LES CRISPRS, INSTRUMENTS D'ETUDE DE L'EVOLUTION INTER ET INTRA-SPÉCIFIQUE CHEZ LES MICROORGANISMES

Ibtissem Grissa

UNIVERSITÉ PARIS XI
UFR SCIENTIFIQUE D'ORSAY

THÈSE

Présentée pour obtenir

Le GRADE de DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS XI ORSAY
Mention BIOINFORMATIQUE

PAR

Ibtissem GRISSA

Titre de la thèse :

LES CRISPRS, INSTRUMENTS D'ETUDE DE L'EVOLUTION INTER ET INTRA-SPÉCIFIQUE CHEZ LES MICROORGANISMES

Soutenue le 24 Juin 2008, devant la commission d'examen :

M.	Patrick	FORTERRE	Président des jury
M.	Richard	CHRISTEN	Rapporteur
M.	Alexandre	LECLERCQ	Rapporteur
Mme.	Christine	POURCEL	Examinatrice
M.	Philippe	HORVATH	Examinateur
M.	Gilles	VERGNAUD	Directeur de thèse
M.	Bernard	LABEDAN	Invité

J'essaie toujours de faire ce que je ne sais pas faire,
c'est ainsi que j'espère apprendre à le faire.
- Pablo Picasso

A Inès...

Liste des abréviations

ADN : acide désoxyribonucléique
AFLP : amplified fragment length polymorphism
ARN : acide ribonucléique
ARNm : ARN messager
BLAST : basic local alignment search tool
CASS : CRISPR et gènes associés
CNR : Centre National de Référence
CRISPR : clustered regularly interspaced short palindromic repeat
CRT : CRISPR Recognition Tool
DR : direct repeat
FDA : Food and Drug Administration
IS : insertion sequence
JGI : Joint Genome Institute
MLST : multilocus sequence typing
MLVA : multilocus VNTR analysis
NCBI : national center for biotechnology information
ORF : open reading Frame
pb : paire de base
PCR : polymerase chain reaction
RISC : RNA-Induced Silencing Complex
RAPD : random amplified polymorphic DNA
RFLP : restriction fragment length polymorphism
SNP : single nucleotide polymorphism
SSR : short sequence repeat
TIGR : the institute for genomic research
TRF : tandem repeats finder
VNTR : variable number of tandem repeats

Remerciements

Je tiens tout d'abord à remercier Monsieur Richard Christen et Monsieur Alexandre Leclercq d'avoir accepté d'être rapporteurs de ce manuscrit. Leurs remarques m'ont permis de le perfectionner et de donner une vision plus claire et plus synthétique du travail que j'ai réalisé durant ces années de recherches.

Je remercie M. Patrick Forterre et M. Philippe Horvath de m'avoir fait l'honneur de juger ce travail ainsi que M. Bernard Labedan, d'avoir accepté de faire partie de ce jury de thèse et d'avoir suivi l'avancement de mon travail pendant trois ans.

Je remercie Gilles Vergnaud, de m'avoir accordé sa confiance, d'avoir accepté de diriger ma thèse et de m'avoir aidé dans le développement des outils bioinformatiques.

Un grand merci à Christine Pourcel, de m'avoir encadrée et soutenu pendant ma thèse. Grâce à sa grande connaissance en biologie, elle m'a offert des avis éclairés et l'opportunité de m'exprimer dans ce domaine. Sa grande pédagogie et sa rigueur ont contribué à améliorer ce manuscrit pour le rendre plus clair et complet. Sa disponibilité m'a toujours permis d'obtenir des conseils avisés et de précieuses réponses. Je la remercie d'avoir accepté de faire partie de ce jury de thèse.

J'adresse en outre mes remerciements à tous ceux qui ont collaboré de près ou de loin à la réalisation de ce travail : Olivier Lespinet et Marie-Anne Poursat, mes parrains de thèse pour leurs suivi et conseils ; Marie Ragon du CNR des *Listeria* à l'Institut Pasteur de Paris, pour les échanges fructueux sur le projet *Listeria* ; Philippe Bouloc et Yanura Noomane pour avoir testé les fonctionnalités du site web. Merci à tous ceux que j'ai côtoyés à l'IGM ou à l'université Paris XI : Marie-Christine pour sa serviabilité et son affabilité, Mazyar, Yolande, Samina, Gaëlle, Gabriela, Ibtissame et en particulier Jinghua pour les échanges culturels ainsi que ma soeur Intissar pour les quatre mois agréables qu'elle a passés dans notre équipe.

Merci à Giuseppe Baldacci et Marie-Josée Daboussi de l'école doctorale "Gènes, Génomes, Cellules" sans qui cette thèse n'aurait pas été possible dans ce laboratoire.

Merci à mes amis qui m'ont soutenue durant cette thèse : Manel, Dorsaf, Riham, Khaled, Naima et surtout Sourour pour la lecture du manuscrit.

J'adresse ma profonde reconnaissance à ma famille, en commençant par ma chère Tata Souad, mes beaux parents, et Ahmed pour leur soutien continu, Ramy pour l'impression des articles, mais surtout mes frères Mohamed et Amine et mes soeurs Imen, Intissar, Afef et Fatma Ezzahra pour leur confiance et leur appui moral.

Je souhaite également faire part de toute ma gratitude à mon père qui voyait en moi un grand chercheur depuis ma tendre enfance et à ma mère qui m'a appris à être patiente et persévérante. Je ne sais pas si je peux les récompenser de leur attente par ces quelques mots de remerciement. Et enfin, merci à Nader pour les conseils techniques en informatique, les leçons de Linux, la recherche de coquilles dans le manuscrit, et surtout pour sa patience, son soutien et son humour en toute circonstance. Le dernier mot est adressé à ma fille Inès pour avoir rendu agréable la période de rédaction de ce rapport.

Table des matières

Table des figures

Liste des tableaux

Avant propos

Le domaine des sciences de la vie est très pluridisciplinaire et la recherche dans ce domaine requiert le concours de scientifiques de diverses compétences tels que les biologistes naturellement, mais aussi les mathématiciens, les chimistes ou les physiciens. Le monde du vivant est tellement vaste à explorer que nous n'en sommes qu'aux prémices des découvertes potentielles que cette science peut révéler.

Avec les nouveaux progrès faits tant au niveau matériel que des méthodes d'analyse, les évolutions technologiques, couplées à une diminution du coût de leur mise en oeuvre, ont conduit à une production en masse des données dont les volumes ne cessent de croître, et à un traitement à haut débit du matériel biologique. La science de la vie ou communément appelée biologie, est devenue alors un domaine qui brasse énormément de données. Il s'agit de la biologie à grande échelle dont les données sont de nos jours difficilement exploitables telles qu'elles et requièrent des outils logiciels toujours plus performants et toujours plus puissants pour permettre des analyses rigoureuses. Ceci sous-entend d'avoir accès aux informations utiles et de réaliser des traitements sur les données.

La bioinformatique, discipline née dans les années 1980, parallèlement à la création des premières banques de séquences d'ADN, est devenue alors incontournable dans le paysage scientifique. Conçue à l'origine pour le stockage et la gestion des données sur ordinateur, elle est aujourd'hui une composante à part entière de la recherche. Les applications développées fournissent aux scientifiques des outils "*in silico*" non seulement pour l'interprétation de leurs résultats (assemblage de séquences, annotation de génomes, modélisation d'interactions moléculaires, etc), mais en plus pour l'exploration et la découverte des secrets cachés dans les séquences génomiques, ouvrant ainsi la voie à la biologie "*in vivo*" pour les valider. Ainsi, de nombreuses études expérimentales, actuellement menées, découlent de découvertes issues d'analyses bioinformatiques dont les résultats sont dits "putatifs" (potentiels) car pas encore vérifiés biologiquement. Le cycle de recherche aujourd'hui nécessite une vraie interaction entre l'analyse "*in silico*" et l'analyse "*in vivo*". La première intervient dans la transformation des données brutes en connaissances. Elle propose de nouvelles hypothèses de travail issues des prédictions. La deuxième permet de tester ces propositions par une validation ou invalidation expérimentale générant ainsi un nouveau flot de données qui sera stocké dans des banques dont l'interrogation sera effectuée grâce à des méthodologies informatiques tout en tenant compte du sens biologique des données.

Un exemple concret de ce processus de recherche est la découverte d'un nouveau système de défense chez les archées et les bactéries contre les agressions virales. C'est le système CASS, système composé d'une structure génétique particulière, le CRISPR et d'un ensemble de gènes l'accompagnant. La découverte de ce système est fondée sur des analyses "*in silico*" et son rôle a été confirmé par des manipulations expérimentales. C'est un système qui existe chez presque la moitié des génomes procaryotes étudiés par la communauté scientifique et dont plusieurs ont souvent été employés pour diverses analyses biologiques à travers des manipulations expérimentales. Cependant, il n'a été découvert que durant les dix dernières années grâce à l'interaction entre la biologie et la bioinformatique. Il s'agit d'un sujet de recherche encore récent et qui nécessite encore beaucoup d'investigations de la part des biologistes ainsi que les bioinformaticiens. C'est dans ce cadre que se situe mon travail de thèse durant laquelle j'ai été amenée à créer une base de données des CRISPRs et des outils associés facilitant l'investigation de ces structures et leur exploitation dans le typage bactérien. Afin d'accomplir ces tâches, il fallait commencer par la création d'un programme de détection automatique des CRISPRs. Toute la difficulté du travail a résidé dans la modélisation informatique précise de ces structures tout en tenant compte, au maximum possible, des variations biologiques. A partir d'observations sur des séquences biologiques, il fallait définir de façon formelle, un schéma figé qui représente n'importe quel CRISPR en évitant d'inclure d'autres structures.

Dans la première partie de ma thèse, je présenterai cette structure fascinante qui est le CRISPR tout en priant le lecteur d'excuser le manque de précision sur certains aspects biologiques que ma formation "matheuse" ne me permet pas d'approfondir. Le premier chapitre servira à présenter principalement les bactéries, auxquelles notre laboratoire s'intéresse du point de vue de l'épidémiologie et de la phylogénie, en insistant sur les aspects utiles pour la compréhension du rôle biologique du CRISPR et des applications qui peuvent en découler. Le second chapitre disséquera les différents éléments du système CASS et tracera l'histoire de leurs découvertes. Dans le troisième chapitre, les outils informatiques utilisés dans l'étude des CRISPRs seront analysés.

La deuxième partie est consacrée à l'exposé des programmes informatiques que j'ai élaborés pour la manipulation *in silico* du CRISPR, ainsi que quelques applications de ces outils. Elle sera divisée en deux chapitres. Le premier présentera le programme CRISPRFinder pour l'identification automatique des CRISPRs et son utilisation pour générer la base de données CRISPRdb ainsi qu'une investigation des CRISPRs de *Yersinia pestis* et *Yersinia pseudotuberculosis*. Le deuxième chapitre s'intéressera aux outils créés pour assister la procédure de typage bactérien en utilisant le CRISPR, à leur mode d'utilisation et à une application au cas particulier de *Listeria monocytogenes*.

Enfin, je conclurai en présentant les perspectives de ce travail du point de vue informatique et biologique.

Première partie

Les CRISPRs : contexte et définitions

La première partie de cette thèse est articulée sur trois chapitres introductifs présentant l'état de l'art sur les CRISPRs. Le premier chapitre est un chapitre d'ordre général décrivant les données biologiques sur les procaryotes, nécessaires pour appréhender le système CASS qui sera exposé dans le second chapitre avec ses propriétés principales et les grandes étapes survenues dans sa découverte. Le troisième chapitre a pour ambition de placer le CRISPR dans le contexte bioinformatique en décrivant les outils utilisés dans son investigation et de montrer l'intérêt de développer des outils spécifiques qui feront l'objet de cette thèse.

Chapitre 1

Les archées, les bactéries et l'identification de souches

Aborder le sujet du CRISPR, ou plus généralement le système CASS nécessite d'abord d'introduire quelques notions biologiques sur les organismes qui le portent et les applications qui peuvent y être liées, notamment le typage moléculaire. Ce chapitre présentera dans un premier temps les archées et les bactéries en insistant sur les virus qui les attaquent. Dans un deuxième temps, il énumérera les différentes techniques de typage moléculaire utilisées dans l'identification des souches bactériennes et donnera les limites de chacune.

1.1 Les archées

1.1.1 Les archées

Le vivant est constitué d'organismes cellulaires qui se divisent en deux types :

1. **les eucaryotes**, dont les cellules ont un noyau bordé d'une membrane (Callen, 1999). Le plus souvent, ils contiennent aussi des membranes internes qui cloisonnent la cellule en y délimitant des organites qui ont des fonctions biologiques spécialisées.
2. **les procaryotes**, qui sont unicellulaires, sont des organismes dépourvus de noyau et bordés d'une membrane. Dans ce travail, on s'intéressera uniquement aux procaryotes.

En 1977, l'Américain C. Woese découvre que les procaryotes ne forment pas un groupe homogène (Woese et Fox, 1977). Pionnier dans ce domaine, il trouve que certains génomes

procaryotes sont très différents des autres. Il les subdivise donc en deux groupes : les eubactéries (communément appelés bactéries) et les archées. Ces travaux, parfois qualifiés de "révolution woesienne", suggèrent que le vivant comprend trois grandes lignées : bactéries, archées et eucaryotes.

La lignée la moins étudiée est celle des archées. Les archées constituent un groupe très hétérogène, regroupant peu d'espèces connues. Elles constituent des niches écologiques qu'elles sont souvent seules à occuper, en particulier dans des environnements extrêmes en température ($\geq 80^{\circ}C$), en acidité (pH < 3) ou en salinité.

Les archées sont très diverses dans leur forme et les milieux qu'elles colonisent. Certaines sont extrêmophiles, c'est à dire que leurs conditions de vie normales sont mortelles pour la plupart des autres organismes. *Pyrolobus fumarii* par exemple, est capable de proliférer à $113^{\circ}C$. D'autres archées sont méthanogènes : elles se développent dans les milieux privés d'oxygène (marais, appareil digestif des ruminants et des termites) et libèrent du méthane. Celles des marais ont ainsi contribué à la création des réserves de gaz naturel. D'autres encore sont halophiles : elles croissent dans les sols saturés en sels, les lacs salés ou les marais salants. Certains biologistes pensent que les archées sont proches de la forme de vie ancestrale qui, sur la Terre, serait à l'origine de toutes les autres (baptisée LUCA pour Last Universal Common Ancestor, le dernier ancêtre commun universel).

Je ne développerai pas plus la description de ces organismes mais il est important de savoir qu'une controverse existe toujours quant aux relations existant entre archées et bactéries et leur place respective dans l'évolution du vivant par rapport aux cellules eucaryotes plus complexes et plus spécialisées. Les travaux de P. Forterre (Forterre, 1995) suggèrent que, loin d'être les exemples vivants de nos jours des organismes qui existaient au tout début de l'apparition de la vie en milieu extrême, les archées pourraient être un résultat plus moderne de l'adaptation à ces milieux. L'hypothèse de la "thermoréduction", est l'adaptation aux températures très élevées conduisant une lignée évolutive à acquérir les caractéristiques propres aux procaryotes : petite taille, renouvellement rapide des macromolécules, élimination du noyau et en conséquence couplage de la transcription et de la traduction.

1.1.2 Les virus d'archées

Un virus est composé d'une capsule contenant de l'ADN. Cependant, comme il ne peut se reproduire sans une cellule hôte, on le considère comme un parasite chimique très évolué et non comme un être vivant. De plus, aucun intermédiaire entre virus et bactérie n'est connu.

Si les archées ont été peu étudiées, les virus associés à ces organismes ancestraux l'ont été moins encore (Prangishvili, 2006). Parmi environ 5.100 virus connus, seulement 36 ont

été isolés chez les archées. Les virus des archées présentent une diversité morphologique et génétique beaucoup plus grande que celle des virus qui infectent les bactéries (Rice, 2004).Ils se classent actuellement en six familles virales différentes, qui infectent principalement des archées hyperthermophiles aérobies des genres *Sulfolobus* et *Acidianus*. Ces espèces représentent 75% de la totalité des virus hyperthermophiles actuellement connus.

1.2 Les bactéries

1.2.1 Bactéries pathogènes vs. bactéries non pathogènes

Une bactérie est un micro-organisme unicellulaire sans noyau (procaryote) dont le génome est constitué d'ADN (Perry, 2004). Elle contient généralement un seul chromosome et éventuellement des plasmides. Les bactéries sont ubiquitaires et sont présentes dans tous les types de biotopes rencontrés sur Terre. Elles peuvent être isolées du sol, des eaux douces, marines ou saumâtres, de l'air, des profondeurs océaniques, de la croûte terrestre, sur la peau, dans l'intestin, etc.

Certaines provoquent des maladies chez l'homme, l'animal ou les plantes et sont dites pathogènes. Cependant, la très grande majorité n'est pas pathogènes. Certaines bactéries remplissent des fonctions vitales dans le corps humain (bactéries de l'intestin), d'autres sont utiles dans l'agroalimentaire lors de la fabrication des yaourts ou du fromage. De plus, certaines bactéries sont utilisées dans la production industrielle de nombreux composés chimiques, dans des processus de traitement des eaux usées et dans le développement de biopesticides, etc.

Les bactéries pathogènes

Les bactéries pathogènes sont responsables de maladies et causent des infections. Les organismes infectieux peuvent être distingués en trois types (Singleton, 2005) : les pathogènes obligatoires, accidentels ou opportunistes. Un pathogène obligatoire ne peut survivre en dehors de son hôte. Parmi les bactéries pathogènes obligatoires, *Corynebacterium diphtheriae* entraîne la diphtérie, *Treponema pallidum* est l'agent de la syphilis, *Mycobacterium tuberculosis* provoque la tuberculose, *Mycobacterium leprae* la lèpre, *Neisseria gonorrhoeae* la gonorrhée par exemple. Les Rickettsia à l'origine du typhus sont des bactéries parasites intracellulaires. Un pathogène accidentel présent dans la nature peut infecter l'homme dans certaines conditions. Par exemple, *Clostridium tetani* provoque le tétanos en pénétrant dans une plaie. *Vibrio cholerae* entraîne le choléra suite à la consommation d'une eau contaminée. Un pathogène opportuniste infecte des individus affaiblis

ou atteints par une autre maladie. Des bactéries comme *Pseudomonas aeruginosa*, des espèces de la flore normale, comme des *Staphylococcus* de la flore cutanée, peuvent devenir des pathogènes opportunistes dans certaines conditions. On rencontre ce type d'infection par exemple en milieu hospitalier (infection nosocomiale).

Les bactéries non pathogènes : utilisation dans l'industrie et les biotechnologies

1. **L'industrie agroalimentaire**

   Les bactéries lactiques transforment les hydrates de carbone en acide lactique et se trouvent partout dans la nature (Singleton, 2005) et chez l'homme (la peau, le système digestif et la muqueuse vaginale) où elles accomplissent de nombreuses fonctions en créant notamment un environnement hostile (milieu acide grâce à la production d'acide lactique) aux bactéries pathogènes. Les espèces bactériennes comme *Lactobacillus casei*, *Lactococcus lactis* ou *Streptococcus thermophilus*, combinées aux levures et moisissures interviennent dans l'élaboration d'aliments fermentés comme les fromages, les yaourts, la bière, le vin, la sauce de soja, le vinaigre et la choucroute.

   Par ailleurs, les bactéries acétiques telles que *Acetobacter aceti* et *Gluconobacter thailandicus* peuvent produire de l'acide acétique à partir de l'éthanol. Elles sont rencontrées dans les jus alcoolisés et sont utilisées dans la production du vinaigre. Elles sont également exploitées pour la production d'acide ascorbique (vitamine C) à partir du sorbitol transformé en sorbose.

2. **Les bactéries probiotiques**

   Les probiotiques sont des micro-organismes ingérés vivants, capables d'exercer des effets bénéfiques sur la santé -définition de la FAO (Food and Agriculture Organization) et de l'OMS (Organisation mondiale de la santé). Il s'agit de "bonnes" bactéries que l'on retrouve notamment dans les flores intestinale et vaginale. En se multipliant dans l'intestin, ces bactéries permettraient de réduire par simple compétition la population bactérienne potentiellement pathogène. Le concept des probiotiques provient d'un chercheur et Prix Nobel Russe, Elie Metchnikoff, qui avait pour théorie que la longévité des paysans bulgares était directement liée à leur consommation de laits fermentés. Différentes études ont suggéré l'effet bénéfique des probiotiques et leur innocuité liée au caractère non pathogène des souches utilisées. Les probiotiques utilisés sont des souches micro-organiques vivantes, productrices d'acide lactique tels que les Lactobacilles, certains Streptocoques, et les Bifidobactéries, notamment *Bifidobacterium bifidum*.

3. **La bioremédiation**

   La bioremédiation est le processus de nettoyage de milieux pollués par des micro-organismes. En effet, certaines bactéries hétérotrophes [1] ont une capacité à dégrader une large variété de composés organiques. Ces bactéries sont utilisées dans les pro-

[1] L'hétérotrophie est la nécessité pour un organisme vivant de se nourrir de constituants organiques préexistants, d'origine animale ou végétale.

cessus de traitement des déchets et dans les fosses septiques pour en assurer l'épuration. D'autres bactéries sont capables de dégrader des hydrocarbures du pétrole, elles peuvent être utilisées lors du nettoyage d'une marée noire.

4. **La biolixiviation**

   Certaines bactéries peuvent être utilisées pour récupérer des métaux d'intérêt économique à partir de minerais. C'est la biolixiviation [2]. L'activité de bactéries est ainsi exploitée pour la récupération du cuivre.

5. **L'agriculture**

   Certaines bactéries peuvent être utilisées à la place de pesticides en lutte biologique (les biopesticides[2]) pour combattre des parasites des plantes limitant ainsi l'usage d'insecticides chimiques qui sont souvent peu spécifiques et tuent indifféremment les insectes nuisibles et les bénéfiques (pollinisateurs). Par exemple, pour lutter contre le borer, insecte ravageur du maïs, *Bacillus thuringiensis* qui produit une protéine (Bt) toxique pour certains insectes ravageurs des cultures peut être utilisée.

   Des bactéries comme *Bradyrhizobium sp. ORS278* et *Rhizobium leguminosarum* sont des bactéries du sol capables de fixer de l'azote atmosphérique et fournir chaque année entre 50 et plus de 300 kg d'azote par hectare aux cultures. Elles peuvent ainsi améliorer la productivité agricole, la fertilité des sols et la production fourragère.

6. **Les outils de la microbiologie expérimentale**

   En raison de leur capacité à se multiplier rapidement et de leur relative facilité à être manipulées, certaines bactéries comme *Escherichia coli* sont des modèles et des outils très utilisés en microbiologie, génétique et biochimie. Les scientifiques peuvent déterminer la fonction de gènes, d'enzymes ou identifier des voies métaboliques nécessaires à la compréhension du vivant et permettant également de mettre en oeuvre de nouvelles applications en biotechnologie. De nombreuses enzymes utilisées dans divers processus industriels ont été isolées de micro-organismes. Les enzymes des détergents sont des protéases issues de certaines souches de *Bacillus*. Des amylases capables d'hydrolyser l'amidon sont très utilisées dans l'industrie alimentaire. La Taq polymérase utilisée dans les réactions de polymérisation en chaîne (PCR)[3] pour l'amplification de l'ADN provient d'une bactérie thermophile *Thermus aquaticus*.

7. **L'industrie pharmaceutique**

   Les bactéries génétiquement modifiées sont très utilisées pour la production de produits pharmaceutiques. C'est le cas par exemple de l'insuline, l'hormone de croissance, certains vaccins, des interférons, etc. Certaines bactéries comme *Streptomyces ambofaciens* sont très employées pour la production d'antibiotiques.

[2] source `http://www.courseweb.uottawa.ca/EVS3520/Data/ChapitreC.pdf`

[3] La PCR (Polymerase Chain Reaction) est une méthode qui permet, *in vitro*, d'amplifier, c'est-à-dire de copier en grand nombre, un fragment d'ADN spécifique et de longueur définie à partir d'une faible quantité d'acide nucléique.

Imaginée par K. Mullis en 1985 (Prix Nobel 1993), la technique est automatisée grâce à l'utilisation (vers 1988) d'une ADN polymérase résistante aux températures élevées (la Taq polymérase).

1.2.2 Bactériophages et attaque virale

Les phages

Les bactériophages ou phages, découverts par F.W Twort en 1915 (Twort, 1915), puis redécouverts par F. d'Hérelle en 1917 (D'Herelle, 1917), qui leur donna le nom de bactériophage ou "mangeur de bactéries" , sont des virus infectant les bactéries. Comme tous les virus, ils se caractérisent par la possession d'un seul acide nucléique et par un parasitisme intracellulaire obligatoire. Ils sont considérés comme étant les organismes les plus nombreux dans la biosphère (ils sont présents dans le sol, dans l'eau, sur les plantes, dans les cavités naturelles de l'homme et des animaux, dans les aliments, etc) puisqu'on considère généralement que chaque organisme vivant peut être la cible de différents virus, parfois très spécifiques.

Pratiquement toutes les espèces bactériennes peuvent être infectées par des phages spécifiques. En effet, le phage se fixe sur des récepteurs de la bactérie présentant une étroite spécificité vers des composants phagiques [4]. Trois éventualités peuvent avoir lieu suite à une agression phagique :

- La bactérie n'est pas infectée ou résiste à l'infection. Cette résistance résulte notamment soit de l'absence de récepteurs permettant la fixation du phage soit de la présence d'enzymes de restriction capables de dégrader l'acide nucléique phagique.
- Le phage ou uniquement son acide nucléique pénètre dans la bactérie et le phage se multiplie à l'intérieur de la cellule bactérienne. Le cycle est productif et le phage est qualifié de virulent. Selon les bactériophages, le cycle productif peut ou non conduire à une lyse de la bactérie.
- Le phage ou uniquement son acide nucléique pénètre dans la bactérie et l'acide nucléique phagique s'intègre dans le génome bactérien où il persiste à l'état latent sous forme de prophage. Le cycle est qualifié de lysogénique et le bactériophage est appelé phage tempéré.

Cependant, les bactéries réagissent aux attaques virales et développent des mécanismes de résistance aux phages qui sont par exemple les enzymes de restriction et l'infection abortive (Sturino et Klaenhammer, 2006). Le système CASS est le système de défense bactérien contre les phages le plus récemment découvert (Barrangou, 2007).

[4] Source : site internet Abrégé de Bactériologie Générale et Médicale à l'usage des étudiants de l'École Nationale Vétérinaire de Toulouse, `http://www.bacteriologie.net`, créé et maintenu par J.P. Euzéby

Une menace pour les bactéries

Les bactériophages sont très abondants dans la nature, mais leur développement, leur reproduction et leur évolution dépendent étroitement de la présence des bactéries. Ils ne sont présents dans un milieu que dans la mesure où celui-ci héberge des bactéries hôtes. Ils constituent alors une vraie menace pour la survie des bactéries. Du point de vue humain, ceci peut être vu sous deux aspects différents, un aspect positif car c'est un moyen de lutter contre les bactéries pathogènes (la phagothérapie) et un aspect négatif en raison de la disparition non désirée de bactéries non pathogènes lors de leur utilisation dans un processus industriel.

1. **La phagothérapie**
   Les phages, comme agents thérapeutiques (Summers, 1993b,a), sont inoffensifs pour l'humain et sont hautement spécifiques pour leur hôte bactérien ce qui évite la destruction non spécifique des bactéries de la flore normale. Ils s'amplifient de façon exponentielle au site d'infection et tuent rapidement la cellule cible. L'avantage majeur des phages par rapport aux antibiotiques est qu'ils évoluent constamment au même titre que les bactéries hôtes qu'ils infectent. Ils sont aptes à répondre à l'évolution rapide de nombreux pathogènes. Dès la découverte des bactériophages, d'Herelle (D'Herelle, 1917) les a employés pour traiter la dysenterie bacillaire. Pour cet auteur, la phagothérapie devait permettre de lutter contre de nombreuses maladies infectieuses et des essais ont été réalisés lors de fièvre typhoïde, de choléra, de peste, ou encore lors d'infections staphylococciques. L'arrivée des antibiotiques au milieu des années 1940, ainsi qu'un manque de connaissances fondamentales sur la biologie des phages à cette époque ont contribué au déclin de l'intérêt pour la phagothérapie, du moins dans les pays occidentaux. Ce n'est qu'en Europe de l'Est (Barrow et Soothill, 1997) et particulièrement dans différents pays de l'ex URSS que la phagothérapie a continué à être pratiquée et l'est encore aujourd'hui, comme c'est le cas au fameux George Eliava Institute of Bacteriophage, Microbiology and Virology (IBMV), à Tbilisi en Géorgie et dans l'"Institute of Immunology and Experimental Therapy" à Wroclaw en Pologne. De nos jours, la résistance des bactéries aux antibiotiques fait à nouveau envisager l'utilisation de la phagothérapie comme une alternative thérapeutique prometteuse. Cependant, la phagothérapie se heurte à deux obstacles majeurs : (i) la sélection rapide de mutants résistants et (ii) l'apparition d'anticorps neutralisants entravant la phase d'adsorption. Cette technique pourrait même être dangereuse dans le cas de bactéries qui libèrent des toxines quand elles meurent, ou de phages porteurs de gènes de virulence ou de sécrétion de toxines.
2. **La menace sur les chaînes de production industrielles**
   Les bactériophages peuvent avoir des répercussions industrielles majeures pour les industries de fermentation (Singleton, 2005) qui utilisent des souches bactériennes (industries laitières, production d'antibiotiques, etc.). Les phages peuvent en effet contaminer et détruire les souches bactériennes utilisées. L'application de mesures strictes de désinfection des locaux et de l'appareillage, ainsi que l'utilisation de

souches résistantes aux phages permettent dans une certaine mesure de contenir ces risques (Sanders, 1988). Le phénomène de la résistance est utile dans ce cas car les infections par des phages dans un contexte industriel peuvent arrêter les chaînes de production et causer d'importantes pertes économiques.

Il y a coévolution entre bactériophages et bactéries. Il suffit de les mélanger pour voir l'apparition de phages mutants. Ceux-ci peuvent ensuite infecter une bactérie anciennement résistante. L'évolution peut être très rapide, il est possible d'obtenir des bactéries résistantes aux phages en quelques jours, mais en quelques jours aussi, les phages deviennent capables de réinfecter ces bactéries.

1.3 Le typage moléculaire et l'identification de souches

L'identification bactérienne consiste à déterminer l'espèce d'une bactérie d'intérêt ; le typage vise à distinguer les souches au sein d'une même espèce bactérienne. Nous avons distingué dans la section précédente 1.2, deux types de bactéries : les pathogènes et les non pathogènes. Le typage bactérien se révèle d'un intérêt particulier dans les deux cas.

Dans le cas des bactéries pathogènes, face à une infection, l'identification de l'agent microbiologique en cause est indispensable pour la prise en charge appropriée et immédiate du patient. Il est également essentiel, dans un deuxième temps, de chercher l'origine de l'agent infectieux, afin si possible de "tarir" l'éventuelle source. Les sources peuvent être l'environnement médical (matériel médical et personnel soignant), l'environnement urbain (canalisations d'eau, tours de refroidissement dans le cas de la légionellose par exemple), ou l'environnement privé dans des pays aux conditions sanitaires mal contrôlées (épidémies de choléra, réémergence de peste, ou en fin d'année 2002, souches de *M. tuberculosis* très résistantes aux antibiotiques venant de Chine). Cette enquête *a posteriori* requiert le typage des souches. En effet, pour une surveillance épidémiologique efficace, il est nécessaire de pouvoir identifier avec le plus de précision possible les souches bactériennes responsables d'épidémies à l'échelle planétaire ou locale. A l'échelle planétaire, la connaissance de l'origine de souches responsables de maladies telles que la tuberculose contribue à la mise en place des mesures sanitaires appropriées dans les pays concernés, voire même peut permettre d'exercer une pression internationale sur ces pays pour qu'ils améliorent leur prise en charge des maladies infectieuses. A l'échelle locale, le typage de souches permet de prendre des mesures internes (mesures sanitaires dans les hôpitaux, identification de porteurs, identification de foyers d'accueil à risque, identification de systèmes de canalisations contaminés).

Quant aux bactéries non pathogènes, le typage bactérien est également une procédure indispensable dans les processus biotechnologiques puisqu'il contribue par exemple à leur protection en terme de propriété industrielle. Par exemple, dans l'industrie laitière, le

choix du ferment a un impact sur le développement des propriétés et caractéristiques du produit telles que la saveur, l'arôme et la texture.

Le typage des souches peut se faire de différentes façons (Onteniente, 2004). On distingue deux classes de méthodes : les méthodes phénotypiques et les méthodes génotypiques.

1.3.1 Les méthodes phénotypiques

Les techniques de phénotypage étudient les propriétés exprimées par les bactéries :

1. test de la résistance aux antibiotiques : détermine par comparaison avec des données de référence si la souche est résistante ou sensible à certains antibiotiques.
2. sérotypage : compare le comportement de la surface membranaire de différentes souches par des tests d'agglutination caractérisant des antigènes bactériens (capsulaires, flagellaires).
3. lysotypage : étudie la sensibilité d'une bactérie à la lyse par un panel de bactériophages connus (Sutter, 1965).
4. MLEE (MultiLocus Enzymes Electrophoresis) : détecte les différences de mobilité électrophorétique d'enzymes identifiées selon leur composition en acides aminés (Selander, 1986).
5. tests biochimiques : étudient les propriétés métaboliques inhabituelles ou spécifiques au sein d'une espèce comme par exemple l'utilisation des sucres.

Les méthodes phénotypiques sont en général ou bien peu discriminantes ou lourdes à mettre en place et peu reproductibles.

1.3.2 Les méthodes génotypiques

Les méthodes de génotypage (analyse de l'ADN), souvent associées à des sigles (RFLP, VNTR, RAPD, AFLP, PFGE etc.) s'adressent à l'analyse du génome bactérien : plasmides, chromosome. Elles prennent une part croissante par rapport aux techniques de phénotypage, du fait d'une part des problèmes de "typabilité" de certaines souches et d'autre part des problèmes de reproductibilité rencontrés avec certaines techniques. L'arrivée de la technique de PCR en 1985 (Saiki, 1985) et la disponibilité des données de séquençage des génomes complets depuis la publication de la première séquence complète d'un génome bactérien, celui d'*H. influenzae*, en 1995 (Fleischmann, 1995) ont encore accéléré cette tendance.

Le principe des méthodes de génotypage consiste à rechercher des polymorphismes qui peuvent être de différentes natures et d'échelles variables allant des grands réarrangements

génomiques à la mutation ponctuelle en passant par des délétions ou insertions de courtes séquences. Suivant le niveau de variabilité au sein d'une espèce bactérienne, différents tests peuvent être appliqués pour différencier les souches. L'analyse de l'ADN bactérien permet d'associer à chaque souche un profil génétique caractéristique. On distingue trois types :

1. techniques basées sur l'analyse de fragments d'ADN génomique : PFGE (Pulse Field Gel Electrophoresis) et RFLP (Restriction Fragment Length Polymorphism). La technique PFGE consiste à digérer l'ADN génomique en fragments de très grande taille par des enzymes de restriction appropriées et à les analyser par électrophorèse tandis que la technique RFLP consiste à digérer l'ADN génomique en fragments plus courts qui sont alors analysés par électrophorèse et hybridation avec des sondes spécifiques. La taille et le nombre des fragments sont caractéristiques de l'isolat étudié.

   Ces techniques sont manuelles, longues et délicates à mettre en oeuvre. Les profils générés ne peuvent pas s'exprimer en un code simple et sont difficiles à comparer d'une expérience à l'autre.

2. techniques basées sur l'analyse d'un profil aléatoire de fragments d'ADN : la RAPD (Random Amplified Polymorphic DNA) ou AP-PCR (Arbitrarily Primed PCR) et l'AFLP (Amplified Fragment Length Polymorphism). La RAPD repose sur l'amplification par PCR de différentes portions d'un génome sans connaître sa séquence en utilisant des amorces aléatoires courtes (9 à 10 paires de base). Le nombre et la localisation des sites d'hybridation des amorces varient d'une souche à l'autre dans une même espèce. L'AFLP repose sur la digestion de l'ADN génomique par deux enzymes de restriction et la ligation d'adaptateurs aux fragments générés suivies par une amplification sélective de fragments de restriction.

   Les deux techniques ne requièrent pas la connaissance de données de séquence génomique. Le reproche majeur fait à la technique RAPD est le manque de reproductibilité et de standardisation. Des variations ont été observées d'une machine PCR à l'autre (Meunier et Grimont, 1993). Par contre, la technique AFLP est très puissante, beaucoup plus robuste et reproductible que la précédente, mais elle exige de l'ADN de bonne qualité, et le respect de procédures rigoureuses.

3. techniques basées sur l'identification de mutations ponctuelles : typage des SNPs (Single Nucleotide Polymorphism) et MLST (Multi Locus Sequence Typing). Cette méthode utilise les informations générées par des séquences de plusieurs génomes d'une même espèce. Le terme SNP désigne un changement d'un seul nucléotide à une position du génome. Les mutations prennent trois formes (Brookes, 1999) : substitution, insertion ou suppression. Quant au MLST, c'est une technique qui consiste à séquencer environ 500 pb de fragments internes de gènes de ménage. En général 7 gènes sont séquencés par isolat. Ensuite les séquences sont comparées aux séquences déjà rencontrées et les isolats classés en types de séquences (ST pour "sequence types").

   Le typage des SNPs est une technique adaptée à des espèces jeunes et également à des études phylogénétiques. La technique MLST est encore coûteuse (2 fois 7

séquences à effectuer) pour des analyses de routine dans des laboratoires d'analyse hospitaliers (Maiden, 1998). L'énorme avantage de cette approche est qu'elle est en théorie parfaitement reproductible quel que soit le laboratoire puisqu'il s'agit de séquencer et de classer les séquences. Il existe d'ailleurs des bases de données en ligne pour le typage MLST (Chan, 2001) avec des données en quantité croissante pour plusieurs bactéries et il est possible de soumettre des séquences sur le site `www.mlst.net`.

4. Analyse du polymorphisme des répétitions en tandem :

   La plupart des bactéries possède des structures génomiques constituées de la répétition en tandem d'une séquence nucléotidique, de 1 à 8 nucléotides pour les microsatellites, 9 et plus pour les minisatellites. D'une souche à l'autre, le nombre de répétitions peut varier constituant ainsi un marqueur génétique polymorphe multiallélique (ces structures sont alors souvent appelées Variable Number of Tandem Repeats ou VNTR). L'analyse de plusieurs marqueurs de ce type (de 10 à 20) permet d'obtenir une "empreinte génétique" d'un clone bactérien. Il s'agit de la technique MLVA (Multi-Locus VNTR Analysis). Les variations de taille des VNTRs peuvent être analysées par PCR en choisissant des amorces situées de part et d'autre des micro- ou minisatellites polymorphes puis en mesurant la taille des produits par électrophorèse.

   Le typage MLVA présente les avantages de la reproductibilité, comparabilité entre laboratoires en plus de son pouvoir discriminant important et de son coût peu élevé. Il a démontré son intérêt depuis 1998 pour le typage de souches de *M. tuberculosis* et est depuis considéré comme technique de référence pour de nombreuses espèces bactériennes (Frothingham et Meeker-O'Connell, 1998; Pourcel, 2007; Le Flèche, 2006; Valjevac, 2005; Vu-Thien, 2007). Le laboratoire GPMS joue un rôle majeur dans le développement de cette technique. Des bases de données de répétitions en tandem (Denoeud et Vergnaud, 2004; Grissa, 2008b) ont été mises au point et rendues disponibles via internet [5] [6] afin de faciliter le développement de marqueurs pour le typage MLVA.

1.3.3 Comparaison des différentes méthodes de typage

Le tableau A.1 de l'annexe A fait le bilan des avantages et inconvénients des différentes techniques de typage présentées dans le paragraphe précédent. Aucune méthode ne peut être considérée comme une méthode parfaite. Cependant certaines méthodes sont très lourdes à mettre en place, très coûteuses et surtout les résultats sont non comparables entre laboratoires. La meilleure façon actuelle d'aborder les problèmes de typage est la combinaison de deux ou plusieurs techniques en même temps. Ce choix dépendra fortement du but de l'étude effectuée. Dans les études épidémiologiques, il suffit en général d'avoir

[5] `http://minisatellites.u-psud.fr/`

[6] `http://mlva.u-psud.fr/`

une méthode qui permette de reconnaître les souches identiques et de les différencier des souches sans lien épidémiologique. L'analyse du polymorphisme des CRISPRs qui fera l'objet de cette thèse représente une nouvelle méthode de typage bactérien facile à mettre en place, peu coûteuse et dont les données sont échangeables et comparables entre différents laboratoires. Elle peut être utilisée essentiellement en complément à d'autres méthodes. Elle ne peut pas être appliquée chez toutes les espèces bactériennes.

1.4 Conclusion

Le CRISPR est une structure portée uniquement par les archées et les bactéries qui sont des petits organismes sans noyau, en perpétuelle guerre avec des virus qui leur sont spécifiques.

Différencier ou comprendre la filiation des souches d'une même espèce est nécessaire que ce soit pour la lutte contre les bactéries pathogènes ou l'exploitation de celles utiles à l'homme. Pour cela différentes techniques de typage ont été mises en place et sont utilisées de façon combinée.

Chapitre 2

Le système CASS (CRISPR et gènes associés)

Ce chapitre sera dédié à la description du système CASS : structure, fonctionnement et rôle. Je commencerai par présenter quelques généralités sur le contexte biologique qui entoure les CRISPRs en décrivant l'interférence ARN chez les eucaryotes qui est probablement un bon modèle du fonctionnement du CRISPR. La deuxième partie de ce chapitre servira de recueil bibliographique décrivant l'histoire des CRISPRs dans l'ordre chronologique des événements liés à leur découverte.

2.1 Généralités

2.1.1 L'interférence ARN chez les eucaryotes

En 1998, les docteurs Craig Mello et Andrew Fire ont découvert le mécanisme de l'interférence ARN (Fire, 1998). Cette découverte a été récompensée en 2006 par le prix Nobel de médecine. Le mécanisme mis en évidence permet d'empêcher l'expression d'un gène en détruisant l'ARN messager, intermédiaire indispensable à l'expression du gène en protéine.

Dans un système d'interférence ARN chez les eucaryotes, les ARN double brins pré-

sents dans une cellule sont tout d'abord pris en charge par une ribonucléase appelée Dicer, l'"éminceuse". Celle-ci clive l'ARN double brin des virus toutes les 21 à 25 pb en petits ARN interférents (siRNAs). La protéine Dicer transfère alors les siRNA à un gros complexe multiprotéique, le complexe RISC (RNA-Induced Silencing Complex). Un des brins du siRNA, dit passager, est éliminé tandis que l'autre (appelé "guide") dirige le complexe RISC vers les ARNm viraux possédant une séquence complémentaire au brin guide. Si la complémentarité entre le siRNA et l'ARNm cible est parfaite, le complexe RISC clive l'ARNm cible qui est alors dégradé par une autre nucléase appelée Slicer (Sontheimer, 2005) et n'est donc plus traduit en protéine. Quelques bases non complémentaires suffisent pour empêcher le clivage. Ce mécanisme est donc très spécifique de la séquence du siRNA et de sa cible, l'ARNm. Dans certains cas, on peut choisir un siRNA capable de cliver un ARNm porteur d'une mutation ponctuelle sans affecter l'ARNm sauvage.

2.1.2 Le transfert horizontal

Le transfert génétique consiste en l'échange de matériel génétique entre deux organismes. L'échange de matériel génétique entre bactéries (transfert horizontal) est connu depuis de nombreuses années. Il influe fortement sur l'évolution des génomes de procaryotes (Kurland, 2000; Philippe et Douady, 2003). En particulier, de nombreuses bactéries deviennent résistantes à des antibiotiques en acquérant les gènes de résistance à partir d'espèces éloignées phylogénétiquement. Il existe trois mécanismes principaux pour expliquer les transferts horizontaux entre organismes :

1. transformation : intégration d'ADN libre, se trouvant dans le milieu extérieur d'un organisme donné et résultant de la mort d'un autre organisme. L'ADN pénètre dans la cellule puis est intégré au génome.
2. conjugaison : deux cellules entrent en contact et s'échangent du matériel génétique via un système particulier utilisant entre autre une structure dite "pilus".
3. transduction : l'ADN est transféré d'une espèce à une autre via des virus ou des phages. Le phage peut amener par erreur une partie du matériel génétique d'un premier hôte et le transférer à une autre espèce hôte.

Les signes permettant parfois de détecter *a posteriori* le transfert horizontal sont l'existence d'un arbre phylogénétique incongruent, une composition nucléotidique anormale, une distribution taxonomique anormale ou une grande similarité de séquences entre espèces éloignées.

2.1.3 La recombinaison homologue de l'ADN

Les recombinaisons génétiques naturelles sont un des mécanismes de l'évolution des espèces, et sont à l'origine de la diversité des individus d'une population. Tous les orga-

nismes dépendent de la recombinaison pour le maintien de la stabilité de leur génome ainsi que pour la production de la variabilité génétique. La recombinaison homologue résulte d'une série d'interactions entre deux séquences d'ADN homologues, présentes sur une ou deux molécules d'ADN, et produit une séquence mixte dérivée des séquences parentales (Smith, 1988). Elle consiste en l'invasion de la terminaison 3'-OH d'une séquence d'ADN simple brin par un deuxième ADN double brin (Smith, 1988). L'appariement initial peut se produire à n'importe quelle position de la région homologue. La réaction d'échange entre brins commence quand les deux molécules sont alignées et que l'extrémité de l'ADN est libre. La protéine RecA chez les procaryotes contrôle la fidélité de la recombinaison.

La recombinaison homologue intervient en général dans les mécanismes de réparation des cassures simple et double chaîne de l'ADN. Elle contribue également au rétablissement de la synthèse d'ADN après blocage de la fourche de réplication et elle joue un rôle central dans le maintien de l'intégrité du génome. En dépit de son importance, la recombinaison peut dans certains cas se révéler dangereuse en générant des réarrangements chromosomiques nuisibles pour la cellule ou créer des intermédiaires toxiques.

L'aspect qui nous intéresse dans ce rapport concerne les réarrangements qui vont permettre la synthèse ou la délétion de motifs[1] du CRISPR à partir de la recombinaison homologue de deux motifs identiques, les DR (voir paragraphe 2.2.1).

2.2 Description du système CASS : CRISPR-Cas

Fig. 2.1 – Éléments du système CASS. Les DR sont représentés par des boites jaunes, les spacers sont représentés par des boites de même taille mais de couleurs différentes. La séquence leader est représentée en bleu et les gènes *cas* en vert.

Le système CASS est un système retrouvé uniquement chez les procaryotes avec une présence quasi-systématique pour les archées et partielle pour les bactéries (40% des bactéries séquencées). Différentes études ont montré qu'au moins une de ses fonctions consiste en la défense contre les agressions virales (Barrangou, 2007) agissant probablement par un mécanisme similaire à l'interférence ARN chez les eucaryotes. Ce complexe est constitué de deux composants principaux, une structure génétique dite "CRISPR" pour Clustered Regularly Interspaced Short Palindromic Repeat et une cassette de gènes accompagnant cette structure et nécessaire à son fonctionnement.

[1]Nous utiliserons le terme motif ou unité pour décrire un ensemble DR + spacer.

Fig. 2.2 – CRISPR 1 de *S. thermophilus*. Les DR sont représentés par la couleur jaune et les spacers par des couleurs différentes. Les positions relatives de début et de fin des séquences sont affichées à droite et à gauche. Il s'agit de la représentation issue de CRISPRFinder (voir partie 4.1 page 64).

Le CRISPR (figure 2.2 et figure 2.3) se caractérise par la présence d'une série (au moins deux) de répétitions directes ou DR (Direct Repeat) de taille 23 à 47 pb, régulièrement espacées de séquences uniques de taille similaire appelées "spacers" (Nakata, 1989; Groenen, 1993; Mojica, 1995, 2000). En général, l'une des extrémités du CRISPR présente un DR dégénéré et l'autre extrémité est flanquée par une séquence riche en AT de taille allant de 8 jusqu'à plus de 500 pb appelée le "leader". Il est à noter que le leader est généralement défini comme étant la portion de séquence située entre le premier DR et le début de la première séquence codante. Cette définition ne permet pas de définir le "leader" en termes fonctionnels.

Dans le voisinage du CRISPR, on peut observer une famille de gènes appelés *cas* (Jansen, 2002b; Haft, 2005) codant pour des protéines dont certaines présentent des homologies avec les protéines du système d'interférence ARN chez les eucaryotes (Makarova, 2006).

Fig. 2.3 – Quelques exemples de CRISPR (représentation de CRISPRFinder (voir partie 4.1 page 64).

Une même bactérie peut posséder plusieurs systèmes CASS ou un seul système CASS associé chacun à un à plus de dix CRISPRs et un seul jeu de gènes *cas*.

Ce système assez curieux et riche en propriétés a commencé au cours des trois dernières années à intéresser les biologistes qui ont développé des applications à vocation épidémiologique et industrielle se basant sur son mode d'évolution et ses fonctions. Dans cette section, je vais décrire les éléments constituants le CRISPR, donner un aperçu sur son rôle et sur les applications qui peuvent en découler.

2.2.1 Les répétitions directes ou DR

Le DR est une séquence qui varie d'un CRISPR à un autre mais en conservant quelques propriétés communes. Il est toujours assez bien conservé au sein d'un même CRISPR : la taille est la même et la séquence est dans la plupart des cas conservée à 100% (figure 2.3). On observe parfois des différences entre les DR au niveau de quelques nucléotides dues probablement à des mutations ponctuelles, hormis bien sûr le DR dégénéré, situé à l'une des extrémités. Ce dernier peut présenter des dissemblances pouvant aller jusqu'à 50% des nucléotides de la séquence (Barrangou, 2007). Le DR, comme l'indique le terme "palindromic" dans l'acronyme CRISPR, se présente dans la plupart des cas sous une forme semi palindromique c'est à dire un palindrome court de 5 à 7 pb. Sa terminaison est caractérisée assez souvent par une signature particulière (Jansen, 2002a) : (C/G)AA(A)(C/G) pouvant constituer un site de reconnaissance de protéines Cas. Grâce à ces palindromes, la séquence DR peut se présenter sous forme d'une structure secondaire stable ("hairpin" (Jansen, 2002a; Kunin, 2007)). Kunin *et col.* ont montré que malgré la divergence des DR entre différentes espèces, ils peuvent être classés selon leur structure secondaire en 12 groupes principaux. Cette classification a été confirmée par Horvath *et col.* (Horvath, 2008b) qui ont trouvé la même agrégation aussi bien en se basant sur l'alignement des DR que sur celui des gènes *cas*, suggérant ainsi que le DR est spécifique d'un locus CRISPR donné fonctionnant avec un jeu particulier de *cas*. Ils ont ensuite remarqué que chacun des trois DR de *S. thermophilus* est associé à un cluster différent, ce qui est en accord avec une activité différente et indépendante pour chaque locus.

2.2.2 Les spacers

Comme l'indique le paragraphe précédent, le DR et les gènes *cas* sont associés. Un CRISPR donné peut être caractérisé par son DR. Le leader et les gènes *cas* sont figés et agissent avec ce DR. Les éléments qui vont différer au sein d'une même espèce dans le même CRISPR sont les spacers. Ces éléments sont uniques sauf quelques exceptions où l'on observe une duplication (Grissa, 2007b). D'ailleurs, les spacers peuvent être considérés comme des marqueurs pour l'identification des souches comme par exemple dans *M. tuberculosis* (Groenen, 1993), *S. thermophilus* (Horvath, 2008b), etc (voir paragraphe 2.3.3).

La description des spacers repose sur deux aspects importants :

1. leur origine : d'où proviennent ces petites séquences ?
2. leur évolution ou mode d'acquisition

En ce qui concerne l'origine, il est évident que la recherche de similarité entre les spacers et d'autres séquences donne un résultat biaisé puisqu'il est relatif à l'état actuel des bases de données de séquences. Cinq équipes ont publié des résultats de "blast" de spacers : Les travaux de Mojica *et col.* (Mojica, 2005), Pourcel *et col.* (Pourcel, 2005), Bolotin *et col.* (Bolotin, 2005) et Horvath *et col.* (Horvath, 2008b) (résumés dans le tableau 2.1) ainsi que le travail de Lillestøl *et col.* (Lillestøl, 2006) s'intéressant aux archées. Ces études montrent que les spacers reconnus sont en majorité d'origine extrachromosomale (essentiellement des séquences de phages ou prophages ou également de plasmides) et dérivent des deux brins du phage (codant et non-codant (Mojica, 2005; Lillestøl, 2006)). Ceci explique que seuls 2% des spacers soient identifiés dans (Mojica, 2005) ou (Bolotin, 2005) quand les spacers de génomes variés sont analysés, mais les données relatives aux procaryotes résultent majoritairement du séquençage de bactéries ou de phages d'intérêt épidémiologique ou industriel. D'ailleurs, le cas de *S. thermophilus* s'est révélé assez significatif comme le montre le tableau 2.1 ; parmi 349 spacers, 124 ont été retrouvés dans les banques (Bolotin, 2005) ou plus récemment 500 parmi 952 dans (Horvath, 2008b). En effet, il s'agit d'une bactérie importante dans la production de produits laitiers, ceci explique l'existence de grandes banques de données phagiques relatives à cette bactérie (huit génomes de phages isolés et séquencés). Cependant, il n'est pas évident de trouver dans les banques les séquences des phages attaquant des bactéries de moindre intérêt surtout si on tient compte de la prépondérance des phages dans la nature (Edwards et Rohwer, 2005). Barrangou *et col.* (Barrangou, 2007; Deveau, 2008; Horvath, 2008b) ont démontré que les spacers sont acquis chez *S. thermophilus* en réponse aux attaques virales auxquelles la bactérie est soumise. En effet, quand la bactérie possède un spacer exactement identique à une région du phage attaquant (proto-spacer d'après (Deveau, 2008)), elle devient immunisée contre ce phage. Le système CASS acquiert apparemment ce morceau du phage en reconnaissant une signature particulière à proximité du proto-spacer (AGAA pour le CRISPR1 et GGNG pour le CRISPR3 de *S. thermophilus*). Plus la bactérie possède de tels spacers, plus elle devient résistante (Bolotin, 2005; Deveau, 2008). Ainsi le CRISPR représente une sorte d'archive des rencontres de la bactérie. Cependant, comme une certaine proportion des spacers est d'origine chromosomale (cas de *Y. pestis* par exemple (Pourcel, 2005)), le CRISPR pourrait avoir une autre fonction que la résistance aux phages.

