonas mobilis subsp. mobilis ZM4 (chr 1)	2056416	AE008692
Spirochaetes	Borrelia burgdorferi B31 (chr 1)	910724	AE000783
Spirochaetes	Borrelia garinii PBi (chr 1)	904246	CP000013
Spirochaetes	Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130 (chr 1)	4277185	AE016823
Spirochaetes	Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130 (chr 2)	350181	AE016824
Spirochaetes	Leptospira interrogans serovar lai str. 56601 (chr 1)	4332241	AE010300
Spirochaetes	Leptospira interrogans serovar Lai str. 56601 (chr 2)	358943	AE010301
Spirochaetes	Treponema denticola ATCC 35405 (chr 1)	2843201	AE017226
Spirochaetes	Treponema pallidum subsp. pallidum str. Nichols (chr 1)	1138011	AE000520
Thermotogae	Thermotoga maritima MSB8 (chr 1)	1860725	AE000512

ARCHAEA

Crenarchaeota	Aeropyrum pernix K1 (chr 1)	1669695	BA000002
Crenarchaeota	Pyrobaculum aerophilum strain IM2 (chr 1)	2222430	AE009441
Crenarchaeota	Sulfolobus acidocaldarius DSM 639 (chr 1)	2225959	CP000077
Crenarchaeota	Sulfolobus tokodaii str. 7 (chr 1)	2694756	BA000023
Crenarchaeota	Sulfolobus solfataricus - (chr 1)	2992245	AE006641
Euryarchaeota	Archaeoglobus fulgidus DSM 4304 (chr 1)	2178400	AE000782
Euryarchaeota	Haloarcula marismortui ATCC 43049 (chr 1)	3131724	AY596297
Euryarchaeota	Haloarcula marismortui ATCC 43049 (chr 2)	288050	AY596298
Euryarchaeota	Halobacterium sp. NRC-1 NRC-1; ATCC 700922 (chr 1)	2014239	AE004437
Euryarchaeota	Methanocaldococcus jannaschii DSM 2661 (chr 1)	1664970	L77117
Euryarchaeota	Methanococcus maripaludis strain S2 (chr 1)	1661137	BX950229
Euryarchaeota	Methanopyrus kandleri AV19 (chr 1)	1694969	AE009439
Euryarchaeota	Methanosarcina acetivorans str. C2A (chr 1)	5751492	AE010299
Euryarchaeota	Methanosarcina barkeri str. fusaro (chr 1)	4837408	CP000099
Euryarchaeota	Methanosarcina mazei strain Goe1 (chr 1)	4096345	AE008384
Euryarchaeota	Methanosphaera stadtmanae DSM 3091 (chr 1)	1767403	CP000102
Euryarchaeota	Methanospirillum hungatei JF-1 (chr 1)	3544738	CP000254
Euryarchaeota	Methanothermobacter thermautotrophicus str. Delta H (chr 1)	1751377	AE000666
Euryarchaeota	Natronomonas pharaonis DSM 2160 (chr 1)	2595221	CR936257
Euryarchaeota	Picrophilus torridus DSM 9790 (chr 1)	1545895	AE017261
Euryarchaeota	Pyrococcus furiosus DSM 3638 (chr 1)	1908256	AE009950
Euryarchaeota	Pyrococcus horikoshii OT3 (chr 1)	1738505	BA000001
Euryarchaeota	Pyrococcus abyssi GE5 (chr 1)	1765118	AL096836
Euryarchaeota	Thermococcus kodakarensis KOD1 (chr 1)	2088737	AP006878
Euryarchaeota	Thermoplasma acidophilum DSM 1728 (chr 1)	1564906	AL139299
Euryarchaeota	Thermoplasma volcanium GSS1 (chr 1)	1584804	BA000011
Nanoarchaeota	Nanoarchaeum equitans Kin4-M (chr 1)	490885	AE017199

Monsieur **CHOULET Frédéric**

DOCTORAT DE L'UNIVERSITE HENRI POINCARE, NANCY 1

en GENETIQUE MOLECULAIRE

VU, APPROUVÉ ET PERMIS D'IMPRIMER N° 1894

Nancy, le 23/11/06

Le Président de l'Université

UNIVERSITÉ HENRI POINCARÉ, NANCY

J.P. FINANCE

Université Henri Poincaré, Nancy I
24-30 rue Lionnois - B.P. 3069 - 54013 NANCY Cédex
Tél. : 03 83 68 20 00 - Fax : 03 83 68 21 00

Résumé

L'analyse et la comparaison des génomes permettent d'appréhender les forces évolutives à la base de leur dynamique et les contraintes qui s'opposent à la variabilité. Chez les bactéries, le transfert horizontal joue un rôle majeur dans leur diversification. Ce travail traite de l'évolution du génome des *Streptomyces* qui sont des bactéries du sol responsables de la biosynthèse d'une très grande diversité de composés actifs. Leur intérêt applicatif est considérable et touche des domaines variés (médecine, agroalimentaire, biotechnologies). Leurs caractéristiques génomiques sont tout à fait originales : leur ADN chromosomique est linéaire, de taille importante (8-11 Mb) et présente une composition extrême en bases G+C (70-73%). Des phénomènes d'instabilité chromosomique spectaculaires affectent les régions terminales. Chez *Streptomyces ambofaciens*, ces régions terminales, qui représentent un quart du génome, sont délétables en conditions de laboratoire.

Le séquençage partiel et l'annotation du chromosome de *S. ambofaciens* ont été entrepris afin d'étudier l'organisation génétique et les forces évolutives modelant son génome. La génomique comparée de *S. ambofaciens* avec les génomes de trois autres espèces de *Streptomyces* (*S. coelicolor, S. avermitilis* et *S. scabies*), confère une vision dynamique de l'évolution. Elle a révélé une compartimentation génomique marquée où la variabilité est concentrée dans les régions terminales (10% à 20% du génome selon les espèces comparées). La région centrale du chromosome est, quant à elle, fortement synténique et contraste avec la dynamique des régions terminales.
Des échanges d'extrémités de réplicons linéaires, observés par comparaison au niveau intraspécifique, seraient responsables d'une partie de la diversification des régions terminales, établissant ainsi un lien entre linéarité chromosomique et variabilité. La taille des régions terminales spécifiques augmente avec la distance phylogénétique qui sépare les espèces comparées. De plus, entre la région centrale très conservée et les extrémités spécifiques, la synténie dégénère de façon progressive vers les extrémités par fixation d'une multitude d'événements d'insertion/délétion de gènes. Les régions terminales sont donc des loci privilégiés d'acquisition de gènes par transfert horizontal.

Ces résultats suggèrent une organisation génétique et un mode d'évolution innovants. En effet, la fréquence des flux de gènes suivrait un gradient d'intensité croissante vers les extrémités chromosomiques. Le taux d'apparition de cassures double-brin pourrait augmenter dans les régions terminales (cassures liées à l'arrêt de la fourche de réplication ou au mécanisme de conjugaison bactérienne). Alternativement, ces cassures pourraient être prises en charge par des mécanismes différents selon la localisation chromosomique.

Ce mécanisme d'évolution par flux de gènes d'intensité croissante vers les extrémités serait le moteur de la diversification des *Streptomyces* et jouerait un rôle majeur dans l'adaptation à l'écosystème sol. Ces régions chromosomiques sont par exemple enrichies en gènes d'adaptation comme ceux impliqués dans la biosynthèse de métabolites secondaires, des gènes de résistance et d'autres codant des enzymes de dégradation de polymères complexes.

Mots clés : *Streptomyces*, évolution, chromosome linéaire, plasticité, génomique comparative.