En ce qui concerne le deuxième aspect relatif au mode d'évolution du CRISPR, l'analyse d'un grand nombre d'isolats de *Y. pestis* a montré que ces éléments sont acquis de façon polarisée du côté du leader (Pourcel, 2005). Cette proposition a été confirmée ultérieurement par des observations chez *Sulfolobus solfataricus* (Lillestøl, 2006), *S. thermophilus* (Barrangou, 2007) et *Leptospirillum* (Tyson et Banfield, 2008). Il est à noter que la perte d'un ou plusieurs motifs de façon interstitielle est fréquente comme en témoigne surtout le CRISPR de *M. tuberculosis*. Ceci suggère que la recombinaison (voir paragraphe 2.1.3) est un événement très probable au sein du CRISPR favorisant ainsi les

Etude	Souches	Spacers uniques	Protospacers identifiés	Origine phagique	Origine plasmidique	Origine chromosomale
Mojica *et col.* (Mojica, 2005)	67 (génomes sur NCBI)	4500	88 (2%)	47 (54%)	10 (11%)	31 (35%)
Pourcel *et col.* (Pourcel, 2005)	109 (souches de *Y. pestis*)	45	32 (71%)	24 (75%)	0	8 (25%)
Bolotin *et col.* (Bolotin, 2005)	198 (génomes sur NCBI)	2156	44 (2%)	29 (66%)	0	15 (34%)
	24 (22 *S. thermophilus* + 2 *Streptoccocus vestibularis*)	349	124 (36%)	(75%)	(20%)	(5%)
Horvath *et col.* (Horvath, 2008b)	124 (souches de *S. thermophilus*)	952	500 (56%)	384 (77%)	80 (16%)	36 (7%)

Tab. 2.1 – Spacers analysés dans la littérature, recherche d'homologies avec des séquences connues

pertes (voire parfois les duplications) de motifs (Bolotin, 2005).

L'acquisition polarisée des spacers implique que la présence d'un spacer commun à une position donnée dans la succession des spacers, chez deux souches de la même espèce ou d'espèces voisines, renseigne sur un ancêtre commun à ces deux souches. Ainsi, le polymorphisme induit par le nombre et la nature des spacers chez une espèce bactérienne donnée fait de la structure CRISPR un outil original de différenciation de souches et éventuellement un outil phylogénétique pour les études intra-espèce (Pourcel, 2005).

2.2.3 Le leader

Le leader est une séquence riche en AT (Jansen, 2002b,a) de taille variable selon le CRISPR (entre 100 et 550 pb). Chez les archées, le leader le plus court recensé à ce jour, a une taille de 132 pb (*Haloarcula marismortui*) et le plus long est de taille 564 pb chez *Methanopyrus kandleri* (Lillestøl, 2006). La séquence leader est directement adjacente au premier DR du CRISPR (voir figure 2.1). C'est une séquence qui ne contient pas de phase ouverte de lecture (Jansen, 2002b; Tang, 2002) et qui diffère d'un CRISPR à un autre. Il semble qu'elle fait partie de la structure minimale nécessaire à la création d'un nouveau CRISPR à partir d'un autre sur le même génome étant donné que sur un même chromosome, si plusieurs CRISPRs ont le même DR, ils ont aussi le même leader (Lillestøl, 2006; Grissa, 2007b). La position du leader peut être déterminée par l'alignement des flanquantes de CRISPRs situés sur le même chromosome ou appartenant à deux espèces différentes mais possédant le même DR. Le leader sera défini comme étant la séquence commune entre ces flanquantes (voir annexe B). Dans certains cas, il est possible de définir les bornes du leader par l'identification de séquences connues à son voisinage (des protéines par exemple). D'ailleurs, c'est ainsi qu'on peut voir que certains CRISPRs sont dépourvus de leader. Par exemple, la souche *Aeropyrum pernix K1* possède trois CRISPRs sur son chromosome dont le deuxième (position : 786657-789355, DR : "GCATATCCCTAAAGGGAATAGAAAG") est flanqué par deux protéines Cas (à moins qu'une séquence codante ne joue également le rôle de leader).

Le rôle du leader est toujours inconnu. Cependant, il a été observé d'abord chez *Archaeoglobus fulgidus* (Tang, 2002) (22 petit ARN non messagers transcrits), ensuite chez *S. solfataricus* (Tang, 2005) et plus récemment chez *Myxococcus xanthus* (Viswanathan, 2007) que le locus CRISPR est transcrit en un long ARN processé par la suite en micro ARN (smRNA). Au vu de sa position à côté du dernier spacer acquis, le leader est soupçonné être le promoteur de la transcription. Lillestol *et col.* ont remarqué la présence d'une TATA box sur le leader du côté du premier DR (Lillestøl, 2006) pouvant jouer un rôle dans l'initiation de la transcription.

2.2.4 Les gènes *cas*

Les gènes *cas* forment un ensemble de gènes trouvés souvent dans le même ordre, se situant à proximité du CRISPR, à une distance moyenne de 1000 pb en tenant compte du leader (Haft, 2005). Ces gènes sont présents uniquement sur les génomes porteurs de cette structure, d'où leur nom "*cas*" pour "CRISPR associated" (Jansen, 2002b). Cependant, la réciproque n'est pas toujours vraie, un CRISPR peut être présent sur un génome sans les gènes *cas*. Tel est l'exemple des deux archées *Thermoplasma acidophilum* et *Pyrococcus abyssi* (Lillestøl, 2006) ou de la bactérie *L. monocytogenes* EGD-e. Sur certains génomes, un seul jeu de gènes *cas* semble être associé à plusieurs éléments de CRISPRs ayant le même DR et le même leader comme par exemple sur les archées *Pyrobaculum aerophilum*, *S. solfataricus*, *Sulfolobus tokodaii* et *H. marismortui* étudiés par Lillestøl et *col.* (Lillestøl, 2006). Cette cassette de gènes servirait à accomplir les fonctions assurées par le CRISPR (voir paragraphe suivant). Elle serait donc impliquée (Haft, 2005) dans la maintenance du locus CRISPR (Jansen, 2002b), la capture de nouveaux spacers et leur insertion dans le CRISPR (gain de motifs) (Mojica, 2005; Pourcel, 2005), dans la neutralisation d'un virus ou d'une séquence étrangère dont une séquence est présente sous forme de spacer (Barrangou, 2007), mais aussi la propagation d'une structure minimale de CRISPR (leader + DR) sur un même chromosome (Bult, 1996) (Grissa, 2007b) et le transfert horizontal du système CASS entier entre différentes espèces (Nelson, 1999; Makarova, 2002; Mongodin, 2005). Il est à noter que l'implication effective des gènes *cas* dans le fonctionnement du CRISPR a été prouvée par Barrangou *et col.* (Barrangou, 2007). Ces travaux ont démontré que chez *S. thermophilus*, le gène *cas5* joue un rôle dans l'immunité contre les phages et le gène *cas7* dans l'acquisition de nouveaux spacers. Le gène *cas2* a fait l'objet d'une analyse expérimentale qui a montré qu'il s'agit d'une ribonucléase (Beloglazova, 2008).

Le nombre des gènes *cas* identifiés à ce jour s'élève au total à 45 répartis en plusieurs groupes dont essentiellement un noyau composé de 6 gènes *1-6* (Haft, 2005), parmi lesquels *cas1* est présent dans presque toutes les cassettes. Pour cette raison, il a été utilisé pour le tracé d'arbres phylogénétiques mettant en évidence l'histoire évolutive des cassettes *cas*.

Les différents groupes de gènes ont été définis par Haft *et col.* en baptisant 53 modèles de Markov cachés (HMMs) déposés dans la base de données TIGRFRAMs (`www.tigr.org/TIGRFRAMs`). Les familles de protéines forment des groupes conservés dans différents génomes. C'est pourquoi elles ont été classées en dix types ("subtype") dont le noyau et le module RAMP pour "Repair Associated Mysterious Protein" (découvert par Makarova *et col.* sans faire le lien avec le CRISPR). L'acronyme a été conservé en remplaçant le terme "Repair" par "Repeat" pour montrer la corrélation avec les CRISPRs et ne pas faire allusion à la fonction de réparation qui est fausse. Les huit autres modules ont été nommés d'après le génome pour lequel il s'agit d'un locus CRISPR unique (exemple : Apern pour *Aeropyrum pernix*).

2.2.5 Le rôle du CRISPR

Le rôle du CRISPR était complètement ignoré lors de sa découverte, puis quelques hypothèses ont été avancées. La première (Mojica, 1995; She, 1998) suggérait que le CRISPR est impliqué, chez les archées, dans la répartition des copies de génomes au cours de la réplication car des similitudes avec certains mécanismes de partition chez les bactéries ont été remarquées par Mojica *et col.* chez *Haloferax mediterranei*. La deuxième hypothèse de perturbateur chromosomique est inspirée de la facilité des recombinaisons chez le CRISPR, sa localisation en particulier sur le génome de *Thermotoga maritima* et sa variabilité entre les souches qui corroborent l'hypothèse de l'implication de cette structure dans les réarrangements et la mobilité de l'ADN (Mongodin, 2005; DeBoy, 2006). De plus, le CRISPR aurait peut être un rôle dans la réplication du chromosome puisqu'il est associé à des mouvements de transfert latéral (Nelson, 1999; Makarova, 2002; Mongodin, 2005; Haft, 2005). La présence de la protéine SSB, connue pour son rôle dans la réplication, la recombinaison et la réparation de l'ADN, dans le CRISPR DRB de *C. diphtheriae* (Mokrousov, 2005) corrobore l'hypothèse d'un rôle du CRISPR dans le métabolisme de l'ADN.

Makarova *et col.* ont avancé en 2002, l'hypothèse de l'implication des gènes *cas* dans un système de réparation de l'ADN (Makarova, 2002), hypothèse rejetée ensuite (Makarova, 2006). L'unique hypothèse concernant le rôle du CRISPR qui soit prouvée expérimentalement est sa fonction immunitaire contre les agressions virales (Barrangou, 2007). Cette fonction serait probablement remplie par un système d'interférence ARN analogue à celui des eucaryotes (Makarova, 2006). Comme certains spacers ont pour origine de l'ADN chromosomique non-viral, il serait judicieux de penser que le CRISPR sert également à réguler certains gènes du chromosome (Sorek, 2008). Cette hypothèse n'a pas encore été testée et validée expérimentalement en dehors des observations faites par l'équipe de D. Lovely sur le CRISPR de *Pelobacter carbinolicus*. En effet, ce CRISPR comprend 111 spacers dont un est exactement identique à *hisS-1*, un gène histidyl-tRNA synthetase. Il a été difficile d'évaluer l'effet de la transcription du CRISPR sur les transcrits de *hisS-1* et la physiologie de *P. carbinolicus*. Cependant, plusieurs gènes riches en histidine semblent être perdus ou modifiés chez cette espèce par rapport aux espèces parentes comme *Desulfuromonas acetoxidans* ou *Geobacter sulfurreducens*, ce qui explique que *P. carbinolicus* n'oxyde pas l'acétate bien qu'elle possède un cycle tricarboxylique acide complet. Ce travail n'a pas été encore publié, il a été uniquement présenté en poster dans la conférence de la Société Américaine de Microbiologie en 2007 (Aklujkar, 2007), mais il s'agit de la première suggestion d'un effet du CRISPR sur l'évolution génomique.

La régulation du développement du corps de fructification chez *M. xanthus* a également été proposée (Thöny-Meyer et Kaiser, 1993; Boysen, 2002) comme rôle du CRISPR.

2.3 Historique : grandes étapes dans la compréhension du système CASS

La découverte et la compréhension du système CASS ont commencé par hasard en 1987 et ont été suivies de grandes étapes réalisées indépendamment par différentes équipes s'intéressant à des domaines variés.

2.3.1 Premières observations

1987 : La découverte

La structure CRISPR a été observée pour la première fois chez *E. coli* en **1987** (Ishino, 1987) en amont du gène *iap*. Elle a été décrite comme étant la succession de 5 DR de taille 29 pb séparés par des séquences de 32 nucléotides. La forme semi-palindromique du DR a été remarquée, mais aucune suggestion sur la fonction physique du CRISPR n'a été proposée. Deux ans plus tard, en **1989**, cette structure a été séquencée et décrite chez *E. coli* K-12 (Nakata, 1989) comme étant un arrangement inhabituel de séquences nucléotidiques sans aucune hypothèse sur son rôle. La présence du même DR a été signalée également chez *Shigella dysenteriae* et *Salmonella typhimurium*.

1993 : Observation chez les archées

En **1993**, 6 ans après la première découverte du CRISPR, Mojica *et col.* signalent la présence de cette structure dans le groupe des Euryarchaeota (Mojica, 1993),(Mojica, 1995), en particulier chez deux génomes de *H. mediterranei* et *H. volcanii*. Dans cette étude, le CRISPR a été vu comme étant un type particulier de répétition en tandem appelé *TREPs*. La présence du CRISPR chez les archées a été ensuite remarquée dans les plasmides de *Sulfolobus* (She, 1998; Greve, 2004).

2.3.2 La nomenclature CRISPR

Depuis sa découverte en 1987, le CRISPR fût désigné par différentes appellations selon l'équipe qui l'étudiait. Le premier acronyme **TREP** pour Tandem REPeat a été proposé par le groupe de Mojica (Mojica, 1995) en 1995. Les chercheurs qui s'intéressent plutôt

aux archées ont choisi l'appellation **SRSR** pour Short Regularly Spaced Repeats (Mojica, 2000; Peng, 2003). Cette appellation est toujours utilisée pour désigner les CRISPR chez les archées (Greve, 2004). Dans le complexe *M. tuberculosis*, l'appellation adoptée est **DVR** pour Direct Variable Repeat locus (van Embden, 2000). Le CRISPR a été également désigné **LCTR** pour Long Clusters of Tandem Repeats (She, 2001) ou **SPIDR** (Spacers Interspersed Direct Repeats) (Jansen, 2002a). Les auteurs de ce dernier article (Jansen *et col.*) se sont mis d'accord avec le groupe de Mojica *et col.* pour une nouvelle nomenclature (Jansen, 2002b) désignant cette famille de répétitions. L'appellation **CRISPR** comprend une description de toutes les caractéristiques de cette structure sans la regrouper avec une autre classe de répétitions telles que par exemple les SSRs (Short Sequence Repeats) communément présents chez les bactéries (van Belkum, 1998).

2.3.3 Le CRISPR du complexe *M. tuberculosis*

1991 : Le CRISPR du complexe *M. tuberculosis*, utilisation dans l'identification et la caractérisation de souches cliniques

Le complexe *M. tuberculosis* est constitué des espèces *M. bovis*, *M. tuberculosis*, *M. africanum* et *M. microti*. Les séquences transposables (van Soolingen, 1991) ont été beaucoup utilisées pour étudier la diversité génétique du complexe. En **1991**, lors de la caractérisation de la séquence IS987 de la souche BCG de *M. bovis* et de ses flanquantes dans le but d'examiner les raisons de son incapacité de transposition par rapport à l'élément IS6110, Hermans *et col.* (Hermans, 1991) ont découvert que l'IS987 était intégré au niveau du 30ème DR d'un CRISPR. Ce CRISPR, présent chez tous les membres du complexe *M. tuberculosis*, est constitué de répétitions directes de 36pb séparées par des spacers de 35-41 pb. L'appellation "DR region" a été attribuée au CRISPR, un motif DR + spacer étant appelé DVR pour "Direct Variable Repeat". Ces appellations sont toujours utilisées pour le CRISPR de *M. tuberculosis*. Dans ce premier article, les auteurs ont observé que le CRISPR présente un polymorphisme dû à la présence de séquences d'insertion chez les isolats du complexe *M. tuberculosis*, absentes chez les autres Mycobacterium. Ils ont conclu que le CRISPR représente un point chaud pour l'intégration d'éléments IS. De plus, on observe un polymorphisme dû à la perte de DVRs.

En **1993**, le même groupe de chercheurs propose cette région DR comme marqueur génétique assez polymorphe pour différencier les souches de *M. tuberculosis*. Un premier article publié en Août (van Soolingen, 1993) a permis de montrer que le typage par RFLP du DR possède un pouvoir discriminant plus important que quatre autres marqueurs étudiés. Dans un deuxième article publié en Décembre (Groenen, 1993), le séquençage des spacers a permis de proposer une nouvelle méthode de typage pour *M. tuberculosis* basée sur une seule PCR appelée direct variable repeat polymerase chain reaction (DVR-PCR).

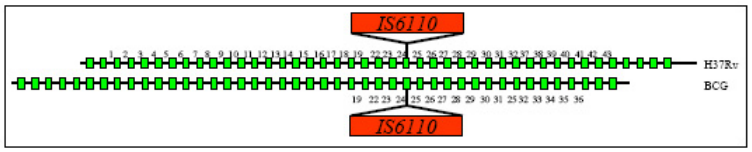


Fig. 2.4 – Structure du locus DR de *M. tuberculosis H37Rv* et *M. bovis BCG P3*. Les boites vertes désignent les DR et les boites rouges les séquences IS.

Cette technique fût un excellent outil de différenciation de souches de cette espèce à des fins épidémiologiques.

La région CRISPR de *M. tuberculosis* a eu par la suite un très grand succès auprès des épidémiologistes grâce à l'invention de la technique de spoligotypage (Kamerbeek, 1997) et des bases de données relatives.

1996 : Le spoligotypage

Le spoligotypage ou "spacer oligotyping" (Kamerbeek, 1997), créé en **1996** pour le typage épidémiologique de *M. tuberculosis*, utilise les DRs comme cible d'amplification de l'ADN. Une membrane d'oligonucléotides est créée à partir de 43 spacers dérivés de *M. tuberculosis* H37Rv (37 spacers) et *M. bovis* BCG (6 spacers supplémentaires). L'ADN amplifié est directement utilisé pour une hybridation sur cette membrane (voir figure 2.5). Le profil obtenu de présence/absence des spacers représente le spoligotype qui peut être différent d'un isolat à l'autre, ce qui permet de distinguer les souches entre elles. Cette méthode est rapide (moins de deux jours pour l'appliquer à des échantillons cliniques), reproductible, facile à mettre en place, comparable entre laboratoires (il suffit de comparer un mot de 43 lettres formé de a pour absence du spacer, b pour sa présence et i pour indéterminé dans le cas des faibles réactions) et possède un pouvoir de résolution important dans *M. tuberculosis* grâce à la présence du CRISPR dans toutes les souches de cette espèce et au polymorphisme de présence-absence des spacers.

Le spoligotypage a été par la suite validé pour le typage de souches de *M. bovis* (Aranaz, 1996), pour la détection rapide de transmission nosocomiale (Goyal, 1997), pour le clustering comparé à d'autres techniques (Sola, 1998) et pour le typage de *M. microti* (van Soolingen, 1998). Depuis, le spoligotypage est devenu une technique de référence dans le typage du complexe *M. tuberculosis*, utilisée surtout pour caractériser les grandes familles phylogénétiques ou clades. Plusieurs équipes de recherche appliquent le spoligotypage en supplément à d'autres méthodes de typage. L'aspect le plus intéressant pour cette technique est que depuis sa première application, tous les spoligotypes obtenus sont recueillis pour alimenter des bases de données. La coordination de cette collecte a été le fait d'une équipe française de l'institut Pasteur de Guadeloupe qui a recueilli des données

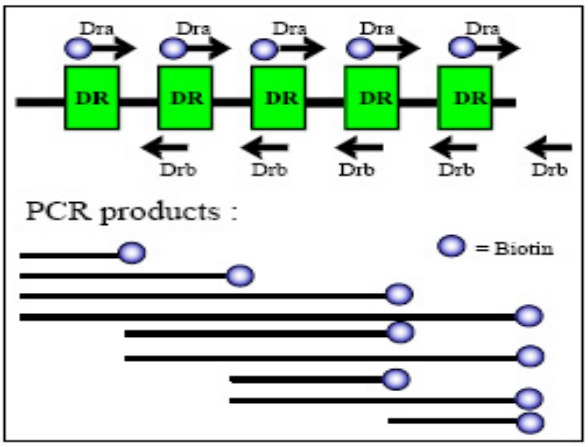


FIG. 2.5 – Amplification du DR pour le spoligotypage.

de multiples équipes dans le monde pour créer la première version (Filliol, 2002) d'une base internationale regroupant 11.708 spoligotypes. La dernière mise à jour de cette base a été faite en 2006 (version Spoldb4 (Brudey, 2006)). Cette base rend possible un travail de fouille de données permettant d'identifier différents groupes ou "grappes" de profils de spoligotypes provenant de 140 pays différents.

Le CRISPR de *M. tuberculosis* semble uniquement perdre des motifs de manière interstitielle. L'ordre des spacers est conservé dans les allèles qui ont été séquencés sans brassage et duplication des motifs.

2.3.4 Les gènes *cas* à proximité du locus CRISPR

La comparaison des gènes qui se situent à proximité de la structure CRISPR a parfois montré une homologie forte entre espèces éloignées (Jansen, 2002b). La première analyse de ces gènes faite en **2002** a révélé la présence d'un groupe de quatre gènes (appelés gènes *cas* pour gènes associés au CRISPR) numérotés de un à quatre. Ces gènes sont absents des génomes ne contenant pas un CRISPR. Le gène *cas1* est le plus fréquent dans les génomes analysés par Jansen *et col.*. Les quatre protéines Cas en question correspondent à quatre protéines appartenant chacune à un COG "Cluster of Orthologous Groups of proteins". La protéine Cas1 correspond à COG1518, Cas2 correspond à COG1343, Cas3 correspond à COG1203 et Cas4 correspond à COG1468.

Certaines des protéines Cas semblent être capables d'interagir avec l'ADN comme le montre la présence de motifs caractéristiques de ces protéines. Cas1 serait impliquée dans la réparation de l'ADN, Cas3 serait une hélicase (Jansen, 2002b) et Cas5 serait une nucléase car elle contient un motif du type HNH spécifique des nucléases.

Les gènes *cas* ont fait l'objet d'études ultérieures qui ont étendu le nombre de quatre gènes à vingt, dans un premier temps en **2002** (étude de Koonin *et col.* sur 40 génomes (Makarova, 2002)). L'étude la plus exhaustive a été réalisée en **2005** sur plus de 200 génomes (Haft, 2005) par un groupe américain du TIGR (The Institute for Genomic Research) montrant que la liste des gènes *cas* peut s'étendre jusqu'à 45 qu'ils ont classés en dix catégories.

Plus récemment, en **2008**, Beloglazova *et col.* (Beloglazova, 2008) ont fourni la première étude expérimentale d'un gène *cas*. Il s'agit de la caractérisation biochimique et structurale du gène *cas2* chez cinq procaryotes. Ils ont démontré que les protéines Cas2 forment une famille de ribonucléases spécifiques des ARN simple brin qui servirait à la dégradation de l'ARN phagique.

2.3.5 Gènes *cas* et transfert horizontal

Des observations sur la présence de DR similaires entre des espèces bactériennes éloignées renseignent sur la possibilité d'acquisition des CRISPRs par transfert horizontal. Un exemple est celui de *E. coli* et *M. avium* (Jansen, 2002b) portant des CRISPRs dont le DR est TGCTCCCCGCGCAAGCGGGGATGAACC. Depuis 2002, Makarova *et col.* (Makarova, 2002) ont suggéré par une étude phylogénétique que l'ensemble des gènes *cas* est transféré horizontalement entre des génomes taxonomiquement éloignés. En **2006**, Godde *et col.* (Godde et Bickerton, 2006) ont confirmé cette proposition et ont conclu que ce transfert est effectué par conjugaison. Dans ce travail, partant du principe que les gènes *cas* fonctionnent ensemble en une cassette entière (Makarova, 2002), les auteurs ont concaténé les séquences d'acides aminés des quatre gènes *cas* (*cas1*, *cas2*, *cas3*, *cas4*) et les ont alignés pour construire un arbre phylogénétique. Sur l'arbre obtenu, les grands groupes phylogénétiques ne forment pas des clades distincts mais sont mélangés dans l'arbre, ce qui donne une preuve du transfert horizontal des gènes étudiés. De plus, les auteurs montrent que les CRISPRs se propagent horizontalement chez les procaryotes via les mégaplasmides qui sont des visiteurs passagers, pas très stables dans leur hôte. Cette hypothèse est suggérée par quelques exemples de mégaplasmides portant un CRISPR (*Thermus thermophilus* HB8 et HB27 par exemple) et surtout par l'analyse de données de séquençage d'échantillons bactériens de l'environnement provenant de la mer des Sargasses (Venter, 2004).

Par ailleurs, un CRISPR a été également identifié sur un prophage (appartenant à *Clostridium difficile*) et il a été suggéré que le phage utiliserait le CRISPR pour limiter la dispersion des phages concurrents (Sebaihia, 2006).

Chez les archées, Lillestøl *et col.* (Lillestøl, 2006) ont remarqué un mouvement de transfert latéral pour deux complexes CASS entre *M. acetivorans* et *M. barkeri* puisqu'ils ont un jeu de gènes *cas* ayant exactement le même ordre (*cas4*-*cas1*-*cas2*-CRISPR1-

csa2-*cas3*-*cas5*-CRISPR2) et une conservation des séquences à 94%. De plus, les DRs du CRISPR 1 et 2 sont différents mais conservés entre les deux organismes. Aucun spacer n'est commun.

L'hypothèse du transfert latéral a été récemment proposée également chez deux populations de *Leptospirillum* par Tyson et Banfield (Tyson et Banfield, 2008) lors d'une étude de cette espèce dans les biofilms microbiens. De plus, Mongodin *et col.* (Mongodin, 2005) pensent que le transfert horizontal entre les bactéries et les archées illustré dans la bactérie hyperthermophilique *T. maritima* par un transfert de gènes important est associé à la présence du CRISPR.

2.3.6 2005 : Origine extrachromosomale des spacers, acquisition polarisée et rôle suggéré du CRISPR

L'apparition de trois publications importantes traitant des CRISPR en **2005** et les observations qui y sont faites ont déclenché un vrai intérêt pour la structure CRISPR et ses applications.

Le principal résultat rapporté par ces travaux consiste à remarquer que les spacers sont acquis à partir d'éléments génétiques préexistants d'origine extrachromosomale. L'article publié par le groupe espagnol de Mojica *et col.* (Mojica, 2005) a analysé 4.500 spacers appartenant à 67 souches de procaryotes. Cette analyse a montré que 88 de ces spacers sont similaires à des séquences de bactériophages, ou de plasmides ou d'ADN chromosomal. Les auteurs ont proposé le rôle d'immunité spécifique conférée par les spacers puisqu'ils ont remarqué la relation entre la présence de spacers similaires à des séquences du virus SIRV et la résistance de *S. solfataricus* à ce virus alors que les souches dépourvues de ces spacers y sont sensibles.

Le deuxième article (Pourcel, 2005) est français et plus précisément, c'est le produit de recherches de notre laboratoire (GPMS), spécialiste des répétitions en tandem, sur la variabilité génétique de *Y. pestis*. En effet, l'un des CRISPRs de cette espèce a été rencontré par hasard et considéré comme étant une répétition en tandem à 60% d'homologie. C'est le CRISPR YP1 (positions 2.769.301-2.769.820 sur la souche *CO92*), qui a été étudié tout d'abord comme étant un marqueur VNTR "yp2769ms06" (Pourcel, 2004; Le Flèche, 2001). Par la suite, C. Pourcel *et col.* se sont intéressés à l'étude de cette "répétition en tandem" particulière. Ils ont alors étudié les trois CRISPRs de *Y. pestis* et *Y. pseudotuberculosis* en séquençant les spacers correspondants de 109 allèles, ce qui a permis de discuter deux aspects importants. Le premier aspect est venu de l'observation que certains spacers correspondent à des séquences d'un prophage ou de gènes situés sur une autre région du génome. Il a été également suggéré que le CRISPR représente une mémoire des agressions génétiques grâce à l'exploration des CRISPRs de *Streptoccocus*

pyogenes où l'on observe une corrélation entre la présence d'un CRISPR et l'absence d'un prophage particulier. Le deuxième aspect consiste à proposer un modèle d'évolution de cette structure basé sur trois propriétés :

1. possibilité de délétions de un ou plusieurs motifs de façon aléatoire et interstitielle,
2. acquisition polarisée de nouveaux motifs du côté du leader,
3. la présence de spacers identiques renseigne sur un ancêtre commun.

Le troisième article (Bolotin, 2005) est aussi le fruit du travail d'un groupe de chercheurs français de l'INRA portant sur l'espèce *S. thermophilus* et *S. vestibularis*. Cet article rejoint les deux autres sur l'observation d'une corrélation entre la présence de spacers provenant d'un ADN étranger et la résistance à certains phages. Bolotin *et col.* ont remarqué que pour les spacers de *S. thermophilus* ayant une homologie avec d'autres séquences dans la base de données de NCBI, 75% proviennent de phages (les phages de bactéries lactiques étant les plus connus dans la littérature étant donné le danger qu'ils représentent pour certaines industries alimentaires (Desiere, 2002)), 20% proviennent de plasmides et 5% sont liés à des gènes du chromosome contenant le CRISPR. Ils ont également remarqué que plus le nombre de spacers provenant d'un phage donné est grand, plus la souche possédant ces spacers résiste à ce phage.

2.3.7 L'interférence ARN comme modèle de fonctionnement

Le mécanisme d'interférence ARN (RNAi) est bien caractérisé chez les eucaryotes où il constitue un système de défense contre les ARN de virus et les éléments transposables (voir paragraphe 2.1.1). Plusieurs observations chez les CRISPRs comme la transcription en petits ARNs, la présence de spacers uniques homologues à des séquences extrachromosomales et la présence des protéines Cas sont à la base de l'hypothèse du fonctionnement de ce système de défense de façon analogue à l'interférence RNAi. La première démonstration de l'existence des petits ARNs à partir d'un grand transcrit digéré a été fournie par Tang *et col.* (Tang, 2002). Makarova *et col.* ont étudié les séquences des protéines Cas en 2002 (Makarova, 2002) puis en 2006. Après la découverte de l'homologie entre les spacers et des gènes de phages et de plasmides, ainsi que la mise en évidence des petits ARNs par d'autres équipes, Makarova *et col.* proposent un modèle de mécanisme siRNA chez les procaryotes ou psiRNA, analogue et non homologue à celui des eucaryotes (Makarova, 2006). En effet, les auteurs ont essayé de montrer l'analogie entre les protéines composant le système d'interférence ARN des eucaryotes (en particulier le RISCs) et les fonctions prédites des protéines du système CASS. Les trois composants essentiels du RISCs sont associés comme suit aux protéines Cas : le "dicer" correspondrait à l'hélicase COG1203 (Cas3) appelé p-dicer pour prokaryotic dicer, le "slicer" correspondrait soit à la protéine COG1468 (Cas4) codant pour une nucléase RecB soit à la protéine COG1857, tandis que les protéines RNA-binding du système RISCs eucaryote correspondraient aux protéines du module RAMP. Le mode d'action proposé consiste en la transcription des motifs DR-spacer en ARN longs, et cet ARN adopterait une structure secondaire. Ensuite, le p-dicer

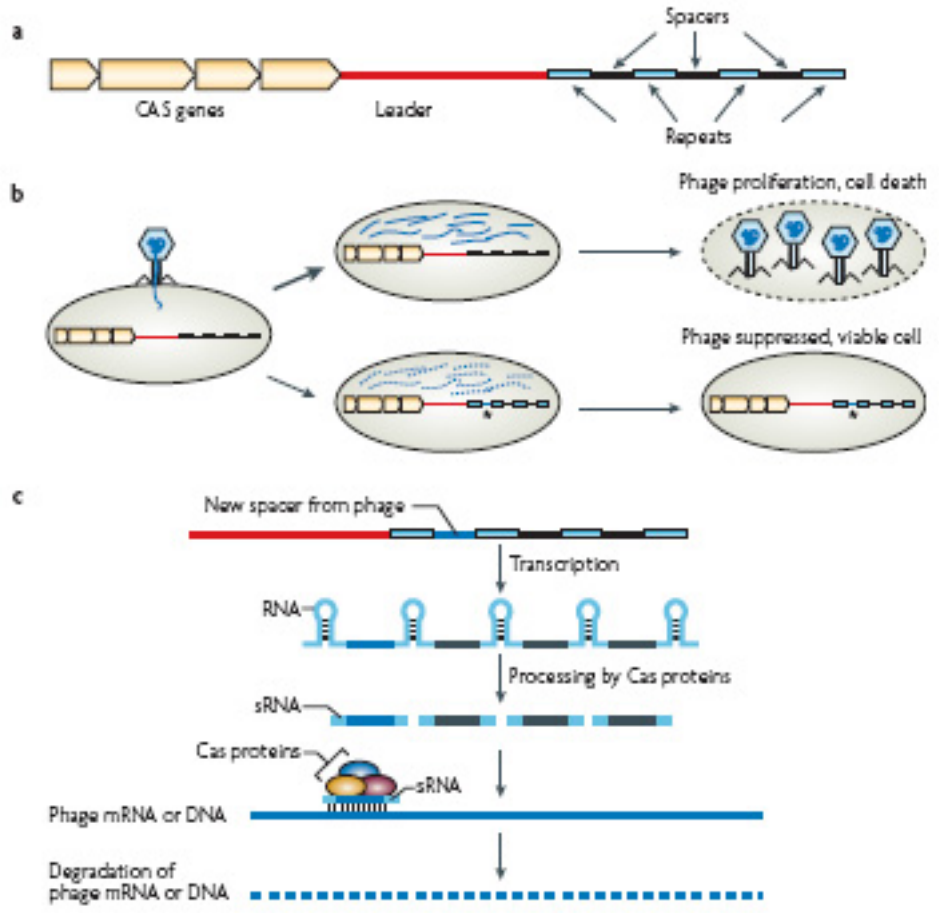


FIG. 2.6 – Modèle d'action du CRISPR (c) (d'après (Sorek, 2008))

reconnaîtrait cette structure et la transformerait en petits ARN (sRNAs), contenant chacun un spacer et une moitié de DR (Sorek, 2008). Le p-slicer agirait par la suite par un appariement entre les sRNAs et les séquences de phages menant à leur dégradation ou à un arrêt de la traduction des ARNm du génome étranger. C'est ce blocage qui réduirait ou annulerait le pouvoir infectieux des phages ou des plasmides.

Une étude plus récente (Beloglazova, 2008) a montré expérimentalement que la protéine Cas2 est une endonucléase et a suggéré qu'elle représente l'analogue du "slicer" chez les eucaryotes.

2.3.8 Première preuve expérimentale du rôle du système CASS

Après les suggestions sur le rôle immunitaire du CRISPR, la première preuve irréfutable de cette hypothèse a été fournie par un groupe de chercheurs français et américains

de Danisco, l'un des géants de l'industrie des ferments lactiques, en collaboration avec des chercheurs canadiens de l'université Laval au Quebec. Cette preuve expérimentale (Barrangou, 2007) a été fournie en **2007** sur la base du modèle de *S. thermophilus* contenant trois CRISPRs dont un, CRISPR1 (positions : 649.125-650.217, cf figure 2.2), est très actif. Les auteurs ont infecté des bactéries par deux phages différents et ont obtenu neuf clones résistants. Le séquençage des CRISPRs de ces bactéries a révélé que les mutants résistants ont acquis indépendamment entre un et quatre nouveaux motifs à proximité de la séquence leader et que les spacers proviennent du phage utilisé dans l'infection. De plus, l'insertion d'un spacer dans un CRISPR ou sa délétion agissent directement sur la sensibilité de la bactérie au phage en question. Ainsi, ils ont confirmé les deux hypothèses précédentes : le CRISPR est un système de défense contre les agressions virales et l'acquisition de nouveaux spacers se fait de façon polarisée du côté du leader. Des observations plus récentes chez des souches de *Leptospirillum* dans les biofilms microbiens ont été faites par Tyson et Banfield (Tyson et Banfield, 2008) montrant que le CRISPR de cette espèce, soumise à une infection phagique importante, évolue très rapidement. Ils montrent surtout que les spacers les plus anciens, partagés par les deux groupes étudiés de *Leptospirillum*, se situent au début du CRISPR alors que les spacers spécifiques de chaque groupe sont plutôt situés du côté du leader, ce sont les spacers acquis plus récemment.

Barrangou *et col.* (Barrangou, 2007) ont également prouvé qu'un seul nucléotide non identique entre le spacer et le proto-spacer (Deveau, 2008) suffit à inhiber le pouvoir protecteur conféré par ce spacer. C'est d'ailleurs ainsi qu'évoluent les phages qui contournent la résistance de la bactérie. Deveau *et col.* (Deveau, 2008) ont mis en évidence la présence d'une signature particulière (motif spécifique à chaque CASS) sur le génome phagique à proximité du proto-spacer. Barrangou *et col.* ont également démontré que les gènes *cas* ont bien un rôle dans le fonctionnement du système CASS. En effet, l'inactivation du gène *cas5*, contenant d'ailleurs un motif endonuclease, a pour conséquence la perte de la résistance même en présence de spacers dérivés du phage attaquant. L'inactivation du gène *cas7* rend impossible l'acquisition de nouveaux motifs DR-spacer. De plus, chaque structure CRISPR semble avoir son jeu de gènes *cas* associés (Horvath, 2008b) spécifiques, ce qui va dans le même sens que le clustering réalisé par Kunin *et col.* (Kunin, 2007) selon la structure du DR.

2.3.9 Le CRISPR dans la coévolution entre les procaryotes et les virus

Le CRISPR, ou plus généralement le système CASS, confère aux procaryotes une immunité acquise contre les agressions virales. Or, les virus évoluent de façon parallèle pour échapper à ce mécanisme de défense. Les travaux de Andersson et Banfield (Andersson et Banfield, 2008) étudient la coévolution des virus et de leur hôte au niveau des communautés de type biofilm. Cette étude a montré que les bactériophages sont capables de muter

pour déjouer ce système de défense bactérien. Ces mutations concernent les proto-spacers permettant ainsi au virus d'échapper à la résistance bactérienne. Les virus recombinent entre eux, ce qui conduit à la formation de spacers différents mais sans induire des changements dans les gènes contenant les proto-spacers. La recombinaison peut modifier des acides aminés mais n'affecte pas l'expression du gène. Lorsque les spacers du CRISPR deviennent inefficaces dans la lutte contre les agresseurs, celui-ci acquiert un nouveau morceau d'ADN du phage sous forme d'un nouveau spacer. Cet article (Andersson et Banfield, 2008) montre qu'au sein de la communauté étudiée, l'évolution est tellement rapide que tous les spacers (à l'exception d'un seul) sont différents au bout de cinq mois, c'est à dire que les CRISPRs analysés au mois de Juin (Tyson et Banfield, 2008) ont perdu tous leurs spacers et ont acquis de nouveaux quelques mois plustard (identifiés au mois de Novembre 2007).

La reconstruction (partielle et parfois totale) des génomes viraux et l'identification des spacers présents dans les génomes bactériens voisins a permis aux auteurs de cet article d'identifier les hôtes de certains virus et surtout de conclure qu'un virus possède un seul type d'hôte alors que ce dernier peut être attaqué par plusieurs virus. Ils ont ainsi reconstruit cinq populations de virus pour quatre archées et une bactérie.

Cette étude présente la première analyse faite sur des souches environnementales et il confirme la proposition du transfert latéral des CRISPRs via les plasmides.

2.4 Applications liées au CRISPR

2.4.1 Typage bactérien et phylogénie intra-espèce

Le typage bactérien permet de reconnaître deux souches identiques dans des cas de contaminations ou d'épidémie et d'identifier des complexes clonaux. L'utilisation du polymorphisme des CRISPRs pour le typage est envisageable uniquement chez des espèces jeunes telles que *M. tuberculosis* ou *Y. pestis* ou dans des complexes clonaux comme au sein de *P. aeruginosa* par exemple.

Par contre, la phylogénie consiste à comprendre la filiation des souches détectées de nos jours et rechercher les ancêtres probables. La reconstruction phylogénétique à l'intérieur d'une espèce donnée peut être faite à l'aide du CRISPR en utilisant les propriétés et la composition de ses spacers. En effet, les spacers sont ajoutés dans le CRISPR de façon polarisée, et tous les spacers d'un organisme gardent le même ordre sur la séquence génomique qui les porte. C'est ainsi que les spacers nouvellement acquis sont toujours situés à proximité du leader (cas de *Y. pestis* par exemple) et que lorsque l'ensemble de spacers chez une espèce est connu, leurs positions relatives sont également connues (cas

des 43 spacers connus de *M. tuberculosis*). Ainsi, il suffit d'observer la nature, le nombre et l'ordre des spacers sur différents isolats d'une espèce pour pouvoir prédire son histoire évolutive : les nouveaux spacers acquis dans le premier cas et les spacers perdus dans le deuxième renseigneront alors sur les liens de parenté entre les différents isolats. En effet, lorsque le CRISPR existe et présente assez de polymorphisme chez les individus étudiés, il peut constituer un assez bon outil phylogénétique. Cependant, le fait qu'il y ait un "turn-over" très important dans certaines espèces efface les traces d'évolution (cas des biofilms acidophiliques par exemple (Andersson et Banfield, 2008)).

Ci-dessous seront cités les cas décrits dans la littérature pour ce genre d'analyse.

Les bactéries lactiques

Chez certaines bactéries, le CRISPR peut constituer un élément d'identification de souches. Russel *et col.* (Russell, 2006), des chercheurs de Danisco, ont déposé un brevet sur la détection et le typage de certaines bactéries lactiques, les *Lactobacillus*, en utilisant les propriétés du CRISPR. Ce brevet fournit par exemple quelques outils pour différencier les espèces de *Lactobacillus* tels que des amorces permettant d'amplifier par PCR uniquement les souches de *Lactobacillus acidophilus* car le CRISPR de cette espèce contient une région qui lui est spécifique. Cette invention permet également de distinguer des souches de *Lactobacillus delbrueckii ssp. bulgaricus* par l'amplification de son CRISPR dans une première étape et par la digestion du produit obtenu dans une deuxième étape. L'analyse des fragments obtenus renseignera sur la nature de la souche.

Le cas de *M. tuberculosis*

La première application liée au CRISPR fût le spoligotypage de *M. tuberculosis* (Kamerbeek, 1997) (voir paragraphe 2.3.3). Cette technique est l'une des méthodes standard utilisées dans le génotypage de cette espèce, permettant essentiellement de classer rapidement des souches, d'intérêt épidémiologique en général, en différentes clades. En effet, le CRISPR de ce complexe est soupçonné de ne plus évoluer que par des pertes de motifs. C'est pourquoi les motifs présents (parmi les quarante trois cherchés par spoligotypage) dans une souche permettent d'identifier la lignée à laquelle elle appartient. Les familles les plus importantes seraient les familles Central Asian (CAS), Beijing, East African Indian (EAI), Haarlem (fréquentes aux Pays-Bas) et Latin American and Mediterranean (LAM). Chacune des ces familles présente une signature spécifique liée à la présence ou l'absence de certains spacers. Notons, par exemple, que la famille Beijing n'a gardé que les neuf derniers spacers en raison d'une large délétion probablement causée par un élément IS. Elle a le spoligotype suivant (nomenclature binaire) :
spoligotype Beijing : **000000000000000000000000000000000000111111111**

La figure 2.7 montre un dendrogramme construit à partir des spoligotypes de 105 isolats cliniques de *M. tuberculosis* indiennes de Delhi.

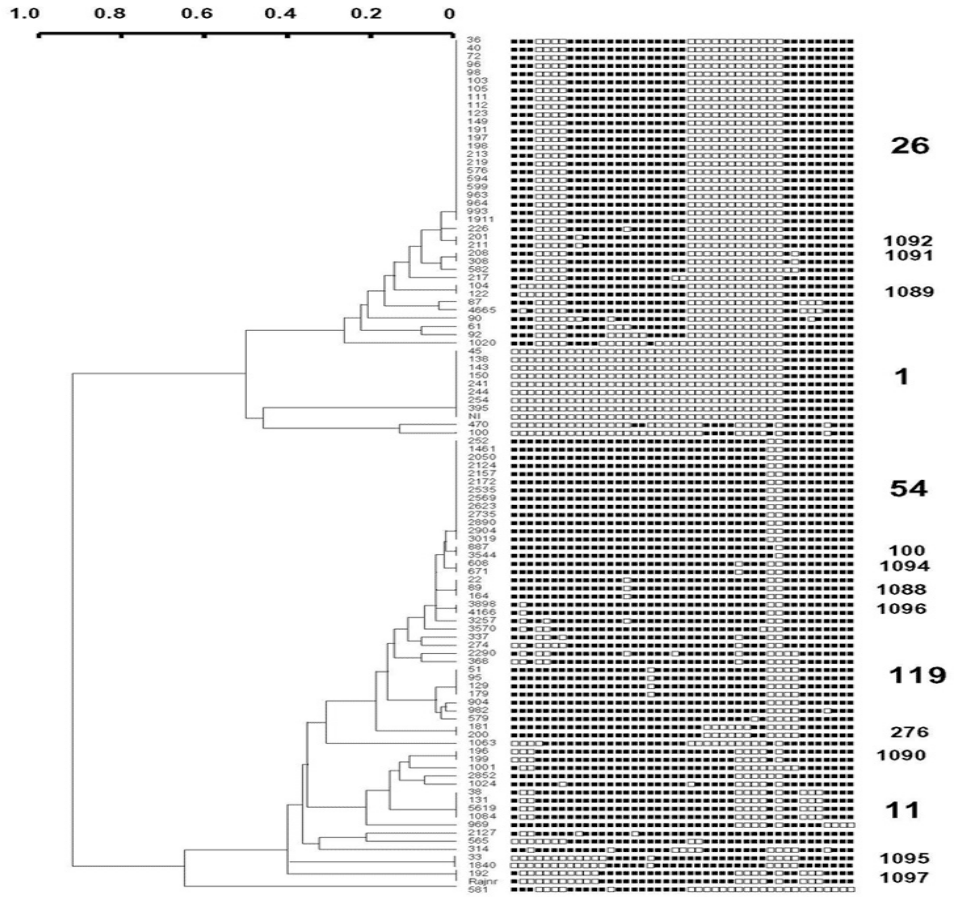


Fig. 2.7 – Dendrogramme de 105 isolats cliniques de *M. tuberculosis* en Inde (Singh, 2004)

Cette agrégation montre la présence de cinq clades principales à Delhi, **1** : la famille Beijing, **26** :les CAS1, **54** : les T1, **119** : famille X1 et **11** : EAI3.

Dans la dernière version de Spoldb4 (Brudey, 2006), 62 clades ont été définies en se basant sur une approche bioinformatique exploitant les profils de spoligotypes enregistrés dans cette base.

Le cas de *Y. pestis* et *Y. pseudotuberculosis*

Y. pestis est un clone récent de *Y. pseudotuberculosis* (Achtman, 1999). Son âge est évalué entre 1.500 et 20.000 ans et il ne présente pas beaucoup de variations avec les techniques de typage telles que MLST. Le typage MLVA possède un pouvoir discriminant important. On distingue par des tests biochimiques trois biovars dans l'espèce *Y. pestis* :

le biovar antiqua soupçonné d'être le plus ancien caractérisant les souches venant d'Asie et d'Afrique, le biovar medievalis et le biovar orientalis associé à la dernière pandémie commencée au XIX^e siècle. Ce pathogène humain est étudié phylogénétiquement par plusieurs équipes dans le monde dans le but de mieux comprendre la structure et la dynamique de cette population. Le CRISPR peut effectivement jouer un rôle dans cette détermination comme le montrent Pourcel *et col.* (Pourcel, 2005). Ils ont, en effet, utilisé le typage du CRISPR pour apporter un supplément d'informativité à la technique MLVA (voir figure C.1) et ont exploité ses propriétés d'évolution unidirectionnelle pour proposer une hiérarchie de l'évolution de *Y. pestis* (figure 2.8).

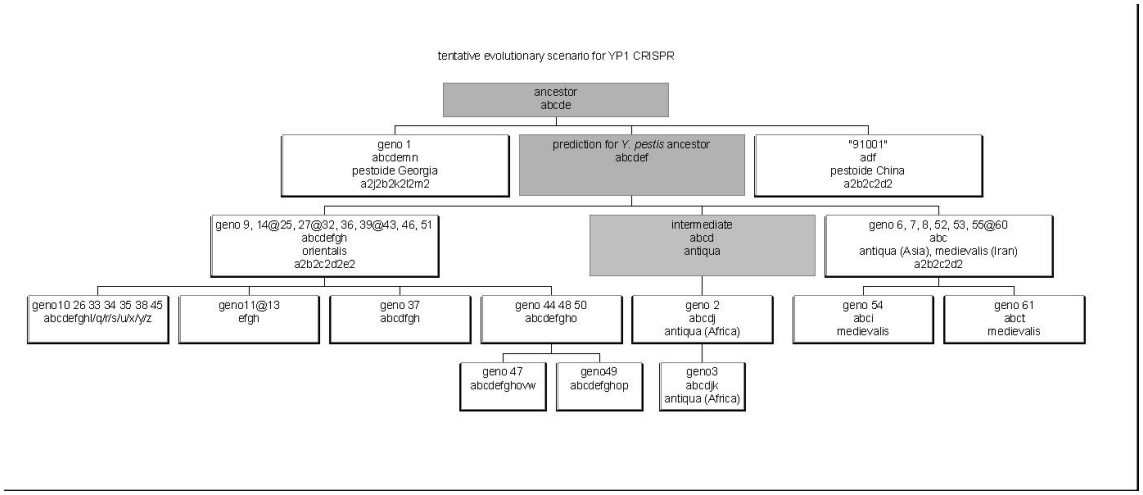


Fig. 2.8 – Scénario évolutif des allèles CRISPR de *Y. pestis* (Pourcel, 2005)

Le cas de *C. diphtheriae*

L'épidémie de diphtérie dans les années 90 en Russie a causé plus de 4.000 décès parmi 140.000 personnes infectées. Cette maladie est causée par *C. diphtheriae* qui, depuis cette date, a fait l'objet de plusieurs études de typage. Les méthodes employées sont essentiellement PFGE, RAPD, MLEE (voir 1.3) qui sont des techniques fastidieuses dont les résultats ne sont pas échangeables entre laboratoires. Or, *C. diphtheriae* contient deux CRISPRs (les locus DRA et le DRB), c'est pourquoi l'équipe de I. Mokrousov, en s'inspirant du spoligotypage de *M. tuberculosis*, a développé la même technique pour cette bactérie (Mokrousov, 2005, 2007). Dans un premier temps, le spoligotypage de *C. diphtheriae* a été réalisé en utilisant 21 motifs du DRB (des oligonucléotide correspondant à 21 spacers) et dans un deuxième temps, le pouvoir de discrimination de la méthode a été augmenté par l'ajout de 6 motifs du locus DRA (Mokrousov, 2007). Ces deux études ont montré l'intérêt du CRISPR, comparé à d'autres techniques, dans les investigations épidémiologiques et la reconstruction phylogénétique dans le cas de *C. diphtheriae*.

Le cas de *C. jejuni*

C. jejuni est la cause principale des maladies intestinales humaines d'origine bactérienne dans les pays industrialisés, via la consommation de viande contaminée, essentiellement la viande de volaille. Dans le cas de cette espèce, la méthode standard de génotypage était le RFLP, coûteux, fastidieux et non comparable entre laboratoires. Plusieurs méthodes ont alors été proposées en alternative parmi lesquelles figure le typage d'un CRISPR (Schouls, 2003). En effet, Schouls *et col.* ont remarqué un polymorphisme important au sein du CRISPR de *C. jejuni* : 170 spacers différents sur 137 souches contenant de un à sept motifs. Ils ont conclu que le CRISPR de cette espèce peut surtout être utilisé comme méthode de génotypage complémentaire à d'autres techniques car le quart des souches étudiées ne contient pas de spacer (aucun ou un seul DR). Le typage d'un CRISPR permet très bien de différencier les souches ayant des spacers, mais d'autres techniques sont nécessaires pour les souches n'en contenant pas (Fouts, 2005). Plus récemment, Price *et col.* (Price, 2007) ont introduit la technique CRISPR HRM (high-resolution melt) pour *C. jejuni* pour améliorer la technique de typage CRISPR.

Le cas de *Thermotogales*

T. maritima MSB8 est une bactérie présentant huit CRISPRs sur son génome. Le typage des CRISPRs au sein de cette espèce s'est révélé intéressant pour l'aspect phylogénétique. Une étude du CRISPR R1 (positions 412-310) (Mongodin, 2005) a permis d'identifier un polymorphisme important dans le nombre de spacers entre les différentes souches de *Thermotoga*, ce qui permet de considérer cette région comme marqueur génétique permettant de prédire la distribution globale des souches. Plus récemment, Deboy *et col.* (DeBoy, 2006) ont analysé cinq locus CRISPR pour classifier des souches de *T. neopolitana* en trois groupes, déduire des relations de parenté entre différents groupes et avancer l'hypothèse de la variation des souches selon la localisation géographique.

Le cas de *S. pyogenes*

Une étude de Hoe *et col.* (Hoe, 1999) s'est intéressée à l'investigation du CRISPR de *S. pyogenes* via l'analyse de 44 isolats. Cette étude a montré que le CRISPR est également utile pour cette espèce dans le clustering et la classification des souches, mais insuffisant pour un génotypage performant des souches (uniquement 13 génotypes différents).

2.4.2 Protection contre les agressions virales

Le CRISPR étant un outil de lutte contre les attaques virales via la composition de ses spacers, il permet alors de prédire le lysotype, c'est à dire connaître la sensibilité et le degré de résistance d'une bactérie à différents types de bactériophages et par la suite avoir une idée sur sa capacité de survivre si elle est soumise à certains phages.

L'application la plus importante de cette propriété du CRISPR est son utilisation pour protéger les cultures de bactéries lactiques dans les chaînes de production industrielles. La démonstration de ce rôle a été faite avec le modèle de *S. thermophilus* et jusqu'à présent, la seule application connue est proposée par Horvath *et col.* dans un brevet (Horvath, 2007) sur ce modèle. En effet, le génome de *S. thermophilus LMD 9* contient 3 locus CRISPR présentant respectivement 16, 3 et 8 motifs et un ensemble de quatre gènes *cas*. Le brevet en question fournit des cultures bactériennes plus résistantes et des outils pour les protéger. Les bactéries "résistantes" peuvent être générées soit naturellement en exposant les bactéries à plusieurs phages ou par manipulation génétique en ajoutant, supprimant ou modifiant des spacers ou même un CRISPR complet. Les outils fournis consistent par exemple en l'utilisation d'un ensemble de gènes ou de protéines Cas pour moduler la résistance contre une cible bien précise. Le brevet couvre également la possibilité de faire des modifications des spacers en fournissant un spacer ou un pseudo spacer pour augmenter la résistance bactérienne.

Cependant, rendre les bactéries résistantes aux phages connus ne présente qu'une solution partielle aux problèmes des attaques virales dans les chaînes de production car les phages et les bactéries évoluent toujours en parallèle. Un autre aspect consiste alors en la génération en amont de souches résistantes en les soumettant à des challenges phagiques au laboratoire permettant ainsi de prévoir l'évolution des phages et obtenir des bactéries qui seraient plus résistantes.

2.4.3 Régulation de l'expression des gènes, outil de biotechnologies ?

Étant donné que les spacers présentent parfois une similarité non seulement avec des séquences phagiques (voir tableau 2.1 page 26) mais également avec des séquences de gènes (25 % chez *Y. pestis* (Pourcel, 2005)), il a été suggéré qu'ils pourraient réguler ou inhiber l'expression des gènes dont ils sont issus. Le CRISPR pourrait alors être manipulé dans le but de créer une immunité contre des gènes de résistance aux antibiotiques, les éléments transposables, les séquences d'insertion, des éléments virulents, des séquences nouvelles ou tout type de séquences non désirées. Il serait possible d'introduire au sein d'un CRISPR, un spacer dont la séquence provient d'un plasmide pour empêcher le transfert de plasmides dans cette bactérie.

Le système CASS est un système dont le fonctionnement inclut plusieurs aspects comme la reconnaissance de séquence étrangère, la sélection et l'insertion de nouveaux spacers, la dégradation d'un génome étranger, etc. Les mécanismes utilisés dans chacun de ces aspects ne sont pas encore connus ; leur compréhension permettra peut être de découvrir plusieurs outils génétiques qui pourraient représenter des avancées dans l'étude de la génétique bactérienne.

2.5 Conclusion

Le système CASS comprend le CRISPR et une cassette de gènes associés. Son rôle, par le biais d'un système d'interférence ARN, consiste en la défense contre les agressions extrachromosomales et peut être même dans la régulation de certains gènes. Le CRISPR, objet du présent rapport, évolue par délétions ou acquisition polarisée de motifs. Le polymorphisme qu'il induit est utilisable dans certains cas pour le typage bactérien ou la reconstitution phylogénétique à l'intérieur d'une même espèce ou dans des espèces voisines.

Chapitre 3

Les CRISPRs en informatique

Lors de la découverte du CRISPR, seule sa structure particulière a été observée. Il a alors été classé comme étant une sorte de répétition existant uniquement chez les procaryotes (voir paragraphe 2.3.2). Peu ou pas d'informations étaient disponibles sur son fonctionnement, son rôle et même ses propriétés caractéristiques. Pour mieux connaître et comprendre cette répétition assez curieuse, il a été évident pour les premiers investigateurs de faire une recherche dans les génomes séquencés disponibles. La comparaison de ces CRISPRs devait permettre de découvrir les caractéristiques principales de la structure. C'est ainsi que le rôle de la bioinformatique s'est révélé crucial dans cet axe de recherche car la découverte du CRISPR correspond bien à l'essor des techniques de séquençage et à l'augmentation spectaculaire du nombre de génomes séquencés. La séquence génomique d'un procaryote peut aller de quelques centaines de kilobases à presque 10 millions pb (9.965.640 pb dans *Solibacter usitatus Ellin6076*) et la taille d'un CRISPR d'environ 80 pb à plus de 21.000 pb (21.637 pb chez *Chloroflexus aurantiacus J-10-fl*). La bioinformatique devait fournir les outils nécessaires pour détecter les CRISPRs dans des séquences génomiques de façon rapide et efficace, puis pour les comparer (alignements, statistiques) et explorer leur voisinage (leader, gènes *cas*, etc). Ce chapitre s'intéressera dans une première partie à décrire les différentes répétitions génomiques. Ensuite, les programmes informatiques utilisés pour l'identification des CRISPRs seront présentés. La troisième partie sera consacrée à la présentation des travaux effectués *in silico* sur les CRISPRs. La dernière partie de ce chapitre sera dédiée à la présentation du but de la thèse.

3.1 Les répétitions

Une répétition sera définie, dans ce manuscrit, comme une séquence d'ADN présente sous des formes similaires au moins deux fois dans un même génome. Cette similarité est caractérisée par un pourcentage d'identité entre les copies.

L'ADN n'étant constitué que de quatre bases différentes (Adénine, Cytosine, Guanine et Thymine), il faut envisager que de nombreuses répétitions apparaissent par de simples mutations ponctuelles. Ces répétitions seront appelées répétitions fortuites. Si la fréquence de A, C, G et T est équiprobable et indépendante, alors dans une séquence de taille L, la probabilité de trouver k copies d'un mot de taille n est donnée par la loi de Poisson $P(k)$, dont la moyenne est $\lambda = \frac{L}{4^n}$. La probabilité de trouver au moins deux copies d'un mot est donc

$$1 - P(0) - P(1) = 1 - e^{-\lambda}(1 - \lambda); \tag{3.1}$$

Le nombre de mots de taille n est 4^n. On a donc environ $4^n(1 - e^{-\lambda}(1 - \lambda))$ mots répétés de taille n.

Considérons le plus petit génome bactérien séquencé, celui de *Mycoplasma genitalium*, composé d'environ 0,58 millions pb. Si on ne tient pas compte de la fréquence de chacun des nucléotides, aucune répétition exacte (totalement identique) de 20 pb ou plus n'est attendue de façon fortuite dans un génome de cette taille. Cependant, malgré leur petite taille, les génomes bactériens possèdent une grande variété de séquences répétées qui ne sont par conséquent certainement pas fortuites. Différents types de séquences répétées ont été découvertes dans les génomes bien avant même leur séquençage complet. Les génomes bactériens et archéens sont, à l'inverse des eucaryotes, connus pour leur compacité, où la quasi totalité de la séquence d'ADN serait codante. Pourtant, contrairement aux attentes, la proportion de séquences répétées n'est pas négligeable chez ces organismes. *S. solfataricus P2* par exemple possède un génome de 3 Mb dont 10% est composé de répétitions (Brügger, 2004; Blount et Grogan, 2005). Les séquences répétées peuvent avoir deux origines :

- une origine interne, par duplication de portions du génome
- une origine externe, par insertions multiples d'éléments génétiques mobiles (virus, plasmides et transposons)

En informatique, une répétition est un objet mathématique simple. Si on considère une séquence ADN comme une chaîne S de taille n, une répétition serait une sous-chaîne w se produisant deux fois dans S. Si on représente une séquence allant de la position i à la position j dans S par $S[i, j]$, alors on a les définitions suivantes :

- Une **Répétition exacte** est une paire des sous chaînes $S[i_1, j_1]$ et $S[i_2, j_2]$, si et seulement si $(i_1, j_1) \neq (i_2, j_2) et S[i_1, j_1] = S[i_2, j_2]$.
- Une répétition est dite **maximale** (Gusfield, 1997) si et seulement si elle n'est pas contenue dans une autre répétition, (i.e. que l'on ne peut étendre ni à gauche ni à droite) :

$$S[i_1 - 1, j_1] \neq S[i_2 - 1, j_2],\ S[i_1, j_1 + 1] \neq S[i_2, j_2 + 1],$$

$$S[i_1 - 1, j_1] \neq S[i_2, j_2 + 1] \text{ et } S[i_1, j_1 + 1] \neq S[i_2 - 1, j_2].$$

Les définitions précédentes évoquent les cas particuliers où les répétitions sont exactes c'est à dire toutes les copies sont identiques. Cependant, dans les séquences biologiques de telles répétitions n'apparaissent pas souvent : le processus de copie n'est pas toujours exact et durant l'évolution les séquences accumulent des mutations, les copies sont alors fortement similaires mais pas identiques.

Les répétitions peuvent être classées principalement en deux catégories, des répétitions en tandem à un seul locus ou des répétitions dispersées dans le génome.

3.1.1 Les répétitions en tandem

Une répétition en tandem est une succession de motifs d'ADN répétés les uns derrière les autres, par opposition aux répétitions "dispersées" dont les unités répétées sont dispersées dans le génome. Une répétition en tandem de taille k s'écrit comme suit :

$$S[i, i + k] = S[i + k + 1, i + 2k + 1] = ...$$

Les différentes unités formant la répétition en tandem ne sont pas nécessairement identiques entre elles : le degré de similarité au sein d'une répétition en tandem est très variable. Selon la taille des copies, une répétition en tandem est classée en deux catégories :

Les microsatellites

Les plus simples répétitions rencontrées dans les génomes sont constituées de séquences de faible complexité, souvent nommées microsatellites (d'autres appellations existent). Elles sont formées d'un motif très simple (une convention souvent utilisée est d'arrêter la taille des microsatellites aux motifs de 8 pb) et sont répétées en tandem des dizaines (voire des centaines) de fois.

Les minisatellites

Les répétitions minisatellites (Vergnaud et Denoeud, 2000) sont définies comme des répétitions tête-à-queue dont la taille unitaire est supérieure à 9 pb. Elles sont présentes dans tous les génomes procaryotes et eucaryotes et ont été particulièrement étudiées initialement dans les génomes des mammifères.

3.1.2 Les répétitions dispersées sur le génome

Les séquences d'insertion (IS)

Les IS sont des éléments génétiques mobiles de moins de 2.5kb en général. Elles jouent un rôle majeur dans la plasticité des génomes procaryotes. Ces séquences, de moins de 2.5kb en général, codent des fonctions impliquées dans leur translocation dans le même génome et entre différents génomes. L'insertion d'une séquence IS dans un gène peut provoquer une inactivation de celui-ci. Ces éléments génétiques mobiles peuvent donc être à l'origine de réarrangements du génome tel que des délétions, des inversions. Les séquences IS peuvent être utilisées comme sonde pour le typage de souches par Southern blot et aussi pour des analyses RFLP, comme par exemple les séquences IS6110 très étudiées chez *M. tuberculosis* (van Embden, 2000).

Les séquences REP (Repetitive Extragenic Palindromic sequences)

Les séquences REP découvertes en 1984 (Stern, 1984) sont aussi appelées PU pour Palindromic Units (Gilson, 1984). La séquence REP a une longueur d'environ 35pb et possède une répétition inversée, elle peut être unique ou en copies multiples adjacentes.

Les séquences ERIC (Enterobacterial Repetition Intergenic Consensus)

Les séquences ERIC (Hulton, 1991) sont aussi appelées IRUs pour Intergenic Repeats Units (Sharples et Lloyd, 1990). Ce sont des éléments de 126 pb qui possèdent dans la région centrale une répétition inversée fortement conservée. Les séquences ERIC permettent la formation de structures secondaires stables de type tige-boucles. Certaines séquences ERIC sont transcrites.

3.1.3 Les CRISPRs : répétitions particulières ?

Les CRISPRs peuvent être classés dans les deux types précédents de répétition. Il est possible de considérer un CRISPR comme une répétition en tandem où chaque copie est formée par le DR et le spacer. Les copies ne sont pas identiques mais similaires (jusqu'à environ 60% de similarité : 100% sur la portion DR et en moyenne 25% sur la portion spacer).

Les CRISPRs sont également des répétitions dispersées particulières puisqu'il existe souvent plusieurs copies du locus dans un même génome.

3.2 Programmes de détection des CRISPRs

La recherche des répétitions en général constitue un volet important dans la bioinformatique. Une multitude de programmes a vu le jour depuis les grands projets de séquençage. Ils mettent en oeuvre des méthodes heuristiques, exactes, probabilistes et utilisent différents types de techniques informatiques (arbre de suffixes, programmation dynamique, etc.).

Une forte proportion de CRISPRs est spontanément identifiée par différents programmes. Son identification plus précise est faite via l'adaptation de certains programmes de détection des répétitions énumérés dans la partie 3.2.1. L'année 2007 a vu l'essor de quelques programmes spécifiques dans la recherche des CRISPRs dont CRISPRFinder, l'un des objets de cette thèse (partie 3.2.2).

3.2.1 Avant 2007 : adaptation de programmes de recherche de répétitions

PATSCAN

PATSCAN, que l'on trouve aussi sur certain serveurs sous le nom de "scan for matches", fût le premier programme utilisé dans l'identification des CRISPRs (Mojica, 1995). Il s'agit d'un programme polyvalent pour la recherche de motifs.

Le principe de la détection de motifs dans les séquences se base sur l'utilisation d'indices en prédictions, ce qui nécessite la définition d'un motif et d'un seuil discriminant au mieux les motifs recherchés des séquences qui présenteraient "par hasard" une similitude avec ce motif. Celui-ci est déterminé par un apprentissage basé sur un ensemble de séquences connues pour posséder le motif (exemples) et des séquences connues pour ne pas contenir le motif (contre-exemples). La méthode est validée sur un ensemble d'exemples et de contre exemples indépendants des séquences utilisées lors de l'apprentissage.

PATSCAN (Dsouza, 1997) [1] a surtout été utilisé pour la recherche de motifs ARN.

[1] `http://www-unix.mcs.anl.gov/compbio/PatScan/HTML/patscan.html`

Il se base sur la recherche d'une signature particulière (propriétés du motif cherché) et utilise une grammaire bien définie selon le type de motifs cherchés. Il ne tient pas compte des séquences chevauchantes.

PATSCAN a été utilisé dans tous les travaux de l'équipe Mojica sur les CRISPRs (Mojica, 1995, 2000, 2005) ainsi que les travaux de l'équipe de Jansen (Jansen, 2002a,b). Les CRISPRs cherchés avec ce programme dans ces travaux ont au moins quatre motifs. Dans le travail de Godde (Godde et Bickerton, 2006), les CRISPRs cherchés ont au moins trois motifs en utilisant les deux algorithmes :
$algorithme1 : p1 = 15...7015...70p115...70p1$, modifié ensuite pour donner
l'$algorithme2 : p1 = 21...3915...45p115...45p1$.
(voir annexe : supplementary files article NAR)

Tandem Repeat Finder : TRF

Definition 3.2.1. ***liste chaînée des k mots*** *Un k-mot est un fragment de texte de taille k. La liste des k-mots permet d'obtenir, pour n'importe quel k-mot donné, toutes les positions de ses occurrences sur le texte.*

TRF [2], logiciel d'utilisation conviviale, est un outil de référence pour la détection de répétitions en tandem dans les génomes (Yeramian et Buc, 1999). Il a été utilisé, par exemple, lors de l'analyse de la séquence brouillon du génome humain par le consortium public (Lander, 2001) et il nourrit les bases de données "minisatellites Database" [3], "Tandem Repeats DataBase" [4] et "Genome Browser" [5]. G. Benson (Benson, 1999) se base dans ce programme sur un algorithme de type probabiliste, heuristique, c'est-à-dire qu'il n'est pas garanti de trouver toutes les répétitions en tandem existantes, mais il a l'avantage, par rapport aux algorithmes exacts, de ne pas nécessiter un temps de calcul trop long ou de rentrer des paramètres trop précis sur les répétitions en tandem recherchées. Ce programme détecte les répétitions en tandem dans les séquences d'ADN, sans avoir à spécifier le modèle ou la taille, ce qui est souvent imposé par d'autres programmes. TRF permet, à partir de critères statistiques, et en utilisant au départ la liste chaînée des k-mots, de localiser des répétitions en tandem avec un pourcentage d'erreur maximal toléré entre les copies. Les critères sont calculés selon un certain nombre de paramètres donnés par l'utilisateur concernant les répétitions en tandem recherchées. L'algorithme construit, à partir de la liste des k-mots, une liste des distances. Il s'agit en fait d'un ensemble de listes : chaque liste D_d a pour rôle de conserver les couples de k-mots rencontrés

[2] `http://tandem.biomath.mssm.edu/trf/trf.html`

[3] `http://minisatellites.u-psud.fr`

[4] `http://tandem.bu.edu/cgi-bin/trdb/trdb.exe`

[5] `http://genome.ucsc.edu/`

dont la distance intra-couple était d. Cette liste des distances, lorsqu'elle est mise à jour (ajout d'un couple à D_d), est utilisée pour détecter une éventuelle répétition en tandem de période approchant d à l'aide des critères calculés. TRF indique la taille du motif, le nombre de copies de ce motif ainsi que sa position dans la séquence soumise.

TRF a été utilisé dans l'identification du CRISPR par les équipes qui ont l'habitude de travailler avec les répétitions en tandem, en particulier l'équipe de G. Vergnaud (Pourcel, 2005) et l'équipe de I. Mokrousov (Mokrousov, 2005) en considérant un motif CRISPR (DR+spacer) comme étant une copie d'une répétition en tandem conservée de 50 à 80% entre les copies. (voir annexe : "supplementary files" article NAR)

REPuter et Vmatch

Definition 3.2.2. ***Arbre de suffixes*** *Un arbre de suffixes est une structure de données arborescente qui permet, entre autre, la détection rapide des répétitions (figure 3.1). La construction se fait en ajoutant à l'arbre des suffixes (séquences) de plus en plus courts. Après chaque insertion dans l'arbre, le nouveau suffixe considéré est le suffixe précédent moins sa première lettre. Si un suffixe partage un préfixe commun avec un autre suffixe, les deux suffixes sont branchés sur leur partie commune. La fin est indiquée par un symbole non utilisé dans la séquence (dans le schéma un \$). Quand toutes les séquences suffixes sont intégrées à l'arbre, la recherche de répétition se fait aisément en parcourant l'arbre et en cherchant les noeuds de l'arbre.*

On peut construire un arbre de suffixes en temps $O(n)$ où n est la longueur de la chaîne de caractères. Cependant, l'arbre de suffixes peut requérir beaucoup d'espace de stockage.

Definition 3.2.3. ***Tableau de suffixes***

L'approche par tableaux de suffixes est une approche différente de celle des arbres. Au lieu de construire un arbre avec les suffixes, on fait le tri (en ordre "alphabétique") des suffixes. L'astuce utilisée se base sur le fait qu'il y ait une correspondance unique entre les entiers de 1 à n où n est la longueur de la chaîne. Pour stocker l'ensemble des suffixes, il

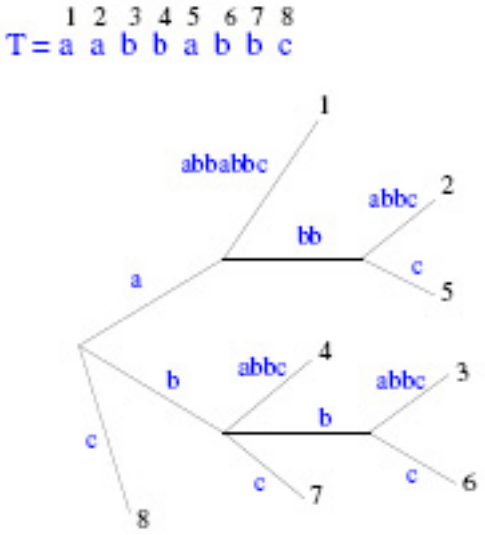


Fig. 3.1 – Construction d'un arbre de suffixes

est possible donc de se contenter de stocker des entiers et ainsi économiser de l'espace.

Reputer (Kurtz et Schleiermacher, 1999) était à l'origine un logiciel de recherche de répétitions exactes basé sur la construction de l'arbre des suffixes (il n'autorisait pas d'erreurs entre les copies). L'arbre des suffixes (définition 3.2.2) permet d'obtenir l'ensemble des répétitions du texte et se construit en temps et en espace linéaire. Un deuxième module a été ajouté afin de compléter la recherche et pouvoir considérer les répétitions approchées (Kurtz, 2000, 2001). Ce module se base sur les "graines" (répétitions exactes) trouvées à l'aide du premier module. Il essaie d'étendre ces graines selon un certain nombre de critères pour détecter des répétitions approchées de taille supérieure aux graines. L'extension utilise deux algorithmes :

- Le premier (MMR) tente d'étendre les graines sur les extrémités afin de découvrir éventuellement une répétition ayant k erreurs selon la distance de Hamming. Il n'autorise donc pas les insertions et les suppressions d'éléments mais uniquement les substitutions.
- Le deuxième (MDR) a aussi pour but de trouver des extensions à certaines graines mais en autorisant les insertions et les suppressions d'éléments : il utilise donc la distance d'édition, et se base sur un algorithme de programmation dynamique effectué sur un domaine limité.

Plus récemment, les modules de REPuter ont encore été améliorés en remplaçant l'utilisation de l'arbre de suffixes par celle des tableaux de suffixes (définition 3.2.3) dans la nouvelle version du programme appelée Vmatch (Abouelhoda, 2004).

REPuter a été appliqué dans la recherche des CRISPRs par le groupe Haft *et col.* (Haft, 2005).

Locating Uniform poly-Nucleotide Areas : LUNA

LUNA est un programme pour l'identification des répétitions dans les séquences. Ce programme est utilisé par la communauté des scientifiques travaillant sur les archées pour la recherche des répétitions dans ces génomes (Peng, 2001, 2003; Lillestøl, 2006). Il n'a pas été publié et n'est disponible qu'après une demande auprès de son développeur K. Brügger et une explication des besoins de son utilisation. Il nécessite l'écriture de scripts pour l'obtention d'un résultat approprié.

PILER

PILER (Edgar et Myers, 2005) est un logiciel pour identifier et classer les répétitions en général. Il est disponible à l'adresse suivante `http://www.drive5.com/piler`. C'est un programme basé sur les alignements des séquences contre elles mêmes pour trouver les répétitions.

PYGRAM

PYGRAM (Durand, 2006) est une méthode de visualisation adaptée à l'analyse et l'identification des structures de répétitions. C'est un navigateur des répétitions dans un génome complet montrant les répétitions et les palindromes. L'algorithme est basé sur la recherche des répétitions maximales à travers l'utilisation des arbres de suffixes, mais il ne permet de trouver que les répétitions exactes.

3.2.2 Après 2007 : outils informatiques dédiés aux CRISPRs

L'utilisation des outils d'identification des répétitions pour trouver les CRISPRs était satisfaisante pour la première phase de découverte de cette structure. Cependant, cette utilisation s'est révélée insuffisante et surtout fastidieuse car elle requiert une intervention manuelle et visuelle pour adapter et ajuster le résultat trouvé à la requête.

Les éléments à identifier dans un CRISPR sont le DR consensus et les spacers. A première vue, cette tâche a l'air assez simple. Cependant, le monde de la biologie n'étant pas parfait, plusieurs difficultés s'imposent sur trois niveaux :

- **L'identification du DR consensus :** il est très important de définir avec exactitude les extrémités du DR car un nucléotide en plus ou en moins signifiera un nucléotide supplémentaire ou manquant dans le spacer qui doit être identifié avec

exactitude vue les propriétés de ces séquences (voir partie 2.2.2 et l'identité avec le proto-spacer).
Il arrive parfois que les DR ne soient pas tous identiques au sein d'un même CRISPR à cause de mutations ponctuelles ou d'erreurs de copie. Une erreur fréquente des programmes informatiques dans ce cas est de fournir un DR consensus plus court. D'autre part, il arrive parfois que deux spacers successifs commencent ou se terminent par un ou deux nucléotides identiques. Le DR consensus fourni peut alors être plus long que le DR réel. Ces deux erreurs sont beaucoup plus fréquentes dans le cas des petits CRISPRs (moins que dix motifs).

- **Le DR dégénéré :** plus il est dégénéré, plus il est facile à omettre par un programme informatique, mais il ne faut pas non plus étendre le CRISPR en ajoutant un faux DR dégénéré.
- **L'identification de tous les CRISPRs :** un bon programme ne doit pas manquer des CRISPRs, en particulier les plus courts. Or, pour qu'une structure possède tous les éléments composant un CRISPR, il faut qu'elle contienne au moins deux DR et un spacer. Ainsi le CRISPR le plus court est défini comme étant un CRISPR à un seul motif. Les CRISPRs à moins de trois motifs n'étaient pas considérés dans les premières investigations de cette structure, mais certains chercheurs pensent que ce sont des éléments clé dans la compréhension du mode de propagation des CRISPRs et du mécanisme d'insertion de nouveaux motifs.

Les programmes décrits dans le paragraphe ci dessus n'arrivent pas à palier efficacement ces différents problèmes. Il s'est révélé alors nécessaire de construire des programmes spécifiques aux CRISPRs permettant une détection rapide sans intervention supplémentaire de l'utilisateur, c'est à dire sans avoir à ajouter des scripts et des filtres informatiques. Depuis 2007, la communauté scientifique a vu la naissance de quelques programmes qui seront énumérés dans la paragraphe qui suit.

PILER-CR

Le premier travail publié sur un programme dédié à la recherche des CRISPRs est le programme PILER-CR (Edgar, 2007), qui est une adaptation du programme PILER. C'est un programme rapide basé sur un algorithme élégant qui consiste essentiellement à aligner une séquence génomique avec elle même en vue d'obtenir des piles ayant les propriétés d'un CRISPR. PILER-CR est un programme rapide et efficace, cependant il est possible de le critiquer sur trois aspects. Le premier concerne la non identification des CRISPRs à moins de trois motifs, ensuite les extrémités du DR ne sont pas toujours bien définies et enfin le DR dégénéré est omis dans certains cas. Cependant, PILER-CR est basé sur des paramètres modifiables dont l'ajustement peut améliorer le résultat trouvé.

PILER-CR a été utilisé par Kunin et Sorek (Kunin, 2007) pour la recherche de structures secondaires dans les DRs.

CRISPRFinder

CRISPRFinder[6] (Grissa, 2007a) est l'un des produits de cette thèse. Il est basé sur l'utilisation du programme Vmatch. Ce programme sera détaillé dans la partie Résultats du manuscrit. CRISPRFinder a été utilisé par Horvath *et col.* (Horvath, 2008b,a) pour étudier la diversité et l'évolution du CRISPR de *S. thermophilus*.

CRISPRFinder est décrit dans l'article (Grissa, 2007a), mais plus récemment des améliorations y ont été introduites pour mieux détecter le DR dégénéré. Il s'agit de l'ajout d'une étape supplémentaire une fois qu'un CRISPR est détecté : elle consiste en un "blast" des deux moitiés du DR consensus contre les flanquantes du CRISPR. Si un fragment est trouvé et est à une distance égale à la taille d'un spacer, il est considéré comme étant le DR dégénéré et est ajouté au CRISPR.

CRISPRFinder a l'avantage de trouver les CRISPRs même les plus petits, cependant les structures courtes qu'il détecte ne sont pas des CRISPRs dans la plupart des cas ; ils sont notés comme "questionables" et nécessitent une vérification supplémentaire avant d'être validés comme CRISPRs (voir partie 6.1 pour plus de détails).

CRISPR Recognition Tool : CRT

CRT (Bland, 2007) est un algorithme qui cherche directement dans la séquence, sans faire des alignements ou utiliser un arbre de suffixe, une série de répétitions exactes courtes séparées par une distance similaire. Ensuite, il étend ces répétitions jusqu'à la taille réelle du DR. Il a ainsi l'avantage par rapport aux autres programmes de trouver avec exactitude les extrémités du DR. Cependant, le défaut majeur du programme est qu'il n'élimine pas le bruit de fond engendré (répétitions en tandem par exemple).

Ab Initio Motif Identification Environment : AIMIE

AIMIE est un service web (Mrázek, 2008) permettant l'analyse de génomes procaryotes dans le but d'identifier les répétitions dispersées telles que les séquences REP pour Repeated Extragenic Palindrome ou les CRISPRs. L'utilisation de ce programme est basée sur deux phases. La première permet de détecter un motif particulier (motifs distribués de façon anormale sur le génome par exemple) alors que la deuxième phase fournit quelques outils pour aider à la compréhension du rôle biologique de ce motif.

Ce programme n'est pas dédié à la recherche des CRISPRs et son utilisation dans ce

[6] `http://crispr.u-psud.fr/Server/CRISPRfinder.php`

publication	génomes analysés	génomes contenant des CRISPRs	programme utilisé
(Mojica, 2000)	-	19	PATscan
(Jansen, 2002b)	148	39	PATscan
(Mojica, 2005)	-	67	PATscan
(Haft, 2005)	-	54	REPuter
(Godde et Bickerton, 2006)	370	148	PATscan
(Lillestøl, 2006)	28 archées	27	LUNA
(Kunin, 2007)	439	195	PILER-CR

Tab. 3.1 – Analyses *in silico* des CRISPRs

but devient alors compliquée.

3.2.3 Analyses *in silico* des CRISPRs

La bioinformatique a joué un rôle pionnier dans la découverte du CRISPR. L'utilisation des programmes cités ci-dessus a permis de réaliser plusieurs analyses *in silico* essentiellement sur les génomes procaryotes (voir tableau 3.1). Chaque étude visant à découvrir l'une des caractéristiques ou des propriétés du CRISPR a été réalisée sur l'ensemble des génomes publics. Cependant, toutes ces études sont statiques et restreintes aux génomes disponibles lors de leur réalisation. La première était celle de Mojica (Mojica, 2000) réalisée en 2000 sur 19 génomes possédant des CRISPRs. Les plus récentes sont celle de Godde sur 370 génomes en 2006 (Godde et Bickerton, 2006) et celle de Kunin et Sorek sur 439 génomes en 2007 (Kunin, 2007)).

3.3 But de la thèse

Cette thèse a pour objectif de fournir des outils informatiques facilitant d'une part l'investigation et la compréhension des CRISPRs et d'autre part leur utilisation dans le

typage bactérien.

Le premier volet concerne la création d'une base de données dynamique. Cette base est importante pour donner un aperçu général sur les CRISPRs et faire des études comparatives. D'ailleurs, plusieurs bases de données statiques ont été créées à l'occasion des analyses *in silico* décrites ci dessus. Pour créer cette base, j'ai été confrontée au début de ma thèse, à l'absence d'un outil spécifique à la recherche des CRISPRs, c'est pourquoi la première étape était de créer CRISPRFinder (en utilisant un autre programme REPuter, qui était à ce moment là le plus efficace et le plus facile à manipuler) pour alimenter la base de données. Le deuxième volet concerne la création de CRISPRcompar, un outil pour assister les biologistes dans les études de micro-évolution en utilisant le CRISPR.

Deuxième partie

Résultats

Chapitre 4

Outils d'investigation des CRISPRs

Le CRISPR est une structure dont la découverte est récente et dont les propriétés représentent un mystère pour plusieurs aspects. Ce travail de thèse fournit des outils d'investigation des CRISPRs pour faciliter d'une part leur détection et d'autre part permettre d'avoir une vision globale de l'ensemble des CRISPRs et ainsi faciliter leur analyse. Ces outils comportent deux volets :

1. CRISPRfinder, un programme disponible en ligne pour trouver les CRISPRs dans une séquence génomique fournie par l'utilisateur ; l'interface obtenue est assez simple, mais permet de décortiquer le CRISPR en colorant différemment ses composants (DRs et spacers) et en précisant leurs coordonnées sur le génome soumis. Cette ressource donne également la possibilité de chercher l'origine des spacers trouvés en effectuant un blast de séquences similaires contre Genbank. De plus, elle permet de télécharger des fichiers pré-calculés résumant les informations relatives à chaque structure détectée. Il est également possible d'obtenir de façon dynamique les séquences flanquantes et la séquence entière du CRISPR (possibilité d'ajuster les positions de début et de fin à la guise de l'utilisateur) tout en permettant de "blaster" également les flanquantes, ce qui facilite l'identification du leader (existence chez d'autres CRISPRs du même génome) ou des gènes *cas*. Un accès à la base de données CRISPRdb est aussi possible pour vérifier l'existence du DR chez des espèces procaryotes enregistrées dans cette base.
2. CRISPRdb, une base de données qui contient, pour des chromosomes ou des plasmides parmi les génomes procaryotes séquencés, l'ensemble des CRISPRs détectés par le programme CRISPRFinder. La base permet d'effectuer des requêtes sur cette collection de CRISPRs en dissociant DR et spacers. Interroger l'ensemble des DR permet de vérifier si un nouveau DR a des homologues déjà stockés dans la base et par la suite appartient à une classe donnée, ce qui renseignera par exemple sur les gènes *cas* qui lui sont associés. D'autre part, la consultation du catalogue des spacers peut être faite pour chercher si un spacer existe chez une autre lignée (surtout dans les études de micro-évolution). Une autre utilisation encore plus intéressante de ce catalogue de spacers (actuellement près de 20.000 séquences) consisterait dans

la recherche d'homologie avec une séquence virale nouvellement identifiée ou potentiellement d'origine virale (métagénomes).

L'outil Flankalign est associé à la base de données, il permet de faire l'alignement des flanquantes de CRISPRs à partir de ceux stockés dans CRISPRdb. L'intérêt se voit surtout dans la détermination de la séquence leader et la comparaison de CRISPRs appartenant à un même génome ou à des lignées voisines.

Dans les deux premières parties de ce chapitre seront exposées les deux ressources citées ci-dessus alors que la troisième partie présentera leur application dans un cas particulier, celui de l'investigation des CRISPRs de *Y. pestis* et *Y. pseudotuberculosis*.

4.1 Présentation du programme CRISPRFinder

CRISPRFinder a été créé dans le but d'avoir un outil fiable qui permette une détection rapide et automatique des CRISPRs. Il a été rendu accessible sur internet, dans l'espoir que cette ressource sera également utile à d'autres utilisateurs, surtout pour ceux qui veulent faire une investigation rapide de l'existence de CRISPRs dans un génome particulier sans avoir à installer un programme ou à gérer un bruit de fond.

L'article suivant (Grissa, 2007a), intitulé "CRISPRFinder : a web tool to identify clustered regularly interspaced short palindromic repeats" (CRISPRFinder : un outil web pour identifier les CRISPRs) présente la méthode d'implémentation du programme et les fonctionnalités de l'interface web relative.

Résumé :

Les CRISPRs constituent une famille particulière de répétitions en tandem dans plusieurs génomes de procaryotes (la moitié des bactéries et presque toutes les archées). Ils consistent en la succession de régions très bien conservées (DR) dont la taille varie de 23 à 47 pb et qui sont séparées par des séquences uniques ayant une taille similaire et ayant en général une origine virale. Un CRISPR est flanqué à une extrémité d'une séquence riche en AT appelée leader et supposée être un promoteur de transcription. Des études récentes suggèrent que cette structure représente un système d'interférence ARN. Dans cet article, nous décrivons CRISPRFinder qui est un service en ligne fournissant des outils pour (i) détecter les CRISPRs y compris les plus courts (un ou deux motifs) ; (ii) définir les DR et extraire les spacers ; (iii) obtenir les séquences flanquantes pour la détermination du leader ; (iv) faire le blast des spacers contre Genbank et (v) vérifier si le DR se retrouve chez un autre génome séquencé de procaryote. CRISPRFinder est accessible librement à l'adresse `http://crispr.u-psud.fr/Server/CRISPRfinder.php`.

W52–W57 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm360

CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats

Ibtissem Grissa[1,*], Gilles Vergnaud[1,2] and Christine Pourcel[1]

[1]Univ Paris-Sud, Institut de Génétique et Microbiologie, UMR 8621, Orsay, F-91405 and [2]Centre d'Etude du Bouchet, 5 rue Lavoisier, 91710 Vert le Petit, France

Received January 25, 2007; Revised April 6, 2007; Accepted April 25, 2007

ABSTRACT

Clustered regularly interspaced short palindromic repeats (CRISPRs) constitute a particular family of tandem repeats found in a wide range of prokaryotic genomes (half of eubacteria and almost all archaea). They consist of a succession of highly conserved regions (DR) varying in size from 23 to 47 bp, separated by similarly sized unique sequences (spacer) of usually viral origin. A CRISPR cluster is flanked on one side by an AT-rich sequence called the leader and assumed to be a transcriptional promoter. Recent studies suggest that this structure represents a putative RNA-interference-based immune system. Here we describe CRISPRFinder, a web service offering tools to (i) detect CRISPRs including the shortest ones (one or two motifs); (ii) define DRs and extract spacers; (iii) get the flanking sequences to determine the leader; (iv) blast spacers against Genbank database and (v) check if the DR is found elsewhere in prokaryotic sequenced genomes. CRISPRFinder is freely accessible at http://crispr.u-psud.fr/Server/CRISPRfinder.php.

INTRODUCTION

Genomic structures corresponding to CRISPRs were observed first in 1987 in *Escherichia coli* (1) and were subsequently reported in other organisms under different names [TREP (2), SRSR (3,4), DRVs (5), LCTR (6), SPIDR (7)] until the CRISPR acronym was proposed by Jansen *et al*. (8). The direct repeat sequences carry in general a low level of palindromic symmetry; they are remarkably well conserved within a species (up to 248 exact copies in *Verminephrobacter eiseniae EF01-2*). However, one of the flanking DRs is frequently truncated or diverged (see Supplementary Data). The DR size varies from 24 to 47 bp whereas the spacer sequence is generally within the range of 0.6–2.5× the DR size. The originality of spacers is that they apparently derive from conjugative plasmids or bacteriophages (2,9–11). A prokaryotic genome may harbour up to 16 CRISPR clusters with the same or a different DR. In a genome, a single CRISPR is generally associated with a family of genes called *cas* for CRISPR-associated (8,12), encoding proteins showing functional similarity with components of the eukaryotic RNA interference (RNAi) systems (13). In addition, it was demonstrated in two archaea, *Archaeoglobus fulgidus* (14) and *Sulfolobus solfataricus* (15), that the CRISPR locus is transcribed into small RNAs (smRNA) probably from one of the flanking regions, the leader, acting as a promoter. These observations and the viral origin of spacers have led to the hypothesis that the CRISPR-associated system (CASS) is a prokaryotic defence mechanism against genetic aggressions (10,13,16). Within species, CRISPRs may be present in a subset of strains, where they sometimes show polymorphism. The DR and the order of the spacers are well conserved, but the number of motifs (DR + spacer) differs from strain to strain. To better understand the mechanisms underlying the CRISPRs' evolutionary scenario, three evolution rules were proposed by Pourcel *et al*. (10) and confirmed by Lillestol *et al*. (15): (i) polarized acquisition of spacers near the leader sequence; (ii) random loss of motifs and (iii) shared ancestry when spacers are identical.

CRISPRs' *in silico* analyses started in 1995 (2) but no specific stand-alone CRISPR software tool was created. Several software were used by different authors to identify these particular repeats but usually a manual discard of background was necessary, and generally some CRISPR clusters were missed or neglected, especially the shortest one (less than three motifs). This is the case, for example, of Tandem Repeat Finder (17) when considering a motif (DR + spacer) as a degenerate repeat (10,18), or Locating Uniform poly-Nucleotide Areas (LUNA), a program for finding degenerate repeats in microbial genomes on a desktop computer. The repeats can be filtered using several parameters including length, distance and level of conservation. LUNA was used especially for finding CRISPRs in archaea (4,15). Another program, Patscan (19) a pattern-matching tool that searches sequences fitting the introduced pattern, was applied to

*To whom correspondence should be addressed. Tel: 33 1 69 15 30 01; Fax: 33 1 69 15 66 78; Email: Ibtissem.Grissa@igmors.u-psud.fr

identify CRISPRs containing at least three (20) or four exact direct repeats (8). PYGRAM (21) is a visualization program browsing all the repeats in the submitted genomic sequence and showing perfectly conserved palindromic repeats as pyramids. The PYGRAM program is mostly efficient in visually displaying large CRISPRs (CRISPRs with as many as seven motifs are considered as being very short in this work) since they will be recognized as a concentration of horizontal bars referring to a group of co-occurring repeats that differ by only a few nucleotides. Finally, Haft *et al*. (12) used REPfind (http://bibiserv.techfak.uni-bielefeld.de/reputer/), a part of the REPuter package (22–24) and BLASTN to identify smaller repeat clusters.

These programs are the most used tools in CRISPR detection, although none of them is especially conceived for this purpose. They require further manual manipulations to eliminate background data (tandem repeats for example) and importantly, do not define accurately the DR consensus (due to errors on the boundaries). Recently, two CRISPR- dedicated software tools were proposed, CRT (http://www.room220.com/crt) and PILER-CR (25). Both of them run fast and perform well in finding CRISPRs. However, CRT results in a considerable background since tandem repeats are considered as putative CRISPRs and in addition, the same CRISPR is sometimes detected more than once with different consensus DRs. PILER-CR has also some drawbacks since it often misidentifies the DR boundaries and omits the truncated DR.

In addition, there is no user-friendly dedicated web site. A specialized program to automatically identify CRISPRs seems to be mandatory for their optimum, rapid exploration and in-depth analysis, in order to increase the efficiency of CRISPRs investigations. CRISPRFinder is a web service offering fundamental tools for CRISPR detection, including the shortest ones, allowing an accurate definition of the DR consensus boundaries and extraction of the related spacers. It offers also additional tools to analyze the CRISPR loci: (i) obtain the CRISPR and the flanking sequences according to flexible size; (ii) make a blast of selected spacers or flanking sequences against the Genbank database and (iii) check if the DR is found elsewhere in prokaryotic sequenced genomes. The CRISPRFinder web interface is accessible through http://crispr.u-psud.fr/Server/CRISPRfinder.php

METHODS AND IMPLEMENTATION

CRISPRFinder core routines were developed in Perl under Debian Linux. The input of the web tool is a genomic query sequence of length up to 67 Mb in 'FASTA' format. Possible locations of CRISPRs (consisting of at least one motif) are detected by finding maximal repeats. A maximal repeat (26) is a repeat that cannot be extended in either direction without incurring a mismatch. The total number of maximal repeats in a sequence of size n is linear (less than n) which is interesting since the computation may be done in linear time using a suffix-tree-based algorithm. A CRISPR pattern of two DRs and a spacer may be considered as a maximal repeat where the repeated sequences are separated by a sequence of approximately the same length.

The operation of the program can be divided into four main steps summarized in Figure 1: (Step 1) browsing the maximal repeats of length 23–55 bp interspaced by sequences of 25–60 bp, (Step 2) selecting the DR consensus according to a defined score taking into account the number of occurrences of the candidate DR in the whole genome and privileging internal mismatches between the DRs rather than mismatches in the first or the last nucleotides, (Step 3) defining candidate CRISPRs after checking if they fit CRISPR definition, (Step 4) eliminating residual tandem repeats.

In the first step, maximal repeats are found by the software Vmatch (http://www.vmatch.de/), the upgrade of REPuter (22–24). Vmatch is based on a comprehensive implementation of enhanced suffix arrays (27) which provides the power of suffix trees with lower space requirements. A one nucleotide mismatch is allowed permitting minimal CRISPRs with a single nucleotide mutation between DRs to be found. Hereafter, the obtained maximal repeats are grouped to define regions of possible CRISPRs with a display of consensus DR candidates related to each cluster.

The second step is aimed at retrieving the DR consensus of each cluster. The difficulty resides especially in the identification of boundaries, which is very important to extract the correct spacers and compare DRs. In fact, the consensus DR is selected as the maximal repeat which occurs the most in the whole underlying genome sequence with respect to the forward and the reverse complement directions (since two CRISPRs having the same DR consensus may be in opposite directions). Thus, ambiguity in the choice of a DR will be eliminated in the case of presence of similar DRs in other CRISPRs of the related genomic sequence. However, if occurrence numbers are equal, more than a single DR consensus candidate are kept and later compared. Given a candidate consensus DR, the pattern search program fuzznuc of the EMBOSS package (28) is applied to get DRs' positions in the related cluster. As the first or the last DR in a CRISPR may be diverged/truncated, a mismatch of one-third of the DR length is allowed between the flanking DRs and the candidate consensus DR, whereas smaller nucleotide differences are allowed between the other DRs to take into account possible single mutations. In case of multiple DR candidates, a score is computed and the best one (minimum) is picked. This score favours candidates which are encountered more frequently, rather than consensus DR showing less internal mismatches.

Once the DR consensus is determined, the corresponding spacers (Step 3) are extracted according to the DR boundaries determined previously. The spacer length is not allowed to be shorter than 0.6 or longer than 2.5 times the DR length. These sizes are in the range of CRISPRs described in the literature.

The last step consists in discarding false CRISPRs. Therefore, tandem repeats are eliminated by comparing the consensus DR with the spacer if there is only one spacer, or by comparing spacers between each other.

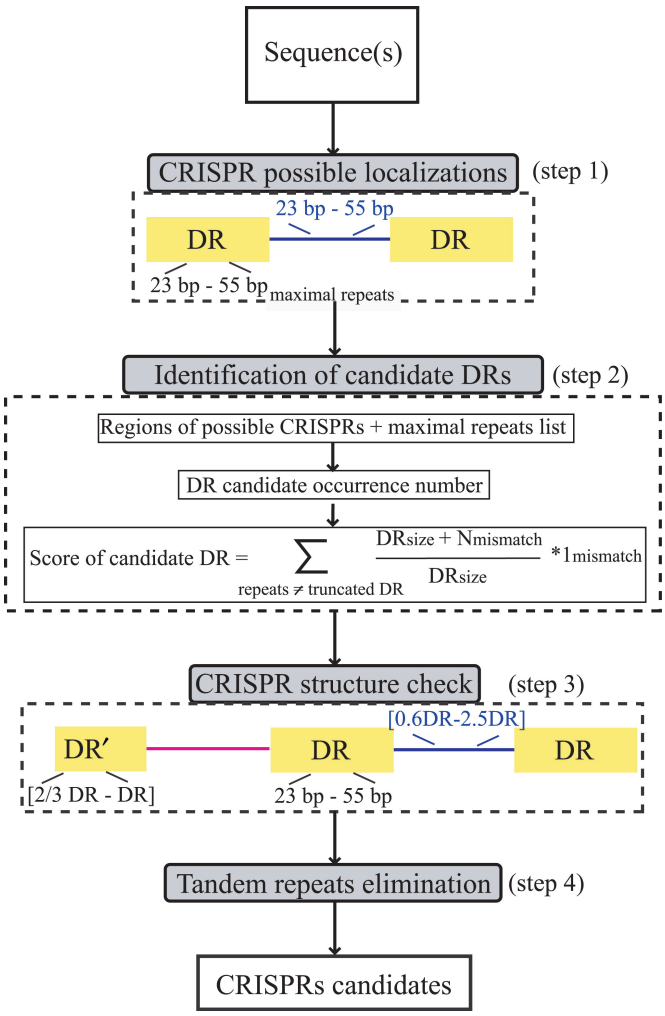


Figure 1. CRISPR Finder flow chart. (Step 1) Browsing the maximal repeats to get possible CRISPR localizations using the Vmatch program. (Step 2) Consensus DR selection according to candidate occurrences and a score computation: the score privileges internal mismatches between direct repeats of a cluster rather than boundary mismatches. (Step 3) DR and spacers size check. (Step 4) Tandem repeats elimination using ClustalW for aligning spacers.

The comparison is done with the CLUSTALW program (29) and the percentage of identity between spacers is not allowed to exceed 60%. Finally, candidates having at least three motifs and at least two exactly identical DRs are considered as confirmed CRISPRs. The remaining candidates are considered as questionable. These should be critically investigated by, for example, checking for intraspecies size variation of the locus.

INPUT AND OPTIONS

The query sequence must be in 'FASTA' format. Ns characters are accepted, IUB/GCG letters (MRWSYKVHDBX) will be converted to Ns and considered as mismatches but any other characters will be deleted. One can either paste the genomic sequence into the input field or upload it from a file on the local machine. Multisequence files are also allowed by the program and will be treated independently. Users may use the default version or click on the 'advanced version' link to set and modify all the program parameters, which may be especially useful for fixing the DR size.

OUTPUT

After querying a genomic sequence by CRISPRFinder, results are summarized in a table with the number of confirmed and questionable CRISPRs (Figure 2A). A CRISPR locus is presented according to a colour code showing DRs in yellow and spacers in different colours. The respective positions are displayed, in addition to links to two files: a summary of the displayed properties (number of motifs, DR consensus, positions, etc.) and a fasta file containing the list of spacers. In addition, a PNG (Portable Network Graphics) figure displays the different candidates' location in the analysed sequence.

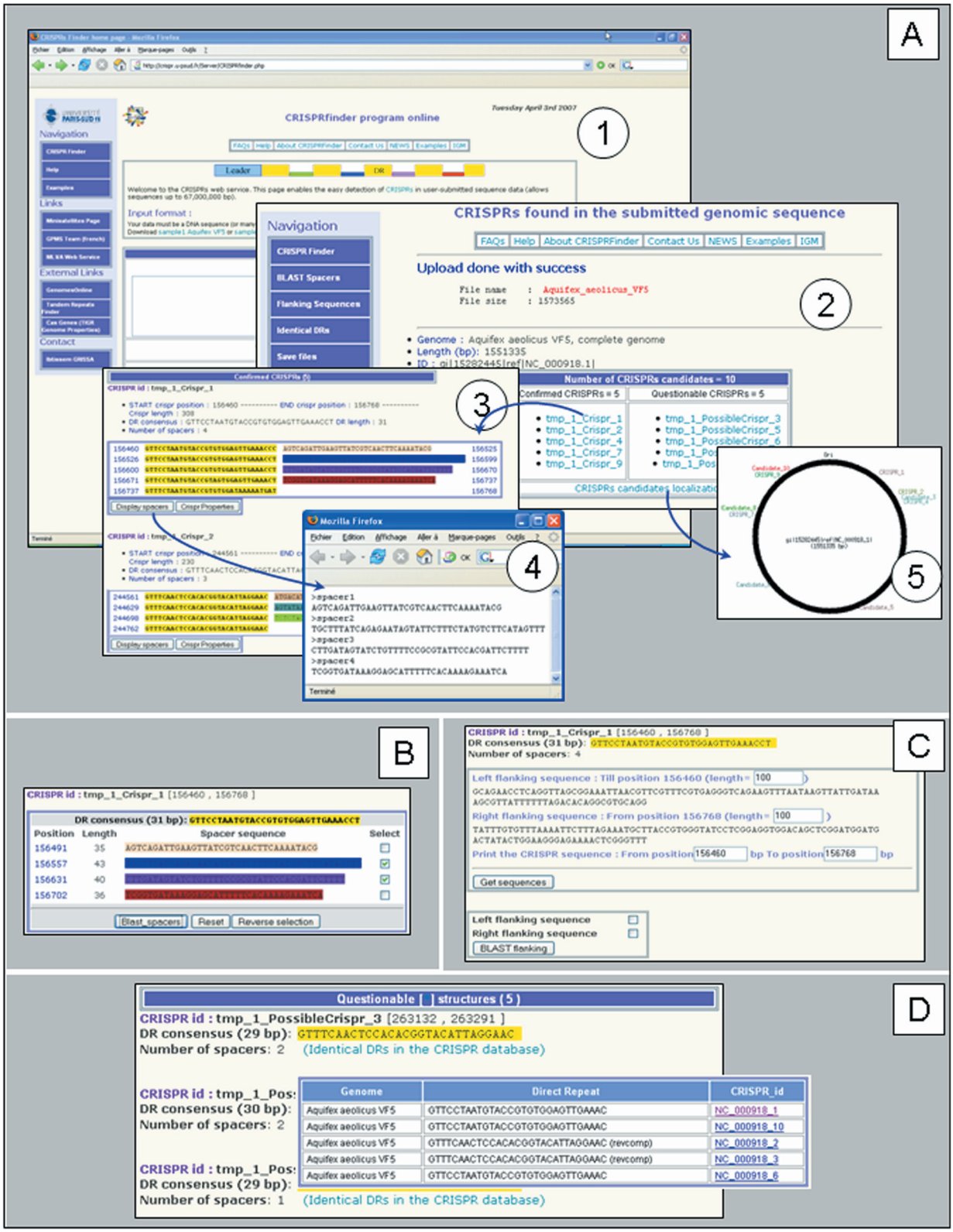


Figure 2. An example of CRISPRFinder output using the Aquifex aeolicus VF5 genomic sequence (Refseq: NC_000918). (Panel **A**) (1) Home page where the genomic sequence is submitted. (2) Table listing the detected CRISPRs candidates (questionable and confirmed) providing links to each one. (3) CRISPRs details, the DR is showed in yellow and the spacers in different colours. (4) A fasta file displaying the first CRISPR spacers. (5) Figure showing the Aquifex circular chromosome with CRISPRs positions. (Panel **B**) One or several spacers may be blasted against NCBI databases by clicking on the blast_spacers button. (Panel **C**) The flanking and the CRISPR sequences may be viewed by clicking on the Get sequences button. The sequences boundaries may be modified by the user. (Panel **D**) The list of consensus DRs for all CRISPRs is shown with a link to identical DRs in the CRISPR database.

In the case of presence of CRISPR clusters, further analysis may be done through three hyperlinks in the left menu: (i) blast spacers against the Genbank databases with a cutoff of 0.1 for the E-value and a matching length of at least 70% the queried spacer size (Figure 2B); (ii) obtain CRISPR and flanking sequences which are especially useful to define the leader sequence. As the size of the leader sequence depends on the species (it varies from 100 to 500 bp), the retrieved sequence may be manually modified by the user (Figure 2C) and (iii) display identical DRs in other known CRISPR loci (Figure 2D). This utility corresponds to a link to CRISPR database (Grissa *et al*. submitted for publication).

DISCUSSION AND CONCLUSIONS

CRISPRFinder is a program that allows the identification of structures with the principal characteristics of CRISPRs, the smaller being composed of a truncated or diverged DR, a spacer and a complete DR. In their analysis, Godde *et al*. (20) using Patscan had chosen to retain only CRISPRs with at least three exact repeats (eliminating CRISPRs constituted of a first truncated repeat plus two exact repeats) thus ignoring most CRISPRs containing less than three spacers. Similarly in the work by Durand et *al*.(21), the PYGRAM program is mostly efficient in visually displaying large CRISPRs. Such stringent criteria were appropriate in order to avoid ambiguities in early investigations which were essentially describing these new structures. However, it is now important, in order to better understand the evolution and spreading of CRISPRs, to provide tools which will not eliminate the smallest CRISPRs. This is what we chose to achieve with CRISPFinder. The major drawback is that when looking for the shortest structures, such as those with a unique spacer, it is clear that the background of spurious candidates can be very high. The output of Patscan and CRT also contains a large quantity of noised data that needs a manual treatment.

CRISPRFinder is accessible on the web and submission is very simple. We provide several samples on the website as demonstrators. Upon submission of the complete genome of *Aquifex aeolicus* VF5 (sample1), five confirmed and five possible CRISPRs are displayed in the following pages. On the contrary, while using the webservice for Patscan (http://www-unix.mcs.anl.gov/compbio/PatScan/), it is necessary to first define a pattern (which is not straightforward) and it is not possible to seek for CRISPRs in a single genomic sequence but rather in an entire predefined database. In addition, Patscan requires a Sun machine for local implementation. Similarly, PYGRAM only runs on linux systems and its installation requires advanced skills. CRT requires either to install JRE (Java Runtime Environment) or compile the source files, and PILER-CR needs to be compiled before use. A comparison between layouts of available online programs (REPuter, Patscan, TRF) and of CRISPRFinder is provided in the Supplementary Data.

To check that CRISPRFinder was efficient in recovering all the CRISPRs from a genome, we compared the results to other available studies on CRISPRs (15,20). The data were generally in good agreement, the differences being always in the DR boundaries' identification (more accurate with CRISPRFinder) or in the number of motifs found, as the truncated DR is sometimes neglected or short clusters are not detected with other programs. Interestingly, some strains were claimed to be devoid of CRISPRs by Godde and colleagues but proved to have short CRISPRs with CRISPRFinder, such as in different *Shigella* sp. (*S. sonnei* Ss046, *S. flexneri* 2a str. 301, *S. flexneri* 2a str), or even long CRISPRs such as in *Pseudomonas aeruginosa UCBPP-PA14*. The latter example is shown on the CRISPRfinder website (sample2), and as can be seen by using the BLAST spacer function, six spacers out of thirty six at two different CRISPR loci, correspond to a bacteriophage sequence (bacteriophages F116, B3, D3112, DMS3 and phi CTX).

The tools developed here will assist in future CRISPRs' analysis. Furthermore, the possibility to identify CRISPRs containing one or two motifs may help understand how new CRISPRs are created. The very small candidates will need to be typed across different isolates within the same species or very closely related species to search for variations. For instance, as shown with the sample file provided on the website (YP1 Yersinia), five *Yersinia pestis* strains possess at the same CRISPR locus two to eight spacers, some being unique and others shared by two or more strains (10). This strain-dependent polymorphism is especially interesting for epidemiological and phylogenetic studies (30,31). A tool to easily create a dictionary of spacers from different strains is proposed in a CRISPR-dedicated web database (http://crispr.u-psud.fr/crispr/).

The CRISPRFinder web server is an interface to extract with precision and to further analyse CRISPRs from genomic sequences. Four main advantages may be cited: (i) short CRISPR-like structures are detected, they are labelled questionable but may be of great interest if later confirmed; (ii) DRs are accurately defined to single base pair resolution; (iii) summary files may be uploaded (CRISPR properties summary and spacers file in Fasta format) and (iv) flanking sequences or spacers can be easily extracted and blasted against different databases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The CNRS and Université Paris Sud 11 have funded this project. Work on the development of genotyping tools for bacterial pathogens is supported by the Délégation Générale pour l'Armement (DGA). Funding to pay the Open Access publication charges for this article was provided by ADGE.

Conflict of interest statement. None declared.

REFERENCES

1. Ishino,Y., Shinagawa,H., Makino,K., Amemura,M. and Nakata,A. (1987) Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J. Bacteriol.*, **169**, 5429–5433.
2. Mojica,F.J., Ferrer,C., Juez,G. and Rodriguez-Valera,F. (1995) Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. *Mol. Microbiol.*, **17**, 85–93.
3. Mojica,F.J., Diez-Villasenor,C., Soria,E. and Juez,G. (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol. Microbiol.*, **36**, 244–246.
4. Peng,X., Brugger,K., Shen,B., Chen,L., She,Q. and Garrett,R.A. (2003) Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in *Sulfolobus* genomes. *J. Bacteriol.*, **185**, 2410–2417.
5. van Embden,J.D., van Gorkom,T., Kremer,K., Jansen,R., van Der Zeijst,B.A. and Schouls,L.M. (2000) Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J. Bacteriol.*, **182**, 2393–2401.
6. She,Q., Singh,R.K., Confalonieri,F., Zivanovic,Y., Allard,G., Awayez,M.J., Chan-Weiher,C.C., Clausen,I.G., Curtis,B.A. *et al.* (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proc. Natl Acad. Sci. USA*, **98**, 7835–7840.
7. Jansen,R., van Embden,J.D., Gaastra,W. and Schouls,L.M. (2002) Identification of a novel family of sequence repeats among prokaryotes. *OMICS*, **6**, 23–33.
8. Jansen,R., Embden,J.D., Gaastra,W. and Schouls,L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.*, **43**, 1565–1575.
9. Bolotin,A., Quinquis,B., Sorokin,A. and Ehrlich,S.D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, **151**, 2551–2561.
10. Pourcel,C., Salvignol,G. and Vergnaud,G. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, **151**, 653–663.
11. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.
12. Haft,D.H., Selengut,J., Mongodin,E.F. and Nelson,K.E. (2005) A Guild of 45 CRISPR-Associated (Cas) Protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.*, **1**, e60.
13. Makarova,K.S., Grishin,N.V., Shabalina,S.A., Wolf,Y.I. and Koonin,E.V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct*, **1**, 7.
14. Tang,T.H., Bachellerie,J.P., Rozhdestvensky,T., Bortolin,M.L., Huber,H., Drungowski,M., Elge,T., Brosius,J. and Huttenhofer,A. (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl Acad. Sci. U.S.A*, **99**, 7536–7541.
15. Lillestol,R., Redder,P., Garrett,R. and Brugger,K. (2006) A putative viral defence mechanism in archaeal cells. *Archaea*, **2**, 59–72.
16. Mojica,F.J., Diez-Villasenor,C., Garcia-Martinez,J. and Soria,E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.*, **60**, 174–182.
17. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
18. Mokrousov,I., Narvskaya,O., Limeschenko,E. and Vyazovaya,A. (2005) Efficient discrimination within a *Corynebacterium diphtheriae* epidemic clonal group by a novel macroarray-based method. *J. Clin. Microbiol.*, **43**, 1662–1668.
19. Dsouza,M., Larsen,N. and Overbeek,R. (1997) Searching for patterns in genomic data. *Trends Genet.*, **13**, 497–498.
20. Godde,J.S. and Bickerton,A. (2006) The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J. Mol. Evol.*, **62**, 718–729.
21. Durand,P., Mahe,F., Valin,A.S. and Nicolas,J. (2006) Browsing repeats in genomes: Pygram and an application to non-coding region analysis. *BMC Bioinformatics*, **7**, 477.
22. Kurtz,S., Choudhuri,J.V., Ohlebusch,E., Schleiermacher,C., Stoye,J. and Giegerich,R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, **29**, 4633–4642.
23. Kurtz,S., Ohlebusch,E., Schleiermacher,C., Stoye,J. and Giegerich,R. (2000) Computation and visualization of degenerate repeats in complete genomes. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 228–238.
24. Kurtz,S. and Schleiermacher,C. (1999) REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics (Oxford, England)*, **15**, 426–427.
25. Edgar,R.C. (2007) PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, **8**, 18.
26. Gusfield,D. (1997) *Algorithms on Strings, Tree and Sequences*. Cambridge University Press, NY.
27. Abouelhoda,M., Kurtz,S. and Ohlebusch,E. (2004) Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, **2**, 53–86.
28. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
29. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
30. Kamerbeek,J., Schouls,L., Kolk,A., van Agterveld,M., van Soolingen,D., Kuijper,S., Bunschoten,A., Molhuizen,H., Shaw,R. *et al.* (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.*, **35**, 907–914.
31. Hoe,N., Nakashima,K., Grigsby,D., Pan,X., Dou,S.J., Naidich,S., Garcia,M., Kahn,E., Bergmire-Sweat,D. *et al.* (1999) Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains. *Emerg. Infect. Dis.*, **5**, 254–263.

4.2 La base de données CRISPRdb

L'article suivant (Grissa, 2007b) intitulé "The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats" (La base de données CRISPRdb et des outils pour l'affichage des CRISPRs et la génération de dictionnaires des spacers et des répétitions) présente les fonctionnalités de la base de données utiles à l'investigation des CRISPRs. Il montre l'utilité de ces ressources et discute des caractéristiques des CRISPRs à partir du recueil de données constitué au moment de la rédaction. Cet article confirme la manière unidirectionnelle par laquelle le CRISPR acquiert de nouveaux éléments via l'observation de quelques cas de CRISPRs. Il présente également l'outil créateur de dictionnaires de spacers ou CRISPRtionary dans sa première version. Ce dernier sera détaillé dans la partie 5.1.2.

Résumé :

Contexte : Les éléments répétés appelés CRISPRs chez les archées et les bactéries sont soupçonnés être des acteurs dans la défense contre les virus. Dans le modèle actuel, les motifs comprenant les spacers et les répétitions peuvent cibler un ADN envahisseur et conduire à sa dégradation par un mécanisme similaire à l'interférence ARN. L'analyse du polymorphisme intra espèce montre qu'un nouveau motif (un spacer et un élément répété) sont ajoutés de façon polarisée. Malgré la description de leurs caractéristiques principales, il reste beaucoup à apprendre sur la manière dont les CRISPRs sont créés et sur leur mode d'évolution. Étant donné que de nouvelles séquences deviennent disponibles, il paraît nécessaire de développer des outils automatiques pour publier les informations relatives aux CRISPRs et faciliter leur investigation.

Résultats : Nous avons produit un programme, CRISPRFinder, permettant l'identification et l'extraction des séquences répétées et uniques. En utilisant ce logiciel, une base de données a été produite et est mise à jour automatiquement tous les mois à partir des nouveaux génomes séquencés. Des outils complémentaires ont été créés pour permettre d'aligner les séquences flanquantes pour la recherche de similarité entre différents locus et pour construire des dictionnaires de séquences uniques. Actuellement, plus de six cent CRISPRs ont été identifiés dans 475 génomes publiés. Deux archées parmi trente-sept et près de la moitié des bactéries ne possèdent pas de CRISPRs. Des analyses fines des séquences répétées confirment le point de vue actuel sur l'addition des motifs à l'extrémité du CRISPR adjacente au promoteur putatif.

Conclusions : On peut espérer que la disponibilité d'une base de données publique, régulièrement mise à jour et qui peut être interrogée sur le web aidera à disséquer et comprendre l'évolution de la structure CRISPR et de ses régions flanquantes. L'analyse du polymorphisme intra espèces du CRISPR sera facilitée par l'utilisation de CRISPRFinder et du créateur de dictionnaires. CRISPRdb est accessible à l'adresse `http://crispr.u-psud.fr/crispr`.

BMC Bioinformatics

Database

Open Access

The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats

Ibtissem Grissa[1], Gilles Vergnaud[1,2] and Christine Pourcel*[1]

Address: [1]Univ Paris-Sud, Institut de Génétique et Microbiologie, UMR 8621, Orsay, F-91405, France; CNRS, Orsay, F-91405, France and [2]Centre d'Etudes du Bouchet, 5 rue Lavoisier, 91710 Vert le Petit, France

Email: Ibtissem Grissa - ibtissem.grissa@igmors.u-psud.fr; Gilles Vergnaud - gilles.vergnaud@igmors.u-psud.fr; Christine Pourcel* - christine.pourcel@igmors.u-psud.fr

* Corresponding author

Published: 23 May 2007

BMC Bioinformatics 2007, **8**:172 doi:10.1186/1471-2105-8-172

Received: 5 January 2007
Accepted: 23 May 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/172

Abstract

Background: In Archeae and Bacteria, the repeated elements called CRISPRs for "clustered regularly interspaced short palindromic repeats" are believed to participate in the defence against viruses. Short sequences called spacers are stored in-between repeated elements. In the current model, motifs comprising spacers and repeats may target an invading DNA and lead to its degradation through a proposed mechanism similar to RNA interference. Analysis of intra-species polymorphism shows that new motifs (one spacer and one repeated element) are added in a polarised fashion. Although their principal characteristics have been described, a lot remains to be discovered on the way CRISPRs are created and evolve. As new genome sequences become available it appears necessary to develop automated scanning tools to make available CRISPRs related information and to facilitate additional investigations.

Description: We have produced a program, CRISPRFinder, which identifies CRISPRs and extracts the repeated and unique sequences. Using this software, a database is constructed which is automatically updated monthly from newly released genome sequences. Additional tools were created to allow the alignment of flanking sequences in search for similarities between different loci and to build dictionaries of unique sequences. To date, almost six hundred CRISPRs have been identified in 475 published genomes. Two Archeae out of thirty-seven and about half of Bacteria do not possess a CRISPR. Fine analysis of repeated sequences strongly supports the current view that new motifs are added at one end of the CRISPR adjacent to the putative promoter.

Conclusion: It is hoped that availability of a public database, regularly updated and which can be queried on the web will help in further dissecting and understanding CRISPR structure and flanking sequences evolution. Subsequent analyses of the intra-species CRISPR polymorphism will be facilitated by CRISPRFinder and the dictionary creator. CRISPRdb is accessible at http://crispr.u-psud.fr/crispr

Background

Clustered regularly interspaced short palindromic repeats (CRISPRs) have been described in a wide range of prokaryotes, including the majority of Archaea and many Bacteria. They consist in the succession of 24–47 bp repeated sequences (often called direct repeats or DR) separated by unique sequences of a similar length (spacers) [1-4]. *Bona fide* CRISPRs possess at one end a partial DR and at the other end after the last DR a sequence of about 200 bp called the leader [5]. The origin of the spacers is still largely unknown but several recent studies identified some of them as fragments of foreign DNA mostly of viral origin [6-9]. Analysis of a large number of *Yersinia pestis* isolates has shown that these elements are sequentially added in a polarised fashion next to the leader [8]. This suggestion was further confirmed by observations in *Sulfolobus solfataricus* and in *Streptococcus thermophilus* [9,10]. A cluster of genes called *cas* (CRISPR-associated) are often found in the vicinity of CRISPRs [5]. When several CRISPRs with the same DR are present, only one is associated with *cas* genes. The exact number of *cas* genes is not known and apparently varies from one strain to another. However, a core of 4 genes is regularly identified, which appears to encode proteins involved in DNA modification and repair [11]. Phylogenetic studies performed on the CAS proteins suggest that CRISPRs are acquired by horizontal transfer [12,13]. This is consistent with their presence on megaplasmids [12]. CRISPRs are non-coding regions but different observations suggest that they are transcribed into small RNAs (smRNA) possibly from the leader acting as a promoter, and that they might play a role as siRNA (small interfering RNA) to block the entry of foreign sequences [10,11,14].

In order to gain further insight into the organisation and behaviour of CRISPR loci it is necessary to perform extensive analyses of the available sequenced genomes. Several studies have been performed, the most extensive being that made on 370 prokaryotic genomes [12]. However, these studies are static and considering the amount of ongoing sequencing projects they are rapidly becoming obsolete. The TIGRFAM database [15] provides information on CAS associated CRISPR loci but it is not dedicated to CRISPR identification and will not report CRISPR structures devoid of neighbouring *cas* genes.

For the algorithmic detection of CRISPR patterns, several methods were empirically applied previously, making use of REPuter [13,16], PatScan [12,17], TRF [8,18], LUNA [10], PYGRAM [19]. These programs are designed to find repeats and are not especially conceived for CRISPR patterns finding, so they may provide the CRISPR location but do not define accurately the consensus DR. The output of such tools requires significant manual discard to eliminate background, and post-processing to define the consensus DR and the spacers. Recently, a CRISPR dedicated software tool called PILER-CR was described [20]. PILER-CR is based on an elegant algorithm that consists mainly in producing piles meeting the CRISPR properties from local alignments of the query sequence to itself. The software tool has the advantage of being rapidly executed but it sometimes misidentifies the DR boundaries and omits the truncated DR.

Finally, using the available programs, "short" or "quite short" CRISPRs (defined as containing less than three, three or seven spacers [5,12,19]) are not considered.

Since future insights into the evolution of CRISPRs may result from the investigation of these very small CRISPRs, some of which may be newly emerging structures, it is important to facilitate access to this enlarged, but much more difficult to define, group.

We have developed tools to identify CRISPRs, select DR and store spacers into dictionaries, and a database which can be queried online at http://crispr.u-psud.fr/crispr. The CRISPRdb is automatically updated; in the May 2007 version, 475 published microbial genomes have been processed.

Construction and content

Database and software design and implementation

CRISPRdb and associated web services are implemented in Perl version 5.8.8 [21] and take advantage of some BioPerl [22] modules for manipulating sequences. They run on an Apache 2.0 web server [23] with a Linux operating system (debian Sarge 3.1) [24]. The core application consists of two main programs: CRISPRFinder to detect CRISPRs and extract them from a genomic sequence, and Database Tools for downloading prokaryotic genomes from the NCBI ftp site [25], saving CRISPRs and making updates.

The first program is a full command line tool written in-house in Perl. It is used to process published genome sequences and feed the CRISPR database. It can also be run interactively through the web interface for submission and analysis of users sequence data [26].

The second program is a set of Perl scripts. Downloading of genomic sequences, CRISPRs detection and motifs extraction are fully automated.

A web resource is built on top of these programs via PHP [27] and Perl CGI scripts. This preserves platform independence across multiple operating systems and allows the user to interact with the different CRISPR tools programs without computer programming or (shell) scripting skills.

The CRISPRs database (CRISPRdb)

CRISPRdb is a relational database implemented using mysql 4.1 [28]. It utilizes the CRISPRFinder program to identify putative CRISPRs and additional tests to further screen for the smallest CRISPRs in a polyphasic approach. Indeed the CRISPRFinder program is conceived to authorize the largest number of possible CRISPRs, especially the shortest ones, containing one or two spacers. The main idea of the program is to first find possible CRISPR localizations in a genomic sequence and then check if these regions contain a cluster that possess the characteristics of "obvious" CRISPR, i.e. containing at least three repeats. Finding possible CRISPR localizations is achieved using the Vmatch package to detect maximal repeats [29], that is a repeat that cannot be extended in either direction without incurring a mismatch [16,30]. Reported matches must have a size within 23 to 55 bp with one possible mismatch, and the gap size between two instances of a repeat must be within 25 to 60 bp. The maximal repeats are clustered according to their position in the genome. In each "cluster", the maximal repeat which is the most frequent in the genome being processed is selected and "blasted" against the cluster. Such a maximal repeat is a candidate DR sequence, and when additional candidate DRs are identified, a score is computed to select the DR resulting in the minimum number of mismatches towards its boundaries. This step is probably instrumental to achieve a very precise identification of proper DR consensus compared to other programs. The related matches are then extracted and tested as putative DRs of a CRISPR, so that the first or the last match is allowed to be degenerated with a maximal number of errors equal to half the match length. This allows the efficient identification of the first, often truncated, DR. The other matches must be globally conserved at least to 80%. Finally two filters are added to check the CRISPR candidates' structure. The first one eliminates clusters for which spacers length are not within the range of 0.6× and 2.5× the DR length. In addition, CRISPR candidates with more than 60% of similarity between spacers (or between DR and spacer) are considered as tandem repeats and are eliminated by the second filter. The selected criteria described above imply that the minimal structure of a putative CRISPR detected by CRISPRFinder should consist in at least two successive direct repeats (one spacer) with a maximum of one mismatch. CRISPRs of more than 2 spacers with three or more perfect repeats are considered "confirmed CRISPR" whereas the shorter CRISPRs are considered "questionable".

Currently, CRISPRdb is composed of 5 tables (Figure 1). For storage in CRISPRdb (Figure 2), several additional tests are applied to the questionable CRISPRs in order to validate a maximum of them. First, a comparison of their DR to previously identified DRs is performed (for example, CRISPR NC_006155_4 in *Yersinia pseudotuberculosis* IP 32953 with 2 spacers has the same DR as CRISPRs NC_006155_6 and NC_006155_7 in the same genome, comprising respectively 4 and 16 motifs; CRISPR NC_003272_3 in Nostoc sp. PCC 7120 with only one spacer, has the same DR as the CRISPR NC_007413_19 of *Anabaena variabilis* ATCC 29413 comprising 33 spacers). Then, a second filter is added to discard some of the non significant short CRISPRs, consisting in a restriction on the spacer allowed length, when the corresponding DR has no classical flanking nucleotides such as GTTT or GAAC.

Authorizing small CRISPR-like structures in the database leads to an important amount of questionable data. Therefore a colour code is being used to differentiate the "confirmed CRISPR" shown in pink to the questionable structures shown in grey. However, and importantly, each time the database is updated, and new genomes are processed, DRs from all questionable structures are rechecked against the updated DR database.

CRISPRs loci are identified from finished microbial genome sequences (as listed by the Genome Online Database [31] and accessed from Genbank) and stored into the database. This procedure is repeated monthly to update the database.

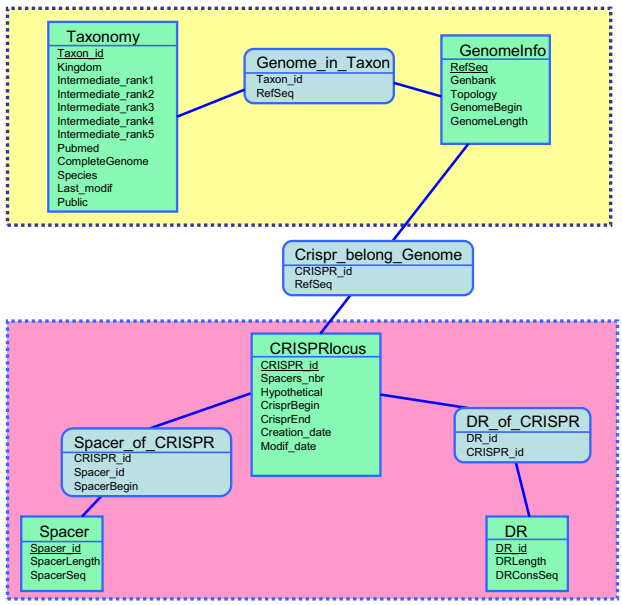


Figure 1
An entity-relationship diagram for the CRISPR database. The downloaded data are represented in the yellow box: on the left the taxonomy report information and on the right the "GenomeInfo" report information about species replicons (chromosome or plasmid). The pink box represents tables related to the CRISPR clusters: a table for the cluster locus, a table for the DR consensus and a table for the spacers.

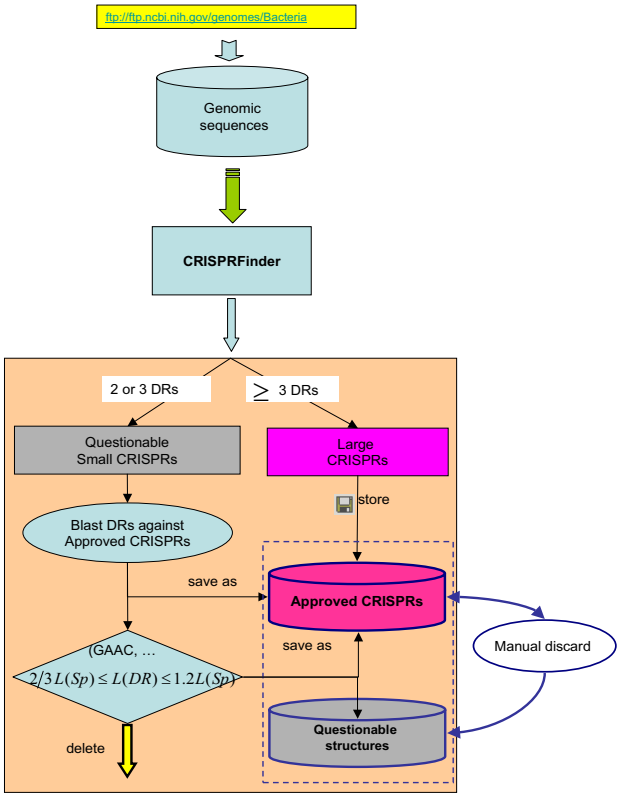


Figure 2
The database construction: from genomes to CRISPRs. The first step consists in downloading prokaryotic genomes which are then submitted to the CRISPRFinder program. The detected clusters are divided into two groups: confirmed CRISPRs (>=3DRs) are stored in the database; small questionable clusters (2 or 3 DRs) are analyzed by blasting their conserved region (DR) against the approved DRs; clusters with already identified DRs are added to the CRISPR database. Remaining questionable CRISPRs are analysed for classical flanking nucleotides and spacers length compared to the DR length. Clusters that do not fit these criteria are deleted, the remaining are kept as questionable. Manual discard of some sequences can be performed by the database curator. Colour code: programs are shown in blue, confirmed CRISPRs are in pink and questionable ones are in grey.

Utility

CRISPRdb: construction and content

Figure 3 details some of the pages which can be viewed when browsing the database [32]. On the home page (extract, top left) is displayed an alphabetical list of Bacteria and Archaea strains for which genome sequence is published, and a colour code indicates whether a CRISPR has been detected or not: species without a CRISPR are coloured in yellow, and species having at least one CRISPR are coloured in pink. The list can also be sorted according to taxonomic order, or according to database processing date. This last option makes it easy to quickly browse the latest entries. The page which appears after selecting a genome (step 1) indicates how many CRISPRs have been found and on which replicon (chromosome or plasmid) they are located. In the following page (step 2) the CRISPR id is indicated together with its position on the genome, the number of spacers and the consensus DR sequence. Querying a CRISPR locus (step 3) leads to a page containing detailed characteristics together with sequence retrieval tools: the DR consensus is shown in yellow, the spacers are shown in different colours, together with their position in the genome, the flanking sequences and the whole CRISPR locus sequence (using the flanking sequence button). Flanking sequences are displayed with flexible positions that may be modified from the 100 bp default value. Spacers can be automatically compared to public sequences databases using blastn. From this page one can access a flanking sequence CLUSTALW multiple alignment tool (FlankAlign) which is used for defining the presence of a leader and searching for homologous sequences in other genomes.

Furthermore, the ability to upload pre-calculated files (such as a summary of selected CRISPR properties or list of spacers in Fasta format, step 4) makes the tool very flexible, as the output can be analysed with other bioinformatics resources.

The CRISPR utilities page [33]

This page provides a global overview of CRISPRs present in the database, focusing on DRs and spacers (Figure 4). Firstly, all identified DRs are listed with their size expressed in base-pairs (bp), and the occurrences in the database of DRs with similar sequences is indicated as shown on the left panel of Figure 4. Selected DRs can be aligned using CLUSTALW and a dendrogram is produced (Figure 4 right panels).

Secondly a list of spacers encountered more than once provides an easy way to identify for instance the relatively rare occurrences of internal duplications within a CRISPR. A BLAST (blastn) can be run using selected spacers against public sequence databases (GenBank, EMBL, DDBJ, PDB) with a cutoff of 0.1 for the E-value and a matching length of at least 70% the queried spacer size. Thirdly, this page provides a classification of CRISPRs according to the number of motifs. The CRISPR id provides the related strain name on mouse-up and links to the page describing the CRISPR properties. Links are also provided to the corresponding pre-computed lists of DRs and spacers which can be downloaded as text files.

The BLAST CRISPRs page [34]

This page will be of use to try and validate a questionable CRISPR. From this page, a candidate DR region (or spacer) can be compared to all DRs (or spacers) characterised so

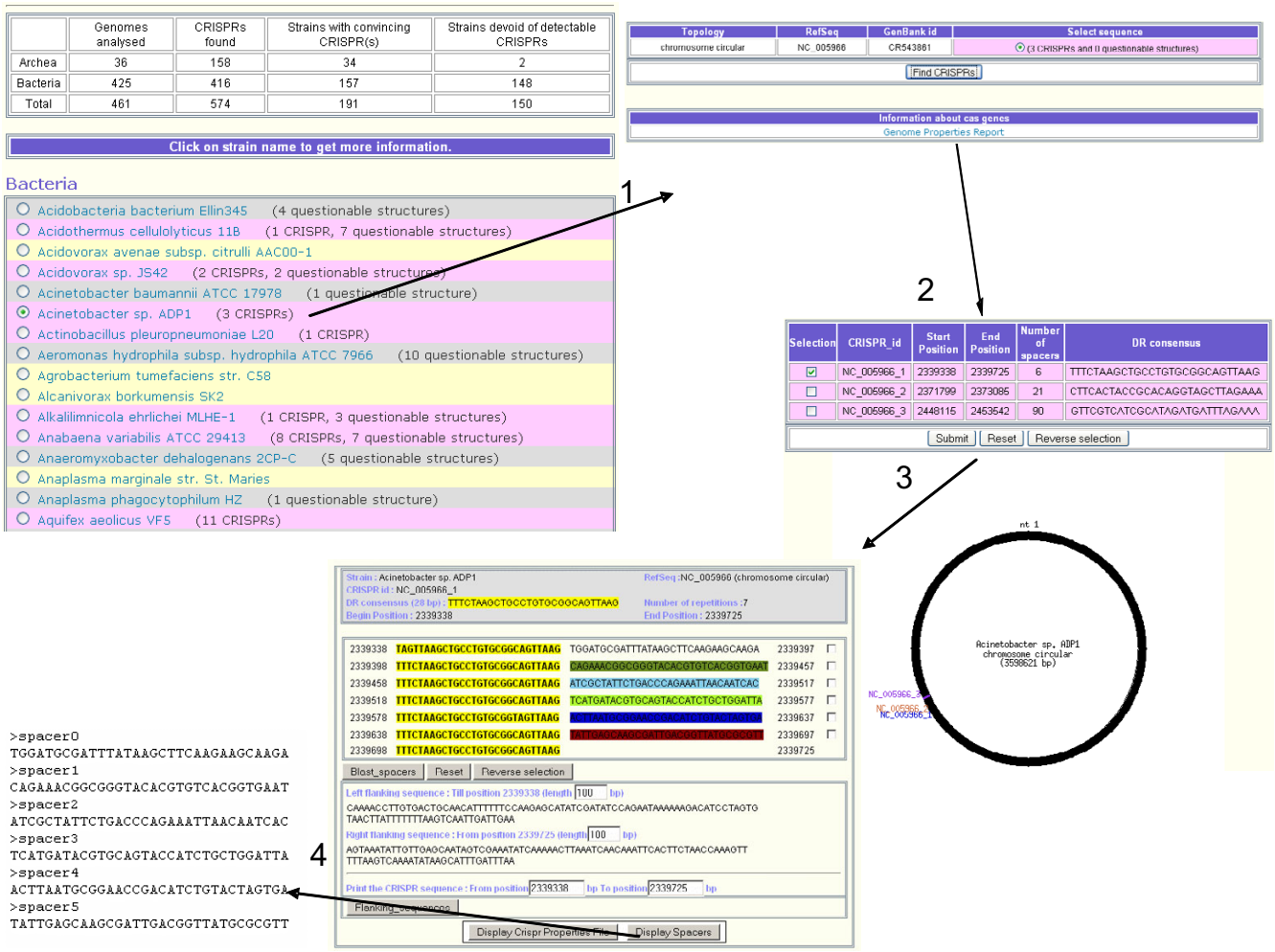


Figure 3

Screenshots of the CRISPRs web-service. 1. The opening page of the prokaryotic strains: strains in pink have at least one CRISPR, strains in grey have only questionable CRISPRs and strains in yellow have no CRISPR. 2. General information on the CRISPR clusters and their location. 3. Detailed information on the clusters: DRs are in yellow, spacers are in random colours. 4. Link to the spacers fasta file.

far from clear-cut CRISPR structures present in the database.

The Spacers Dictionary Creator page [35]

The analyses of CRISPRs in different strains of a species has shown that polymorphism exists in the number and nature of spacers [8-10,36,37]. This can be used to assess the degree of polymorphism inside the species thus providing additional information for epidemiological analyses. For this reason, it is important to be able to extract spacers from a sequence, and to store them into a database that can be queried when new sequences are produced. Upon submitting CRISPR sequences into the Spacer Dictionary Creator page, spacers are extracted and stored into an Excel file, either predefined or newly created. When a spacer is already present in the dictionary, its number appears in the output whereas a new spacer will be given a new number and will be added into the Excel file.

Discussion

Sensitivity and selectivity of CRISPRFinder

To build the CRISPRdb we have used a new program, CRISPRFinder, specifically created to identify CRISPRs. We checked that all the CRISPRs described in the literature were detected with CRISPRFinder and, in addition, we found that CRISPRFinder performs better than other CRISPR finding tools in particular in defining the DR boundaries and in identifying short CRISPRs. Among available programs, we found that PILER_CR is the most efficient. However, in the chromosome of *Aquifex aeolicus*

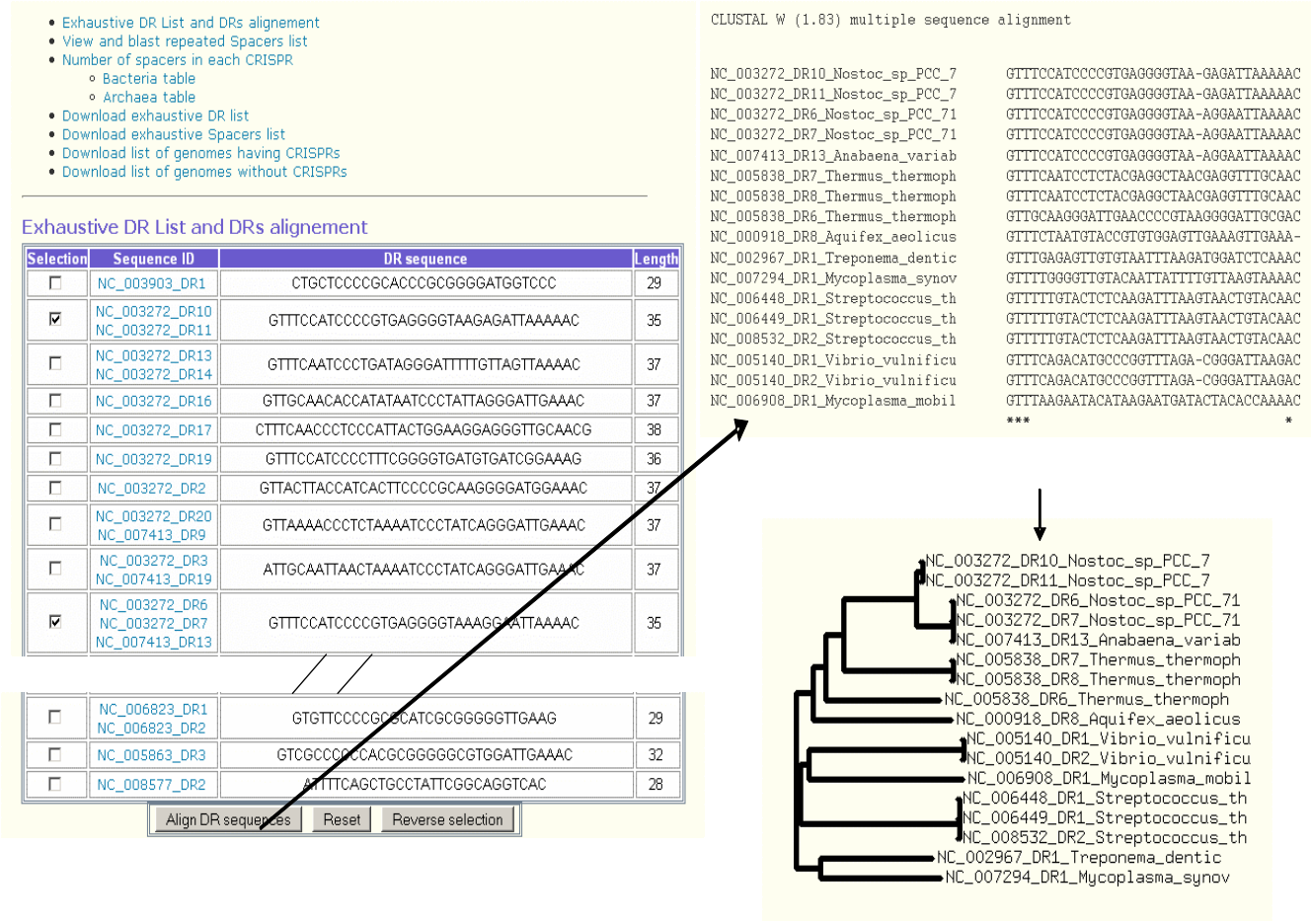


Figure 4
The DR comparison tool. Screenshot from the Utilities page showing the list of DRs with an alignment example.

VF5 (NC_000918) for instance, PILER_CR (default parameters: minarray 2, mincons 0.7, minid 0.85) detects 9 CRISPRs, three of which have misidentified DR boundaries and three are missing the truncated DR. In addition, one CRISPR locus is missed because only CRISPRs of at least three repeats are detected. CRISPR NC_000918_6 (one spacer) in the CRISPRdb was not detected by PILER_CR although it has the same DR as CRISPRs NC_000918_1, NC_000918_2, NC_000918_3 and NC_000918_10 containing respectively 5, 4, 3 and 3 repeats). Furthermore, CRISPRFinder is capable of detecting CRISPRs which DRs contain multiple differences such as NC_009009_1 and NC_009009_2 in *Streptococcus sanguinis*. Using the default parameters of PILER_CR no CRISPR was detected in this bacterium. When parameters were changed, only part of the CRISPRs were found. It will be interesting in the future to check whether these exceptional CRISPRs and *cas* genes are functional. Conversely, CRISPRFinder occasionally fails to exclude some false positives. We manually analysed all the CRISPRs identified in the current version of the database and eliminated a few false positive structures, principally tandem repeats with a low internal conservation. We estimate these cases to be less than 1% of "confirmed" CRISPRs.

Characteristics of CRISPRs

CRISPRdb has been constructed using public domain genome sequences (unpublished sequences can be submitted to CRISPRFinder to detect CRISPRs and extract the spacers). Sixty three percent (63%) of the structures qualifying as CRISPRs using the defined parameters possess 4 or less than 4 spacers. The majority of these are classified as questionable. Their confirmation or exclusion as *bona fide* CRISPR structures will require additional evidence,

such as the presence of a DR already described in a CRISPR, the presence of *cas* genes in the vicinity or the search for polymorphism within multiple isolates from the same species.

We have chosen to restrict the definition of CRISPRs to comprise DRs 23 to 55 bp-long and spacers 0.6 to 2.5 the DR size because these sizes are in excess of the range of previously described CRISPRs. These parameters do not exclude CRISPRs also containing a subset of much larger spacers as can be seen in *Methanopyrus kandleri* with spacers 51 to 72 bp-long. There are no clear rules defining the limits of a DR or a spacer and we might be missing currently unknown CRISPRs with characteristics outside of the range currently covered, even if the present rules were deduced from the published investigation by various means of more than 300 genomes. Should such CRISPRs be observed in the future, the database, as designed, can be easily adjusted.

Wide differences are observed among the CRISPRs, in the DR sequence, its size and the size of the spacers. Table 1 summarizes the size distribution observed for DRs. Interestingly, in both Archaea and Bacteria, three well-separated size classes are observed: small DRs (24–25 base-pairs), medium-size (28–30 bp) and large (36–37 bp). The smaller DRs group is more represented in Archaea (42% versus less than 2% for this size class in Bacteria) and curiously it is also where the differences between DR and spacer size are the largest. In *Pyrobaculum aerophilum* 7 CRISPRs have a 24 or 25 bp-long DRs whereas the spacer sizes range from 38 to 53 bp. The longer spacers were observed in *Methanopyrus kandleri* which possess 5 CRISPRS with DRs 35 or 36 bp-long and spacers 51 to 72 bp-long, as previously mentioned. In contrast, a remarkably constant spacer length is observed in some bacteria. In *Geobacter sulfureducens* a single CRISPR with a 29 bp DR possess one hundred and thirty eight 32 bp-long spacers and three 33 bp-long spacers. A similar situation is observed in *Mycoplasma mobile* and in *Treponema denticola*. The longest DR presently found is 47 bp-long in the CRISPR of *Bacteroides fragilis*. The associated spacers are either 29 or 30 bp-long. This suggests that the precise mechanisms producing spacers is different from one bacterium or archaeon to another although a common set of CAS proteins is generally associated with all the CRISPRs. The largest CRISPR locus was found in *Verminephrobacter eiseniae* consisting of 245 repeats on one side and 45 repeats on the other side of an IS element (NC_008786_2 and NC_008786_3). The DR is 28 bp-long and the average spacer length is 32 bp. The longest CRISPR previously described was NC_003869_3 from *Thermoanaerobacter tengcongensis* MB4 with 217 repeats.

Mojica and col. [4] observed the existence of terminal and inner-inverted repeats in the DR sequence, and Jansen and col. [5] further suggested that the secondary structure might play an essential biological role. A protein binding on one side of the repeat and producing an opening of the opposite side of the DNA structure was described in *Sulfolobus solfataricus* [38] and might be used in the processing of small RNAs [14]. A future development of our work will be the analysis of all the DRs in search for a common secondary structure that might help in understanding the role of the DR.

Table 1: Summary of the characteristics and number of CRISPRs.

DR length	Number of CRISPRs (percentage %)		
	Bacteria	**Archaea**	**Total**
47	1 (<1)	0	1 (<1)
38	3 (<1)	0	3 (<1)
37	55 (14.7)	14 (8.9)	69 (13)
36	69 (18.4)	9 (5.7)	78 (14.7)
35	10 (2.7)	1 (<1)	11 (2.6)
34	1 (<1)	0	1 (<1)
33	4 (1)	0	4 (<1)
32	31 (8.3)	1 (<1)	32 (6)
31	6 (1.6)	2 (1.3)	8 (1.5)
30	51 (13.6)	46 (29.1)	97 (18.2)
29	68 (18.1)	9 (5.7)	77 (14.4)
28	67 (17.9)	7 (4.4)	74 (13.9)
27	2 (0.53)	2 (1.7)	4 (<1)
26	1 (0.27)	0	1 (<1)
25	6 (1.6)	37 (23.4)	43 (8)
24	0	30 (19)	30 (5.7)
Total Number	375	158	533
Mean Length	32	32	32

Only confirmed CRISPRs are counted. The first column shows the DR length. In the second and the third columns are shown the number of clusters having the corresponding DR length in Bacteria and Archaea respectively (the percentage of CRISPR DR having this length is indicated). Only one strain per species is counted. In the last column, the two populations of CRISPRs are merged. The last two lines are respectively the total number of CRISPRs in each category, and the average DR length.

Inside a species several strains can share a set of spacers, but in a given CRISPR spacers are generally unique except in a few cases where duplications of one to 7 motifs (a DR and a spacer) were observed [33]. Apparently, duplications are more frequently observed in Archaea as described in detail by Lillestol *et al*. [10].

It is important to note that the absence of CRISPR in one strain does not imply that CRISPRs are absent from all the members of the corresponding species. However in some species or genus no CRISPR has been identified yet although a number of strains have been fully sequenced. This is the case for example in *Staphylococcus aureus* and *Burkholderia sp*.

Multiplication of CRISPR

It is believed that CRISPR and associated genes *cas* can be horizontally transferred between bacteria of different species and possibly between Archaea and Bacteria. This is strongly suggested by comparison of CAS protein sequences, but it does not explain how several CRISPRs with a similar DR can be present in a single genome, only one of which being associated with *cas* genes. The small CRISPRs are particularly interesting in this respect to try and elucidate the mechanism of creation of a new CRISPR and of insertion of new motifs in an existing CRISPR. For example in *Clostridium tetani* among eight CRISPRs possessing 1 to 33 motifs, seven are clustered between position 1570766 and 1595950 (spanning 25.184 bp), five of which with exactly the same DR and two with a derivative (6 different nucleotides out of 30). The leaders of the seven clustered CRISPR aligned over about 150 bp with 80% similarity, *cas* genes are present once between CRISPR 5 and CRISPR 6 and no spacer is in common. It is then most likely that starting from an ancestral complete CRISPR and *cas* genes locus, new CRISPRs have been created not by duplication of the complete complex but rather by the insertion of a minimum structure comprising a leader sequence, a DR, and no spacer, which then grows by adding new motifs. This absence of common spacers even when several CRISPRs are present in a single Bacteria or Archea is also suggesting that gene conversion is not a significant process for new motif acquisition.

The CRISPR intra-species polymorphism: insight into the mechanism of acquisition of new motifs

We developed the spacer dictionary tool to facilitate the extraction of spacers and their analysis, principally for phylogenetic studies. To better demonstrate the efficiency of this tool we propose a demonstrator based on the sequences of five *Y. pestis* genomes. An initial dictionary was first created from the 26 published spacers, named using the alphabet from "a" to "z" [8]. The CRISPRs of newly sequenced alleles as could be derived from sequencing the locus in a collection of diverse strains can be submitted to the dictionary tool in fasta format. The spacers which were not already present in the dictionary are given a number and they are added sequentially into the dictionary. The alleles are coded in a convenient way using this dictionary.

In our previous study of three CRISPRs in 180 *Y. pestis* isolates, most of which were genetically very similar, we described the polymorphism at each locus due to different number of motifs [8]. Our observations suggested that one or several motifs could be lost by precise deletion between 2 DRs whereas new motifs were added precisely at the level of the last DR flanking the leader. A similar suggestion was made based upon observations in *S. solfataricus* P1 and in *S. thermophilus* [9,10]. This mechanism is further supported by the analysis of the structure of some CRISPRs in which a first series of motifs containing a particular DR is followed by motifs with a DR differing at a single nucleotide up to the last one near the leader. For example in the CRISPR NC_005085_3 of *Chromobacterium violaceum*, 13 motifs with DR "GTGTTCCCCACGT-GCGTGGGGATGAACCG" are followed by 6 motifs with DR "GTGTTCCCCACGCCCGTGGGGATGAACCG". Another interesting example is found in *Carboxydothermus hydrogenoformans* where two CRISPRs, NC_007503_3 and NC_007503_4 (59 and 84 spacers respectively) share the same 30bp-DR, although in one of them the last 13 repeats adjacent to the leader have a modified DR. The first three bases of the DR are absent whereas the three bases AAC are added to the other end to produce a modified DR (Figure 5). This suggests that at some point the last DR plus 3 bases of a newly added spacer were duplicated to create a new DR which then served as a matrix for subsequent duplications. Alternatively, the AAC addition could be the result of some stuttering since the initial DR ends by AAAAC (and the modified DR by AAAACAAC). These observations are in favour of the model of polarised sequential insertion of new motifs by duplication of the last DR and insertion of a new spacer [8,10], rather than random insertion by homologous recombination as proposed by Makarova *et al*. [11]. If the newly copied last DR contains a mutation, compatible with CRISPR metabolism, then this mutation will be copied in all subsequent motif acquisitions.

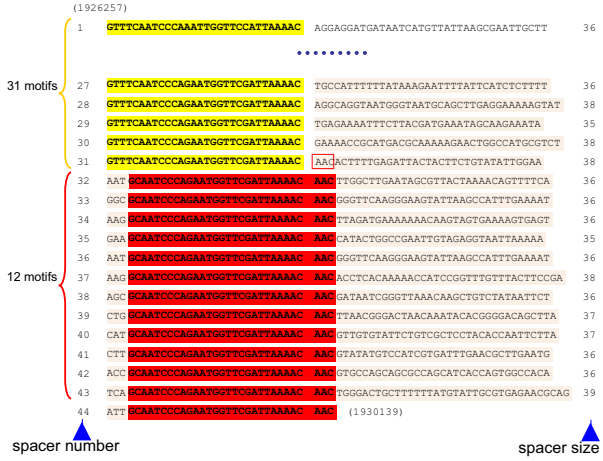


Figure 5
The first and last 17 motifs of CRISPR NC_007503_3 from *Carboxydothermus hydrogenoformans* Z-2901. The DRs shared by the two CRISPR loci NC_007503_3 and NC_007503_4 are shown in yellow and the variant DR observed only in NC_007503_3 is in red. CRISPR units (DR + spacer) are numbered on the left and spacers' length is indicated on the right.

Future developments

Further development of our software will include new parameters to analyse genomes for which only questionable structures were detected. An additional aspect will be the identification of minimum CRISPRs structure, devoid of spacers and comprising only a DR and leader.

Conclusion

The described software and database are exclusively devoted to the identification and the analysis of CRISPRs structures, *i.e*. the succession of motifs made up of DRs and spacers. A database for *cas* gene identification has been developed by TIGR [15]. We have added a link to this web page in order to search for the presence of *cas* genes in the vicinity of a CRISPR.

CRISPRs are fascinating structures, which conceal complex biological mechanisms to account for their transfer, evolution and behaviour. They have probably played an important role in the evolution of Archaea and Bacteria by providing a defence mechanism against foreign DNA. A lot remains to be discovered, and this necessitates the possibility to rapidly investigate newly sequenced genomes, and to be able to easily browse across many different species. The CRISPRdb and associated web service provides all the necessary tools to decipher the organisation of these structures. Several studies have shown that when an origin can be found for a spacer, it is most frequently a virus or a plasmid sequence. Thus the spacer database will serve as a repository of sequences of probable viral or plasmid origin. Finally the intra-species polymorphism of CRISPRs and their evolution mode (organised acquisition and loss of motifs) make them interesting tools for epidemiological studies. The possibility exists that a given spacer be added twice independently into a CRISPR, which could hamper its use for phylogenetic studies. However the polarized addition of motifs, and limited events of recombination insure that their order should be preserved. In *Y. pestis* we believe that they could be used to investigate ancient DNAs (Vergnaud et al. in press).

Availability and requirements

The resource described here is accessible with no restrictions, except for the demand to quote the site [32] (see Creative Commons license on the site).

Authors' contributions

GV and CP designed the study. IG developed the programs and database, and ran initial tests. Additional tests were done by IG, GV and CP together with collaborators. CP, GV and IG wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Bernard Labedan and Olivier Lespinet for their valuable comments. We thank the reviewers for their constructive analysis and comments of the manuscript.

References

1. Nakata A, Amemura M, Makino K: **Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome.** *J Bacteriol* 1989, **171(6):**3553-3556.
2. Groenen PM, Bunschoten AE, van Soolingen D, van Embden JD: **Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method.** *Mol Microbiol* 1993, **10(5):**1057-1065.
3. Mojica FJ, Ferrer C, Juez G, Rodriguez-Valera F: **Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning.** *Mol Microbiol* 1995, **17(1):**85-93.
4. Mojica FJ, Diez-Villasenor C, Soria E, Juez G: **Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria.** *Mol Microbiol* 2000, **36(1):**244-246.
5. Jansen R, Embden JD, Gaastra W, Schouls LM: **Identification of genes that are associated with DNA repeats in prokaryotes.** *Mol Microbiol* 2002, **43(6):**1565-1575.
6. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD: **Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin.** *Microbiology* 2005, **151(Pt 8):**2551-2561.
7. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E: **Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.** *J Mol Evol* 2005, **60(2):**174-182.
8. Pourcel C, Salvignol G, Vergnaud G: **CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies.** *Microbiology* 2005, **151(Pt 3):**653-663.
9. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero D, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes.** *Science* 2007, **315:**1709-1712.
10. Lillestol RK, Redder P, Garrett RA, Brugger K: **A putative viral defence mechanism in archaeal cells.** *Archaea* 2006, **2(1):**59-72.
11. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV: **A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.** *Biol Direct* 2006, **1:**7.
12. Godde JS, Bickerton A: **The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes.** *J Mol Evol* 2006, **62(6):**718-729.
13. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes.** *PLoS Comput Biol* 2005, **1(6):**e60.
14. Tang TH, Bachellerie JP, Rozhdestvensky T, Bortolin ML, Huber H, Drungowski M, Elge T, Brosius J, Huttenhofer A: **Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*.** *Proc Natl Acad Sci U S A* 2002, **99(11):**7536-7541.
15. **The TIGRFAM page** [http://www.tigr.org/TIGRFAMs/]
16. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R: **REPuter: the manifold applications of repeat analysis on a genomic scale.** *Nucleic Acids Res* 2001, **29(22):**4633-4642.
17. Jansen R, van Embden JD, Gaastra W, Schouls LM: **Identification of a novel family of sequence repeats among prokaryotes.** *Omics* 2002, **6(1):**23-33.
18. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.
19. Durand P, Mahe F, Valin AS, Nicolas J: **Browsing repeats in genomes: Pygram and an application to non-coding region analysis.** *BMC Bioinformatics* 2006, **7:**477.
20. Edgar RC: **PILER-CR: fast and accurate identification of CRISPR repeats.** *BMC Bioinformatics* 2007, **8:**18.
21. **The Perl directory** [http://www.perl.org/]
22. **BioPerl** [http://www.bioperl.org/]

23. **The Apache Software Foundation** [http://www.apache.org/]
24. **Debian** [http://www.debian.org/]
25. **The NCBI ftp site for Bacterial and Archaeal genome sequences** [ftp://ftp.ncbi.nih.gov/genomes/Bacteria]
26. **The CRISPRFinder** [http://crispr.u-psud.fr/Server/CRISPRfinder.php]
27. **PHP** [http://www.php.net/]
28. **MySQL** [http://www.mysql.com/]
29. **Vmatch** [http://www.vmatch.de/]
30. Abouelhoda M, Kurtz S, Ohlebusch E: **Replacing suffix trees with enhanced suffix arrays.** *Journal of Discrete Algorithms* 2004, **2:**53-86.
31. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC: **The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide.** *Nucleic Acids Res* 2006, **34(Database issue):**D332-4.
32. **The CRISPR database** [http://crispr.u-psud.fr]
33. **CRISPRUtilities** [http://crispr.u-psud.fr/crispr/CRISPRUtilitiesPage.html]
34. **BLAST CRISPRs** [http://crispr.u-psud.fr/crispr/BLAST/CRISPRsBlast.php]
35. **The CRISPR spacers dictionary** [http://crispr.u-psud.fr/crispr/MultipleAnalysis/CRISPRdetector.php]
36. Schouls LM, Reulen S, Duim B, Wagenaar JA, Willems RJ, Dingle KE, Colles FM, Van Embden JD: **Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination.** *J Clin Microbiol* 2003, **41(1):**15-26.
37. Hoe N, Nakashima K, Grigsby D, Pan X, Dou SJ, Naidich S, Garcia M, Kahn E, Bergmire-Sweat D, Musser JM: **Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains.** *Emerg Infect Dis* 1999, **5(2):**254-263.
38. Peng X, Brugger K, Shen B, Chen L, She Q, Garrett RA: **Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in Sulfolobus genomes.** *J Bacteriol* 2003, **185(8):**2410-2417.

4.3 Investigations des CRISPRs de *Y. pestis* et *Y. pseudotuberculosis*

L'article suivant (Vergnaud, 2007) intitulé "Analysis of the Three *Y. pestis* CRISPR Loci Provides New Tools for Phylogenetic Studies and Possibly for the Investigation of Ancient DNA" (L'analyse des trois locus CRISPR de *Y. pestis* fournit de nouveaux outils pour les études phylogénétiques et la possibilité d'investigation de l'ADN ancien), a constitué la première application des ressources bioinformatiques d'investigation des CRISPRs. Cette étude fournit une vue élargie sur la diversité des CRISPRs de *Y. pestis*. Un catalogue de 70 spacers a été construit pour le CRISPR YP1 et semble couvrir les spacers les plus fréquents. Il est alors possible de penser à élaborer une technique de spoligotypage à partir de ce CRISPR pour améliorer les connaissances actuelles sur l'agent responsable des pandémies de la peste.

Résumé :

La nature précise du pathogène ayant causé les pandémies de peste est incertaine. Bien que *Y. pestis* soit probablement un candidat pour les trois pandémies de la peste, les quelques preuves directes qui peuvent être déduites de l'analyse d'ADN (aADN) ancien sont controversées. De plus, il y a toujours un débat concernant celui des trois biovars, Antiqua, Medievalis ou Orientalis, associé à chacune de ces pandémies. Il est nécessaire d'effectuer des analyses phylogénétiques sur des souches de *Y. pestis* isolées dans des pays où la peste s'est produite et est encore endémique. De plus, il existe des difficultés techniques relatives à l'investigation de l'aADN et une absence de cible génétique. La structure CRISPR récemment décrite pourrait représenter une telle cible. Les locus CRISPRs consistent en une succession de régions très bien conservées séparées par des spacers spécifiques généralement d'origine virale. Pour être utilisables, des données décrivant les mécanismes d'évolution et la diversité des CRISPRs de *Y. pestis*, de ses voisins proches, et d'autres espèces qui peuvent contaminer l'ADN ancien, sont nécessaires. L'investigation d'isolats de *Y. pestis* génétiquement très proches a révélé des événements de mutation récents dans lesquels les éléments qui constituent les CRISPRs sont acquis ou perdus, ce qui a fourni un aperçu essentiel sur leur évolution. Les règles déduites sont fondamentales dans les interprétations ultérieures. Dans cette étude, les locus CRISPRs d'un échantillon représentatif de *Y. pestis* et *Y. pseudotuberculosis* ont été analysés via des amplifications par PCR et des analyses de séquence. L'investigation de ce grand panel de souches comprenant quelques sous espèces ou écotypes à l'intérieur des souches de *Y. pestis* et *Y. pseudotuberculosis* fournit une base de données des spacers existants et aide à la prédiction de la structure CRISPR attendue de l'ancêtre de *Y. pestis*. Cette connaissance ouvrira la voie au développement d'un protocole de spoligotypage dans lequel les spacers pourraient être amplifiés même lorsqu'ils proviennent d'échantillons d'ADN très dégradé. Les données obtenues montrent que l'analyse CRISPR peut fournir un outil de typage puissant, adapté à un génotypage systématique et à grande échelle des isolats de *Y. pestis*

et la création de bases de données internationales pour le typage de cette espèce. De plus, les CRISPRs constituent un outil prometteur et une cible génétique intéressante pour l'investigation d'ADN ancient. Les cibles génétiques correspondantes sont assez petites (<70pb), présentes sous forme de plusieurs copies (généralement plus que 10), hautement conservées et spécifiques. De plus, la production d'une séquence de très grande qualité n'est pas indispensable à l'interprétation des données, ce qui est important dans le cadre de l'étude d'ADN ancien.

30

Analysis of the Three *Yersinia pestis* CRISPR Loci Provides New Tools for Phylogenetic Studies and Possibly for the Investigation of Ancient DNA

Gilles Vergnaud[1,2], Yanjun Li[3], Olivier Gorgé[2], Yujun Cui[3], Yajun Song[3], Dongsheng Zhou[3], Ibtissem Grissa[1], Svetlana V. Dentovskaya[4], Mikhail E. Platonov[4], Alexander Rakin[5], Sergey V. Balakhonov[6], Heinrich Neubauer[7], Christine Pourcel[1], Andrey P. Anisimov[4], and Ruifu Yang[3]

[1] Univ Paris-Sud, Institut de Génétique et Microbiologie, gilles.vergnaud@igmors.u-psud.fr
[2] Division of Analytical Microbiology, Centre d'Etudes du Bouchet
[3] Institute of Microbiology and Epidemiology, Academy of Military Medical Sciences
[4] State Research Center for Applied Microbiology and Biotechnology
[5] Max von Pettenkofer-Institute of Hygiene and Medical Microbiology
[6] Antiplague Research Institute of Siberia and Far East
[7] Friedrich-Loeffler Institute

Abstract. The precise nature of the pathogen having caused early plague pandemics is uncertain. Although *Yersinia pestis* is a likely candidate for all three plague pandemics, the very rare direct evidence that can be deduced from ancient DNA (aDNA) analysis is controversial. Moreover, which of the three biovars, Antiqua, Medievalis or Orientalis, was associated with these pandemics is still debated. There is a need for phylogenetic analysis performed on *Y. pestis* strains isolated from countries from which plague probably arose and is still endemic. In addition there exist technical difficulties inherent to aDNA investigations and a lack of appropriate genetic targets. The recently described CRISPRs (clustered regularly interspaced short palindromic repeats) may represent such a target. CRISPR loci consist of a succession of highly conserved regions separated by specific "spacers" usually of viral origin. To be of use, data describing the mechanisms of evolution and diversity of CRISPRs in *Y. pestis*, its closest neighbors, and other species which might contaminate ancient DNA, are necessary.

The investigation of closely related *Y. pestis* isolates has revealed recent mutation events in which elements constituting CRISPRs were acquired or lost, providing essential insight on their evolution. Rules deduced represent the basis for subsequent interpretation. In the present study, the CRISPR loci from representative *Y. pestis* and *Yersinia pseudotuberculosis* strains were investigated by PCR amplification and sequence analysis. The investigation of this wider panel of strains, including other subspecies or ecotypes within *Y. pestis* and also *Y. pseudotuberculosis* strains provides a database of the existing CRISPR spacers and helps predict the expected CRISPR structure of the *Y. pestis* ancestor. This knowledge will open the way to the development of a spoligotyping assay, in which spacers can be amplified even from highly degraded DNA samples.

The data obtained show that CRISPR analysis can provide a very powerful typing tool, adapted to the systematic, large-scale genotyping of *Y. pestis* isolates, and the creation of

international typing databases. In addition, CRISPRs do constitute a very promising new tool and genetic target to investigate ancient DNA. The corresponding genetic targets are small (<70bp), present in multiple copies (usually more than 10), highly conserved and specific. In addition, the assay can be run in any laboratory. Interpretation of the data is not dependent on accurate sequence data.

30.1 Introduction

In the past few years, and owing in part to the availability of whole genome sequence data from many bacterial species including different strains from the same species, a high number of polymorphisms sources, and consequently of typing methods, has emerged. Regarding *Yersinia pestis*, these include MLST (Multiple Loci Sequence Typing) (Achtman et al. 1999), SNPs (Single Nucleotide polymorphism) (Achtman et al. 2004), MLVA (Multiple Loci VNTR Analysis) (Achtman et al. 2004; Pourcel et al. 2004), and CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) analysis (Pourcel et al. 2005).

These new methods are likely to replace the previous pattern-comparison methods (such as IS typing by southern blotting, or pulsed-field gel electrophoresis) which are more expensive and not fully appropriate to the creation of international databases. One among the emerging methods takes advantage of the polymorphism of particular structures, the CRISPR loci. CRISPRs are well-defined structures (Fig. 1). They are present in many bacteria and in most archaea, sometimes in multiple copies. The CRISPR structure itself is usually surrounded by CRISPRs-associated genes (*cas* genes) (Jansen et al. 2002). CRISPRs have been shown to be transcribed, and the transcription product is processed into micro-RNAs (Tang et al. 2002). New spacers are not synthesized *de novo*, but are copied from existing DNA sequences (Pourcel et al. 2005). The vast majority of known spacers lack any similarity with currently available sequences. However, when similarities exist, they most often correspond to short portions of mobile elements such as phages. These observations have led to the suggestion that CRISPRs were a defense-mechanism against genetic aggressions (Mojica et al. 2005; Pourcel et al. 2005; Lillestol et al. 2006; Makarova et al. 2006).

Simple evolution rules have been proposed for CRISPR which open the way to phylogenetic investigations (Pourcel et al. 2005; Lillestol et al. 2006): (1) new spacers are acquired in a polarized way from one extremity adjacent to the leader sequence which acts as a transcription promoter; (2) losses may occur randomly along the array; (3) the probability of acquisition of the same spacers independently is extremely low.

The analysis of CRISPR has already played an important role in investigating the epidemiology of the major human pathogen, *Mycobacterium tuberculosis* (the corresponding typing method is called “spoligotyping”). A database containing the typing information from thirty thousands isolates has been built (Brudey et al. 2006). Although this represents only a very small fraction of TB isolates worldwide, the database is by far the largest existing typing database for a bacterial pathogen. One reason for this is that the method was sufficiently robust, easy to run at a reasonable cost, so that many laboratories could produce data easy to share and eventually

merge. Another reason was that the resulting data did make sense and enabled the definition of large families of strains. A third reason was the relatively simple situation of the CRISPR locus in *M. tuberculosis*. The locus is apparently inactive, it does not acquire new spacers, so that a fixed and limited set of relevant spacers could be defined to produce "spoligotyping membranes".

The availability of a similar approach for *Y. pestis* would be of use for at least two reasons. The first one is that it would allow for the large-scale screening of *Yersinia pseudotuberculosis* strains in search of *Y. pestis* closest neighbors, as well as for the systematic routine typing of current *Y. pestis* collections and new isolates. The second one is that CRISPRs represent potentially very interesting tools for the investigation of ancient DNA (aDNA). The nature of the early plague pandemics is still controversial, and one reason for this situation may be the lack of appropriate genetic targets for *Y. pestis* aDNA investigation.

Y. pestis CRISPR loci are still active (Pourcel et al. 2005) and able to acquire new spacers (in contrast with the *M. tuberculosis* CRISPR). Therefore, it is necessary to list the repertoire of existing CRISPR spacers. If this repertoire eventually turns out to be very large, as in some bacteria, then the development of such an assay will necessitate the use of DNA chips which are able to deal with a larger number of spacers compared to the current spoligotyping assay format. The first repertoire of

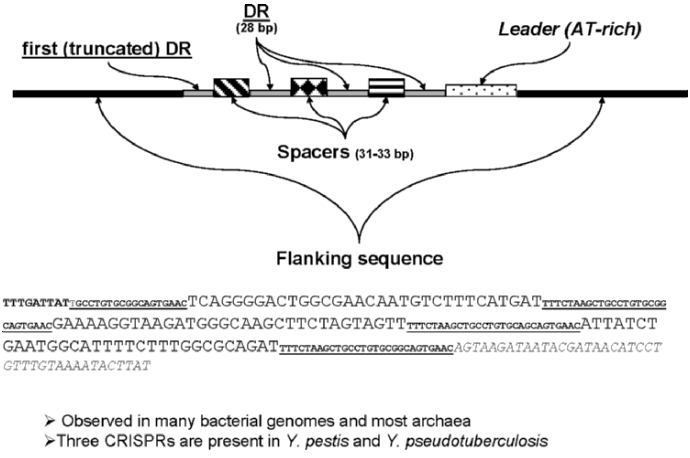


Fig. 1. Organization of the CRISPR region. The CRISPR locus generally begins with a truncated direct repeat (DR) unit, followed by a succession of spacers and DRs. The leader sequence is presumed to act as the promoter region for transcription. In *Y. pestis* and *Y. pseudotuberculosis* CRISPRs, the DR is 28 bp long, and the spacers are 31 to 33 bp long. This may be different in other CRISPRs, and the observed range so far for DRs is 24 to 47 bp. Different spacers are totally unrelated. In the sample sequence, DRs are underlined, spacers are in larger typeset, the start of the leader sequence is in italics.

CRISPR spacers was deduced from the analysis of only a small part of *Y. pestis* genetic diversity (Pourcel et al. 2005). Less than 40 spacers were identified, in the three *Y. pestis* CRISPR loci, suggesting that perhaps the repertoire within a larger collection might remain tractable.

The detailed field investigations of *Y. pestis* natural foci in the former Soviet Union and China provide a relatively complete view of *Y. pestis* diversity for these countries. In addition to the *Y. pestis* main subspecies *pestis*, which is highly pathogenic for humans, five subspecies have been defined based upon biochemical analyses, geographic distribution, and favored host. These subspecies are called *caucasica, altaica, hissarica, ulegeica, talassica* (reviewed by (Anisimov et al. 2004)) and a similar biovar Microtus was proposed for isolates from *Microtus brandti* and *Microtus fuscus* in China (Song et al. 2004). The collective "pestoides" name is also used. Biochemical analysis suggests that the *caucasica* subspecies represents the oldest lineage. In addition a unique example of an African pestoides (the so-called "Angola" strain) has been described and the genome is currently being sequenced. Investigation on this strain using a number of molecular typing methods suggests that it represents a lineage which is even older than the *caucasica* lineage (Achtman et al. 2004).

In the present study, a representative collection of strains was investigated in order to produce a library of the most frequent spacers present within *Y. pestis*.

30.2 CRISPRs as a Potential Tool for Large-scale Screening of *Yersinia pseudotuberculosis* Strains Most Closely Related to *Yersinia pestis*

Whereas *Y. pestis* evolution studies attracts some attention and interest, mainly for biodefense purposes (microbial forensics), similar investigations in *Y. pseudotuberculosis* are much more limited. One reason for this is the huge diversity existing in the *Y. pseudotuberculosis* species, as compared to *Y. pestis*, and the very large number of *Y. pseudotuberculosis* isolates. A MLST assay was used by Achtman et al. to better define the relative position of *Y. pestis* and *Y. pseudotuberculosis* (Achtman et al. 1999). In this study, approximately 2 kb of sequence data were produced for a small number of diverse *Y. pseudotuberculosis* strains. The data obtained showed that all *Y. pestis* strains investigated were identical, whereas *Y. pseudotuberculosis* strains were quite diverse. Figure 2, based on published data, illustrates this finding. Although this work clearly confirms the recent emergence of *Y. pestis*, it also shows the distance between the two species. Many isolates representing intermediate evolutionary steps are missing. Isolating such intermediates will be necessary to understand the emergence of *Y. pestis*. MLST is a very powerful method, however it is not presently adapted to the large-scale and low-cost screening required to undertake

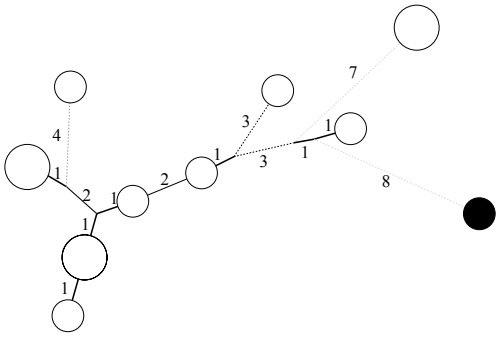


Fig. 2. Twelve *Y. pseudotuberculosis* and 36 *Y. pestis* strains compared by MLST analysis. MLST data (Achtman et al. 1999) were re-analysed and presented in a way which illustrates the close relationship between the two species and within *Y. pestis* in terms of point mutations observed within approximately 2 kb of sequence. The number of point mutations occurring along each segment is indicated (open circles: *Y. pseudotuberculosis* strains; full circle, 36 diverse *Y. pestis* strains).

such a study. CRISPRs typing may offer an alternative but additional data is needed to validate this approach.

30.3 CRISPRs as a Tool for Ancient DNA Investigations

Ancient DNA analysis is probably the only way to prove the role if any of a given pathogen in long past pandemics. However, aDNA investigations raise a number of very challenging technical issues due to DNA degradation, rarity, contamination, and chemical alterations (Prentice et al. 2004; Shapiro et al. 2006). The quality and conditions of corpse conservation are obviously key parameters and varies according to local geographic context (soil nature, average humidity, temperature, etc.). For instance, investigation of aDNA in search of *M. tuberculosis* traces in mummies (Zink et al. 2003) is more favorable than the search for *Y. pestis* traces in remains from collective graves and direct soil burial (Gilbert et al. 2004). A key element in aDNA investigations is the choice of bacterial genetic targets which take into account the characteristics of aDNA (Table 1). One major aspect is that because the sample will be usually contaminated with foreign DNA, the target should have a strong phylogenetic content (Shapiro et al. 2006). The other aspects are the size of the target and the copy number, which should be respectively as small and as high as possible. Finally, data analysis should not be extensively dependent upon the accuracy of sequence

Table 1. Ancient DNA analysis challenges

Wish list: the appropriate target for aDNA analysis	
Because aDNA is:	The appropriate genetic target should be:
Fragmented	Small
Present in low amounts	Present in multiple-copies, stable
Contaminated	With a high phylogenetic content
Subject to erroneous nucleotide incorporation	And interpretation not dependant upon sequence accuracy
Y. pestis CRISPRs satisfy all these criteria	

data. There is no consensus so far on the choice of the genetic target in the case of *Y. pestis*. Plasmid targets are sometimes used, for their relatively high copy number. However, plasmids have a low phylogenetic value as they are usually acquired by horizontal transfer from different species. In a more recent investigation (Drancourt et al. 2004), a polymorphic tandem repeat (VNTR) was targeted but some of the weaknesses of this approach were subsequently demonstrated (Vergnaud 2005). The investigators were able to amplify a particular VNTR allele from ancient remains from the first two pandemics. Among the *Y. pestis* strains investigated so far, only strains from biovar Orientalis possess this VNTR allele. Consequently the authors concluded that all three pandemics were caused by Orientalis. However the phylogenetic evidence is too weak, because the strain collection investigated is by far not representative of the diversity of *Y. pestis*. Indeed studying an enlarged collection, the same "Orientalis" allele was observed in Antiqua and Medievalis strains as well (Yang and colleagues, unpublished data).

30.4 CRISPRs Diversity Within *Y. pestis* and *Y. pseudotuberculosis*

Y. pestis and *Y. pseudotuberculosis* contain three CRISPRs called YP1, YP2 and YP3 (Pourcel et al. 2005) (Fig. 3A). Since CRISPR loci can also be considered to some extent as polymorphic tandem repeats, they have been designated, respectively, ms06, ms76 and ms77 (Le Flèche et al. 2001; Pourcel et al. 2005). The number of motifs (one DR and one spacer) in an allele is easily deduced from the size of the PCR product obtained by using flanking primers. These PCR products can be sequenced to identify the spacers (Fig. 3A-3B).

It is this approach, applied to many very closely related isolates, some of which differed at the CRISPR loci, which led to the current model of evolution for CRISPR (Pourcel et al. 2005) (Fig. 4).

```
RS 28bp                    Sp 32-33bp
VNTRyp2769ms06 CRISPR YP1
GTTACAAAATGCGCTTCCGCTCGCAATTTTGCTCCCCAAATAGCATCAGCACATGGCCCA
tttgattatTGCCTGTGCGGCAGTGAACTCAGGGGACTGGCGAACAATGTCTTTCATGAT      a
TTTCTAAGCTGCCTGTGCGGCAGTGAACGAAAAGGTAAGATGGGCAAGCTTCTAGTAGTT      b
TTTCTAAGCTGCCTGTGCGGCAGTGAACATTATCTGAATGGCATTTTCTTTGGCGCAGAT      c
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCGCCATTCCGTGAACCTGAGCGCGTTCGCGA      d
TTTCTAAGCTGCCTGTGCGGCAGTGAACATATTCTCGAGCGATAGCAATAGCCATTCCAC      e
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCGGTCAAACAAATTTAGGCGACGATTTAACA      f
TTTCTAAGCTGCCTGTGCGGCAGTGAACAAAAAGAATTTGGGATTAAAGTTACCCATCAG      g
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCAATGCCTGAATCTCTGGCGTGATAGCTGCGG     h
TTTCTAAGCTGCCTGTGCGGCAGTGAACAGTAAGATAATACGATAACATCCTGTTTGTAA
AATACTTATTTCGCTAATGGGGAAAAAACCCTTTTTTTAGACCACCGATAACCACAATGT
AAAATCAATGAGTTAGCAGTAGCTAAAAAAATAGGGTCAGAACATAACTCATAATAAAAC

yp2895 CRISPR YP2
CAGGTAGATGCCTTCCGATCTCAATCAGCCACGCTCTGTCTAGTGCAGTCGCTGGTCGTG
GCGTTGGCCTACCAGCAGGAGGCGCAGGCCGGGGCCGCGCTGGCGCACAGACAGTGACCC
tctaTAAGCTGCCTGTGCGGCAGTGAACTCTGTACGCATACCGCCATCTTGCATCAGTCT      a2
TTTCTAAGCTGCCTGTGCGGCAGTGAACAGCAAAAATCTTAATTACATCTGATGATTTCGG     b2
TTTCTAAGCTGCCTGTGCGGCAGTGAACTTTACGGCACGGCGAAAGATTCGGTTCTTGTC      c2
TTTCTAAGCTGCCTGTGCGGCAGTGAACTTCTGGATAGGACAAATAGGATGATTGTATCAG     d2
TTTCTAAGCTGCCTGTGCGGCAGTGAACAACGAACCCACGTAGAATTGCCATCACCGCCGG     e2
TTTCTAAGCTGCCTGTGCGGCAGTGAACAGTAAGATAATACGGGTAACAGACTGTTTGTAA
AATAATTCTTTCGCCAAAGGGTAAAAAATGATTTTTTTTAACCCTCGGTAAGCAGGATAT
AAAATCAATGAGTTAGCCATAGCTAAAAAAATAGGGTCAAAAAATGATTCCCCTGATGCG

yp1773 CRISPR YP3
AATATGCCAAGGGATTAGTGAGTTAATATTTGCAGATAAAACGCCGCCAGAGAGCTGAGA
ttattggGCTGCCTGTGCGGCAGTGAACGTTATACCCCGCGCAGGGAGTGAAGCGTTGAC      a3
TTTCTAAGCTGCCTGTGCGGCAGTGAACTTAAGTTCTTTTTGTCAGCATCTTTAATAAATA     b3
TTTCTAAGCTGCCTGTGCGGCAGTGAACCTGAAATACAAATAAAATAAATCGTCGAACATA     c3
TTTCTAAGCTGCCTGTGCGGCAGTGAACGTAAGATAATACGGATAACCCGATGTTTATCAA
ATGAGCAATGGCGCAAAATGCGTAAAAACCCTTTTTTTAGTGAAATACCTGAGTAGCATA
AAAATCAATACGTTAGTCATAGTGATAAAAAGAGGGTCACAAGAATCGGGGGGGACGTAA
```

3A

Strain 195P

```
tttgattatTGCCTGTGCGGCAGTGAACATATTCTCGAGCGATAGCAATAGCCATTCCAC      e
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCGGTCAAACAAATTTAGGCGACGATTTAACA      f
TTTCTAAGCTGCCTGTGCGGCAGTGAACAAAAAGAATTTGGGATTAAAGTTACCCATCAG      g
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCAATGCCTGAATCTCTGGCGTGATAGCTGCGG     h
TTTCTAAGCTGCCTGTGCGGCAGTGAACAGTAAGATAATACGATAACATCCTGTTTGTAA
```

Strain Java9

```
tttgattatTGCCTGTGCGGCAGTGAACTCAGGGGACTGGCGAACAATGTCTTTCATGAT      a
TTTCTAAGCTGCCTGTGCGGCAGTGAACGAAAAGGTAAGATGGGCAAGCTTCTAGTAGTT      b
TTTCTAAGCTGCCTGTGCGGCAGTGAACATTATCTGAATGGCATTTTCTTTGGCGCAGAT      c
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCGCCATTCCGTGAACCTGAGCGCGTTCGCGA      d
TTTCTAAGCTGCCTGTGCGGCAGTGAACATATTCTCGAGCGATAGCAATAGCCATTCCAC      e
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCGGTCAAACAAATTTAGGCGACGATTTAACA      f
TTTCTAAGCTGCCTGTGCGGCAGTGAACAAAAAGAATTTGGGATTAAAGTTACCCATCAG      g
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCAATGCCTGAATCTCTGGCGTGATAGCTGCGG     h
TTTCTAAGCTGCCTGTGCGGCAGTGAACACGTCATCCTGAAGGCTAGGCAGCTCGGCTTC      0
TTTCTAAGCTGCCTGTGCGGCAGTGAACAGTAAGATAATACGATAACATCCTGTTTGTAA
```

Strain CEB02-449

```
tttgattatTGCCTGTGCGGCAGTGAACTCAGGGGACTGGCGAACAATGTCTTTCATGAT      a
TTTCTAAGCTGCCTGTGCGGCAGTGAACGAAAAGGTAAGATGGGCAAGCTTCTAGTAGTT      b
TTTCTAAGCTGCCTGTGCGGCAGTGAACATTATCTGAATGGCATTTTCTTTGGCGCAGAT      c
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCGCCATTCCGTGAACCTGAGCGCGTTCGCGA      d
TTTCTAAGCTGCCTGTGCGGCAGTGAACATATTCTCGAGCGATAGCAATAGCCATTCCAC      e
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCGGTCAAACAAATTTAGGCGACGATTTAACA      f
TTTCTAAGCTGCCTGTGCGGCAGTGAACAAAAAGAATTTGGGATTAAAGTTACCCATCAG      g
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCAATGCCTGAATCTCTGGCGTGATAGCTGCGG     h
TTTCTAAGCTGCCTGTGCGGCAGTGAACACGTCATCCTGAAGGCTAGGCAGCTCGGCTTC      0
TTTCTAAGCTGCCTGTGCGGCAGTGAACGAAATTGTGGGTGTAGATGTTGCAGACGCCTC      V
TTTCTAAGCTGCCTGTGCGGCAGTGAACTCTGACGTTGCCTGTGTTGCCGCTCTCGTATT      W
TTTCTAAGCTGCCTGTGCGGCAGTGAACAGTAAGATAATACGATAACATCCTGTTTGTAA
```

3B

(Continued)

Previous work by Yang and colleagues, based upon deletion analysis of a significant number of strains from China, led to the currently available view of relationships between the different biovars of *Y. pestis* (Zhou et al. 2004a; Zhou et al. 2004b). This is illustrated in Fig. 5. In this view, the Orientalis lineage branched out of the Antiqua biovar earlier than the Medievalis biovar.

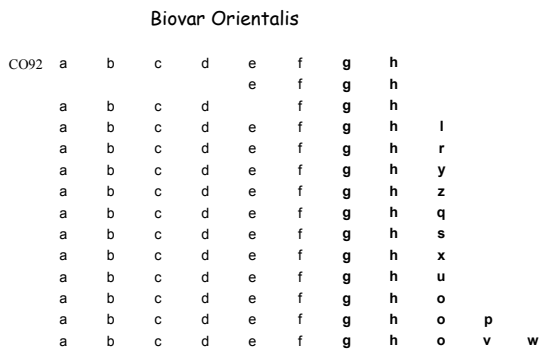


Fig. 4. CRISPR YP1 variations observed within biovar Orientalis. The list of different CRISPR YP1 alleles observed within Orientalis isolates illustrates the pattern of variations and mode of evolution of CRISPR structures. Interstitial losses are compensated by polarized insertions.

Fig. 3. (Continued)

Fig. 3. Sequence of CRISPR alleles. A- the three CRISPR loci in the CO92 genome sequence: CO92 is a *Y. pestis* strain belonging to the Orientalis biovar. The three loci contain, respectively, eight, five and three spacers. CRISPR sequence data can be coded by giving each spacer a name. Following the nomenclature proposed by Pourcel et al. (Pourcel et al. 2005) a combination of letters and figures is used. Spacers from CRISPR locus 2 and 3 are identified by the 2 and 3 added to the spacer name. Spacers are given names as they are discovered. Spacer 'a' from locus YP1 is unrelated to spacer 'a2' from locus YP2 or 'a3' from locus YP3. In the initial report (Pourcel et al. 2005), less than 26 spacers were observed at each locus. Spacers after 'z' are numbered starting at spacer 27. B- CRISPR YP1 in three Orientalis isolates: three different alleles illustrate the main features of CRISPR evolution. Within the Orientalis biovar, the vast majority of isolates shares an identical "abcdefgh" CRISPR YP1 allele. In some rare instances, differences are observed. The independent analysis done by MLVA does not suggest that these rare alleles belong to specific lineages within Orientalis. On the contrary, isolates with an otherwise identical MLVA type may show CRISPR YP1 differences, demonstrating that these mutation events are of very recent origin (Pourcel et al. 2005). Here an interstitial deletion event is observed in strain 195P, resulting in the loss of the 4 contiguous spacers a, b, c, d. Addition of one or more spacers is observed in strains Java9 and CEB02-449.

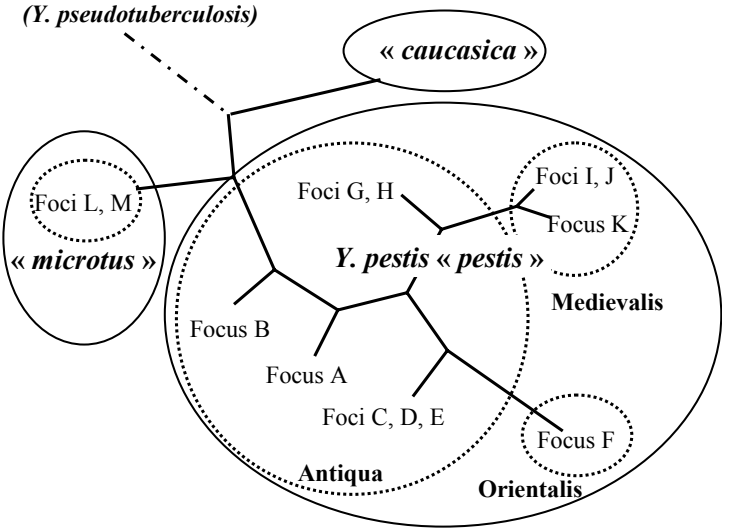


Fig. 5. Current view of relationships between some *Y. pestis* "subspecies". Deletion analysis within *Y. pestis* "*pestis*" with respect to *Y. pestis* "*microtus*" lead to this current view of *Y. pestis* evolution. The relative position of African Antiqua strains is unknown. The naming of Chinese foci is as described in Zhou et al. (Zhou et al. 2004a).

A set of diverse strains from each of the main foci identified in China (Zhou et al. 2004a) and of some foci from the former Soviet Union (Anisimov et al. 2004) including strains from the different *Y. pestis* subspecies (with the exception of *talassica*) was selected accordingly for YP1 sequence analysis. More than 80 representative YP1 alleles were investigated and more than 40 new YP1 spacers were identified. The current total number of YP1 spacers is 71. Interestingly the number of new spacers is relatively low, given the very significant increase in the diversity of *Y. pestis* strains investigated. In particular, the non-*pestis* "subspecies" including the "Angola strain" share a number of previously identified spacers (spacers labeled a, b, c, d, e, f, 37). New lineages were uncovered in particular within the A and B "Antiqua" foci (Fig. 6).

In *Y. pseudotuberculosis* the CRISPR polymorphism is very large with several hundred spacers identified to date (Gorge et al. unpublished). Spacers 'a', 'b', and 'c' were observed in a couple of strains. In agreement with previous findings, the

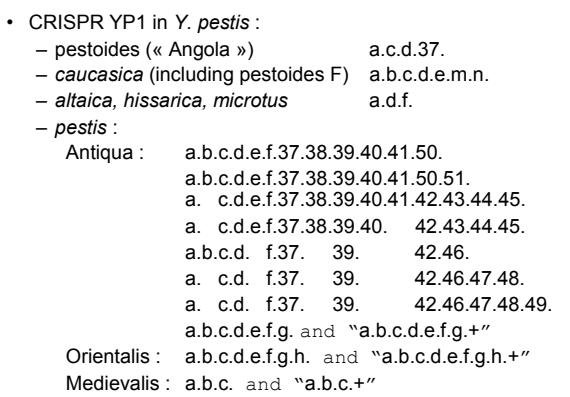


Fig. 6. CRISPR YP1 alleles observed across *Y. pestis.* A few representative alleles are indicated. Allele codes are aligned to illustrate differences resulting from interstitial deletion or progressive addition of spacers from the right end. Spacer 37 is observed in the "Angola strain" which indicates that the combination "a.b.c.d.e.f.37" was already present in the *Y. pestis* ancestor.

majority of the spacers for which an origin could be found corresponds to a prophage. This strengthens the hypothesis that the remaining spacers correspond to presently unknown viruses.

30.5 Conclusions

The present work provides a significantly enlarged view of the diversity of CRISPR spacers within *Y. pestis* intraspecies groups. Seventy CRISPR YP1 spacers have now been uncovered, and these are likely to represent the most frequently occurring spacers. Some very recently acquired and rare spacers present in only a few isolates will probably be identified in the future, but they would not significantly increase the validity of a future spoligotyping assay for *Y. pestis*. Consequently it will be possible to develop a very efficient typing assay when the other two (and less variable) CRISPR loci will have been similarly investigated.

We anticipate that such an assay will help in deciphering the phylogeny of *Y. pestis* and in identifying closely related *Y. pseudotuberculosis* strains. In addition, the investigation of CRISPRs in aDNA should greatly improve our knowledge of the agent responsible for the different pandemics.

30.6 Acknowledgments

Research by SD, MP, SB, AA, was partially supported by the International Science and Technology Center Project #2426.

30.7 References

Achtman, M., Morelli G., Zhu, P., Wirth, T., Diehl, I., Kusecek, B., Vogler, A.J., Wagner, D.M., Allender, C.J., Easterday, W.R., Chenal-Francisque, V., Worsham, P., Thomson, N.R., Parkhill, J., Lindler, L.E., Carniel, E. and Keim, P. (2004) Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc. Natl. Acad. Sci. USA 101, 17837-17842.

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A. and Carniel, E. (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis* [published erratum appears in Proc. Natl. Acad. Sci. USA 2000 97, 8192]. Proc. Natl. Acad. Sci. USA 96, 14043-14048.

Anisimov, A.P., Lindler, L.E. and Pier, G.B. (2004) Intraspecific diversity of *Yersinia pestis*. Clin. Microbiol. Rev. 17, 434-464.

Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuno, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J.T., Fauville-Dufaux, M., Ferdinand, S., Garcia de Viedma, D., Garzelli, C., Gazzola, L., Gomes, H.M., Guttierez, M.C., Hawkey, P.M., van Helden, P.D., Kadival, G.V., Kreiswirth, B.N., Kremer, K., Kubin, M., Kulkarni, S.P., Liens, B., Lillebaek, T., Ho, M.L., Martin, C., Martin, C., Mokrousov, I., Narvskaia, O., Ngeow, Y.F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M.L., Rusch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I.G., Singh, U.B., Somoskovi, A., Skuce, R.A., van Soolingen, D., Streicher, E.M., Suffys, P.N., Tortoli, E., Tracevska, T., Vincent, V. Victor, T.C. Warren, R.M., Yap, S.F., Zaman, K., Portaels, F., Rastogi, N. and Sola, C. (2006) *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6, 23.

Drancourt, M., Roux, V., Dang, L.V., Tran-Hung, L., Castex, D., Chenal-Francisque, V., Ogata, H., Fournier, P-E., Crubézy, E. and Raoult, D. (2004) Genotyping, Orientalis-like *Yersinia pestis*, and plague pandemics. Emerg. Infect. Dis. 10, 1585-1592.

Gilbert, M.T., Cuccui, J., White, W., Lynnerup, N., Titball, R.W., Cooper, A. and Prentice, M.B. (2004) Absence of *Yersinia pestis*-specific DNA in human teeth from five European excavations of putative plague victims. Microbiology 150, 341-354.

Jansen, R., Embden, J.D., Gaastra, W. and Schouls, L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. Mol. Microbiol. 43, 1565-1575.

Le Flèche, P., Hauck, Y., Onteniente, L., Prieur, A. Denoeud, F., Ramisse, V., Sylvestre, P., Benson, G. Ramisse, F. and Vergnaud, G. (2001) A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. BMC Microbiol. 1, 2.

Lillestol, R.K., Redder, P., Garrett, R.A. and Brugger, K. (2006) A putative viral defence mechanism in archaeal cells. Archaea 2, 59-72.

Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. and Koonin, E.V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of

the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol. Direct 1, 7.

Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J. and Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. 60, 174-182.

Pourcel, C., Andre-Mazeaud, F., Neubauer, H., Ramisse, F. and Vergnaud, G. (2004) Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*. BMC Microbiol. 4, 22.

Pourcel, C., Salvignol, G. and Vergnaud, G. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 151, 653-663.

Prentice, M.B., Gilbert, T. and Cooper, A. (2004) Was the Black Death caused by *Yersinia pestis*? Lancet Infect. Dis. 4, 72.

Shapiro, B., Rambaut, A. and Gilbert, M.T. (2006) No proof that typhoid caused the Plague of Athens (a reply to Papagrigorakis et al.). Int. J. Infect. Dis. 10, 334-335; author reply 335-336.

Song, Y., Tong, Z., Wang, J., Wang, L., Guo, Z., Han, Y., Zhang, J., Pei, D., Zhou, D., Qin, H., Pang, X., Zhai, J., Li, M., Cui, B., Qi, Z., Jin, L., Dai, R., Chen, F., Li, S., Ye, C., Du, Z., Lin, W., Yu, J., Yang, H., Huang, P. and Yang, R. (2004) Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. DNA Res. 11, 179-197.

Tang, T.H., Bachellerie, J.P., Rozhdestvensky, T., Bortolin, M.L., Huber, H., Drungowski, M., Elge, T., Brosius, J. and Huttenhofer, A. (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. Proc. Natl. Acad. Sci. USA 99, 7536-7541.

Vergnaud, G. (2005) *Yersinia pestis* genotyping. Emerg. Infect. Dis. 11: 1317-1318; author reply 1318-1319.

Zhou, D., Han Y., Song, Y., Huang, P. and Yang, R. (2004a) Comparative and evolutionary genomics of *Yersinia pestis*. Microbes Infect. 6, 1226-1234.

Zhou, D., Han, Y., Song, Y., Tong, Z., Wang, J., Guo, Z., Pei, D., Pang, X., Zhai, J., Li, M., Cui, B., Qi, Z., Jin, L., Dai, R., Du, Z., Bao, J., Zhang, X., Yu, J., Wang, J., Huang, P. and Yang, R. (2004b) DNA microarray analysis of genome dynamics in *Yersinia pestis*: insights into bacterial genome microevolution and niche adaptation. J. Bacteriol. 186, 5138-5146.

Zink, A.R., Sola, C., Reischl, U., Grabner, W., Rastogi, N., Wolf, H. and Nerlich, A.G. (2003) Characterization of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping. J. Clin. Microbiol. 41, 359-367.

Chapitre 5

Outils pour le typage bactérien

L'une des applications fondamentales de la structure CRISPR est son utilisation pour le typage bactérien 2.4.1. La validation et la mise en place de cette technique obéissent à certaines règles et nécessitent l'utilisation de différentes manipulations informatiques. Le second volet de ce travail de thèse a consisté à élaborer des outils permettant de faciliter cette tâche. Ces outils seront représentés dans la première section de ce chapitre. La deuxième section sera dédiée à la démonstration de l'utilisation de ces outils. La dernière section présentera une application à *L. monocytogenes*.

5.1 Outils bioinformatiques pour le typage CRISPR

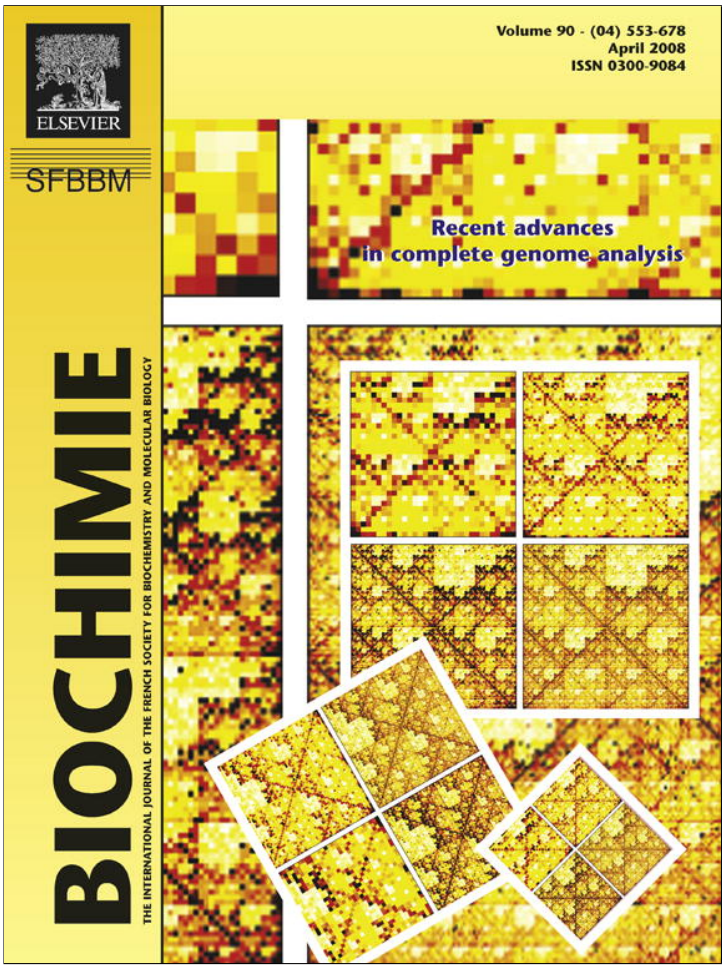
5.1.1 Description du CRISPR web service

L'article suivant (Grissa, 2008b) intitulé "On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing" (Ressources web pour les études de micro-évolution en utilisant le MLVA ou le typage CRISPR) est une revue sur l'utilisation des outils développés au sein de notre laboratoire pour le typage MLVA et le typage CRISPR.

Résumé :

Le contrôle des bactéries pathogènes nécessite le développement d'outils permettant une identification précise des souches au niveau des sous espèces. Il est maintenant largement accepté que ces outils doivent être des techniques basées sur l'ADN (contrairement à l'identification au niveau des espèces, où les techniques biochimiques sont encore très utilisées, même si les bases de données puissantes des séquences ADN 16S existent). Les

techniques de typage doivent 1) ne pas être coûteuses et 2) être propices à l'élaboration de bases de données internationales. Le succès de ces outils de typage sera éventuellement mesuré par la taille des bases de données associées accessibles sur internet. Trois méthodes ont fait leurs preuves sur ce plan, la technique de spoligotypage (*M. tuberculosis*, 40.000 entrées), le Multiple Loci Sequence Typing (MLST ; jusqu'à quelques milliers d'entrées pour plus de 20 espèces bactériennes), et plus récemment le Multiple Loci VNTR Analysis (MLVA ; jusqu'à quelques centaines d'entrées pour plus de vingt pathogènes). Dans cet article, nous allons faire l'état de l'art des outils et ressources que nous avons développés tout au long des sept dernières années dans la mise en place ou l'utilisation de la technique MLVA ou plus récemment pour l'analyse des CRISPRs qui sont à la base de la technique du spoligotypage.



Available online at www.sciencedirect.com
ScienceDirect
Biochimie 90 (2008) 660–668
BIOCHIMIE
www.elsevier.com/locate/biochi

Review

On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing

Ibtissem Grissa [a,*], Patrick Bouchon [a], Christine Pourcel [a], Gilles Vergnaud [a,b]

[a] *Univ Paris-Sud, Institut de Génétique et Microbiologie, Orsay F-91405, France; CNRS, Orsay F-91405, France*
[b] *Division of Analytical Microbiology, Centre d'Etudes du Bouchet, 5 rue Lavoisier, 91710 Vert le Petit, France*

Received 16 June 2007; accepted 19 July 2007
Available online 28 July 2007

Abstract

The control of bacterial pathogens requires the development of tools allowing the precise identification of strains at the subspecies level. It is now widely accepted that these tools will need to be DNA-based assays (in contrast to identification at the species level, where biochemical based assays are still widely used, even though very powerful 16S DNA sequence databases exist). Typing assays need to be cheap and amenable to the designing of international databases. The success of such subspecies typing tools will eventually be measured by the size of the associated reference databases accessible over the internet. Three methods have shown some potential in this direction, the so-called spoligotyping assay (*Mycobacterium tuberculosis*, 40,000 entries database), Multiple Loci Sequence Typing (MLST; up to a few thousands entries for the more than 20 bacterial species), and more recently Multiple Loci VNTR Analysis (MLVA; up to a few hundred entries, assays available for more than 20 pathogens).

In the present report we will review the current status of the tools and resources we have developed along the past seven years to help in the setting-up or the use of MLVA assays or lately for analysing Clustered Regularly Interspaced Short Palindromic Repeats called CRISPRs which are the basis for spoligotyping assays.

Keywords: VNTR; MLVA; CRISPR; Tandem repeats; Bacterial pathogens; Molecular epidemiology; Genotyping; Databases; MLVA web service; CRISPRfinder; CRISPRdb

1. Introduction

Although identification of bacteria at the species level is usually sufficient for short-term health care response, the tracing of bacterial pathogens (microbial forensics) or the identification of emerging clones escaping prophylactic or therapeutic strategies require much more precise strain identification tools. Epidemiological investigations rely upon molecular assays providing accurate and rapid differentiation of bacterial strains using some specific sites of genetic variability such as particular repeats or mutations. Various bacterial strain typing techniques have been described in the literature to differentiate strains but not all methods are equally applicable to every species. In addition many techniques cannot provide a portable result and therefore the strain genetic profile cannot be easily coded and stored into databases that can be exchanged between laboratories. This is in particular the problem of methods relying on restriction enzyme polymorphism analysed by gel electrophoresis such as restriction fragment length polymorphism (RFLP) and pulsed field gel electrophoresis (PFGE). Other pattern-producing techniques such as the random amplification of polymorphic DNA (RAPD) or amplified fragment-length polymorphism (AFLP) are respectively notably not reproducible enough or technically too demanding to allow accurate or convenient inter-laboratory comparisons of profiles. Multiple loci sequence typing (MLST) is a highly accurate, reproducible

* Corresponding author: Tel.: +33 1 6915 3001; fax: +33 1 6915 6678.
E-mail address: ibtissem.grissa@igmors.u-psud.fr (I. Grissa).

0300-9084/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.biochi.2007.07.014

and portable method but is not adapted to the typing of the most highly homogenous species such as *Mycobacterium tuberculosis* and its current cost prevents its systematic use.

Multiple loci variable number of tandem repeats (VNTR) analysis (MLVA) is a typing technique based on the polymorphism of certain tandemly repeated DNA sequences. VNTRs consist in consecutive occurrences of a DNA repeat unit, and they are found in all organisms, prokaryotes as well as eukaryotes. Although their biological function and evolution mechanism is not fully understood, they have diverse practical applications including strain identification in bacterial epidemiology. In a typical MLVA assay, a few to more than twenty VNTRs, distributed over the entire bacterial genome, are analysed, and a code corresponding to the number of repeats at each locus is determined. This code is easily stored into databases and can be used for strain clustering and epidemiological studies.

MLVA is nowadays increasingly replacing or at least completing traditional genotyping methods, providing a different or complementary point of view in *M. tuberculosis*, *Bacillus anthracis*, *Yersinia pestis*, *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Legionella pneumophila* investigations thanks to its design easiness, low cost and portability (see refs. [1,2] for reviews). MLVA is best applied within a highly homogeneous group of strains, typically with genomes showing an average similarity well above 98%. Some MLVA assays have been developed in species with an internal genome homogeneity in the 95%–98% range (as illustrated for instance in *L. pneumophila* [3]), but the design of primers, and the level of homoplasy at VNTR loci, introduce specific technical challenges.

In parallel, the particular polymorphic structures called CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat); or TREP (Tandem REPeat) [4], SRSR (short regularly spaced repeats) [5,6], DVRs (direct variant repeats) [7], LCTR (long clusters of tandem repeats) [8], SPIDR (spacers interspersed direct repeats) [9] have been used in some bacteria for genotyping (Fig. 1). They were firstly detected in *Escherichia coli* [10] and then in about 40% of bacterial genomes and almost all archaea. A CRISPR consists in exact repeat sequences (DR for direct repeat) of approximately 24–48 bases long separated by unique spacers of similar length [11]. It has been shown that CRISPRs may be used for evolutionary studies due to their mechanism of acquisition/deletion of motifs (a repeat and a spacer) [12,13]. The CRISPR region was widely used to genotype *M. tuberculosis* strains [14–18] and to a much lesser extent for *Streptococcus pyogenes* [19] *Y. pestis* [12], *Corynebacterium diphtheriae* [20,21] and *Campylobacter jejuni* [22,23]. Presumably because of strong structural constraints, the DR sequence is more conserved than the surrounding genomic elements, so that a CRISPR typing assay (in which a single PCR amplification with primers corresponding to the DR sequence is sufficient to amplify the whole set of spacers) might in theory be applied on a larger evolutionary scale than MLVA.

In this review, we describe the current set of bioinformatics tools helping in MLVA and CRISPR assay setting-up and analysis. Before starting bench work, biologists need to accomplish *in silico* pre-processing and post-processing phases for developping the typing assay and analyzing the results [2]. The pre-processing phase consists in the identification of genetic markers from sequence data, which requires the use of bioinformatics resources. Software for sequence analysis must be developed or adapted to detect correctly and exclusively the markers of interest from completed or unfinished genomes. In addition, dedicated databases archiving the markers must also be created and regularly updated from publicly available sequenced genomes and publications. Such databases provide platforms gathering information about the markers and allowing their comparison in related species to facilitate the assay design. After the genotyping has been completed, additional tools are necessary to store, analyze and compare the findings.

We will describe the current state of three databases (Minisat_db 3.0 [24], the Genotyping page version 2.0 [25] and CRISPRdb 1.0 [13]) and some web-accessible tools (CRISPRFinder [26], Spacers Dictionary Creator…). The databases have been developed using MYSQL 4.1 (http://www.mysql.com). The administration and querying process use PHP (http://www.php.net) and Perl CGI scripts.

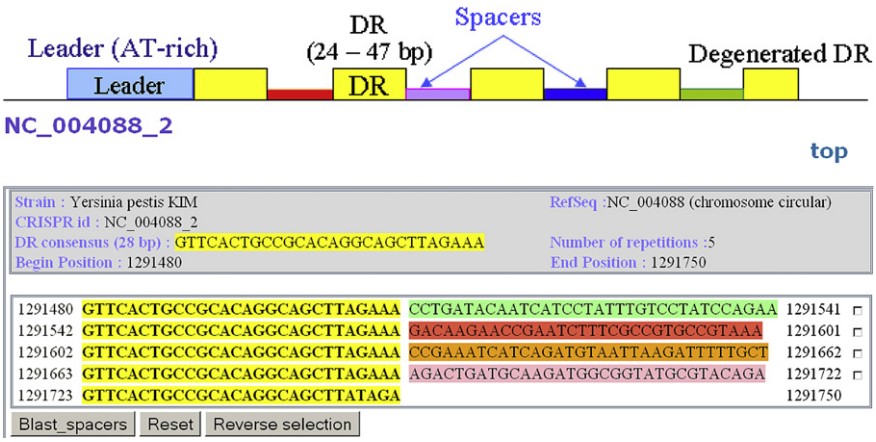


Fig. 1. Schematic representation of a CRISPR locus and output of the CRISPR database for one of the three *Y. pestis* CRISPR loci [12].

2. MLVA

2.1. The tandem repeats database

Tandem repeats (TRs) have been shown to be a source of very informative markers for epidemiological purposes not only in eukaryotes but also in bacteria. Currently, several programs have been developed to find TRs in a genome sequence. Tandem Repeat Finder (TRF) which detects efficiently perfect and also imperfect TRs [27] is probably used most often. Other programs include STRING [28], STAR [29,30], REPuter [31], mreps [32], ARTHunter [33]. None of these programs is capable of predicting if a given TR will be polymorphic in different members of a species, and usually only a fraction of all TRs will eventually turn out to be polymorphic after testing a relevant population of strains.

The TR database [24] is a platform to assist biologists in the identification of VNTRs, firstly by computing the TRs from available sequenced bacterial genomes and secondly by providing an interface to query the database using different parameters including the repeat unit length, the number of repeats and the percentage of internal conservation. In addition, information about published markers and PCR products is provided, as well as a tool to ''Blast'' sequences against the TR database which is particularly useful when selecting PCR primers. TRF is used to identify TRs and thereafter filters are added to eliminate some irrelevant structures mostly on the basis of stringency level. The version 1.0 of the TR database was first presented in early 2001 [34]. At that time, 36 genome sequences had been published. The reason for developing this database was to help projects within the laboratory, and the database was made publicly available as a side-project. It was expected then that this initiative would likely be relayed and replaced by other products proposed by dedicated bioinformatics and genomics centres. However, no such alternative appeared, and in 2004 we added complementary tools, once again in order to assist in-house projects. These tools were described in [35]. The main addition was the strain comparison tool. By then approximately 100 new genome sequences had been released, and in some cases, more than one strain was available for a given species. The rational for the strain comparison tool had been previously explored [36]. Since this 2004 upgrade (database version 2.0), the rate of genome release has raised to more than 20 per month. We consequently have now made a second major upgrade which addresses two issues: in terms of database management, genome import has been largely automated; in terms of database content, tandem repeats with short repeat units (also called microsatellites) are now included. For instance, a dinucleotide repeat with less than 12 perfect copies or a trinucleotide with less than 8 perfect copies would not have been included in the initial database. In the current version (3.0) the following major modifications have been made:

(i) The database building technique was switched from Microsoft development tools to freeware (from Microsoft Access 2000 to Mysql and from ASP to PHP). This technical improvement increases the storage capacity and makes the program more platform independent.

(ii) A semi-automatic import tool was developed taking advantage of the information table provided at http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1, so that updating the database requires a few minutes of work by the database curator.

(iii) Microsatellites (repeat units shorter than 9 bp) which were not taken into account in the first version and represent a potentially interesting class of markers are now included into the database. The current version is based on a new procedure consisting in running TRF four times with different criteria. In order to detect most if not all relevant TRs, the lowest-stringency alignment parameters ''2,3,5'' provided with the TRF package are still used but the maximum period size and the minimum alignment score are put respectively to (2000, 50), (15, 40), (10, 30) and (5, 20). In contrast, the first version of the database was filtering out tandem repeats with a TRF score below fifty, which was the reason for excluding a significant proportion of microsatellites. Using this new procedure, perfect mononucleotide repeats comprising 10 or more repeat units are present in the database; the threshold for perfect dinucleotide repeats is 5 repeat units.

(iv) Treatment of redundancies: Tandem Repeats Finder usually proposes different views of a given tandem repeat. These views most often differ in terms of repeat unit length: sometimes, a tandem repeat locus can be viewed as for instance a 6 bp repeat unit element, or as a 18 bp repeat unit element, perhaps reflecting different evolutionary mechanisms. The current version of the database considers two different categories of tandem repeats, microsatellites with a repeat unit size up to 8 base-pairs, and minisatellites with a repeat unit size of 9 bp and more. When a locus is presented by TRF both as a microsatellite and a minisatellite, the database will keep one output from each category. Within each category, the alignment presented by default is the one with the highest TRF defined score (the TRF score is a value which depends of the total length and the percent match). A link to alternative alignments, if any, is also provided.

(v) A simple tool allowing the identification of TR families, such as the MIRUs (mycobacterial interspersed repetitive units) in *M. tuberculosis* [37] or the SIRUs in *S. aureus* [38], which possess highly similar repeated unit and are dispersed throughout the chromosome, often in intergenic regions.

(vi) In order to avoid the duplication of work by independent groups and to limit the giving of different names to the same locus (as recalled by, for instance, Le Flèche et al. [39]), the database includes links to tandem repeats which have already been investigated and given names in the literature, and to the corresponding publication. The database is a repository of different

MLVA assays achieved on bacterial genomes including also the PCR products accessible in the result of primers blast search. This part of the database requires the scanning of the literature, so that input from the community is most welcome, in order to ensure that it is up to date.

2.2. The bacterial genotyping page

One key feature of the MLVA typing is the possibility to compare fingerprints between different laboratories owing to the high portability of the typing assay. Therefore, it is important to build a repository of published typing results and to provide additional tools to compare genotypes and identify a strain using its MLVA code. The Bacterial Genotyping page (as proposed in the first prototype by ref. [39]) is a gateway allowing the comparison of a newly typed strain to existing data. The first version was based upon a proprietary commercial software (Bionumerics from Applied-Maths and the BNserver associated tool), which represents a significant cost, in comparison with the very simple requirements associated with MLVA data querying. Once the tandem repeats have been typed by PCR, and the allele sizes have been converted to repeat numbers at each VNTR locus, the MLVA profile is expressed as a code in a predetermined order. The query process includes the choice of a VNTR panel and the submission of the corresponding code. The result consists in a table indicating related strains, and a clustering dendrogram showing proximity of the queried strain to related strain collections. The current version 2.0 database and web page is based upon open source software (PHP-MySQL) and includes links to databases for *L. pneumophila* [3], *B. anthracis* [40], *Brucella* [41], *M. tuberculosis* [42] among others (Fig. 2). In the next version, which will be released in a near future (version 3), tools will be added to enable the setting up of private databases, shared by groups of users developing new MLVA databases in collaboration across the internet.

2.3. CRISPR-based epidemiological studies

Groenen et al. [14] invented in 1993 a genotyping technique for the *M. tuberculosis* complex, called spoligotyping, which makes use of the polymorphism of a particular repeated sequence called the DR. The *M. tuberculosis* DR is a CRISPR, a structure that has now been identified in a growing number of bacteria. In some species the diversity of the CRISPR unique sequences or ''spacers'' can be used to perform epidemiological studies. The spoligotyping is a macroarray technique and to date about 40,000 profiles have been generated and stored into the SpolDB4 database dedicated to the *M. tuberculosis* complex [18].

In addition there is a large interest for the behaviour of these particular structures which appear to play a role in defence against foreign DNA attacks as initially predicted [12] and recently demonstrated [43]. Therefore there is a need for tools to identify CRISPR in sequenced genomes and to store them and analyse their composition.

2.4. CRISPRFinder: (http://crispr.u-psud.fr/Server/CRISPRfinder.php)

Detecting CRISPRs was generally achieved with classical repeat and pattern finding programs such as REPuter, Patscan, or TRF which require fastidious manual investigation and post-processing. Indeed, using repeat-finding programs CRISPRs are detected as particular patterns or as particular tandem repeats which results in detecting also a high proportion of false positives.

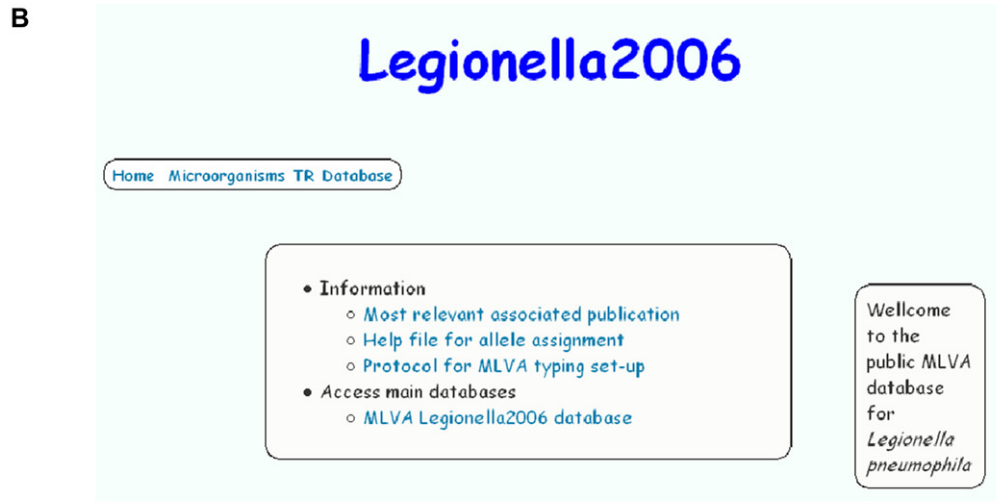
One reason for this is the current absence of an algorithmic definition of the CRISPR structure. The CRISPR is defined in terms of sequence as a succession of repeated sequences (DR) separated by unique sequences of similar size. This definition remains insufficient and the main difficulty is to determine the consensus DR and its boundaries. Indeed although the DRs are remarkably conserved inside a CRISPR even when more than a hundred motifs exist, in some cases mutations are observed. In addition an identification based on comparison between DRs can be efficient only when at least three DRs are present.

Consequently, there is at present no easy and clear-cut way to perform an objective evaluation and comparison of different CRISPR finding methods. Recently, some specific programs for CRISPR finding have been described. Pygram [44] is a visualizing program to browse repeats in genomic sequences. As a visualization tool, its use remains more interesting for long CRISPR sequences (more than 7 repeats). PILER-CR [45] is a dedicated software for detecting CRISPRs, based on an adaptation of the PILER program. This software is efficient in CRISPR detection and the possibility of modifying parameters enhances its performance in some particular cases. The software has the advantage of being rapidly executed and producing only few false positives, but it sometimes misidentifies the DR boundaries and omits the degenerated repeat present at one end of the CRISPR. PILER-CR has been used by Kunin et al. to extract CRISPRs in order to analyze the secondary structure of the direct repeats [46].

The CRISPRFinder program [26] is based on an adaptation of the Vmatch package [47] to find accurately the CRISPRs and especially the short ones (less than four spacers). This program is available as a web tool at http://crispr.u-psud.fr/Server/CRISPRfinder.php offering complementary tools to download the results, get the flanking sequences and blast spacers against GenBank database and CRISPRdb database.

The main purpose of CRISPRFinder is to provide an easy-to-use program for detecting CRISPRs without any requirements in programming or computer skills or even any previous knowledge of CRISPRs. The default parameters have been scrupulously chosen after several empirical tests to detect the maximum number of CRISPRs, reducing the false positives and selecting the best consensus DR. Nonetheless, an advanced flexible version of the program with a detailed tutorial is provided for skilled users to modify the parameters.

CRISPRFinder was used for setting up the CRISPRdb database [13].



C

Here are the results in the Legionella2006 MLVA database

Key	Strain	Serogroup_mAB-subgroup_SBT	Geographic origin	Distance from queried strain	Other strains with identical typing data
Leg105	EUL030	1_France _3,4,1,14,14,9	France	0	Genotype 6
Leg136	EUL073	1_Philadelphia_3,4,1,1,14,9	England & Wales	0	Genotype 12
Leg168	EUL118	1_Philadelphia_3,4,1,1,14,9	Germany	0	Genotype 16
Leg099	EUL020	1_Benidorm _3,10,1,3,14,9	Scotland	0	none
Leg132	EUL069	1_Philadelphia_3,4,1,1,14,9	England & Wales	0	none
Leg156	EUL099	1_Bellingham _3,13,1,25,14,9	Sweden	0	none
Leg122	EUL052	1_Philadelphia_3,10,3,1,14,9	Spain	1	none
Leg137	EUL074	1_Philadelphia_1,4,1,1,14,9	England & Wales	1	none
Leg154	EUL097	1_Knoxville _3,10,1,3,14,9	Sweden	1	none
Leg155	EUL098	1_Knoxville _3,10,1,3,14,9	Sweden	1	none
Leg098	EUL019	1_Knoxville _3,10,1,3,14,9	Scotland	2	none
Leg097	EUL018	1_Bellingham _2,6,21,12,12,8	Scotland	3	none
Leg151	EUL091	1_Bellingham _9,6,3,10,22,15	Denmark	3	none
Leg158	EUL101	1_Benidorm _7,6,17,3,13,11	Sweden	3	none
Leg164	EUL111	1_Benidorm _2,6,17,15,12,8	Germany	3	none

Access dendrogram

Fig. 2. The bacterial genotyping page A: part of the home page; B: the *L. pneumophila* page [3]; C: result of a query using the genetic profile from the Philadelphia1 reference strain; a list of best matches is provided.

2.5. CRISPRdb: a database for CRISPRs (http://crispr.u-psud.fr/crispr/CRISPRdatabase.php)

The CRISPRdb [13] is a comprehensive relational database, regularly updated, that browses the CRISPRs present in all published genomes (http://www.genomesonline.org/). It is the first database dedicated to these structures providing a repository of CRISPRs and tools to further analyse them. The database is automatically updated and presently contains more than 500 analyzed prokaryotic genomes (Fig. 3).

In addition to listing the CRISPRs, the CRISPRdb web service provides complementary tools to analyze and compare them. They allow: (i) downloading pre-calculated files; (ii) blasting any queried sequence against the stored direct repeats and spacers; (iii) blasting spacers against public databases; and (iv) aligning flanking sequences which helps in primers selection for typing.

CRISPRs probably represent important elements for the archaea life cycle or for survival in their natural environment as only two out of 38 lack a CRISPR. Indeed it was suggested that some archaea lost it upon culture in the absence of invading extrachromosomal elements [48]. In contrast, only about 40% of bacteria possess a CRISPR. When browsing the CRISPRdb, an important diversity in CRISPR number and structures can be observed. Within a species, different strains do not necessarily harbour the same CRISPRs. This is demonstrated for example, with the three available *L. pneumophila* genome sequences. The *Philadelphia1* strain has no CRISPR, the Lens strain possesses one CRISPR on the circular chromosome and one on a plasmid with the same DR, and the Paris strain has a different CRISPR (with a different DR), but has a common spacer with the Lens CRISPR.

Similarly, among the twelve sequenced strains of *Streptococcus pyogenes*, four strains have "CRISPR1" and "CRISPR2" of this species, three of them have only the CRISPR2, whereas five strains are deprived of both of them.

2.6. Spacers Dictionary Creator (http://crispr.u-psud.fr/crispr/CRISPRdetector.php)

When a similar CRISPR is present in different strains within a species, polymorphism in the number of motifs is often observed. It is due to interstitial deletions or polarised insertions

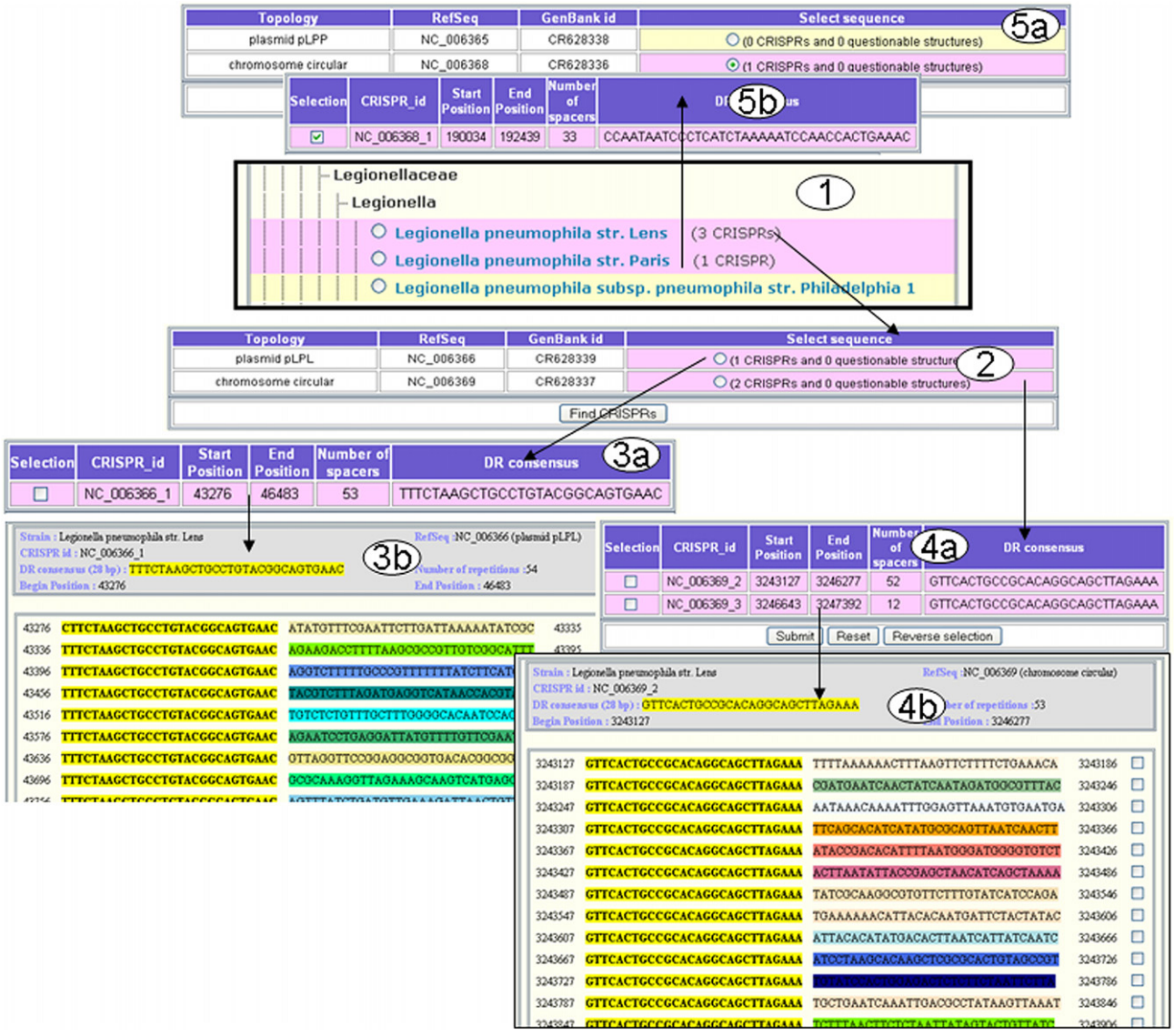


Fig. 3. Illustration of the CRISPRdb web page: 1-the strains are listed according to taxonomy; among the three *Legionella* strains, two contain CRISPR loci. They can be visualised in detail by going through 2, then 3a, 3b, or 4a, 4b. The color selection is a random process, an other selection can be obtained by reloading the page.

of entire motifs (DR + spacer). These properties have led to the use of the CRISPR loci as a strain-typing tool. In addition to the spoligotyping assay in the *M. tuberculosis* complex, (construction of a database of 39,295 profiles in spolDB4 [18]) CRISPRs have been used for example to differentiate yoghurt strains (*Lactobacillus acidophilus*, *Streptococcus thermophilus*), *S. pyogenes* [19] and *Corynebacterium diphtheriae* [20] and as a complementary typing technique for *Y. pestis* [12].

The Spacers Dictionary Creator is a very interesting tool for intra-species studies allowing to create and store an exhaustive catalogue of spacers that can be queried and updated when new spacers are identified. Each spacer is numbered, and this number can then be used to automatically code CRISPR alleles in a compact way. Up to now, three sample dictionaries are available online with this tool: *Y. pestis* dictionary of 26 spacers, *C. jejuni* dictionary with 59 with spacers and *S. thermophilus* dictionary with 328 spacers.

3. Discussion

The minisatellite database [34] was the first Tandem repeats database dedicated to prokaryotic tandem repeats. TRbase [49] is relating tandem repeats to gene locations and disease genes of the human genome. The ABCC GRID database [50] is devoted to the human microsatellites. PlantSat [51] is specialized for tandem repeats in plants. Several databases considers only microsatellites (MICdb [52], MRD [53]).

More recently, the VNTRDB has been reported by Chang et al. [54] as a VNTR prediction database built by comparing several bacterial genome sequences using TRF and ATRHunter. TRDB [55] provides in addition to a tandem repeats repository for some genomes, a private working space to store analysis results and permits collaborators to privately share their data.

Microorganisms Tandem Repeats Database, the VNTRDB and the TRDB may be used by biologists in a complementary fashion.

To date, more than 500 bacterial genomes and 40 Archaea genomes have been analysed and their TRs are included in the database. The automated administration permits quick and frequent updates, as the number of sequencing projects is growing constantly. More than 40 comparisons have been created. The selection of genomes appropriate for the making of comparisons is not automated yet. Species, and even genus names, are not good indicators of the relevance of doing a genome comparison in search for polymorphic tandem repeats, so that a specific sequence-based procedure will need to be implemented. Finally, more than 300 published markers from different species have been added to the database.

The bacterial genotyping page is very interesting for rapid strain identification especially for molecular epidemiology and forensic analysis of pathogens. For example, this web-based service helped, through exchanges of MLVA codes between laboratories in different countries, to show that some lineages of *Pseudomonas aeruginosa* infect preferentially cystic fibrosis patients (Vu-Thien et al., in press). In addition to the importance for clinical epidemiology, the possibility to quickly check the identity of a strain is also very important for the maintenance of strain collections, in particular when dangerous pathogens or precious strains are involved.

The CRISPR web service is the first dedicated online service for detecting and analyzing CRISPRs. A database of the Cas genes has been described by Haft et al. [56] but did not provide any information about the CRISPR. In the current version of CRISPRdb, 491 genomes have been analyzed, 616 CRISPRs have been detected. Usually, the CRISPRs are found

Table 1
Overview of CRISPRs loci located on plasmids

Taxon id	Ref Seq	Kingdom	Species	CRISPRs number	Size
348780	NC_007427	Archaea	*Natronomonas pharaonis* DSM 2160 (plasmid PL131)	1	2
348780	NC_007428	Archaea	*Natronomonas pharaonis* DSM 2160 (plasmid PL23)	1	7
272569	NC_006392	Archaea	*Haloarcula marismortui* ATCC 43049 (plasmid pNG400)	3	(4;25;51)
272569	NC_006391	Archaea	*Haloarcula marismortui* ATCC 43049 (plasmid pNG300)	1	47
349163	NC_009468	Bacteria	*Acidiphilium cryptum* JF-5 (plasmid pACRY04)	1	3
349163	NC_009468	Bacteria	*Acidiphilium cryptum* JF-5 (plasmid pACRY02)	1	39
224324	NC_001880	Bacteria	*Aquifex aeolicus* VF5 (plasmid ece1)	1	3
76114	NC_006823	Bacteria	*Azoarcus sp.* EbN1 (plasmid 1)	2	(8;14)
319795	NC_008010	Bacteria	*Deinococcus geothermalis* DSM 11300 (plasmid 1)	1	66
882	NC_005863	Bacteria	*Desulfovibrio vulgaris subsp. vulgaris str.* Hilden…(megaplasmid)	1	27
297245	NC_006366	Bacteria	*L. pneumophila str.* Lens (plasmid pLPL)	1	53
338966	NC_008607	Bacteria	*Pelobacter propionicus* DSM 2379 (plasmid pPRO1)	1	3
298386	NC_005871	Bacteria	*Photobacterium profundum* SS9 (plasmid pPBPR1)	1	64
338969	NC_007901	Bacteria	*Rhodoferax ferrireducens* T118 (plasmid1)	1	24
100226	NC_003903	Bacteria	*Streptomyces coelicolor* A3 (plasmid SCP1)	1	4
1148	NC_005230	Bacteria	*Synechocystis* sp. PCC 6803 (plasmid pSYSA)	3	(38;49;56)
262724	NC_005838	Bacteria	*Thermus thermophilus* HB27 (plasmid pTT27)	8	(3;6;7;9;13;15)
300852	NC_006462	Bacteria	*Thermus thermophilus* HB8 (plasmid pTT27)	9	(1;2;3;9;12;20;23) Mean = 20

on circular chromosome, but 38 CRISPRs (Table 1) are localized on plasmids having a size from 1 motif in *Thermus thermophilus HB8* which harbours 9 CRISPRs to 66 motifs in *Deinococcus geothermalis DSM 11300* with a mean of 20 motifs. Among the 15 species having CRISPRs on plasmids, only three have no CRISPR on their chromosomes: *Acidiphilium cryptum JF-5*, *Desulfovibrio desulfuricans G20* and *Haloarcula marismortui ATCC 43049*.

4. Concluding remarks and perspectives

It is relatively easy, through collaborative efforts, to significantly expand the content of the current databases. It is hoped that these databases, or databases developed along these lines, will constitute an easy-to-use high resolution classification resource which will then help address medical and epidemiological issues. We are grateful to those who participated in this resource development and improvement and we welcome suggestions and contributions.

References

[1] B.A. Lindstedt, Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria, Electrophoresis 26 (2005) 2567–2582.

[2] G. Vergnaud, C. Pourcel, Multiple locus VNTR (variable number of tandem repeat) analysis, in: E. Stackebrandt (Ed.), Population Structure of Prokaryotes, Springer-Verlag, Berlin Heidelberg, 2006, pp. 83–104 Molecular Identification, Systematics.

[3] C. Pourcel, P. Visca, B. Afshar, S. D'Arezzo, G. Vergnaud, N.K. Fry, Identification of variable-number tandem-repeat (VNTR) sequences in *Legionella pneumophila* and development of an optimized multiple-locus VNTR analysis typing scheme, J. Clin. Microbiol. 45 (2007) 1190–1199.

[4] F.J. Mojica, C. Ferrer, G. Juez, F. Rodriguez-Valera, Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning, Mol. Microbiol. 17 (1995) 85–93.

[5] F.J. Mojica, C. Diez-Villasenor, E. Soria, Juez GBiological significance of a family of regularly spaced repeats in the genomes of Archaea Bacteria and mitochondria, Mol. Microbiol. 36 (2000) 244–246.

[6] X. Peng, K. Brugger, B. Shen, L. Chen, Q. She, R.A. Garrett, Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in *Sulfolobus* genomes, J. Bacteriol. 185 (2003) 2410–2417.

[7] J.D. van Embden, T. van Gorkom, K. Kremer, R. Jansen, B.A. van Der Zeijst, L.M. Schouls, Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria, J. Bacteriol. 182 (2000) 2393–2401.

[8] Q. She, R.K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M.J. Awayez, C.C. Chan-Weiher, I.G. Clausen, B.A. Curtis, A. De Moors, G. Erauso, C. Fletcher, P.M. Gordon, I. Heikamp-de Jong, A.C. Jeffries, C.J. Kozera, N. Medina, X. Peng, H.P. Thi-Ngoc, P. Redder, M.E. Schenk, C. Theriault, N. Tolstrup, R.L. Charlebois, W.F. Doolittle, M. Duguet, T. Gaasterland, R.A. Garrett, M.A. Ragan, C.W. Sensen, J. Van der Oost, The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 7835–7840.

[9] R. Jansen, J.D. van Embden, W. Gaastra, L.M. Schouls, Identification of a novel family of sequence repeats among prokaryotes, Omics 6 (2002) 23–33.

[10] Y. Ishino, H. Shinagawa, K. Makino, M. Amemura, A. Nakata, Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product, J. Bacteriol. 169 (1987) 5429–5433.

[11] R. Jansen, J.D. Embden, W. Gaastra, L.M. Schouls, Identification of genes that are associated with DNA repeats in prokaryotes, Mol. Microbiol. 43 (2002) 1565–1575.

[12] C. Pourcel, G. Salvignol, G. Vergnaud, CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies, Microbiology 151 (2005) 653–663.

[13] I. Grissa, G. Vergnaud, C. Pourcel, The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats, BMC Bioinformatics 8 (2007) 172.

[14] P.M. Groenen, A.E. Bunschoten, D. van Soolingen, J.D. van Embden, Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method, Mol. Microbiol. 10 (1993) 1057–1065.

[15] Z. Fang, N. Morrison, B. Watt, C. Doig, K.J. Forbes, IS6110 transposition and evolutionary scenario of the direct repeat locus in a group of closely related *Mycobacterium tuberculosis* strains, J. Bacteriol. 180 (1998) 2102–2109.

[16] I. Filliol, J.R. Driscoll, D. van Soolingen, B.N. Kreiswirth, K. Kremer, G. Valetudie, D.A. Dang, R. Barlow, D. Banerjee, P.J. Bifani, K. Brudey, A. Cataldi, R.C. Cooksey, D.V. Cousins, J.W. Dale, O.A. Dellagostin, F. Drobniewski, G. Engelmann, S. Ferdinand, D. Gascoyne-Binzi, M. Gordon, M.C. Gutierrez, W.H. Haas, H. Heersma, E. Kassa-Kelembho, M.L. Ho, A. Makristathis, C. Mammina, G. Martin, P. Mostrom, I. Mokrousov, V. Narbonne, O. Narvskaya, A. Nastasi, S.N. Niobe-Eyangoh, J.W. Pape, V. Rasolofo-Razanamparany, M. Ridell, M.L. Rossetti, F. Stauffer, P.N. Suffys, H. Takiff, J. Texier-Maugein, V. Vincent, J.H. de Waard, C. Sola, N. Rastogi, Snapshot of moving and expanding clones of *Mycobacterium tuberculosis* and their global distribution assessed by spoligotyping in an international study, J. Clin. Microbiol. 41 (2003) 1963–1970.

[17] A.R. Zink, C. Sola, U. Reischl, W. Grabner, N. Rastogi, H. Wolf, A.G. Nerlich, Characterization of *Mycobacterium tuberculosis* complex DNAs from Egyptian mummies by spoligotyping, J. Clin. Microbiol. 41 (2003) 359–367.

[18] K. Brudey, J.R. Driscoll, L. Rigouts, W.M. Prodinger, A. Gori, S.A. Al-Hajoj, C. Allix, L. Aristimuno, J. Arora, V. Baumanis, L. Binder, P. Cafrune, A. Cataldi, S. Cheong, R. Diel, C. Ellermeier, J.T. Evans, M. Fauville-Dufaux, S. Ferdinand, D. Garcia de Viedma, C. Garzelli, L. Gazzola, H.M. Gomes, M.C. Guttierez, P.M. Hawkey, P.D. van Helden, G.V. Kadival, B.N. Kreiswirth, K. Kremer, M. Kubin, S.P. Kulkarni, B. Liens, T. Lillebaek, M.L. Ho, C. Martin, C. Martin, I. Mokrousov, O. Narvskaia, Y.F. Ngeow, L. Naumann, S. Niemann, I. Parwati, Z. Rahim, V. Rasolofo-Razanamparany, T. Rasolonavalona, M.L. Rossetti, S. Rusch-Gerdes, A. Sajduda, S. Samper, I.G. Shemyakin, U.B. Singh, A. Somoskovi, R.A. Skuce, D. van Soolingen, E.M. Streicher, P.N. Suffys, E. Tortoli, T. Tracevska, V. Vincent, T.C. Victor, R.M. Warren, S.F. Yap, K. Zaman, F. Portaels, N. Rastogi, C. Sola, *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology, BMC Microbiol. 6 (2006) 23.

[19] N. Hoe, K. Nakashima, D. Grigsby, X. Pan, S.J. Dou, S. Naidich, M. Garcia, E. Kahn, D. Bergmire-Sweat, J.M. Musser, Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains, Emerg. Infect. Dis. 5 (1999) 254–263.

[20] I. Mokrousov, O. Narvskaya, E. Limeschenko, A. Vyazovaya, Efficient discrimination within a *Corynebacterium diphtheriae* epidemic clonal group by a novel macroarray-based method, J. Clin. Microbiol. 43 (2005) 1662–1668.

[21] I. Mokrousov, E. Limeschenko, A. Vyazovaya, O. Narvskaya, *Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci, Biotechnol. J. 2 (7) (2007) 901–906.

[22] L.M. Schouls, S. Reulen, B. Duim, J.A. Wagenaar, R.J. Willems, K.E. Dingle, F.M. Colles, J.D. Van Embden, Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination, J. Clin. Microbiol. 41 (2003) 15–26.

[23] E.P. Price, H. Smith, F. Huygens, P.M. Giffard, High-resolution DNA melt curve analysis of the clustered regularly interspaced short palindromic repeat locus of *Campilobacter jejuni*, Appl. Environ. Microbiol. 73 (10) (2007) 3431–3436.
[24] The Microorganisms Tandem Repeats Database, <http://minisatellites.u-psud.fr>.
[25] The MLVA web service, <http://bacterial-genotyping.igmors.u-psud.fr>.
[26] I. Grissa, G. Vergnaud, C. Pourcel, CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats, Nucleic Acids Res. 35 (2) (2007) W52–W57.
[27] G. Benson, Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Res. 27 (1999) 573–580.
[28] V. Parisi, V. De Fonzo, F. Aluffi-Pentini, STRING: finding tandem repeats in DNA sequences, Bioinformatics 19 (2003) 1733–1738.
[29] O. Delgrange, E. Rivals, STAR: an algorithm to Search for Tandem Approximate Repeats, Bioinformatics 20 (2004) 2812–2820.
[30] A. Krishnan, F. Tang, Exhaustive whole-genome tandem repeats search, Bioinformatics 20 (2004) 2702–2710.
[31] S. Kurtz, C. Schleiermacher, REPuter: fast computation of maximal repeats in complete genomes, Bioinformatics 15 (1999) 426–427.
[32] R. Kolpakov, G. Bana, G. Kucherov, mreps: Efficient and flexible detection of tandem repeats in DNA, Nucleic Acids Res. 31 (2003) 3672–3678.
[33] Y. Wexler, Z. Yakhini, Y. Kashi, D. Geiger, Finding approximate tandem repeats in genomic sequences, J. Comput. Biol. 12 (2005) 928–942.
[34] P. Le Flèche, Y. Hauck, L. Onteniente, A. Prieur, F. Denoeud, V. Ramisse, P. Sylvestre, G. Benson, F. Ramisse, G. Vergnaud, A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*, BMC Microbiol. 1 (2001) 2.
[35] F. Denoeud, G. Vergnaud, Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a Web-based resource, BMC Bioinformatics 5 (2004) 4.
[36] F. Denoeud, G. Vergnaud, G. Benson, Predicting human minisatellite polymorphism, Genome Res. 13 (2003) 856–867.
[37] P. Supply, E. Mazars, S. Lesjean, V. Vincent, B. Gicquel, C. Locht, Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome, Mol. Microbiol. 36 (2000) 762–771.
[38] B. Ghebremedhin, W. Konig, W. Witte, K.J. Hardy, P.M. Hawkey, B. Konig, Subtyping of ST22-MRSA-IV (Barnim epidemic MRSA strain) at a university clinic in Germany from 2002 to 2005, J. Med. Microbiol. 56 (2007) 365–375.
[39] P. Le Flèche, M. Fabre, F. Denoeud, J.L. Koeck, G. Vergnaud, High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing, BMC Microbiol. 2 (2002) 37.
[40] F. Lista, G. Faggioni, S. Valjevac, A. Ciammaruconi, J. Vaissaire, C. le Doujet, O. Gorge, R. De Santis, A. Carattoli, A. Ciervo, A. Fasanella, F. Orsini, R. D'Amelio, C. Pourcel, A. Cassone, G. Vergnaud, Genotyping of *Bacillus anthracis* strains based on automated capillary 25-loci Multiple Locus Variable-Number Tandem Repeats Analysis, BMC Microbiol. 6 (2006) 33.
[41] P. Le Fleche, I. Jacques, M. Grayon, S. Al Dahouk, P. Bouchon, F. Denoeud, K. Nockler, H. Neubauer, L.A. Guilloteau, G. Vergnaud, Evaluation and selection of tandem repeat loci for a *Brucella* MLVA typing assay, BMC Microbiol. 6 (2006) 9.
[42] M. Fabre, J.L. Koeck, P. Le Fleche, F. Simon, V. Herve, G. Vergnaud, C. Pourcel, High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of hsp65 gene polymorphism in a large collection of "*Mycobacterium canettii*" strains indicates that the *M. tuberculosis* complex is a recently emerged clone of "*M. canettii*, J. Clin. Microbiol. 42 (2004) 3248–3255.
[43] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D.A. Romero, P. Horvath, CRISPR provides acquired resistance against viruses in prokaryotes, Science 315 (2007) 1709–1712.
[44] P. Durand, F. Mahe, A.S. Valin, J. Nicolas, Browsing repeats in genomes: Pygram and an application to non-coding region analysis, BMC Bioinformatics 7 (2006) 477.
[45] R.C. Edgar, PILER-CR: fast and accurate identification of CRISPR repeats, BMC Bioinformatics 8 (2007) 18.
[46] V. Kunin, R. Sorek, P. Hugenholtz, Evolutionary conservation of sequence and secondary structures in CRISPR repeats, Genome Biol. 8 (2007) R61.
[47] M. Abouelhoda, S. Kurtz, E. Ohlebusch, Replacing suffix trees with enhanced suffix arrays, J. Discrete Algorithms 2 (2004) 53–86.
[48] R. Lillestol, P. Redder, R. Garrett, K. Brugger, A putative viral defence mechanism in archaeal cells, Archaea 2 (2006) 59–72.
[49] T. Boby, A.M. Patch, S.J. Aves, TRbase: a database relating tandem repeats to disease genes for the human genome, Bioinformatics 21 (2005) 811–816.
[50] J.R. Collins, R.M. Stephens, B. Gold, B. Long, M. Dean, S.K. Burt, An exhaustive DNA micro-satellite map of the human genome using high performance computing, Genomics 82 (2003) 10–19.
[51] J. Macas, T. Meszaros, M. Nouzova, PlantSat: a specialized database for plant satellite repeats, Bioinformatics 18 (2002) 28–35.
[52] V.B. Sreenu, V. Alevoor, J. Nagaraju, H.A. Nagarajaram, MICdb: database of prokaryotic microsatellites, Nucleic Acids Res. 31 (2003) 106–108.
[53] S. Subramanian, V.M. Madgula, R. George, R.K. Mishra, M.W. Pandit, C.S. Kumar, L. Singh, MRD: a microsatellite repeats database for prokaryotic and eukaryotic genomes, Genome Biol. 3 (2002) PREPRINT0011.
[54] C.H. Chang, Y.C. Chang, A. Underwood, C.S. Chiou, C.Y. Kao, VNTRDB: a bacterial variable number tandem repeat locus database, Nucleic Acids Res. 35 (2007) D416–D421.
[55] Y. Gelfand, A. Rodriguez, G. Benson, TRDB—the Tandem Repeats Database, Nucleic Acids Res. 35 (2007) D80–D87.
[56] D.H. Haft, J. Selengut, E.F. Mongodin, K.E. Nelson, A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes, PLoS Comput. Biol. 1 (2005) e60.

5.1.2 Description de CRISPRcompar pour la comparaison de souches

L'article suivant (Grissa, 2008a), intitulé "CRISPRcompar : a website to compare clustered regularly interspaced short palindromic repeats" (CRISPRcompar : un site web pour la comparaison des CRISPRs), décrit deux outils d'analyse et de comparaison des CRISPRs ainsi que la procédure de leur utilisation. CRISPRcomparison permet essentiellement de comparer les CRISPRs enregistrés dans la base de données CRISPRdb ou MyCRISPRdb (base privée homologue à CRISPRdb dont le contenu est géré par l'utilisateur lui même). Cet outil permet de retrouver le même CRISPR sur des génomes distincts. CRISPRtionary, est un outil de construction de catalogue de spacers permettant d'identifier et de stocker les spacers à partir de séquences alléliques d'un même CRISPR.

Résumé :

Les CRISPRs forment une famille particulière de répétitions en tandem présente dans les génomes de procaryotes, dans presque toutes les archées et près de la moitié des bactéries. Les CRISPRs participent à un mécanisme d'acquisition de résistance contre les phages. Ils consistent en une succession de répétitions directes (DR) de 24-47 pb séparées par des séquences de taille similaire (spacers). Dans la grande majorité des cas, les répétitions directes sont hautement conservées, alors que le nombre et la nature des spacers sont souvent diversifiés, même à l'intérieur d'une même espèce. De plus, il a été démontré que l'acquisition de nouveaux motifs (DR + spacer) se fait presque exclusivement d'un seul côté du locus. Ainsi, le CRISPR présente un marqueur génétique intéressant pour les analyses comparatives et d'évolution pour des souches bactériennes très proches. CRISPRcompar est une ressource internet créée dans le but d'aider les biologistes dans le processus de typage du CRISPR. Deux outils faciliteront les investigations *in silico* : CRISPRcomparison et CRISPRtionary. Le site web correspondant est accessible à l'adresse `http://crispr.u-psud.fr/CRISPRcompar/`.

Nucleic Acids Research, 2008, 1–4
doi:10.1093/nar/gkn228

CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats

Ibtissem Grissa[1,*], Gilles Vergnaud[1,2] and Christine Pourcel[1]

[1]Univ. Paris-Sud 11, CNRS, UMR8621, Institut de Génétique et Microbiologie, 91405 Orsay and [2]DGA/D4S - Mission pour la Recherche et l'Innovation Scientifique, 7, rue des Mathurins, 00470 Armées, France

Received January 25, 2008; Revised April 6, 2008; Accepted April 11, 2008

ABSTRACT

Clustered regularly interspaced short palindromic repeat (CRISPR) elements are a particular family of tandem repeats present in prokaryotic genomes, in almost all archaea and in about half of bacteria, and which participate in a mechanism of acquired resistance against phages. They consist in a succession of direct repeats (DR) of 24–47 bp separated by similar sized unique sequences (spacers). In the large majority of cases, the direct repeats are highly conserved, while the number and nature of the spacers are often quite diverse, even among strains of a same species. Furthermore, the acquisition of new units (DR + spacer) was shown to happen almost exclusively on one side of the locus. Therefore, the CRISPR presents an interesting genetic marker for comparative and evolutionary analysis of closely related bacterial strains. CRISPRcompar is a web service created to assist biologists in the CRISPR typing process. Two tools facilitates the *in silico* investigation: CRISPRcomparison and CRISPRtionary. This website is freely accessible at http://crispr.u-psud.fr/CRISPRcompar/.

INTRODUCTION

The clustered regularly interspaced short palindromic repeat (CRISPR)-associated system (CASS) comprises the particular repeated element CRISPR itself, the promoter for its transcription (also called the leader) and a set of *cas* genes responsible for its maintenance and function (1,2). It is found in most Archea and 40% bacteria, and is linked to a mechanism of acquired resistance against bacteriophages (3). Some genomes harbour a significant number of CRISPRs [18 in *Methanocaldococcus jannaschii DSM 2661*with three different direct repeats (DRs)] (4). When different CRISPRs with the same DR are present in a genome, they have a very similar leader, generally different spacers, and only one is associated with *cas* genes (5). When CRISPRs from different CRISPR families exist in the same genome, one set of cas genes specific for each family is present. Finally, within a species, different strains may have different CRISPRs. The example of the three sequenced strains of *Streptococcus thermophilus* is very illustrative of this situation, since three CRISPRs were identified in this species but only strain LMD-9 possesses the three of them (4).

CRISPRs evolve either by deletion or acquisition of units (a DR and a spacer) following a mechanism proposed firstly by Pourcel *et al.* (6) and recently confirmed (7–9). In the majority of cases, new units are added at one end of the CRISPR adjacent to the leader, whereas motif deletions can occur randomly. The independent acquisition of the same spacer twice is possible but is not frequent and easily detected. Thus, the presence of identical spacers in the same CRISPR locus in distinct strains reflects shared ancestry.

The polymorphism of CRISPRs can be used for molecular typing. The standard and classical technology developed for *Mycobacterium tuberculosis* typing (10) is the spoligotyping, which consists in detecting the presence/absence of a range of spacers. This technique and other PCR-based typing methods have been applied in CRISPR genotyping to study other bacterial species (6,11–16).

We recently implemented a program (CRISPRFinder) allowing the identification of a CRISPR structure based on a thorough characterization of its components, i.e. the DR and the spacers (17). Using this program, public genome sequences are analysed and the extracted CRISPRs are stored into a database (CRISPRdb) (4). CRISPRFinder and CRISPRdb are accessible on the web together with different tools that assist in recovering spacers and DR sequences, and blasting them against Genbank.

We now report on the development of a new website dedicated to the comparison of CRISPRs between strains and the labelling of spacers when multiple alleles are analysed.

CRISPRcompar is freely accessible at http://crispr.u-psud.fr/CRISPRcompar/index.php.

*To whom correspondence should be addressed. Email: ibtissem.grissa@igmors.u-psud.fr

METHODS AND IMPLEMENTATION

CRISPRcompar is a friendly web resource offering tools to compare CRISPRs between strains of a given species or between closely related species, and to classify the spacers. Its core routines were developed in Perl under Debian Linux. It is composed of two main applications; CRISPRcomparison and CRISPRtionary. CRISPRcomparison identifies and compares the CRISPRs of two or more genomes (complete or partial sequences). It is particularly useful when strains of a species possess several CRISPRs for which positions on the genome might vary, as a result for instance of large-scale genome rearrangements, or of presence–absence polymorphism of CRISPR loci in the genomes of interest. The similarity criteria are based on having an identical consensus DR and similar flanking sequences. The flanking sequences are compared by the ClustalW alignment of the 200 bp adjacent sequences to the CRISPR with a threshold of 90% of similarity. In the majority of cases, when multiple CRISPRs with the same DR are present in a genome, only one flanking sequence is similar, the one corresponding to the leader.

CRISPRtionary lists the spacers from different alleles derived from the same CRISPR locus and annotates them in a polarized fashion. Such data will be produced for instance when investigating the diversity (evolution) of CRISPRs within a species by sequencing the locus in different isolates. This tool can then be used to automatically number spacers, produce a 'dictionary' or repertoire of spacers and code the alleles using this dictionary. CRISPRFinder is used to identify the DR and order the spacers according to the DR sequence. When sequencing PCR products, the first few nucleotides may be missed or the data may be of poor quality. In addition, the first, often partial and degenerated DR (up to 50% of differences have been observed) may be missed by CRISPRFinder in this context. For this reason, a filter exploring the existence of stretches of additional DR in the flanking sequence was added so as to correctly identify the first spacer. It consists in blasting the two halves of the DR against the remaining nucleotides of the allele sequence. Given the mechanism of acquisition of new spacers, we recommend to orientate the CRISPR such that the degenerated DR is located on the left extremity and the leader is on the right. These criteria are convenient to attribute increasing numbers to the spacers from left to right, according to their acquisition order, i.e. the more recently added spacer close to the leader will be given the highest number.

Input

The CRISPRcompar program automatically recovers from CRISPRdb all members of a genus containing a CRISPR and proposes to compare each of them using the alphabetic list of strains harbouring a CRISPR (alternatively, all strains from a given genus can be selected at once using the 'strain taxonomy browser'). To compare unpublished sequences and genomes, a private database on the model of CRISPRdb (4) must first be created (http://crispr.u-psud.fr/CRISPRcompar/private/). Additional sequences from the private database can then be added in the comparison. Once a selection of sequences has been performed, the 'compare' button leads to a page where it is possible to choose the strain that will be used as a reference for the CRISPRs annotation. At this step, it is also possible to remove or add sequences in the comparison. When several alleles of a given locus are present in the submitted sequences, their spacers can be annotated using CRISPRtionary. Fasta files containing sequenced CRISPR alleles can also be directly submitted to CRISPRtionary.

Output

For the CRISPRcomparison application, the result is shown in a table where CRISPRs are grouped. Figure 1 shows the result of the comparison of three *S. thermophilus* strains. Information is given on the CRISPR position and on the number of repeats (Figure 1A). A link to the corresponding CRISPR in CRISPRdb can be activated. When two or more alleles of a given CRISPR are found, the flanking sequences can be aligned and a link is provided to the second application 'CRISPRtionary' to annotate and classify the spacers. By activating the 'compare spacer' button a table is shown in which the CRISPR sequences are provided in fasta format (Figure 1B). At this step, it is possible to upload a previous dictionary of spacers to which the spacers of the new CRISPR alleles will be compared. If no pre-determined dictionary exists, one will be created in the following steps. With the FindCRISPR button, the CRISPRFinder program is used to identify DRs and spacers. Often more than one DR candidate will be proposed for several reasons. One is due to the existence of several possible DRs, especially with short sequences (less than four units) and another is due to the CRISPR orientation on the genome. Indeed, when the submitted alleles are in different orientations, two DR sequences will be proposed. Therefore, the user should select the appropriate consensus DR or introduce a DR sequence. The 'find spacer' button leads to a page where spacers are labelled (Figure 1C) and different files can be recovered: (i) different formats of text and tab-delineated text files representing the corresponding CRISPRs and spacers labels (AnnotFasta, AnnotFasta_CodedAlleles, Fasta_CodedAlleles, Table_Coded_Alleles), (ii) 'Spacers dictionary' which is a tab-delineated text file containing a catalogue of the found spacers and their labels and (iii) 'binary file': a tab-delineated text-file where columns represent the spacer labels and rows represent the queried alleles. For each CRISPR allele, a spacer will be given the '1' value when it exists and '0' when it is absent. The binary file is especially interesting for providing a spoligotyping-like profile of the CRISPR and to visually illustrate the spacer composition in the strains. The different files may be used in further studies such as the evolutionary analysis of the species according to the spacer organization in the different strains or for epidemiological purposes.

The last step may be added to improve the output; this is called the re-annotation step. It might be interesting when a collection of alleles has been analysed to

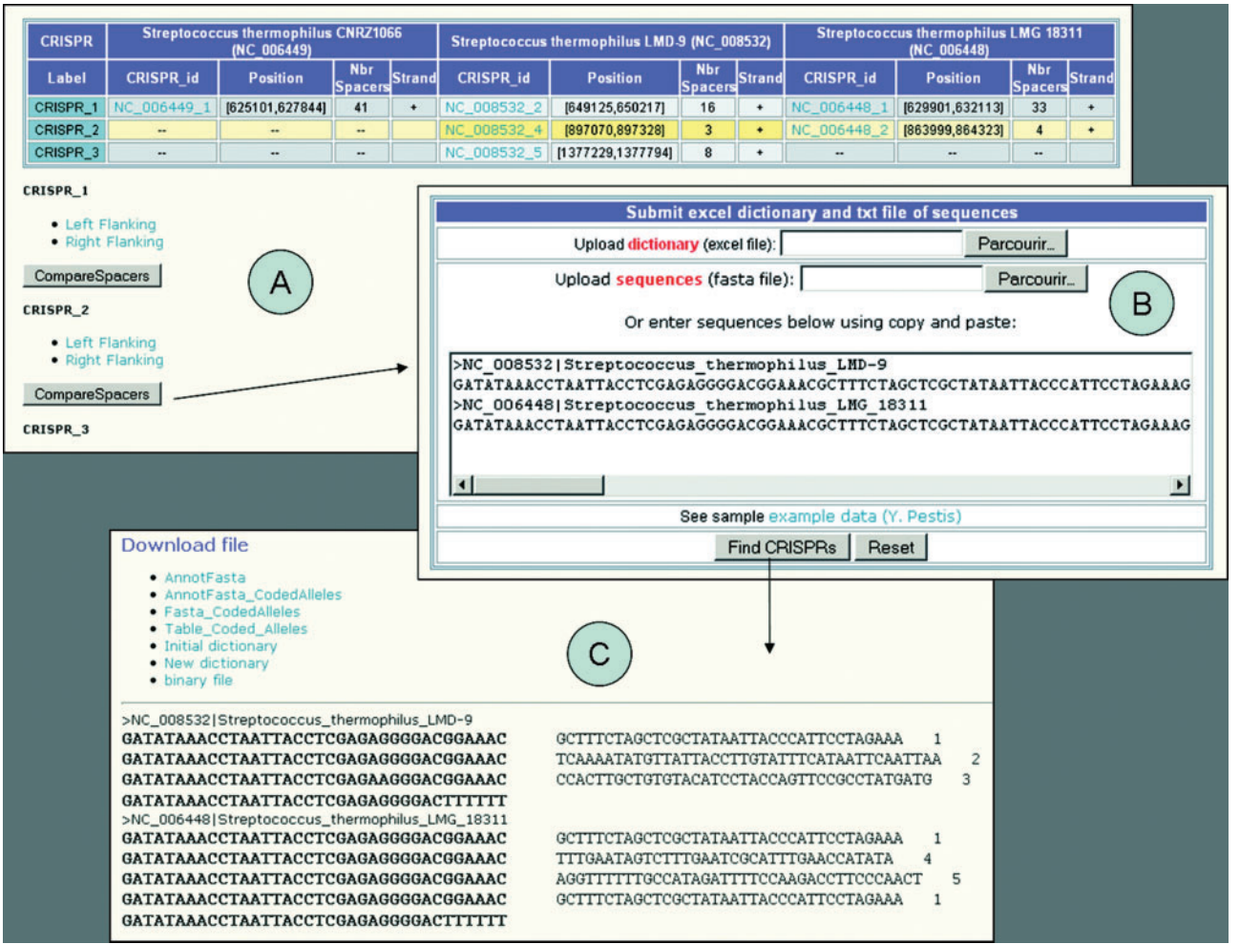


Figure 1. Example of CRISPRcompar and CRISPRtionary output using the three *S. thermophilus* genomic sequences (RefSeq: NC_006449, NC_008532, NC_006448). (**A**) Table showing the classification of the different CRISPRs. Three CRISPRs are identified, of which two are found in two or more strains. (**B**) CRISPR_2 sequences are submitted to the CRISPRtionary program. (**C**) The spacers are labelled and different files can be recovered.

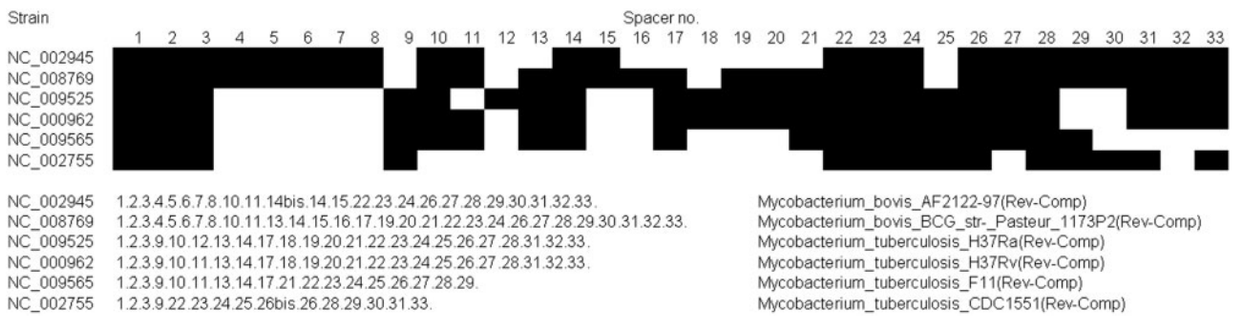


Figure 2. Schematic representation of the CRISPR repeats organization in four *M. tuberculosis* and two *M. bovis* strains. The binary file produced by CRISPRtionary, after the spacers have been re-annotated, is used to produce a figure in which the presence of a spacer is indicated by a dark square. The detail of the spacer composition for each strain is indicated in the bottom part of the figure.

re-annotate the spacers such that numbering is increasing starting from one end of the CRISPR. We propose that the oldest spacer, i.e. the one near the degenerated DR, when the later is identified, be given the label 1 and subsequent ones increasing numbers. The re-annotation tool modifies the labels such that all the labels inside an allele are in an increasing order and a new set of output files is produced. Sometimes, a duplication of one or several spacers may occur and in this case, the term 'bis' is added to the spacer label in the CRISPR code. On Figure 2 is shown the distribution and annotation of spacers in six members of the *M. tuberculosis* complex

(MTBC). The binary file was converted into a diagram for an easy comparison. The profile corresponds to the order of spacers described upon sequencing of a collection of alleles (18). However, and similarly to a real spoligotype, the presence in the same allele of two identical spacers is not indicated.

DISCUSSION AND CONCLUSIONS

The CRISPRcompar web server proposes a set of bioinformatic tools assisting biologists in the development and the setting up of a CRISPR genotyping scheme. In the pre-processing phase, the comparison of CRISPRs is mandatory and may be fulfilled using the CRISPRcomparison tool, which helps in selecting the most appropriate CRISPR loci and associated primers for the PCR amplification.

CRISPRcomparison allows the identification of families of strains that share a CRISPR, inside species with high genetic diversity or the identification of homologous CRISPRs within species containing multiple CRISPR loci. In the post-processing phase, the CRISPRtionary program is very interesting since it allows the user to easily compare multiple alleles of a CRISPR locus investigated in a collection of strains and to obtain pre-calculated files that may be directly used in clustering analysis. Many clustering methods are applicable and may provide a good clustering of the strains even if these methods usually do not take full advantage of the CRISPR rules of evolution, which could be used to better assess—in addition to forming groups of related strains—parental relations between taxa. The primary evolutionary events considered are motifs insertion and deletion. In the case of inactive (in terms of spacer acquisition) CRISPRs, only deletions are possible, and the Camin–Soakal (19) Parsimony model may be considered. In Camin–Soakal parsimony, two states are considered (0 and 1 for example), and no transition from derived state back to ancestral state is allowed. For an inactive CRISPR locus, the ancestral state is the presence of a unit and the derived state is unit absence; thus only deletion changes are allowed. Our future developments of CRISPRcompar will incorporate applications such as the MIX program of the package phylip (Felsenstein), which carries out the Camin–Soakal Parsimony method. It can be applied using the binary file with minor modifications.

ACKNOWLEDGEMENTS

The CNRS and Université Paris Sud 11 have funded this project. I.G. is supported by the TBChina EU project grant LSHPCT-2005-012166. Funding to pay the Open Access publication charges for this article was provided by XXX.

Conflict of interest statement. None declared.

REFERENCES

1. Makarova,K.S., Grishin,N.V., Shabalina,S.A., Wolf,Y.I. and Koonin,E.V. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct*, **1**, 7.
2. Sorek,R., Kunin,V. and Hugenholtz,P. (2007) CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.*
3. Deveau,H., Barrangou,R., Garneau,J.E., Labonte,J., Fremaux,C., Boyaval,P., Romero,D.A., Horvath,P. and Moineau,S. (2007) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J. Bacteriol.*
4. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinform.*, **8**, 172.
5. Jansen,R., Embden,J.D., Gaastra,W. and Schouls,L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.*, **43**, 1565–1575.
6. Pourcel,C., Salvignol,G. and Vergnaud,G. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, **151**, 653–663.
7. Lillestol,R.K., Redder,P., Garrett,R.A. and Brugger,K. (2006) A putative viral defence mechanism in archaeal cells. *Archaea*, **2**, 59–72.
8. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.
9. Tyson,G.W. and Banfield,J.F. (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.*, **10**, 200–207.
10. Groenen,P.M., Bunschoten,A.E., van Soolingen,D. and van Embden,J.D. (1993) Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. *Mol. Microbiol.*, **10**, 1057–1065.
11. Mokrousov,I., Limeschenko,E., Vyazovaya,A. and Narvskaya,O. (2007) *Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci. *Biotechnol. J.*, **2**, 901–906.
12. Mokrousov,I., Narvskaya,O., Limeschenko,E. and Vyazovaya,A. (2005) Efficient discrimination within a *Corynebacterium diphtheriae* epidemic clonal group by a novel macroarray-based method. *J. Clin. Microbiol.*, **43**, 1662–1668.
13. Hoe,N., Nakashima,K., Grigsby,D., Pan,X., Dou,S.J., Naidich,S., Garcia,M., Kahn,E., Bergmire-Sweat,D. and Musser,J.M. (1999) Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains. *Emerg. Infect. Dis.*, **5**, 254–263.
14. Schouls,L.M., Reulen,S., Duim,B., Wagenaar,J.A., Willems,R.J., Dingle,K.E., Colles,F.M. and Van Embden,J.D. (2003) Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. *J. Clin. Microbiol.*, **41**, 15–26.
15. DeBoy,R.T., Mongodin,E.F., Emerson,J.B. and Nelson,K.E. (2006) Chromosome evolution in the Thermotogales: large-scale inversions and strain diversification of CRISPR sequences. *J. Bacteriol.*, **188**, 2364–2374.
16. Vergnaud,G., Li,Y., Gorge,O., Cui,Y., Song,Y., Zhou,D., Grissa,I., Dentovskaya,S.V., Platonov,M.E., Rakin,A. *et al.* (2007) Analysis of the three *Yersinia pestis* CRISPR loci provides new tools for phylogenetic studies and possibly for the investigation of ancient DNA. *Adv. Exp. Med. Biol.*, **603**, 327–338.
17. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **35**, W52–W57.
18. van Embden,J.D., van Gorkom,T., Kremer,K., Jansen,R., van Der Zeijst,B.A. and Schouls,L.M. (2000) Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J. Bacteriol.*, **182**, 2393–2401.
19. Camin,J. and Soakal,R. (1965) A method for deducing branching sequences in phylogeny. *Evolution*, **19**, 311–326.

5.2 Applications des outils de typage CRISPR

5.2.1 Application au typage de bactéries pathogènes

L'article suivant "Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) for the Genotyping of Bacterial Pathogens" (Les CRISPRs pour le génotypage des bactéries pathogènes), est un chapitre du livre "Molecular Epidemiology of Microorganisms" à paraître dans the Methods in Molecular Biology series, Editor Dominique Caugant. Il présente un protocole de typage CRISPR depuis la première phase d'investigation du CRISPR jusqu'à la mise en place et la validation de la technique. Il montre l'utilité des outils dans chacune de ces étapes.

Résumé :

Les CRISPRs sont des séquences d'ADN composées d'une succession de répétitions (23 à 47pb) séparées par des séquences uniques appelées spacers. Le polymorphisme peut être observé dans différentes souches d'une espèce et peut être utilisé dans le génotypage. Nous décrivons des protocoles et des outils bioinformatiques permettant l'identification des CRISPRs à partir de génomes séquencés et la détermination de leurs composants (les répétitions directes et les spacers). Une représentation schématique de l'organisation des spacers peut être produite, permettant une comparaison facile entre les souches.

7

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) for the Genotyping of Bacterial Pathogens

Ibtissem Grissa, Gilles Vergnaud, Christine Pourcel

Abstract

Clustered regularly interspaced short palindromic repeats (CRISPRs) are DNA sequences composed of a succession of repeats (23 to 47bp long) separated by unique sequences called spacers. Polymorphism can be observed in different strains of a species and may be used for genotyping. We describe protocols and bioinformatics tools that allow the identification of CRISPRs from sequenced genomes, their comparison and their component determination (the direct repeats and the spacers). A schematic representation of the spacer organisation can be produced, allowing an easy comparison between strains.

Key Words: CRISPR, genotyping, bacteriophage, database, spacer, phylogeny.

1. Introduction

Clustered regularly interspaced short palindromic repeats (CRISPRs) loci typically consist of the succession of 23-47 bp repeat elements, the direct repeats (DR), separated by variable and non-repetitive sequences called spacers (*see* **Fig. 1**). A CRISPR generally possesses at one end a degenerated DR and at the other end a complete DR immediately followed by a sequence called the leader and acting as a promoter ***(1)***. In a single genome several CRISPRs with the same DR can be found, but only one is associated with a group of 4 to 6 genes called *cas* (for CRISPR-associated) ***(2)***. The CASS system (a CRISPR and several *cas* genes) has been identified in a broad range of prokaryotic species, almost all archaea and 40% of bacteria.

In the majority of cases, the spacers, when identified, happen to be fragments of bacteriophages or plasmids ***(1,3)***. Different observations suggest that the CASS system constitutes a defence system against foreign sequences.

The CRISPR structure is continuously evolving, either through the addition of new motifs (a DR and a spacer) or by interstitial deletion of one or several motifs through recombination between two DRs. New motifs are added to the CRISPR in a polarized manner by duplicating the DR next to the leader and adding a new fragment of DNA ***(3,4)***. The currently sequenced CRISPR structures listed in CRISPRdb ***(5)*** show an important variability in the nature of DRs and the number of motifs (varying from one spacer to 276 in *Chloroflexus aurantiacus* J-10-fl).

Interestingly, within a particular species, comparative analysis of sequences between closely related strains revealed on the first hand, a high degree of polymorphism from strain to strain and on the other hand, an inheritable nature of spacers from parental

strains. Developing a CRISPR typing scheme may, thus, be a good addition to classical typing techniques applicable for strain differentiation, epidemiological investigations and phylogenetic reconstruction. We have tentatively used the polymorphism of one CRISPR to propose a phylogeny for *Yersinia pestis* and we suggested that it could be also used to trace ancient bacterial DNA ***(3,6)***. To date, CRISPR typing have been used in a limited range of species: *Streptococcus pyogenes* ***(7)***, *Campylobacter jejuni* ***(8-10)***, *Y. pestis* and *Yersinia pseudotuberculosis* ***(3,6)***, *Thermotoga maritima* ***(11,12)***, *Corynebacterium diphtheriae* ***(13,14)***, *Streptococcus thermophilus* ***(15,16)***, *Lactococcus casei* ***(17)***, but it has been mainly used in *Mycobacterium tuberculosis* ***(18)***. Indeed, the so called "DR" locus in *M. tuberculosis* is in fact a CRISPR element which diversity inside the species is analysed with the spoligotyping method (see dedicated chapter in this book). Spoligotyping only investigates the presence/absence of known spacers by hybridisation and is well suited for a DR locus which is not acquiring new motifs (such as in *M. tuberculosis*) or when an extensive survey of the CRISPR diversity inside a species has been performed. In species with one or several rapidly evolving CRISPRs, PCR analysis and sequencing of these loci remain the best approach to investigate their diversity.

2. Materials

2.1. DNA Purification

Good quality DNA should be available as CRISPRs may sometimes be large and long-range PCR amplification is required.

1. The Qiagen DNeasy® Tissue kit was successfully used for different bacterial species.

2. The quality and concentration of DNA was measured using a ND-1000 Spectrophotometer (NanoDrop®, Labtech, France).

2.2. PCR Amplification

1. Standard *Taq* polymerase (Qiagen, Roche, Promega or Invitrogen).
2. The Qiagen kit provides the Q solution and corresponding buffer for amplification of GC-rich DNA. Alternatively, 0.5*M* betain (Sigma) can be used in the PCR reaction.
3. dNTPs (Eurogentec or MWG Biotech).
4. Reaction buffer is as recommended by the *Taq* polymerase manufacturer.The concentration of $MgCl_2$ in the reaction is 1.5µM.
5. Oligonucleotides are dissolved at 100 µ*M*, in 10 m*M* Tris-HCl, 1 m*M* EDTA, pH 7.8.

2.3. Agarose Gel Electrophoresis

1. Standard molecular biology grade agarose (from Invitrogen, Sigma, or Q-BIOgene).
2. Tris Borate EDTA (TBE) buffer: the 10X stock solution is 890mM Tris-borate and 20mM EDTA pH 8.3 (Sigma).
3. The DNA size marker is the 100-bp ladder (from Bio-Rad, MBI Fermentas, or Euromedex).
4. Ethidium bromide stock solution 10mg/ml (Sigma)

2.4. Sequencing

1. PCR products are purified using the QIAquick PCR purification kit (Qiagen) or precipitated with a solution of PEG8000 20% (w/v), 2.5 *M* NaCl (*see* **Note 1**) ***(19)***.
2. Sequencing is performed using the primers used for PCR.

2.5. Data Analysis

The web-based tools necessary for CRISPRs identification and comparison are freely accessible at http://crispr.u-psud.fr/CRISPRcompar/. The output of the analysis consist in different excel and text files.

3. Methods

The development of a CRISPR genotyping assay follows three main phases: the pre-processing phase, the typing assay *per se* and the *in silico* post-processing phase.

3.1. Pre-Processing Phase

In the pre-processing phase, an *in-silico* investigation is performed to find CRISPRs which could be potential typing markers. This is achieved by firstly checking whether the studied species harbours at least one CRISPR, using tools designed to identify these particular repeated sequences. It is thus necessary to have access to the genome sequence of a least one strain (even in an unassembled phase). When several CRISPRs are present in a single genome, they should be clearly differentiated, and primers should be designed to amplify specifically each locus. Since their relative positions in the genome may vary

from strain to strain due to large-scale DNA rearrangements, CRISPRs labels are assigned in each strain.

3.1.1. Checking for CRISPR Presence

1. Consult the CRISPRdb database (http://crispr.u-psud.fr/crispr/CRISPRdatabase.php?page=tax). Structures marked in pink colour correspond to confirmed CRISPRs and can be retained for further analyses (*see* **Note 2**).
2. Non public sequenced genomes (or even contigs) can be submitted as a fasta file to the CRISPRFinder program (http://crispr.u-psud.fr/Server/CRISPRfinder.php). The detected CRISPRs will be displayed either as confirmed or questionable, according to characteristics described by Grissa *et al*. ***(20)*** (*see* **Note 3**).

3.1.2. Inter-Species Comparison of CRISPRs

When the genome of several strains of a given species have been sequenced, their CRISPRs, if any, can be classified and identified in each sequenced genome.

1. Go to the CRISPR Comparison Page http://crispr.u-psud.fr/CRISPRcompar/, which analyse CRISPRs present in CRISPRdb.
2. Activate the button "Compare the CRISPRs of two or several genomEs»
3. Select the strains to be compared either by browsing the strain taxonomy list or from the alphabetical list, and click on the comparison button. The related CRISPRs will be labelled and displayed in a table. Each line corresponds to a CRISPR; the CRISPR id, its position and the number of spacers are given in the

columns. An alignment of the flanking sequences is given when a locus is present in two strains.

4. Explore the spacers diversity using the Spacer dictionary Creator tool for which a link is available via the button “compare spacers”. It is now possible to select the CRISPR to be typed according to the spacers polymorphism and their number (*see* **Note 4**).

3.2. CRISPR Genotyping

3.2.1. Assessing the CRISPR Polymorphism

It is necessary to get a rapid idea of a CRISPR potential use as a genotyping tool, as important variations have been observed between different species. Some CRISPRs may be present only in a subset of strains or they might show very little polymorphism. Indexing the markers polymorphism is achieved by amplifying the CRISPR locus in a selected set of strains in order to check if the PCR products vary in size.

1. Design PCR primers from the flanking sequences alignment when several alleles of the same CRISPR are available. Otherwise, 20-30bp-long oligonucleotides can be picked in the flanking sequences, at least 40bp away from the first and last DR (Figure 1, primers F1 and R) (*see* **Note 5**).
2. Alternatively it can be decided to analyse only the CRISPR portion which is growing by addition of new motifs. In this case a primer is chosen in the flanking region containing the leader and the other in one of the spacers (Figure 1, primers F2 and R) (*see* **Note 6**).

Figure 1 Schematic representation of a CRISPR locus

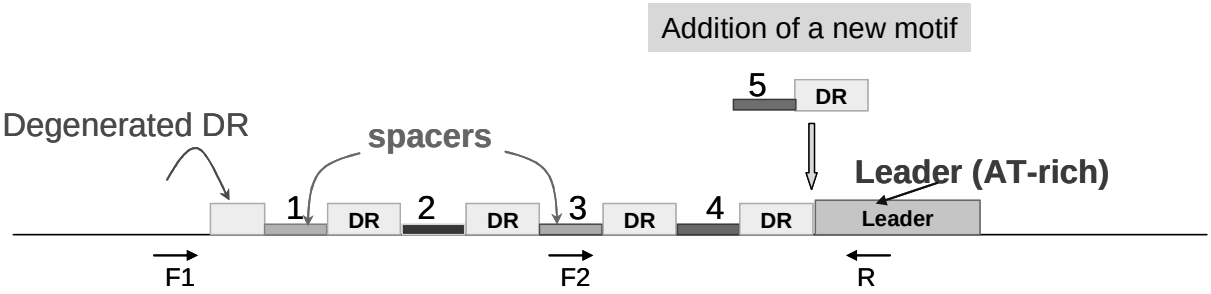


Fig. 1. Schematic representation of a CRISPR locus. The CRISPR is a succession of repeated sequences, the DRs, separated by unique sequences, the spacers. The first DR is often degenerated. Flanking the last DR is a sequence called the leader and acting as a promoter. On this figure the CRISPR is oriented such that the spacer numbered 1 is the oldest whereas the newly added spacer is next to the leader.

3. Check the presence/absence of a CRISPR region and its polymorphism on a representative subset of 10-15 strains including when possible a sequenced reference strain (*see* **Note 7**).

The following PCR conditions are routinely used: PCR reactions are performed in 15 µL containing 1-5 ng of DNA, 1X reaction buffer, 1.5 m*M* $MgCl_2$, 3 U *Taq* DNA polymerase, 200 µ*M* of each dNTP, 0.3 µ*M* of each flanking primer. Amplification is performed using the following conditions: initial denaturation cycle for 5 min at 94°C, 35 cycles of denaturation for 30 s at 94°C, annealing for 30 s at 55°C to 60°C depending on the oligos and elongation for 45 to 60 s at 72°C, plus a final elongation step for 10 min at 72°C. The PCR products are analysed on a 2% (w/v) agarose gel in 0.5X TBE buffer, run at 8 V/cm. A variation in the PCR products size reveals variation in the spacer content between strains, suggesting that this method is promising for strain discrimination of the studied bacterial species (*see* **Note 8**).

3.2.2. Setting a CRISPRGenotyping Assay

The CRISPR alleles of a larger collection of strains can now be analysed by sequencing in order to get an idea of the spacer diversity and provide a catalogue of these spacers.

1. For sequencing of the amplicons, perform a PCR amplification in a total volume of 45 µL by multiplying all the reagents three fold.
2. To assess the efficiency of the PCR reaction, 2 µL of PCR products are run on a 2% agarose gel (*see* **Note 9**). Electrophoresis is performed in 0.5X TBE buffer, run at 8 V/cm.

3. Purification of PCR products can be performed using dedicated kits. Alternatively, PEG precipitation can be performed (*see* **Note 10**). For this, the PCR reaction is transferred into a 1.5 mL eppendorf tube and 0.6 volume of PEG8000/NaCl solution is added. After 10 min at 37°C the tubes are centrifuged for 10 min at 12148 g. The PEG/NaCl is carefully removed by pipetting, avoiding the usually invisible DNA pellet, and 500 µl 80% ethanol is added. Centrifuge for 10 min at 12148 g, pour the ethanol, and dry the pellet.
4. Sequence the DNA using the primers used for PCR amplification. Ten to twenty ng purified PCR products are used in a sequencing reaction for 100bp sequence. Thus to sequence a 500bp PCR product, 50 to 100ng fragment must be used. PCR products and sequencing primers are sent for custom DNA sequencing to a specialized company (MWG biotech, Germany for example).

3.3. In silico *Post-Processing Phase*

In this phase the spacers are identified from the sequenced alleles, annotated, compared to previously known spacers and stored into a database also called dictionary. Appropriate bioinformatics tools are available on the CRISPR web service http://crispr.u-psud.fr/ to analyze the CRISPR sequencing data without any requirements in programming or computer skills ***(6)***. The site is called CRISPRtionary (*see* **Note 11**).

3.3.1. CRISPRtionary, the Spacer Dictionary Creator

1. Go to the corresponding page http://crispr.u-psud.fr/CRISPRcompar/Dict/Dict.php.

2. Submit sequences in fasta format, i.e. for each sequence, the first line starts with a greater than sign “>” and contains a unique identifier per sequence. This is the sequence header which must be in a single line. It is possible to put additional fields in the header separated by a pipeline “|”, these fields will be especially useful in the final output files.
3. If a catalogue of annotated spacers is already available in the literature or in a previous study in the laboratory, it is recommended to use this catalogue as a spacers dictionary, i.e. an excel file fulfilling the following properties (*see* **Note 12**) :

    - the first row should contain columns labels;

    - the first column should contain the spacer labels;

    - the second column may contain alternative labels or information about the spacers;

    - the third column should contain the spacers sequences;

    - the three first columns should not be empty and should not contain skipped rows.
4. If a previous excel dictionary is used, check the appropriate sheet. If no dictionary is uploaded, press the “continue” button and a file will be created.
5. The CRISPRFinder program is applied for each introduced sequence, separating DR and spacers. Several slightly different DR sequences might be proposed showing some nucleotide differences (especially with short arrays) or provided in the reverse complement orientation (when the CRISPR is present on the anti-

sense strand) (**Fig. 2A**). Therefore, the user should select or introduce the **Figure 2:Output of the spacer annotation and dictionary creator tools**

Fig. 2. Output of the spacer annotation and dictionary creator tools. Analysis of a CRISPR in three *Y. pestis* strains: **(A)** shows the output using the CRISPRtionary tool in which several candidate DRs are proposed. (**B)** shows the annotated spacers and a list of files which can be downloaded. **(C)** shows a dictionary of annotated spacers and their position in the three strains.

appropriate DR sequence oriented such that the leader position is on the right. Short CRISPR sequences with degenerate DRs may not be displayed at this step, but they will be recovered latterly.

6. After the DR selection, activate the “Find spacers” button. The CRISPR alleles are coded in a compact way by querying and updating the dictionary (**Fig. 2B** and **Fig. 2C**). When a spacer is already present in the dictionary, its code appears in the output, but when a new spacer is identified, it is numbered and added to the excel file. The second column of the dictionary is also updated by indicating for each spacer, the locus name which it belongs to and its occurrence order in this allele. Different loci names will be separated by an underscore “_” and orders by a colon “:”. For example, the spacer “f” of *Y. pestis* is the sixth spacer in the CO92 strain and is the third in the strain biovar Microtus str- 91001; so it will be coded in the second dictionary column as:

   “f_Yersinia_pestis_CO92:6_Yersinia_pestis_biovar_Microtus_str-_91001:3”.

3.3.2. Collecting the Results of the Analysis

The results are displayed on the screen and are stored in a user-friendly database (downloadable excel and text files) (*see* **Note 13**):

- AnnotFasta: a text file representing the corresponding CRISPRs. Each motif (DR+ spacer) is written on a separate line, the DR and spacer are separated by a tabulation and followed by the spacer label (*see* **Note 14**).

- AnnotFasta_CodedAlleles: the same file as the previous one in addition to the spacers codes in the header separated by dots.

- Fasta_CodedAlleles: the previous file represented in fasta format.

- Table_Coded_Alleles: excel file representing one allele per row. The header information (separated by a pipeline in the submitted sequences) are presented in separate columns. The last column provides the spacers labels separated by dots.

- Initial dictionary: the initial uploaded dictionary.

- New dictionary: the updated dictionary.

- Binary file: excel file where columns represent the spacer labels and rows represent the queried alleles. For each CRISR allele, a spacer will have a value 1 when it exists and 0 when it is absent. The binary file is especially interesting for providing a spoligotyping profile of the CRISPR and to visually illustrate the spacers composition in the strains (**Fig. 3**).

3.3.3. Re-annotating the Spacers

The obtained codes of newly added spacers to the dictionary are usually not ordered in a coherent way because the spacer labels are added according to the introduced alleles order.

1. To re-adjust them and obtain ordered numbers according to spacers acquisition by the CRISPR, use the “Re-annotate spacers” button. This will open the page: http://crispr.u-psud.fr/cgi-bin/crispr/ReannotateSpacers.cgi.
2. Introduce a dictionary and the table coded alleles files or simply use the option “use previous files” and a new annotation of the spacers will be produced. In fact, the first spacer next to the degenerated DR will be coded “one”, the next one “two”, etc... (*see* **Note 15**).

Figure 3 Organization of the CRISPR locus in six *Y. pestis* strains

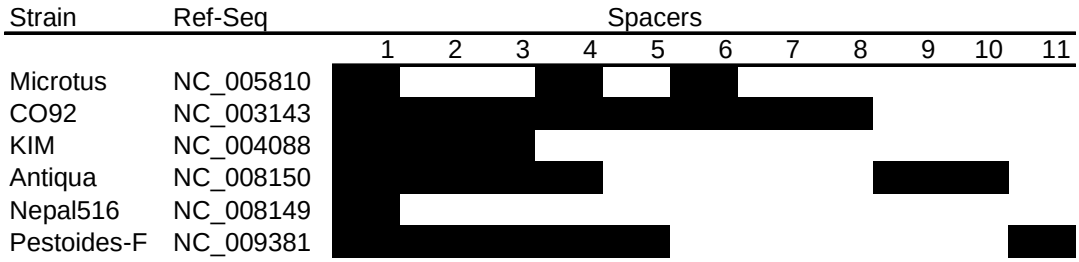


Fig. 3. Organization of the CRISPR locus in six *Y. pestis* strains. Using the binary file of the reannotated spacers a schematic representation of the CRISPR is shown in which a black box represent the presence of the spacer.

4. Notes

1. This solution is very viscous. Cut the end of the pipette tip so that the section is wider, thus facilitating aspiration.
2. The smaller CRISPRs detected by CRISPRfinder consist in a leader and two DRs (a complete and a degenerated one) separated by a spacer. Large ones can contain more than two hundred repeats. The presence of a CRISPR in a strain does not preclude its existence in all the members of the species.
3. Critical examination of the sequences must be performed as in some cases a confirmed CRISPR may, in fact, be a tandem repeat structure. In such “CRISPRs”, the spacers will show a high degree of similarity.
4. It is preferable in a first step to avoid non polymorphic spacers and long CRISPRs (more than 60 spacers) because of technical constraints.

5. It is important for the spacer identification that the DRs on both end of the CRISPR be included into the sequence.
6. Some bacterial species have rapidly evolving CRISPRs as a response to a quickly changing selection pressure associated with phage predation (such as bacterial strains present in food product and dietary supplements ***(15)***). Hence, sequencing the extremity adjacent to the leader could be sufficient to differentiate and identify strains.
7. Carefully select the strain panel, such that the control isolates belong to discernable genotypes (determined by another genotyping method, for example).
8. In some cases, PCR amplification fail with part of the tested strains due to either absence of the locus or high genetic divergence of the primer sequences.
9. Do not add the sample loading solution into the PCR tube as this might interfere with sequencing.
10. This protocol is very efficient for PCR products equal or larger than 300bp. It is rapid and cheap. It was described by ***(19)***.
11. Up to now, three sample dictionaries are available online with this tool: *Y. pestis* dictionary of 26 spacers ***(3)***, *C. jejuni* dictionary with 59 spacers ***(10)*** and *S. thermophilus* dictionary ***(15,21)*** with 328 spacers.
12. This is illustrated with a demonstrator based on the sequences of five *Y. pestis* genomes. An initial spacers catalogue was first created from the 26 published spacers, named using the alphabet from "a" to "z" ***(3)***.
13. The obtained files may be used to store the CRISPRs in a BioNumerics database (Applied Maths).

14. Rev-Comp option is added in the header when the DR orientation corresponds to the sequence on the anti-sense strand.
15. After observing the CRISPR diversity among the strains, the biologist may formulate hypotheses about the locus evolution. If a restricted number of spacers is present in all the strains with many internal absences, it may be postulated that the CRISPR locus evolves only by interstitial deletions. It is apparently the case of the *M. tuberculosis* CRISPR. Otherwise, when there is an important diversity of spacers next to the leader, it may be assumed that the CRISPR is still active and able to acquire new motifs.

Acknowledgement

We thank the CNRS and Université Paris Sud 11.

References

1. Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., and Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**, 174-182.
2. Haft, D.H., Selengut, J., Mongodin, E.F., and Nelson, K.E. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.* **1**, e60.

3. Pourcel, C., Salvignol, G., and Vergnaud, G. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653-663.
4. Lillestol, R.K., Redder, P., Garrett, R.A., and Brugger, K. (2006) A putative viral defence mechanism in archaeal cells. *Archaea* **2**, 59-72.
5. Grissa, I., Vergnaud, G., and Pourcel, C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172.
6. Vergnaud, G., Li, Y., Gorge, O., Cui, Y., Song, Y., Zhou, D., et al. (2007) Analysis of the three *Yersinia pestis* CRISPR loci provides new tools for phylogenetic studies and possibly for the investigation of ancient DNA. *Adv. Exp. Med. Biol.* **603**, 327-338.
7. Hoe, N., Nakashima, K., Grigsby, D., Pan, X., Dou, S.J., Naidich, S., et al.. (1999) Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains. *Emerg. Infect. Dis.* **5**, 254-263.
8. Schouls, L.M., Reulen, S., Duim, B., Wagenaar, J.A., Willems, R.J., Dingle, K.E., et al. (2003) Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. *J. Clin. Microbiol.* **41**, 15-26.
9. Fouts, D.E., Mongodin, E.F., Mandrell, R.E., Miller, W.G., Rasko, D.A., Ravel, J., et al. (2005) Major structural differences and novel potential virulence

mechanisms from the genomes of multiple *Campylobacter* species. *PLoS Biol.* **3**, e15.

10. Price, E.P., Smith, H., Huygens, F., and Giffard, P.M. (2007) High-resolution DNA melt curve analysis of the clustered, regularly interspaced short-palindromic-repeat locus of *Campylobacter jejuni*. *Appl. Environ. Microbiol.* **73**, 3431-3436.
11. DeBoy, R.T., Mongodin, E.F., Emerson, J.B., and Nelson, K.E. (2006) Chromosome evolution in the Thermotogales: large-scale inversions and strain diversification of CRISPR sequences. *J. Bacteriol.* **188**, 2364-2374.
12. Mongodin, E.F., Hance, I.R., Deboy, R.T., Gill, S.R., Daugherty, S., Huber, R., et al. (2005) Gene transfer and genome plasticity in *Thermotoga maritima*, a model hyperthermophilic species. *J. Bacteriol.* **187**, 4935-4944.
13. Mokrousov, I., Limeschenko, E., Vyazovaya, A., and Narvskaya, O. (2007) *Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci. *Biotechnol. J.* **2**, 901-906.
14. Mokrousov, I., Narvskaya, O., Limeschenko, E., and Vyazovaya, A. (2005) Efficient discrimination within a *Corynebacterium diphtheriae* epidemic clonal group by a novel macroarray-based method. *J. Clin. Microbiol.* **43**, 1662-1668.
15. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S.,et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712.

16. Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., et al. (2008) Diversity, activity and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol*. **190,** 1401-1412.
17. Diancourt, L., Passet, V., Chervaux, C., Garault, P., Smokvina, T., and Brisse, S. (2007) Multilocus sequence typing of *Lactobacillus casei* reveals a clonal population structure with low levels of homologous recombination. *Appl. Environ. Microbiol.* **73**, 6601-6611.
18. van Embden, J.D., van Gorkom, T., Kremer, K., Jansen, R., van Der Zeijst, B.A., and Schouls, L.M. (2000) Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J. Bacteriol.* **182**, 2393-2401.
19. Embley, T.M. (1991) The linear PCR reaction: a simple and robust method for sequencing amplified rRNA genes. *Lett. Appl. Microbiol.* **13**, 171-174.
20. Grissa, I., Vergnaud, G., and Pourcel, C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **35**, W52-57.
21. Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S.D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551-2561.

5.2.2 Application à l'exploration des CRISPRs de *L. monocytogenes*

Ce travail se présente dans le cadre d'une collaboration avec le Centre National de Référence (CNR) des *Listeria* à l'Institut Pasteur de Paris dans le but d'étudier la possibilité d'un typage utilisant les CRISPRs. C'est un travail qui est encore en phase de développement et la contribution concerne la première phase de "pre-processing" et l'exploration bioinformatique des CRISPRs de cette espèce.

Contexte de l'étude

Le genre *Listeria* compte 6 espèces proches : *L. ivanovii*, pathogène pour le bétail (avortement), *L. innocua*, *L. seeligeri*, *L. welshimeri*, *L. grayi.* et *L. monocytogenes*. *L. monocytogenes* est une bactérie saprophyte, Gram positif, dont le génome est à bas %GC, mobile à $22^{\circ}C$ (péritriche), immobile à $37^{\circ}C$, non capsulée, non sporulée, mesurant $1 - 4\mu m/0,5\mu m$, en chaînes courtes ou petits amas. Elle est ubiquitaire, largement répandue dans la nature ; *L. monocytogenes* est responsable d'infections sporadiques sévères chez l'homme et les animaux. Son pouvoir invasif lui permet de traverser le placenta et la barrière hémato encéphalique (méningo-encéphalites). C'est une bactérie intracellulaire facultative, capable de survivre et de croître à l'intérieur de la plupart des cellules de l'hôte infecté. La listériose affecte majoritairement des groupes de populations à risque (immuno-déprimés, femmes enceintes, nouveau-nés). Cette maladie reste rare (en France $<$ à 300 cas / an). En dépit de l'antibiothérapie, la mortalité est estimée à 25-30% avec 40% de séquelles neurologiques.

Treize sérovars sont définis pour cette espèce, mais près de 75% des souches humaines appartiennent au sérovar 4b, les autres appartenant au sérovar $1/2a$, $1/2c$, $1/2b$ (24%). L'identification des souches épidémiques est réalisée par l'analyse des profils de restriction du chromosome après électrophorèse en champ pulsé.

Pour les analyses bioinformatiques, nous disposons des séquences du génome complet pour 4 souches, alors que 16 autres génomes sont en cours de séquençage. Ces derniers génomes sont accessibles sur le site suivant : `http://www.broad.mit.edu/annotation/genome/listeria_group/MultiDownloads.html`.

Nature des CRISPRs détectés

L'utilisation de CRISPRFinder a montré que la plupart des souches possède au moins un CRISPR. Deux DR ont été détectés, le DR29 : **GTTTTAGTTACTTATTTTGAAATGTAAAT** et le DR36 : **GTTTTGGTAGCATTCAAAATAACATAGCTCTAAAAC**. Quatre souches parmi les vingt étudiées sont complètement dépourvues de

CRISPR alors que les autres en contiennent de 1 à 3. Dans les cinq séquences de plasmides disponibles, nous n'avons détecté aucun CRISPR. Les données analysées montrent qu'il existe trois CRISPRs chez *L. monocytogenes* : deux portant le DR29 et un portant le DR36.

L'existence de deux types de DR est en conformité avec la détection de deux clusters de gènes *cas*. Le tableau 5.1 donne la liste de ces gènes chez *L. monocytogenes 1/2a F6854* (d'après `http://cmr.tigr.org/tigr-scripts/CMR/shared/MakeFrontPages.cgi?page=genome_property`). Le cluster 1 associé au DR29 contient 7 gènes Cas, tandis que le cluster 2 associé au DR36 est constitué de quatre gènes.

Analyse des DR

Le DR29 est présent dans deux CRISPRs distincts dans les génomes étudiés : le CRISPR29_1 et le CRISPR29_2. Les séquences des génomes n'étant pas encore terminées (multiples contigs), et afin de pouvoir différencier ces deux locus, nous avons procédé de la manière suivante :

1. Alignement des séquences flanquantes ; c'est l'un des éléments clé de notre étude puisqu'il nous a permis d'identifier les amorces d'amplification des CRISPRs pour l'étude expérimentale qui est en cours de réalisation. Cependant dans certains cas, le CRISPR se trouve au début ou à la fin d'un contig et il est alors impossible d'extraire les séquences flanquantes pour les aligner.
2. Alignement des DR ; l'alignement du DR29 consensus de tous les CRISPRs trouvés montre trois groupes de DR (voir figure 5.1). Le DR29_2 est toujours le même alors que le DR29_1 présente deux variants.

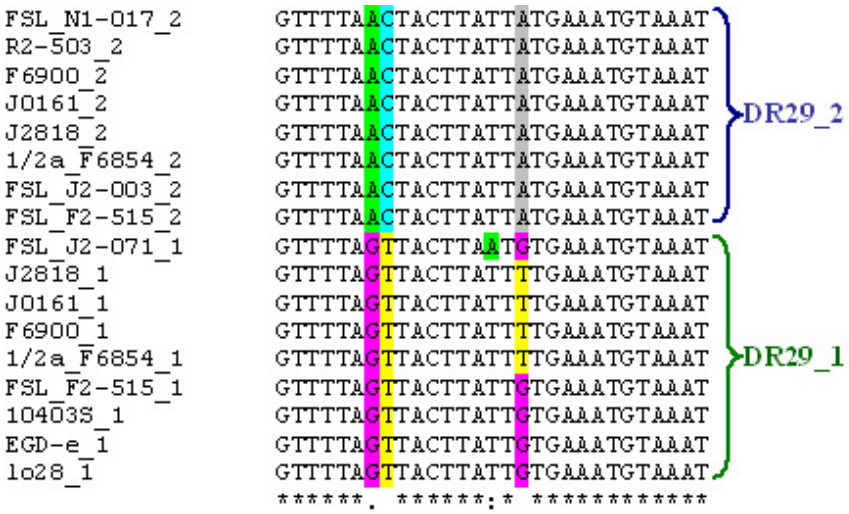


Fig. 5.1 – Alignement du DR29 appartenant aux CRISPRs de *L. monocytogenes*. Les nucléotides colorés ne sont pas toujours conservés entre les CRISPRs 29_1 et 29_2.

gène *cas*	locus	Nom	Cluster
Cas1(Cas1)	LMOf6854_1753.7	hypothetical protein	2
Cas2(Cas2)	LMOf6854_0557	CRISPR-associated protein Cas2	1
	LMOf6854_1753.6	hypothetical protein	2
Cas3 (core, HD domain, Ypest-specific)(Cas3)	LMOf6854_0555	ATP-dependent RNA helicase	1
Cas5(Cas5)	LMOf6854_0553.1	CRISPR-associated protein, TM1800 family	1
Cas6(Cas6)	LMOf6854_0551	CRISPR-associated protein, TM1814 family	1
Nmeni subtype specific proteins(Nmeni)	LMOf6854_1753.5	hypothetical protein	2
	LMOf6854_1753.8	hypothetical protein	2
Tneap subtype specific proteins(Tneap)	LMOf6854_0551.1	hypothetical protein	1
	LMOf6854_0553	CRISPR-associated negative autoregulator, putative	1
	LMOf6854_0553.1	CRISPR-associated protein, TM1800 family	1

Tab. 5.1 – Liste des gènes *cas* trouvés chez *L. monocytogenes 1/2a F6854*

Le tableau 5.2 résume les propriétés des trois CRISPRs trouvés dans les 20 génomes analysés. Le CRISPR29_1 est généralement plus petit (de 4 à 10 spacers) que le CRISPR29_2 (de 20 à 59 spacers).

Hypothèse 5.2.1. *Le CRISPR29_1 est un essaimage du CRISPR29_2.*

L'hypothèse 5.2.1 repose sur quelques observations sur CRISPR29_1 et CRISPR29_2. En effet, les deux DR ne sont pas exactement identiques. Le DR29_1 présente plus de variabilité, mais nous avons surtout observé que dans certains locus, le DR dégénéré du CRISPR29_1 est proche du DR consensus DR29_2 (figure 5.2) et présente donc une forme intermédiaire entre les deux DR. Comme les DR dégénérés sont les plus anciens du CRISPR, nous pensons qu'il est alors probable que le DR29_1 dérive du DR29_2. De plus, le CRISPR29_1 est toujours plus petit que le CRISPR29_2 donc probablement plus récent.

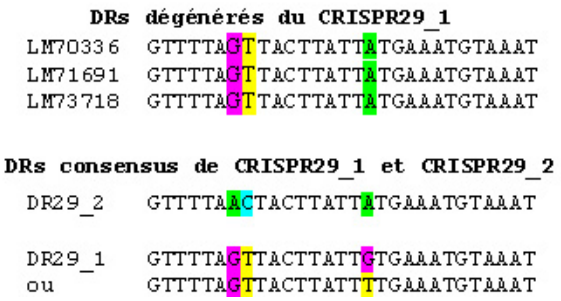


Fig. 5.2 – Variants du DR dégénéré du CRISPR29_1 par rapport aux DR consensus DR29_1 et DR29_2. Les nucléotides colorés représentent les points de différence entre les différents DR. Les noms des souches portant les DR dégénérés sont indiqués à gauche.

Analyse des spacers

Le blast des spacers contre Genbank montre quelques cas de similarité avec des séquences des bactériophages A118 (pour les deux clusters), PSA (pour les CRISPR29 uniquement) et P100 (pour les CRISPR36 uniquement). Ce dernier phage possède un intérêt particulier car il a été reconnu par la FDA (Food and Drug Administration) comme utilisable dans l'industrie agro-alimentaire pour réduire le nombre de souches de *L. monocytogenes* dans les aliments comme les fromages ou les produits carnés.

Il y a également une similarité avec une séquence de protéine (*Dictyostelium discoideum AX4*) qui est retrouvée chez les espèces *L. welshimeri serovar 6b str. SLCC5334* et *L. innocua Clip11262*.

De plus, nous avons noté une similarité entre des tronçons de spacers appartenant à des CRISPRs différents (figure 5.3) :

souche	serovar	CRISPR29_1 [nbr spacer]	CRISPR29_2 [nbr spacer]	CRISPR36 [nbr spacer]
4b	4b H7858	-	-	-
1/2b	J1 175	-	-	contig_2.353(revcomp) (32-529) [7]
1/2b	FSL J1-194	-	-	contig_2.40 (5172-5801) [9]
4b	hpb2262	-	-	-
1/2b	FSL J2-064	-	-	-
4b	4b F2365	-	-	-
4b	FSL N1-017	-	contig_2.20 (49423-50750) [20]	contig_2.68 (5161-7838) [40]
1/2b	FSL R2-503	-	contig_2.17 (49469-50800) [20]	contig_2.44 (5172-7849) [40]
4c	FSL J2-071	contig_2.6 (158451-159120) [10]	-	-
1/2a	10403S	contig_1.1 (190705-191248) [8]	-	contig_1.9 (5179-7196) [30]
1/2a1	FSL F2-515	contig_2.204 (188-672) [7]	contig_2.1377 (1343-3573) [34]	-
1/2a	1/2a F6854	564445 - 564727 [4]	579923 -581958 [31]	2733707-2734995 [19]
1/2a	F6900	contig_2.16 (28937-29219) [4]	contig_2.16 (44418-46453) [31]	contig_2.33 (5165-6453) [19]
1/2a	J0161	contig_1.19 (28933-29215) [4]	contig_1.19 (44413-46448) [31]	contig_1.11(revcomp) (93725-95013) [19]
1/2a	J2818	contig_2.13 (33994-34276) [4]	contig_2.14 (3783-5817) [31]	contig_2.37 (5164-6452) [19]
1/2a1	FSL N3-165	-	-	contig_2.33 (5261-6287) [15]
1/2a	FSL J2-003	-	contig_2.46 (2402-625) [59]	-
1/2c	EGD-e	544375-544663 [4]	-	-
1/2c	LO28	contig_2.72 (9566-9854) [4]	-	-
4a	FSL J1-208	-	-	contig_2.1456 (378-1006) [9]

Tab. 5.2 – Analyse des CRISPRs dans 20 souches de *L. monocytogenes*

- un spacer du CRISPR29_1 et un spacer du CRISPR29_2
- un spacer du CRISPR29_1 et un spacer du CRISPR36

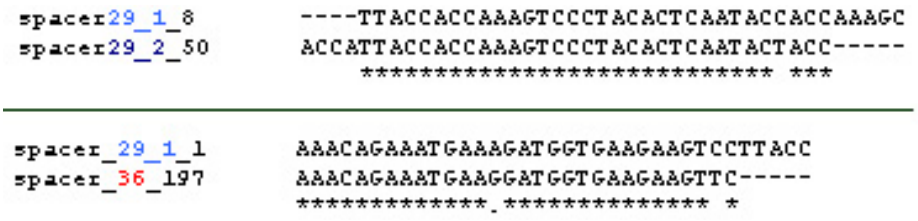


Fig. 5.3 – Alignement de quelques spacers similaires dans des CRISPR différents de *L. monocytogenes*.

Cette observation va dans le même sens que l'hypothèse 5.2.1 selon laquelle le CRISPR29_1 serait un essaimage du CRISPR29_2. En effet, la similarité entre un spacer du CRISPR29_2 et un spacer du CRISPR36 avec des éléments du CRISPR29_1 suggère que ce dernier aurait pour rôle de renforcer la résistance phagique (ou un autre rôle accompli par ces spacers dont l'origine est encore inconnue). Il aurait alors attrapé des morceaux similaires ou en train de se modifier de la séquence génomique à réguler. La création de ce CRISPR serait alors probablement une réponse à une pression extérieure ou un challenge provoqué par l'environnement de la souche pendant une période donnée.

L'application de CRISPRtionary à l'ensemble des trois CRISPRs trouvés a montré l'existence de :

- 30 spacers distincts pour le CRISPR29_1
- 105 spacers distincts pour le CRISPR29_2
- 125 spacers distincts pour le CRISPR36

L'aspect intéressant à noter est que cette analyse confirme bien la conservation de l'ordre des spacers dans les trois CRISPRs. En se basant sur les fichiers binaires produits par CRISPRtionary, nous avons construit quelques arbres phylogénétiques (utilisation du logiciel bionumerics avec les options "average from experiments" et UPGMA), ce qui nous a permis de faire quelques suggestions concernant l'évolution de *L. monocytogenes*. Notons cependant que les analyses suivantes sont sujettes à caution du fait que les séquences utilisées sont parfois incomplètes, et en cours de finition.

Utilisation du polymorphisme des CRISPRs pour la phylogénie de *L. monocytogenes*

A partir des différentes souches pour lesquelles des données génomiques sont disponibles, il a été possible de produire un arbre phylogénétique en comparant les séquences concaténées de 7 gènes (à la base du typage par MLST (Salcedo, 2003)) (Figure 5.4)

Dans l'ensemble des figures suivantes, les génotypes correspondent à l'étude MLST (voir figure 5.4).

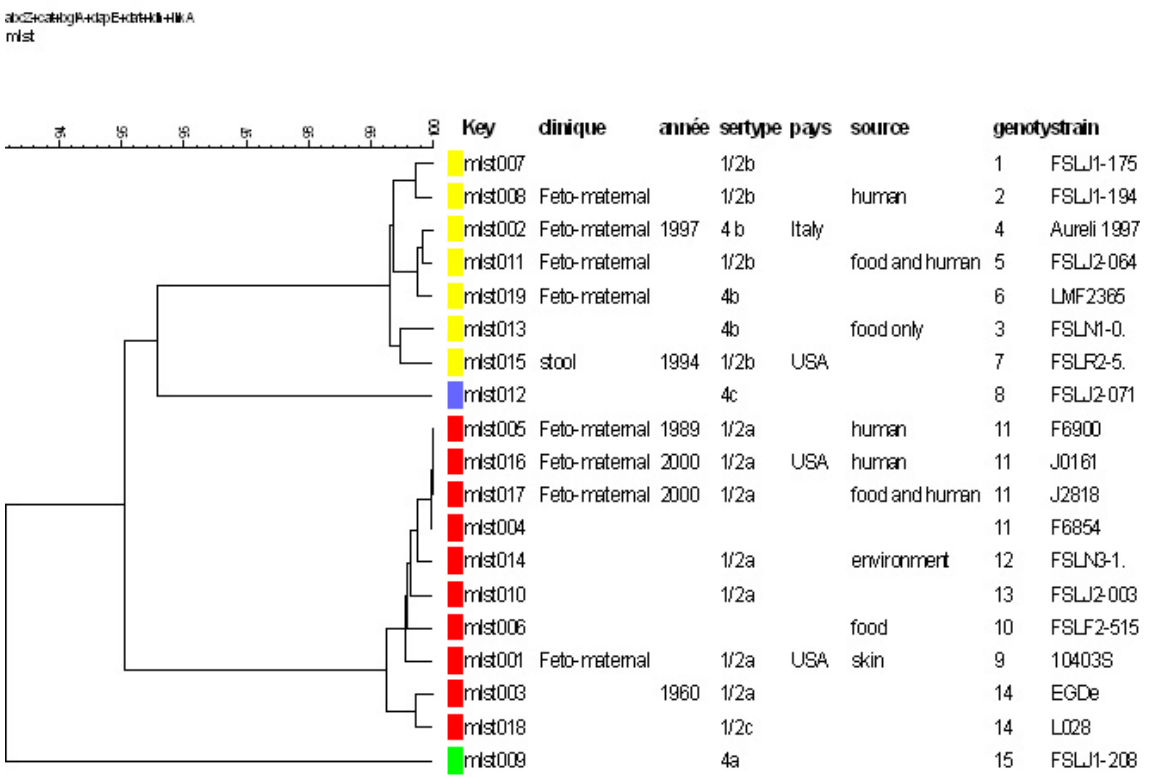


Fig. 5.4 – Arbre MLST de vingt souches de *L. monocytogenes*. Les gènes concaténés sont choisis à partir de (Salcedo, 2003). Cet arbre est construit avec le logiciel bionumerics options "average from experiments" et UPGMA. Le même génotype (de 1 à 14) est attribué aux souches très proches, il est indiqué dans l'avant dernière colonne "genoty". La colonne key désigne une clé arbitraire des souches. Les colonnes source et clinique désignent l'origine d'isolement des souches dont le nom figure dans la dernière colonne "strain". La colonne "sertype" désigne le sérotype de chaque souche. Le code couleur reflète les quatre groupes obtenus par l'arbre MLST.

Analyse du CRISPR29-1 et comparaison avec l'arbre MLST : Les spacers sont numérotés de telle sorte que les plus anciens (du côté du DR dégénéré) aient les numéros les plus petits, et les plus récents (du côté du leader) aient les numéros les plus grands.

La figure 5.5 montre qu'il n'y a pas beaucoup de spacers en commun entre les différentes souches (les spacers s5, s6 et s7 sont en commun entre les génotypes 9 et 14). Le locus CRISPR29_1 n'existe pas chez le groupe jaune. A partir de cet arbre, on pourrait bien penser que l'ancêtre du groupe jaune a perdu ce CRISPR ou ne l'a jamais eu.

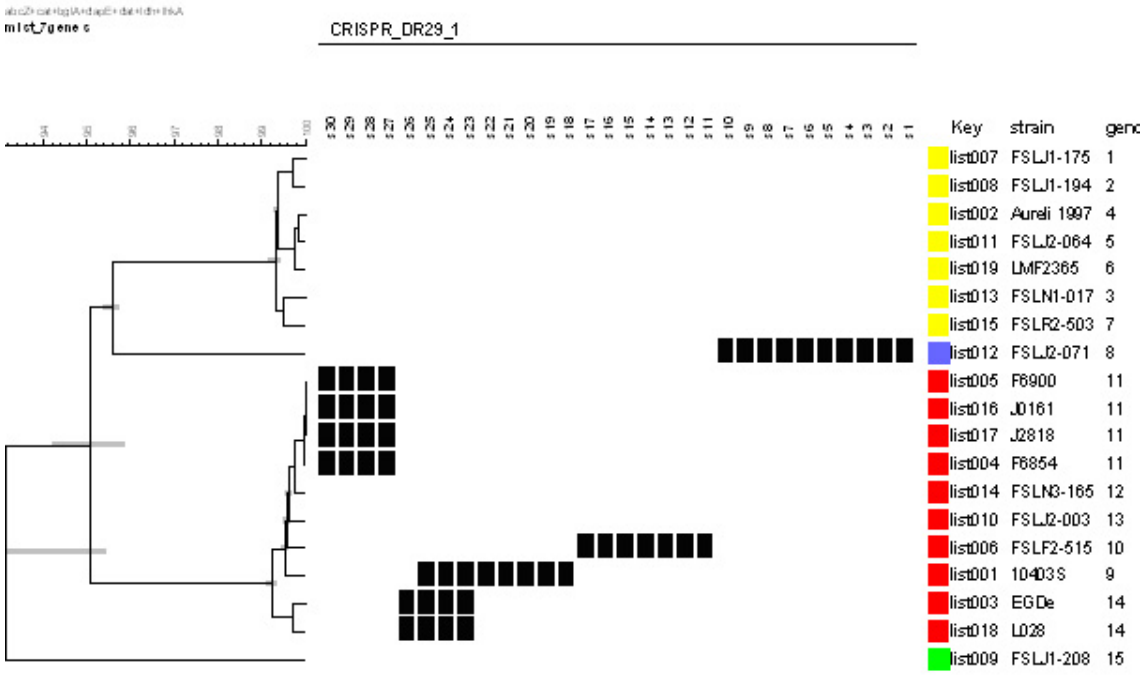


Fig. 5.5 – Arbre MLST et locus CRISPR29_1 de *L. monocytogenes*. Cet arbre est construit avec le logiciel bionumerics options "average from experiments" et UPGMA. Le génotype et le code couleur sont obtenus à partir de l'arbre MLST (figure 5.4). Les boîtes noires représentent la présence des spacers (spoligotype théorique des CRISPRs à partir de la liste des spacers obtenue de toutes les souches étudiées).

En se basant sur l'hypothèse d'acquisition polarisée des motifs et à partir de l'arbre 5.6, on peut penser que l'ancêtre commun possède les soixante dix premiers spacers situés à droite (du côté du DR dégénéré). D'après cet arbre, le génotype 10 est ancestral à 13 qui est lui même ancestral à 12 ancestral à 11, ce qui laisse supposer que 13 a perdu un certain nombre de motifs ancestraux avant d'acquérir les motifs récents.

L'arbre de la figure 5.7 montre l'organisation de CRISPR29_1 et CRISPR29_2. Le génotype 15, en position la plus ancestrale, représenté ici ne possède pas les CRISPR29 et on pourrait donc imaginer que ces locus sont apparus ultérieurement. Bien entendu une étude beaucoup plus poussée devra être conduite pour explorer cette hypothèse.

La figure 5.8 montre que le CRISPR36 existe chez des membres des trois groupes de l'arbre. Il semblerait alors que ce CRISPR est plus ancien que les CRISPR29.

Analyse du CRISPR 29-2 :

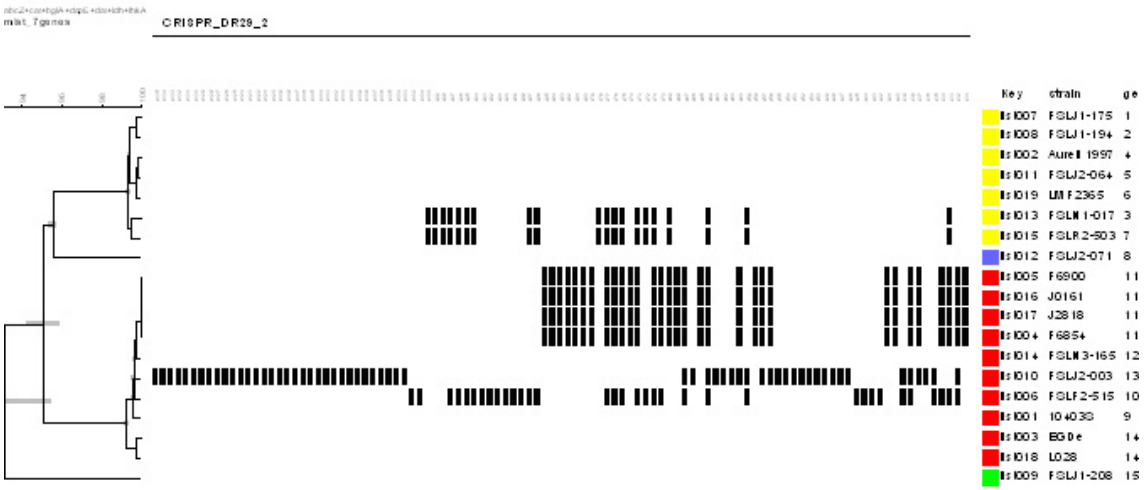


Fig. 5.6 – Arbre montrant le CRISPR29_2 de *L. monocytogenes*. Cet arbre est construit avec le logiciel bionumerics options "average from experiments" et UPGMA. Le génotype et le code couleur sont obtenus à partir de l'arbre MLST (figure 5.4). Les boîtes noires représentent la présence des spacers (spoligotype théorique des CRISPRs à partir de la liste des spacers obtenue de toutes les souches étudiées).

Conclusion

L'analyse MLST classe les 20 génomes disponibles (représentants les sérotypes 1/2a, 1/2b, 1/2c, 4a, 4b, 4c en deux clusters majeurs, le groupe 1/2b-4b d'une part et le groupe 1/2a d'autre part (figure 5.4). Les représentants 4a et 4c (un de chaque) pourraient constituer deux autres branches assez distinctes. Trois CRISPRs sont identifiés, correspondant à 2 types de DR (nommés DR29 et DR36) et 2 familles de gènes *cas*. Deux CRISPRs possèdent le DR29, ils sont nommés ici CRISPR29_1 et CRISPR29_2. CRISPR29_2 est situé à proximité du cluster de gènes *cas*, et selon le modèle proposé notamment par (Grissa, 2007b), CRISPR29_1 serait un sous-produit du locus principal. Ceci est conforté secondairement par l'observation que la séquence du premier DR de CRISPR29_1 est intermédiaire entre les DR consensus de CRISPR29_1 et 2, et aussi par le fait que le DR29_1 est plus petit que DR29_2 (ces observations sont toutefois des arguments faibles, en comparaison de l'argument d'association "gènes *cas*"). CRISPR29_1 ou CRISPR29_2 sont rencontrés dans des souches de chacune des 4 branches, à l'exception de la branche 4a. L'hypothèse la plus parcimonieuse est donc que l'ancêtre commun à ces 3 branches contenait déjà les deux complexes CASS et que les souches dépourvues de l'un ou l'autre l'ont perdu au cours de leur évolution. Le même raisonnement appliqué au CRISPR36 implique que ce locus existait dans un ancêtre commun aux 3 branches (la branche 4c fait exception). La combinaison des deux observations oblige d'ailleurs à penser que l'un des deux sites au moins existait chez un ancêtre commun aux 4 branches, et la phylogénie

Fig. 5.7 – Arbre montrant les CRISPRs CRISPR29_1 et CRISPR29_2 de *L. monocytogenes*. Cet arbre est construit avec le logiciel bionumerics options "average from experiments" et UPGMA. Le génotype et le code couleur sont obtenus à partir de l'arbre MLST (figure 5.4). Les boîtes noires représentent la présence des spacers (spoligotype théorique des CRISPRs à partir de la liste des spacers obtenue de toutes les souches étudiées).

suggérée par l'analyse MLST suggère que CRISPR36 est dans ce cas. Certains spacers sont retrouvés dans deux des branches. Le modèle d'évolution des CRISPRs proposé par (Pourcel, 2005) et maintenant largement accepté suggère alors que peu de spacers ont été acquis depuis la séparation des 4 branches, et que l'évolution des locus résulte principalement de pertes interstitielles (ou complètes) des différents locus, d'une façon qui rappelle la situation de *M. tuberculosis*.

Analyse du CRISPR 36 :

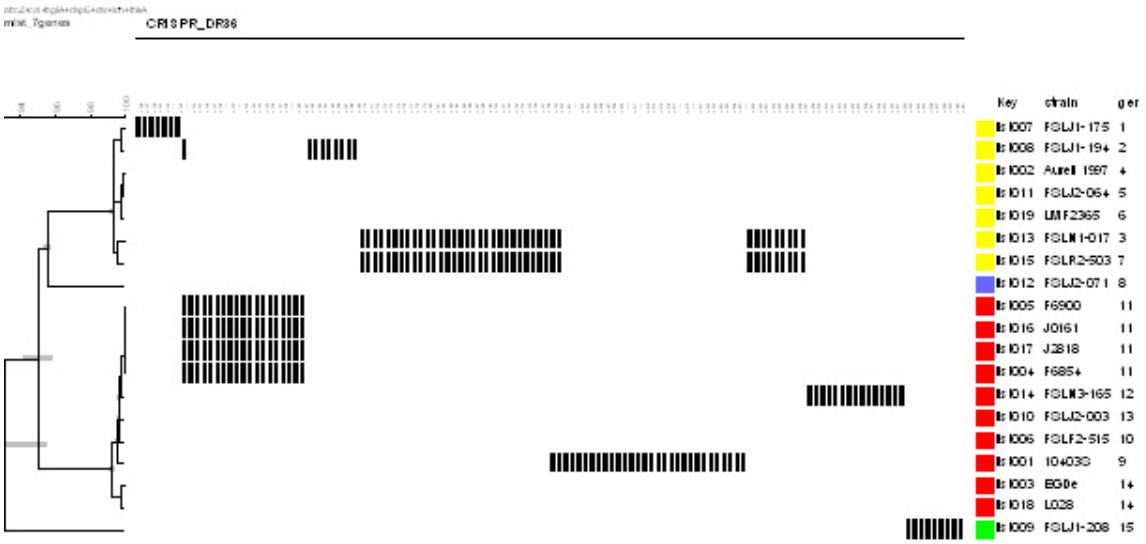


Fig. 5.8 – Arbre montrant le CRISPR36 de *L. monocytogenes*. Cet arbre est construit avec le logiciel bionumerics options "average from experiments" et UPGMA. Le génotype et le code couleur sont obtenus à partir de l'arbre MLST (figure 5.4). Les boîtes noires représentent la présence des spacers (spoligotype théorique des CRISPRs à partir de la liste des spacers obtenue de toutes les souches étudiées).

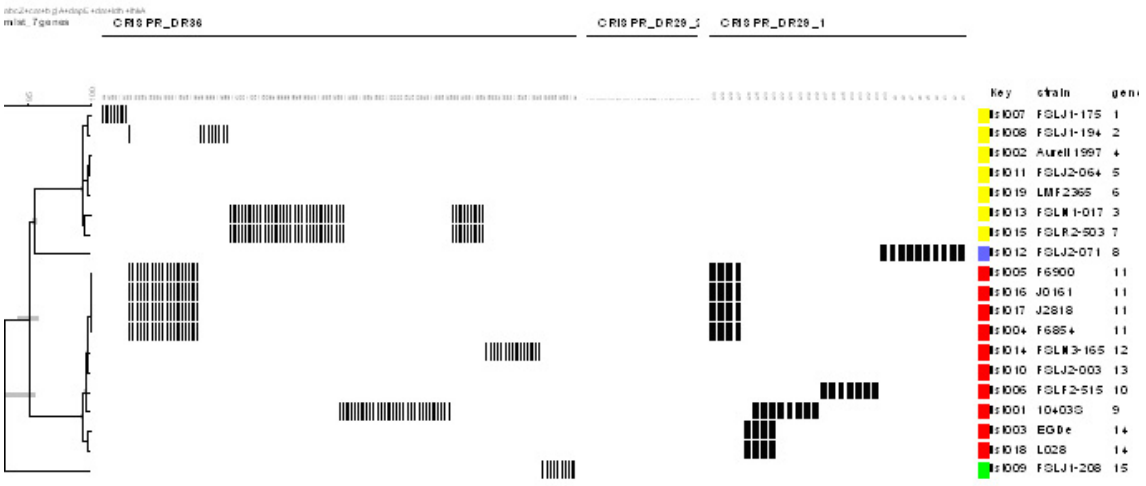


Fig. 5.9 – Arbre montrant le CRISPR36 et les CRISPRs 29_1 et 29_2 de *L. monocytogenes*

Troisième partie

Discussion et perspectives

Chapitre 6

Aspect bioinformatique

6.1 CRISPRFinder

Bien qu'au moment où j'ai commencé cette thèse, il n'y ait eu aucun programme dédié à la recherche des CRISPRs, au jour de la rédaction de ce rapport il en existe trois : CRISPRFinder, PILER-CR et CRT. Leurs performances sont assez proches dans la plupart des cas (voir paragraphe 3.2.2 page 55). PILER-CR rencontre des difficultés avec les séquences CRISPR non parfaites et CRT produit un bruit de fond important étant donné qu'il considère les répétitions en tandem comme des CRISPRs. CRISPRFinder a l'avantage de détecter les CRISPRs les plus courts. Certaines améliorations méritent d'être apportées au programme sur ce point en particulier. En effet, CRISPRFinder considère les structures courtes (1 ou 2 motifs) comme "questionables" et dans la plupart des cas ce ne sont en réalité que des faux positifs. Nous avons choisi de garder ces structures afin de ne pas éliminer les plus petits CRISPRs. Ce défaut pourrait être compensé par un filtrage supplémentaire pour confirmer ou rejeter ces structures. Les critères de confirmation sont par exemple 1) la reconnaissance d'un DR connu (vérification avec le contenu de CRISPRdb) ; cette étape est réalisée pour la validation des CRISPRs enregistrés dans la base, mais pas encore dans le résultat brut de CRISPRFinder 2) l'identification de gènes *cas* à proximité lorsqu'ils ne sont pas associés à un autre CRISPR confirmé et 3) le blast contre les bases de données NCBI qui pourrait permettre d'éliminer les séquences connues comme par exemple certaines protéines.

Par ailleurs, un CRISPR court est composé au moins de deux DR et un spacer, alors que certains génomes portent uniquement un leader et un DR ou l'un des deux. Dans les perspectives d'amélioration des performances de CRISPRFinder, il est également prévu d'identifier le leader (voir modalité au paragraphe 2.2.3 page 27) et d'ajouter une fonctionnalité permettant de chercher les traces de CRISPR (leader seul ou DR seul) quand l'utilisateur le désire, en effectuant un blast des DR et leader connus contre le génome soumis.

Un autre point concerne l'orientation du CRISPR car CRISPRFinder dans sa version actuelle représente les CRISPRs dans le sens du brin direct sans tenir compte de la position du leader qui pourrait être représenté par convention à gauche puisque c'est le promoteur de la transcription (voir figure 2.1 page 21). L'orientation du CRISPR peut être retrouvée soit par l'identification de la position du leader quand il existe, soit par la position du DR dégénéré, soit par l'alignement des séquences flanquantes quand plusieurs CRISPRs ayant le même DR se trouvent sur le même génome. De plus, cette opération permettra de stocker les CRISPRs dans la base CRISPRdb dans la bonne orientation. Il sera ainsi facile de retrouver l'orientation d'un nouveau CRISPR ayant le même DR qu'un autre CRISPR de CRISPRdb en reconnaissant l'orientation de ce DR.

D'autre part, on retrouve parfois à l'intérieur du CRISPR une séquence IS dont la taille est bien supérieure à celle d'un spacer. Bien que de part et d'autre de l'IS il s'agisse du même CRISPR, CRISPRFinder identifie deux CRISPRs différents. Sur la figure 6.1, deux CRISPRs du chromosome circulaire et un CRISPR du plasmide pTT27 de *T. thermophilus HB8* sont partagés en deux par la présence des IS Is1000A et Is1000B. Dans ces trois cas d'ailleurs un spacer a été omis par CRISPRFinder (S1, S2 et S3 sur la figure 6.1). Dans le cas du CRISPR NC_006461_5 de *T. thermophilus HB27* sur la même figure, le spacer S4 et le DR dégénéré DR' n'ont pas été pris en compte par CRISPRFinder.

Il serait alors intéressant d'ajouter à ce programme une nouvelle fonction qui permette de vérifier pour les CRISPRs voisins sur le même génome (plus de 40% des génomes séquencés ayant un CRISPR en contiennent au moins deux, voir figure 7.2) et ayant le même DR s'ils ne sont pas séparés par des séquences d'insertion et dans ce cas les considérer comme une entité unique.

Malgré ces quelques réserves, CRISPRFinder a l'avantage de fournir une liste exhaustive des CRISPRs et nous n'avons noté aucun cas où un vrai CRISPR de 3 motifs ou plus n'ait pas été détecté (par comparaison avec les analyses effectuées par d'autres programmes de détection). Il reste à ce jour le programme le plus simple d'utilisation, le seul accessible sur le web avec une interface graphique agréable et plusieurs utilitaires associés (le blast, les flanquantes, etc.). Nous notons en moyenne une douzaine d'utilisateurs par jour qui s'y connectent et plus de 22.000 séquences soumises à CRISPRFinder depuis Janvier 2007.

6.2 La base de données des CRISPRs

6.2.1 La mise à jour

La mise à jour de CRISPRdb se fait de façon automatique grâce à un "crontab" qui lance la mise à jour une fois par mois, l'espace de travail étant un système linux (Debian Sarge). Les nouveaux génomes sont alors téléchargés à partir du site ftp de NCBI. En-

suite, ils sont analysés par CRISPRFinder et stockés dans une base temporaire. Un mail de notification est envoyé aux administrateurs de la base qui se chargeront de la validation de ces nouvelles données pour les rendre publiques. Durant cette étape, le curateur se connecte sur une page d'administration à partir de laquelle il peut modifier le DR consensus d'un CRISPR si les extrémités ne sont pas définies correctement. Il peut également supprimer des faux positifs ou approuver des structures non confirmées. L'enregistrement de ces données permettra de partager en public les données qu'il a validées.

La version actuelle de CRISPRdb contient près de 700 génomes procaryotes analysés dont à peu près 300 portent des CRISPRs. Les CRISPRs confirmés qui sont archivés dans la base sont de l'ordre de 900. Cependant, le nombre des CRISPRs hypothétiques ou "questionables" est du même ordre de grandeur malgré la confirmation parce qu'ils portent un DR connu, de 130 CRISPRs courts (13% des structures à moins de trois motifs) et l'élimination manuelle de quelques uns. Ce nombre reste très important et nécessite une investigation supplémentaire pour faire le tri selon les critères cités dans le paragraphe précédent. Une collaboration avec R. Sorek nous a permis d'explorer le voisinage des CRISPRs non confirmés dans le but de chercher la présence de gènes *cas*. Dans 0,9% des cas, un "hit" important a été trouvé (selon les critères décrits par Haft *et col.* (Haft, 2005)). Dans la plupart de ces cas (2/3), il s'est avéré qu'un CRISPR confirmé est situé à proximité et n'a pas le même DR. Le plus logique alors serait de penser que les gènes trouvés sont plutôt associés au CRISPR confirmé. Dans les quelques cas restants, lorsque des gènes sont identifiés à proximité d'une structure "questionable" isolée, ce critère a été utilisé pour la confirmer comme étant un vrai CRISPR.

Nous avons également exploré le voisinage des CRISPRs courts confirmés par le critère du DR trouvé ailleurs dans la base. Dans 58% des cas, nous n'avons pas trouvé de gènes *cas* à proximité. Ces observations suggèrent que les petits CRISPRs sont ou bien les traces d'un système CASS devenu inactif qui a alors perdu ses gènes, ou bien le résultat de la formation d'un nouveau CRISPR nouvellement acquis dans un génome, par transfert horizontal par exemple, accompagné de sa machinerie CAS ou bien, un essaimage d'un autre CRISPR sur le même génome utilisant dans ce cas le même jeu de gènes associés.

Cependant, les faux positifs stockés dans CRISPRdb, même s'ils ne représentent probablement pas de vrais CRISPRs sont des structures curieuses à analyser, surtout quand elles ne peuvent pas être identifiées comme structure connue (répétition particulière ou protéine par exemple). Elles pourraient alors faire l'objet d'une recherche qui amènerait peut être à une nouvelle découverte.

6.2.2 My CRISPRdb

Durant les projets de séquençage, la publication d'un génome complet est une étape très longue. La plupart des nouveaux génomes séquencés existent alors sur les bases de données NCBI sous forme de brouillon ou "draft". Ces séquences sont très bien utilisées

par la communauté des biologistes, mais ne peuvent pas être enregistrés dans CRISPRdb puisqu'ils ne sont pas dans leur forme définitive. L'investigation de tels génomes peut être accomplie avec CRISPRFinder, mais cette opération ne sera pas enregistrée et l'utilisateur devra la refaire à chaque fois qu'il a besoin de ces données ; il ne pourra surtout pas utiliser les fonctionnalités de CRISPRcompar dans ce cas. C'est pourquoi, nous avons créé une base privée, à l'image de CRISPRdb mais qui constituera un recueil des données privées de l'utilisateur. Ce dernier pourra ainsi enregistrer des CRISPRs et les comparer. La création d'une base privée nécessite l'emploie d'un identifiant et d'un mot de passe mais ne requière aucune demande d'autorisation.

6.2.3 Développements futurs

La base de données CRISPRdb est la première et l'unique base de données à ce jour qui traite de la structure CRISPR. Elle constitue une référence pour les analyses des CRISPRs. On a noté une connexion quotidienne de plus de trente utilisateurs depuis sa création, elle a surtout la particularité de fournir non seulement la structure CRISPR via une interface attrayante, mais en plus une série d'outils facilitant son analyse. Cette base a beau être un outil indispensable dans l'investigation de la structure CRISPR, elle ne couvre pas tous les éléments du système CASS puisqu'elle n'analyse pas les gènes *cas*, qui sont d'ailleurs traités sur le site du TIGR par le groupe de Haft (Haft, 2005). Nous nous sommes contentés dans une première étape de focaliser notre intérêt sur la structure CRISPR et mettre un lien vers la base du TIGR (`http://cmr.tigr.org/tigr-scripts/CMR/shared/MakeFrontPages.cgi`) pour la recherche des gènes associés. L'étape suivante consistera en la création d'outils propres à CRISPRdb pour la recherche et la description de ces gènes. L'analyse combinée des CRISPRs et des gènes *cas* serait très intéressante surtout pour éclaircir davantage la relation entre le type de DR et les sous ensembles de gènes *cas* qui lui sont spécifiques. Les développements futurs de la base concerneront la détection automatique des gènes associés au CRISPR et surtout la constitution de classes du système CASS caractérisées par le DR, les gènes associés et de façon plus ambitieuse la signature du proto-spacer correspondante. L'identification de l'un de ces éléments facilitera par la suite la recherche des autres éléments surtout dans les cas les plus compliqués.

6.3 CRISPRcompar

Cette thèse illustre l'intérêt du CRISPR comme outil de typage bactérien. L'outil CRISPRcompar a été développé en vue de faciliter cette tâche. Il a été mis à profit au laboratoire pour l'investigation de plusieurs pathogènes humains :

- *Y. pestis* (Vergnaud, 2007),
- *L. monocytogenes* (voir la section 5.2.2 page 75),
- *P. aeruginosa*,

– *L. pneumophila.*
– *E. coli*

Cet outil est encore dans sa première phase de développement. Son développement et l'amélioration de ses performances dépendent étroitement des remarques et suggestions de ses utilisateurs. En effet, comme toute structure biologique le CRISPR n'obéit pas toujours de façon stricte aux critères par lesquels nous le définissons (voir partie 7.1) et il est inutile d'essayer de prévoir tous les cas particuliers ; la découverte des exceptions ne peut se faire que de façon empirique.

6.4 CRISPRcomparison

CRISPRcomparison effectue la comparaison des CRISPRs sur deux critères, l'identité entre les DR dans un premier temps et la comparaison des deux flanquantes dans un deuxième temps. Ces deux critères peuvent être critiqués dans quelques cas particuliers. Le premier concerne la dégénérescence du DR qui peut être parfois muté de quelques nucléotides d'une espèce à l'autre alors qu'il s'agit bien du même CRISPR. Ce problème sera résolu par l'ajout d'une option permettant à l'utilisateur de choisir un ordre de similarité entre les DR à comparer. La valeur par défaut de cette option sera une identité parfaite entre les DR.

Pour le critère de comparaison des flanquantes, sa fiabilité diminue avec la version actuelle de CRISPRFinder qui ne gère pas les insertions de séquences au milieu d'un CRISPR (voir partie 6.1 de ce chapitre). Pour illustrer cet inconvénient, le CRISPR 3 sur le plasmide de *T. thermophilus HB27* (voir figure 6.1) pourrait être un bon exemple. Ce CRISPR s'étend de la position 144.129 à la position 146.983, mais il contient, comme le montre la figure, une séquence IS qui s'étend sur 1.161 pb entre le dixième DR et le spacer S1. Il est alors considéré par CRISPRFinder comme étant deux CRISPRs NC_006462_5 et NC_006462_6 séparés par une séquence de 1.199 pb formée en réalité par la séquence IS et le spacer S1. Le CRISPR NC_006462_6 partage onze spacers avec le CRISPR NC_005838_3 du plasmide *HB8* de la même espèce, mais ils ne sont pas reconnus comme similaires par l'outil CRISPRcomparaison car la séquence flanquante gauche est différente. Or, cette flanquante qui est d'ailleurs le leader du CRISPR est bien conservée lorsque l'on tient compte du fait que NC_006462_5 et NC_006462_6 ne forment qu'un seul CRISPR. Il s'agit du même type d'erreur pour les trois CRISPRs CRISPR1, CRISPR2 et CRISPR4 mentionnés sur la figure 6.1.

La gestion des séquences IS dans les CRISPRs par CRISPRFinder permettra de palier à ce problème. L'exemple précédent est donné à titre illustratif des limites de CRISPR-comparison dans la version actuelle.

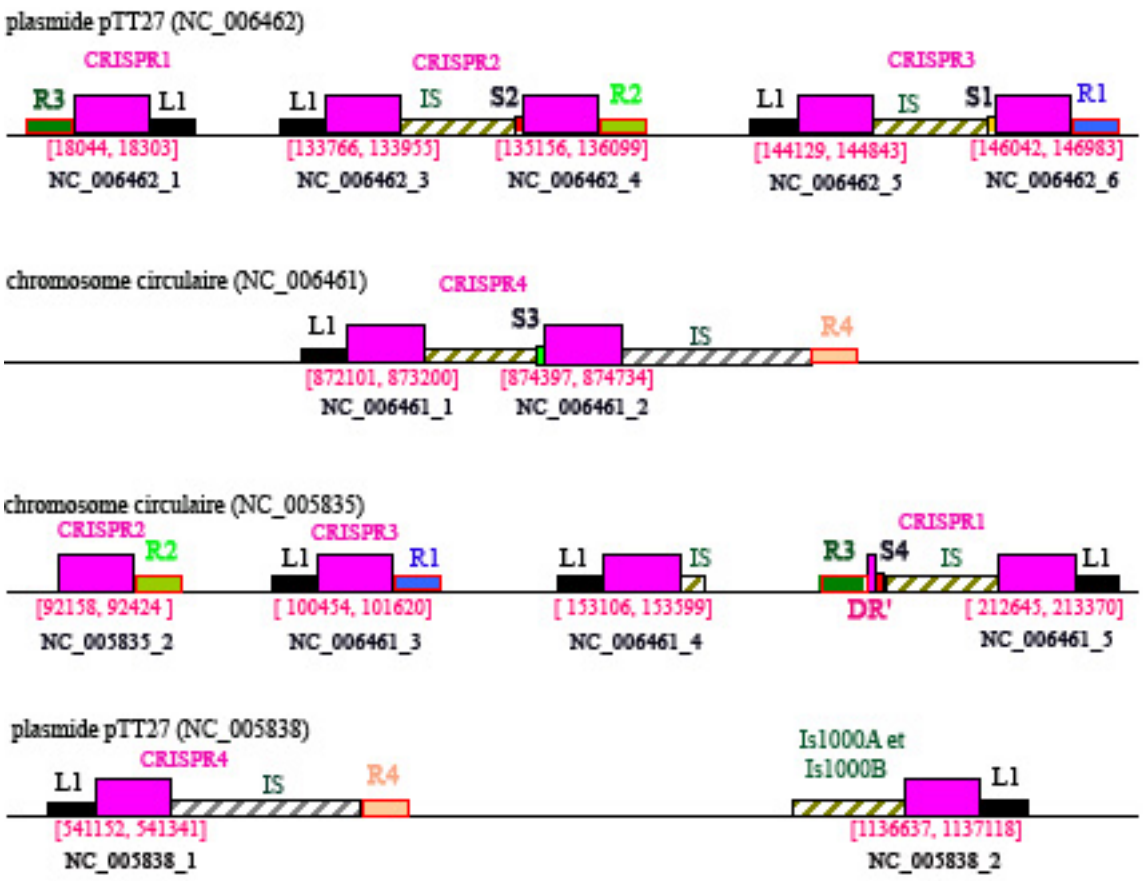


Fig. 6.1 – Les CRISPRs de *T. thermophilus HB27* et *HB8* ayant le DR1. DR1 = **GTCGCAATCCCCTTACGGGGCTCAATCCCTTGCAAC**. Les séquence Is1000A et Is1000B sont insérées dans cinq CRISPRs, trois fois au milieu chez *T. thermophilus HB27* et deux fois à la fin chez *T. thermophilus HB8*. L1 désigne la même séquence leader de taille 216 pb. **R** désigne une flanquante droite. Les séquences S1, S2 et S3 désignent des spacers à côté des séquences IS.

6.5 CRISPRtionay

Pour CRISPRtionary, les développements futurs concerneront deux points de réflexion :

- **Réflexion sur l'ordre des spacers :** il a été observé chez *S. thermophilus* qu'un même spacer peut être inséré deux fois indépendamment chez des mutants de la même espèce et que l'acquisition d'un nouveau motif peut être effectuée parfois de façon non polarisée. Ces deux événements restent assez rare, mais possibles et ils introduisent une confusion dans la nomenclature des spacers.

D'autre part, les phénomènes de recombinaison étant possibles au sein du CRISPR, il arrive de trouver des motifs dupliqués, à la suite l'un de l'autre (en tandem).

FIG. 6.2 – Amplification du neuvième motif dans le CRISPR NC_000909_5 de *Methanocaldococcus jannaschii DSM 2661*.

A titre d'exemple, le neuvième motif de *Methanocaldococcus jannaschii DSM 2661* est répété en tandem sept fois (figure 6.2). Ce cas de figure n'est pas très gênant dans CRISPRtionary du fait qu'un seul motif est répété au même endroit et il n'altère donc pas l'ordre des autres motifs. Cependant, lorsque plusieurs motifs répétés sont dispersés dans le CRISPR, il devient difficile de comprendre comment les motifs ont été acquis. C'est le cas par exemple du CRISPR **NC_000916_7** de *Methanothermobacter thermautotrophicus str. Delta H.* La figure 6.3 indique les motifs qui se répètent. Lorsque les motifs répétés se poursuivent (le bloc **f-g** et le bloc **h-i-j-k-l-m**), il est judicieux de penser qu'il s'agit de duplication à l'intérieur du CRISPR. Cependant, pour les motifs **a** et **e**, il est difficile de conclure s'ils ont été acquis indépendamment deux fois ou s'il s'agit tout simplement de duplications de motifs à l'intérieur du CRISPR. Le problème dans ce genre de situation est que les motifs dupliqués de façon dispersée perturbent la détermination de l'ordre d'acquisition des motifs par l'outil "reannotate". On ne sait plus par exemple si le motif **b** est acquis avant ou après le motif **h** puisque **b** est situé avant et après **h**. Dans ce cas, c'est toute la nomenclature des motifs qui est mise en question.
Dans sa version actuelle, CRISPRtionary néglige les deux phénomènes cités ci-dessus

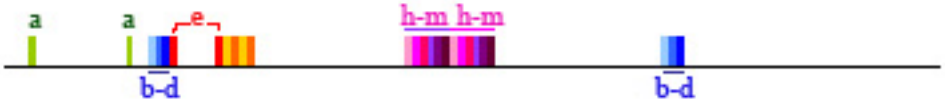


Fig. 6.3 – Répétitions de quelques spacers dans le CRISPR NC_000916_7 de *M. thermautotrophicus str. Delta H.* Seuls les motifs répétés ont été nommés. Une couleur a été attribuée par motif.

et se contente d'ajouter la terminaison "bis" à l'un des motifs répétés. Les motifs répétés plus que deux fois ne sont pas gérés et dans la représentation binaire par exemple, un motif répété ne sera représenté qu'une seule fois. D'ailleurs, il faut noter que le spoligotypage ne tient pas compte non plus des dédoublements de motifs dans le CRISPR de *M. tuberculosis*.

– **Réflexion sur l'exploitation phylogénétique du CRISPR**
L'acquisition polarisée des spacers est une propriété très intéressante dans le CRISPR puisque non seulement les spacers communs à deux souches différentes renseignent sur un ancêtre commun mais en plus, l'ordre des spacers sera toujours conservé et les spacers les plus anciens seront toujours situés d'un même côté. D'autre part, il est possible qu'un CRISPR perde des motifs interstitiels par unités ou par blocs (les deux hypothèses sont envisageables mais aucune n'est prouvée être plus fréquente que l'autre). L'exploitation de ces propriétés dans la création d'un logiciel phylogénétique spécifique à l'analyse des CRISPRs serait une tâche bioinformatique ambitieuse et présenterait un complément important de l'outil CRISPRcompar. Il permettrait de produire des arbres phylogénétiques à partir par exemple de modèles probabilistes obéissant aux propriétés du CRISPR, mais ce programme devra également tenir compte des cas particuliers discutés dans le paragraphe précédent.

6.6 Conclusion

Trois outils bioinformatiques ont été développés durant cette thèse pour l'identification, l'investigation et le typage des CRISPRs. Ces programmes gèrent assez bien la majorité des CRISPRs chez les procaryotes, mais nécessitent encore quelques améliorations surtout pour gérer certains cas particuliers intéressants comme la répétition de motifs ou l'insertion de séquences IS.

Chapitre 7

Aspect biologique

7.1 Le CRISPR, encore un mystère

7.1.1 Diversité du CRISPR chez les procaryotes

Le système CASS en général montre une très grande diversité dans ses constituants que ce soit dans la nature des gènes *cas* ou dans la structure des CRISPRs. La diversité des gènes *cas* et leur regroupement en différentes familles ont été discutés dans le chapitre 2. Concernant la structure CRISPR, elle représente non seulement une diversité importante dans la nature et le nombre de ses spacers à l'intérieur d'une même espèce, mais également une grande diversité entre différentes espèces. Cette diversité concerne toutes les caractéristiques du CRISPR :

- Le DR : il présente une grande diversité d'abord par sa taille variable de 23pb à 47pb et la nature de sa séquence ainsi que la capacité à adopter une structure secondaire particulière (Kunin, 2007). De plus, bien que l'observation générale des DR montre qu'il s'agit de motifs semi palindromiques et qu'ils possèdent une terminaison particulière (C/G)AA(A)(C/G), ces propriétés ne sont pas toujours observées, elles dépendent plutôt de la classe à laquelle appartient le CRISPR (Kunin, 2007). Par ailleurs, le DR est surtout caractérisé par une conservation très forte au sein d'un CRISPR, même le plus grand. Cependant, nous avons remarqué l'existence de quelques cas surprenants où les DR sont très hétérogènes comme dans le cas de la plupart des CRISPRs présents sur un plasmide de l'espèce *Clostridium botulinum* (voir exemple dans la figure 7.1). L'aspect remarquable dans ce CRISPR est que les spacers aussi semblent avoir une taille peu conservée. La présence de gènes *cas* à proximité (*cas5* hmari et *cas6*) confirme qu'il ne s'agit pourtant pas d'un faux CRISPR. Cette dégénérescence est peut être due à la divergence de ces CRISPRs car ils ne seraient plus utiles à l'espèce ou traduit un mode de fonctionnement différent. Cette variabilité de la taille des spacers et la non conservation parfaite du DR

a été également observée par Kunin *et col.* dans les métagénomes (Kunin, 2008).

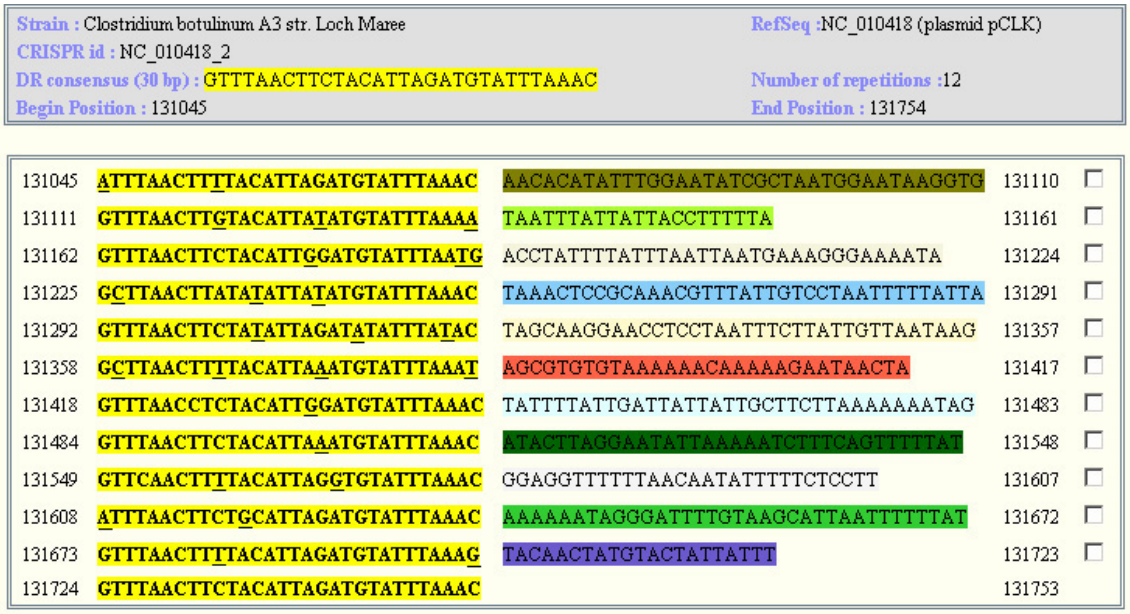


Fig. 7.1 – CRISPR NC_010418_2 du plasmide pCLK de *C. botulinum A3 str. Loch Maree*. Les nucléotides divergeants de chaque DR par rapport au consensus sont soulignés.

- Les spacers : l'existence de spacers communs chez des souches proches est un critère important dans les études de micro évolution car un spacer commun indique un ancêtre commun. Cependant, les spacers ne présentent quasiment pas de similarité d'une lignée à l'autre, même entre taxons proches. Cela suggère un renouvellement rapide à l'échelle évolutive. En corollaire, cela suggère que même entre taxons proches, les phages ou les plasmides les plus courants sont différents et/ou que les phages ou les plasmides dominants se renouvellent très rapidement.
- Le nombre de clusters : le nombre de CRISPRs par génome varie de un seul CRISPR dans plus de 35% des génomes séquencés possédant cette structure à vingt CRISPRs sur le chromosome circulaire de *Methanocaldococcus jannaschii DSM 2661*. Cependant, dans presque 90% des cas le nombre de CRISPRs par génome varie de un à cinq (voir figure 7.2).
- Le nombre de motifs : le nombre de motifs au sein d'un même CRISPR varie de un seul motif (voir même un seul DR, données non gérées par CRISPRdb) pour aller jusqu'à 276 chez *Chloroflexus aurantiacus J-10-fl* ou 293 motifs chez *Verminephrobacter eiseniae EF01-2* qui est un CRISPR partagé en deux (NC_008786_3-NC_008786_3) par la présence d'une séquence IS de à peu près 1000pb. La figure 7.3 représente le nombre de CRISPRs par classe de taille en nombre de motifs. Elle montre que la distribution est presque la même pour les bactéries et les archées et surtout que les CRISPRs petits sont prépondérants. Plusieurs questions surgissent face à cette distribution. Pourquoi certains CRISPRs sont ils si longs et d'autres si petits ? Y a -t-il une limite pour la taille d'un CRISPR et si oui quelle est cette limite ? La réponse à ces questions reste encore un mystère de ce système curieux. Par ailleurs, il semble que certains génomes ont plus de capacité que d'autres à accumu-

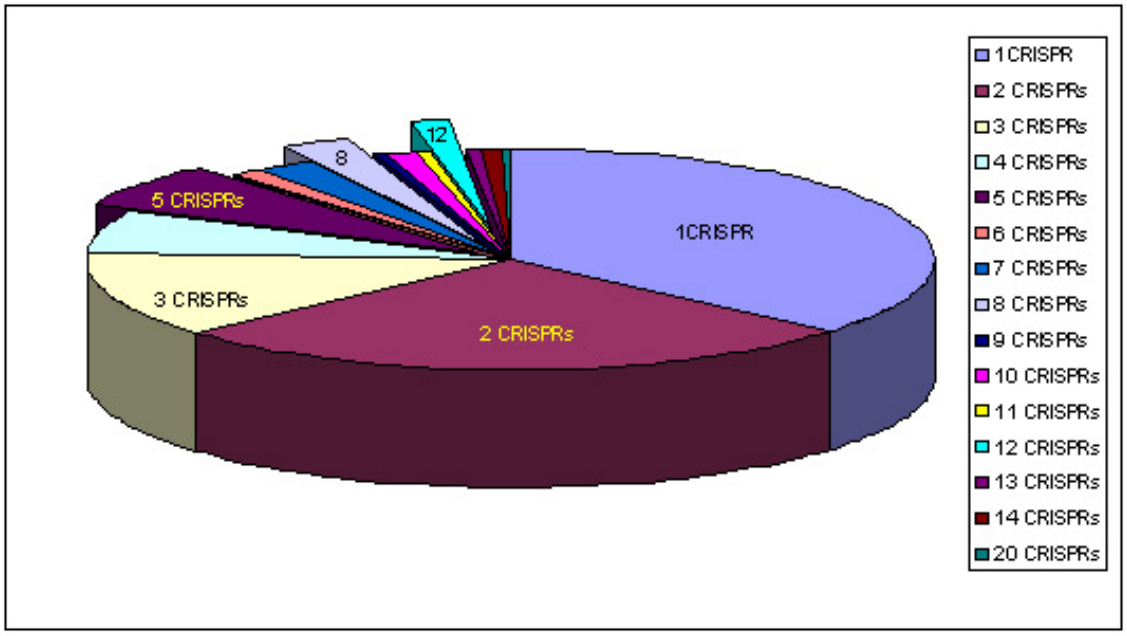


Fig. 7.2 – Distribution du nombre de CRISPRs par génome.

ler les motifs de CRISPR. Dans l'annexe D est reporté un tableau récapitulatif du nombre de clusters, de la taille du plus long CRISPR et du nombre total de motifs par génome. Le nombre de motifs maximal est de 539 chez *Chloroflexus aurantiacus J-10-fl* sur trois CRISPRs. Ce nombre aurait peut être un lien avec l'habitat ou une particularité biologique du génome concerné. Dans une tentative de trouver cette particularité, la figure 7.4 montre la dispersion du nombre de motifs CRISPRs en fonction de la classe taxonomique ("intermediate rank 3" dans CRISPRdb (Grissa, 2007b)) pour quelques cas intéressants (une figure de toutes les classes taxonomiques est fournie en annexe C, figure C.2). Comme les boites de dispersion semblent être assez étroites, il est possible qu'un lien existe entre la classe taxonomique et la capacité à cumuler des motifs de CRISPR. Cependant, ceci reste une simple suggestion et un contre exemple peut être fourni pour les *Clostridiales* dont la valeur maximale des répétitions est de 436 chez *Clostridium thermocellum ATCC 27405* alors que la médiane du nombre de motifs est inférieur à 150 motifs.

Par ailleurs, la diversité du système CASS est remarquable du fait même de sa présence ou absence chez les organismes. Chez les archées, il semble soit que la défense contre les agressions génétiques via ce système est indispensable soit que son rôle ne se limite pas à cette fonction et qu'il accomplit une autre tâche vitale pour ces micro-organismes. En ce qui concerne les bactéries, il n'est présent que chez 40% d'entre elles. Il est alors probablement indispensable chez la bactérie lactique *S. thermophilus*, mais il faut noter que seulement 46,1% des bactéries lactiques (47/102) possèdent un CRISPR (Horvath, 2008a). L'aspect le plus curieux est qu'au sein d'une même espèce, le CRISPR peut être présent chez certaines souches et absent chez d'autres, comme dans les espèces *L. pneumophila* ou *P. aeruginosa* alors que ce sont des bactéries de l'environnement. Cependant les souches dont le génome a été séquencé sont le plus souvent associées à une infection.

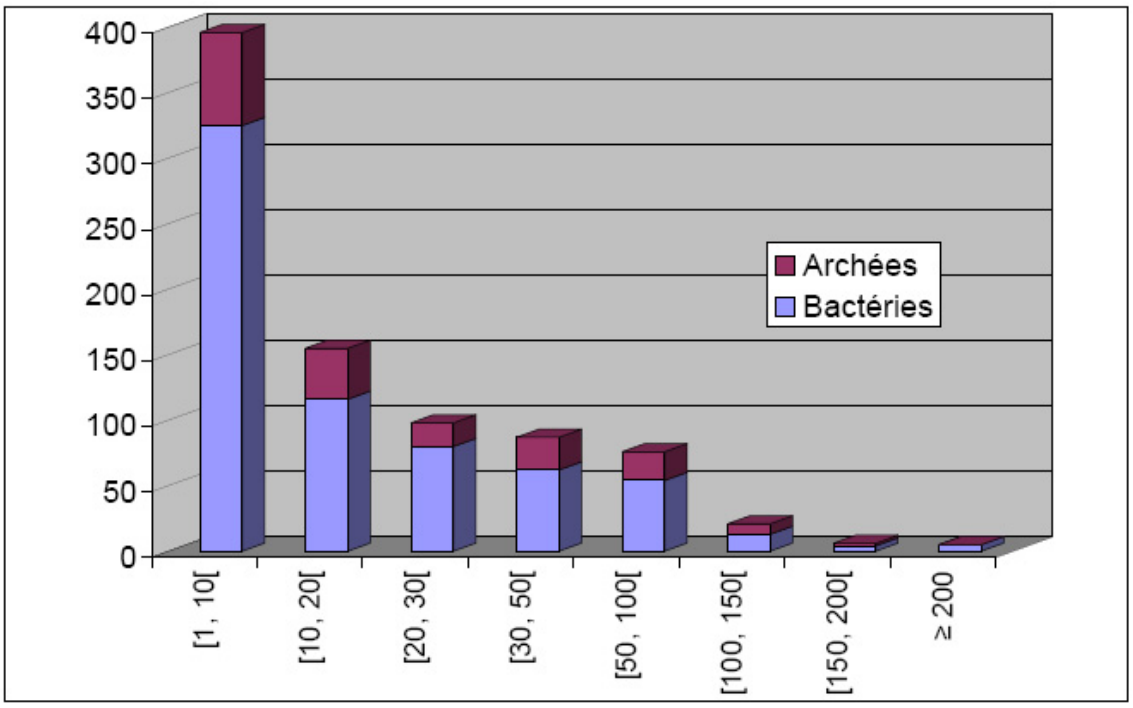


Fig. 7.3 – Distribution du nombre de motifs dans tous les CRISPRs de CRISPRdb.

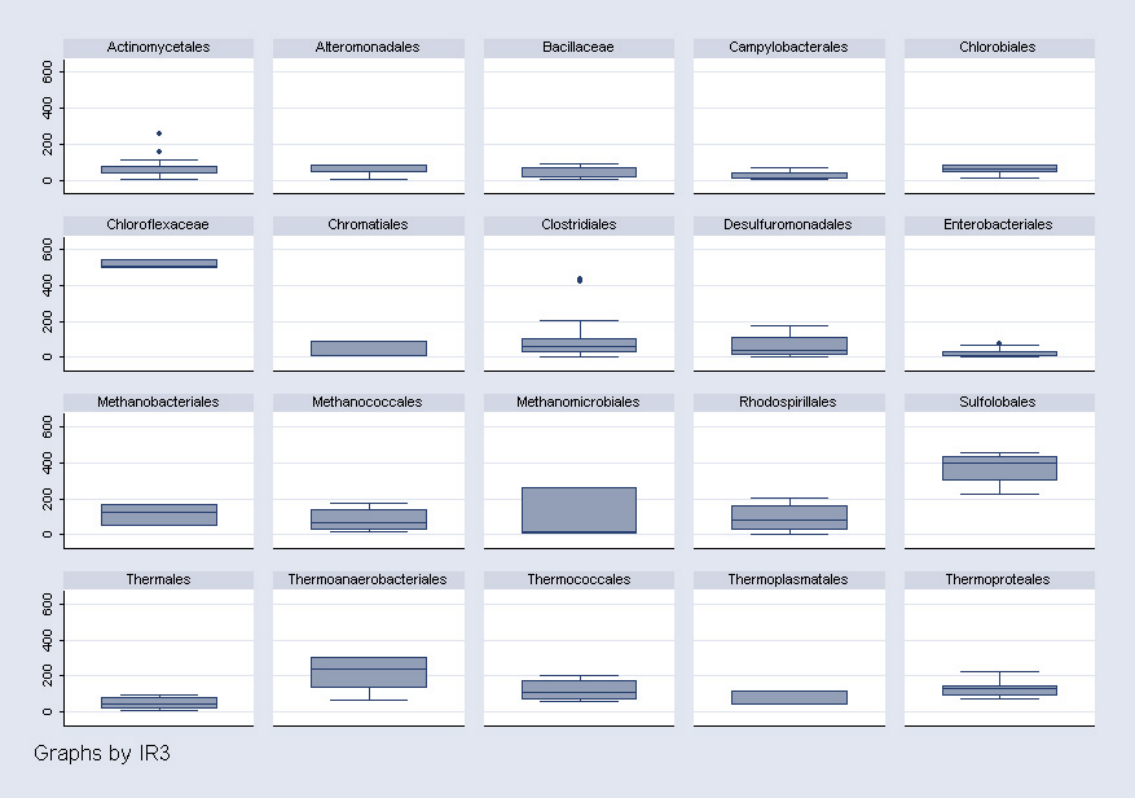


Fig. 7.4 – Dispersion du nombre de motifs dans le cas de quelques classes taxonomiques.

Plusieurs interrogations concernant la nécessité et le rôle de ce système si complexe restent à débattre. L'une des clés de ce débat serait l'analyse des bactéries de l'environnement et de leurs CRISPRs en contenu, taille, activité et vitesse d'évolution à l'instar du travail récent d'exploration des CRISPRs dans un biofilm de drainage minier acide par une équipe du Joint Genome Institute (JGI) (Andersson et Banfield, 2008) ayant montré par exemple une activité intense des CRISPRs au sein de cette communauté.

7.1.2 Mode d'évolution du locus CRISPR

Formation de nouveaux motifs

Le mode d'acquisition des motifs au sein du CRISPR est encore un phénomène biologique non connu. Une première suggestion a été fournie par van Embden (van Embden, 2000) supposant que le CRISPR est à l'origine une répétition en tandem formée d'une centaine de motifs. Les spacers auraient émergé et se seraient accumulés durant l'évolution pour former le CRISPR. Les changements actuels consisteraient alors en des délétions d'unités ou de blocs par recombinaison. Cette proposition a été faite à partir des observations faites chez *M. tuberculosis* et ne peut pas expliquer comment de nouveaux motifs peuvent être acquis par le CRISPR. Ce sont les observations faites par notre laboratoire sur *Y. pestis* (Pourcel, 2005) puis par l'équipe de Horvath et col. sur *S. thermophilus* (Barrangou, 2007) qui ont montré l'ajout de nouveaux motifs presque exclusivement de manière polarisée. Le groupe de E.V. Koonin (Makarova, 2006) a proposé qu'un nouveau motif puisse être inséré par conversion impliquant une recombinaison homologue entre un nouveau DR et un DR du CRISPR (voir figure C.3 dans l'annexe C). Cette proposition ne repose sur aucune donnée expérimentale et ne tient pas compte des observations précédentes. Nous pensons que l'acquisition d'un nouveau motif impliquerait la duplication du dernier DR immédiatement adjacent au leader. En effet, nous avons observé deux cas de figure soutenant cette hypothèse (décrits dans (Grissa, 2007b)).

Stabilité des motifs

Un CRISPR donné peut contenir jusqu'à près de 300 motifs (voir paragraphe 7.1.1), mais tous les CRISPRs ne sont pas aussi grands. Il semble que certains CRISPRs ont tendance à perdre des motifs très régulièrement comme chez *M. tuberculosis* alors que d'autres semblent les accumuler. Cependant, chez *S. thermophilus*, un équilibre paraît s'établir entre les motifs acquis et les motifs perdus au début du CRISPR. Cette remarque est confirmée par les observations de Tyson *et col.* (Tyson et Banfield, 2008) chez *Leptospirillum* à propos d'une importante perte de motifs dans la première partie du CRISPR corrélée à l'accumulation de spacers spécifiques à chaque population. Plus récemment, Andersson et Banfield (Andersson et Banfield, 2008) ont trouvé que chez des souches naturelles d'un biofilm acidophilique, le CRISPR a presque renouvelé tous ses spacers au

bout de cinq mois uniquement. En effet, il semble que le CRISPR réagit rapidement à la modification de la pression de sélection par une accumulation rapide de nouveaux motifs du côté du leader et qu'en parallèle, il se "débarrasse" de motifs anciens correspondant à des phages que la bactérie ne rencontre plus (perte de spacers en l'absence de pression de sélection).

Interruption par l'insertion de séquences IS

La présence d'éléments transposables au milieu d'un CRISPR a été observée dans plusieurs cas :

- *T. thermophilus*, figure 6.1,
- le CRISPR de *M. tuberculosis*,
- deux plasmides d'archées reportés par Lillestøl *et col.* (Lillestøl, 2006) : *M. acetivorans* et *H. marismortui*,
- *Leptospirillum* rapporté par Tyson et Banfield (Tyson et Banfield, 2008), etc.

L'interruption d'un CRISPR par des séquences IS peut être expliqué ou bien par une affinité des transposons pour certains spacers, ce qui limite leur mouvement dans le génome et fournit un mécanisme pour leur excision ou bien par une relation d'interférence entre ce genre de séquences et le CRISPR ou bien par l'absence de pression de sélection en faveur de la conservation d'un locus CRISPR intègre. Il est également probable que les séquences IS servent par exemple à réactiver la fin du CRISPR en permettant d'initier la transcription d'anciens spacers.

7.1.3 Multiplication des CRISPRs au sein d'un même génome

Le nombre de CRISPRs dans un même génome peut s'élever jusqu'à vingt (figure 7.2). Parfois il s'agit de CRISPRs différents avec des DR différents et des gènes *cas* complètement indépendants (le cas des trois CRISPRs de *S. thermophilus*) et parfois plusieurs CRISPRs sont arrangés côte à côte, ont le même DR, le même leader et possèdent un seul jeu de gènes *cas*. On parle dans ce cas de l'essaimage d'un même CRISPR en plusieurs copies. Plusieurs exemples ont été remarqués dans CRISPRdb, citons le cas de plusieurs clusters chez différentes espèces de *Clostridium* telle que *C. botulinum F str. Langeland* pour laquelle dix petits CRISPRs (NC_009699_5- NC_009699_14) ayant de un à quatre motifs sont situés côte à côte entre les positions 2351669 et 2382729pb. Il est possible alors qu'à partir d'un seul CRISPR naissent d'autres CRISPRs ayant une structure minimale composée probablement d'un leader et d'un DR en présence d'un unique ensemble de gènes *cas* agissant en trans. Tous les DR ne sont peut être pas capables d'effectuer cette opération, alors que certains DR le font systématiquement. Par exemple le DR "GTTGAACATTAACATAAGATGTATTTAAAT" est un DR spécifique du genre *Clostridium*, il existe pratiquement chez tous les génomes publiés de ce taxon, avec quelques différences d'une espèce à l'autre, mais le CRISPR correspondant se présente à chaque fois

en plusieurs copies. Le processus de multiplication et son utilité s'ajoutent aux points mystérieux dans le comportement du système CASS.

7.1.4 Origine des spacers

Les spacers du CRISPR semblent provenir majoritairement de séquences virales ou plasmidiques mais parfois également chromosomiques (voir tableau 2.1 page 26). Un spacer donné est rarement acquis deux fois de façon indépendante dans un même CRISPR (observation faite chez *S. thermophilus*). Cependant, Lillestøl *et col.* (Lillestøl, 2006) ont rapporté l'existence de spacers similaires entre différents CRISPRs sur le même génome. En analysant cette possibilité, nous avons trouvé que par exemple chez *S. solfataricus P2*, les CRISPRs NC_002754_3 et NC_002754_3 qui ont le même DR, possèdent des tronçons de spacer en commun (voir l'exemple dans la figure 7.5). En effet, ces spacers ne sont pas identiques à 100% ce qui suggère que même s'ils sont similaires et donc acquis de la même famille de phages, il ne s'agit pas du même spacer. Sachant que le spacer doit être identique à 100% à son homologue phagique pour être efficace, c'est alors un renforcement de la résistance phagique en réponse par exemple à l'existence de plusieurs variants de cet envahisseur ou à l'évolution rapide de son proto-spacer. C'est d'ailleurs le même genre d'observation que nous avons faite pour des spacers de CRISPRs distincts chez *L. monocytogenes* (voir paragraphe 5.2.2 page 78).

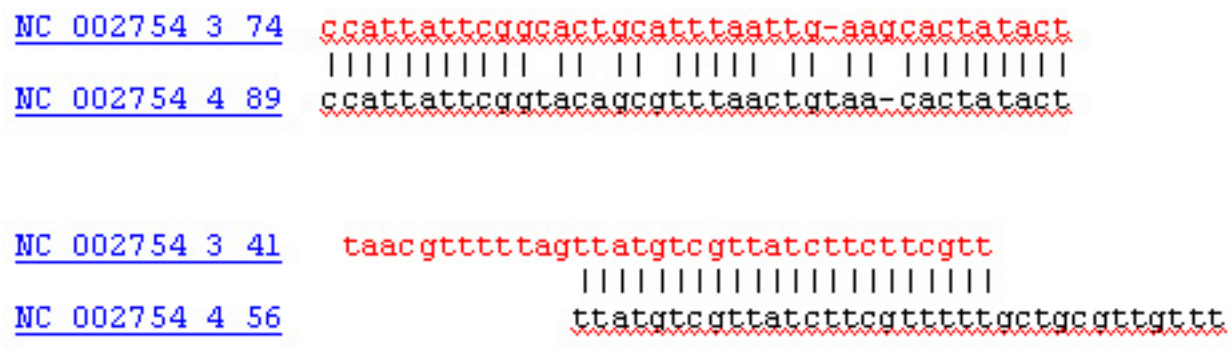


Fig. 7.5 – Similitudes entre spacers de CRISPRs différents chez *S. solfataricus P2* (Lillestøl, 2006)

7.2 Rôle du système CASS et ses applications

Le système CASS est composé d'une structure CRISPR et d'un ensemble de gènes *cas*. Ces deux entités ont des rôles complémentaires dans le fonctionnement du système. Les

gènes mettent ce système en marche, ils interviennent dans le déclenchement du processus de capture de nouveaux éléments et celui de défense contre les agressions étrangères. La structure CRISPR constitue le catalogue où sont stockées les séquences étrangères. Elle représente ainsi une sorte d'archive des rencontres que fait l'organisme. Cette archive est bien ordonnée puisque les spacers sont acquis de façon unidirectionnelle, elle peut alors être lue de droite à gauche (si on place le leader à gauche puisque c'est un promoteur). De plus, ce système garde non seulement la trace des agressions, mais les transmet également à sa descendance, traçant ainsi l'histoire évolutive de l'espèce.

Le système CASS, dans certains cas, n'est plus fonctionnel ou partiellement fonctionnel c'est à dire qu'il ne fait plus que perdre des motifs par exemple ou qu'il n'assure plus son rôle de système immunitaire ; ces deux cas surviennent essentiellement quand une partie ou tous les gènes *cas* sont absents chez l'organisme. En effet, cette situation a été observée plusieurs fois mais le contraire c'est à dire présence d'une cassette de gènes *cas* et absence du locus CRISPR n'a pas été observé. Parfois seuls des traces de la structure CRISPR sont observées : un leader et un DR ou uniquement un DR. Cependant, les applications liées à ce système peuvent concerner la structure CRISPR seule, les gènes *cas*, ou l'ensemble.

7.2.1 Utilité de la structure CRISPR en épidémiologie et phylogénie

La structure CRISPR s'est révélée très intéressante dans certains cas pour le typage bactérien et la reconstruction phylogénétique à l'intérieur d'une espèce (voir section 2.4). En effet, lorsqu'un organisme porte sur son génome un ou plusieurs CRISPRs dont la composition et le nombre de motifs sont assez polymorphes, la structure CRISPR peut être un bon outil pour différencier et classifier les isolats de cet organisme. Cependant, l'intérêt de l'étude dépend étroitement de la nature du CRISPR analysé.

Lorsque les spacers sont très polymorphes et que très peu d'entre eux sont conservés entre les souches, alors le typage ou la différenciation des souches est bien envisageable, mais la reconstruction phylogénétique ou le clustering devient un peu difficile. C'est d'ailleurs le cas de l'espèce *Y. pseudotuberculosis* (Pourcel, 2005) où 139 spacers ont été détectés dans un ensemble de 9 souches uniquement alors que pour son clone *Y. pestis*, espèce plus récente, uniquement 40 spacers ont été identifiés sur 200 souches. Un deuxième exemple est celui de *M. canettii* dont le CRISPR présente une très grande diversité des spacers (Fabre, 2004) par rapport à la soixantaine de spacers connus chez son clone récent *M. tuberculosis*.

Par opposition aux cas où la diversité des spacers est très grande, lorsque le CRISPR n'est pas assez polymorphe entre les différentes souches d'une espèce et surtout lorsqu'il est absent chez une large proportion des isolats, le typage CRISPR n'est pas très informatif ; c'est le cas par exemple d'une étude faite sur le CRISPR de *L. casei* (Diancourt, 2007). Dans cette étude, l'amplification PCR a échoué pour la moitié des souches testées et pour

l'autre moitié, la composition en spacers s'est révélée peu discriminante.

Dans la plupart des cas traités dans la littérature, le typage CRISPR n'est pas suffisant comme méthode de génotypage, il peut plutôt être utilisé comme méthode de typage complémentaire (*C. jejuni* (Schouls, 2003)) ou comme outil de clustering des souches (exemple de *M. tuberculosis*, *Thermotoga* (DeBoy, 2006)). Chez *Y. pestis* (Pourcel, 2005) par exemple, l'utilisation de la technique MLVA est indispensable pour un génotypage précis, mais le CRISPR y apporte des éléments supplémentaires sur la phylogénie de l'espèce.

Cependant, lorsqu'une espèce contient au moins un CRISPR, il est toujours intéressant d'essayer de mettre en place un typage CRISPR et analyser ses résultats surtout lorsque les méthodes de génotypage classiques échouent dans la différenciation des souches.

7.2.2 Utilité des gènes *cas*

Les gènes *cas* sont des gènes dont on ignore encore pour la plupart le rôle et le fonctionnement. A part l'analyse expérimentale de trois gènes (vérification de l'implication des deux gènes *cas5* et *cas7* dans la protection contre les phages (Barrangou, 2007) et l'analyse des propriétés biochimiques du *cas2* (Beloglazova, 2008)), l'étude des gènes accompagnant le CRISPR n'a pas encore permis d'élucider leurs propriétés et leur implication dans le fonctionnement du système CASS.

Les gènes *cas* peuvent être étudiés *in silico* pour des constructions phylogénétiques (Makarova, 2006). Quelques études se sont essentiellement intéressées au gène *Cas1* (Bolotin, 2005) ou aux quatre premiers gènes du noyau [1] (Godde et Bickerton, 2006) montrant l'implication du système CASS dans le transfert horizontal.

Cependant, comprendre le rôle des gènes et des protéines associés au CRISPR permettrait la découverte de nouveaux outils entrant dans la panoplie des outils de la biologie moléculaire afin d'assurer diverses manipulations de l'ADN microbien et pouvant être exploités sur le plan industriel ainsi que thérapeutique.

7.2.3 Utilité du système CASS complet

Le système CASS est probablement un système de régulation polyvalent jouant un rôle non seulement dans la résistance aux phages mais également dans la régulation des gènes, l'inhibition d'autres séquences que les virus et peut être dans d'autres fonctions qui ne sont pas encore connues. Le plus impressionnant dans ce système est qu'à travers la diversité du contenu de ses spacers et leur nombre, il est susceptible d'accomplir plusieurs fonctions en

[1] (voir paragraphe 2.2.4 page 28)

même temps : protéger contre divers virus, bloquer l'expression d'un ou de plusieurs gènes, empêcher certaines séquences étrangères de s'infiltrer dans le génome (des plasmides par exemple). Le fait que ce système soit régi par l'interférence ARN montre que ce mécanisme est extrêmement fondamental puisqu'il a été conservé depuis les organismes unicellulaires, les procaryotes jusqu'aux plantes, la drosophile et les mammifères. La compréhension du système permettrait de développer les psiRNA, pour en faire des outils puissants pour disséquer la fonction des gènes, réguler ou inhiber leur expression si elle est non désirée. Le système CASS pourrait être un outil qui révolutionnerait certaines pratiques du chercheur au même titre que la PCR a révolutionné en son temps l'étude de l'ADN.

Cependant, à ce jour il s'agit encore d'un système plein de mystères qui devront être élucidés et au sujet duquel plusieurs hypothèses ont été formulées et ont besoin d'être confirmées par des approches expérimentales.

7.3 Conclusion

La compréhension du CRISPR et le système dont il fait partie (CASS) est encore dans ses prémices. Cette structure curieuse présente une diversité importante au niveau de ses composants. Plusieurs aspects biologiques concernant son fonctionnement et son rôle demeurent encore inconnus et nécessitent une investigation surtout chez des organismes de l'environnement pour être élucidés.

Quatrième partie

Annexes

Annexe A

Comparaison des différentes techniques de typage

méthode de typage	typabilité	reproductibilité	pouvoir discriminant	facilité de mise en oeuvre	facilité d'interprétation	accessibilité de la méthode	coût
Profils de résistance aux antibiotiques	bonne	bonne	faible	excellente	excellente	excellente	faible
Sérotypage	variable	bonne	variable	bonne	bonne	variable	moyen
Lysotypage	variable	moyen	variable	faible	faible	excellente	moyen
MLEE	excellente	excellente	bon	bonne	excellente	variable	élevé
Tests biochimiques manuels	bonne	faible	faible	excellente	excellente	excellente	faible
Tests biochimiques automatisés	bonne	bonne	faible	bonne	bonne	variable	moyen
PFGE	excellente	bonne	excellent	bonne	bonne	variable	élevé
RAPD / AP-PCR	excellente	faible	faible	bonne	faible	bonne	moyen
AFLP	excellente	bonne	excellent	bonne	moyen	faible	élevé
RFLP	excellente	variable	variable	bonne	moyen	variable	moyen
Ribotypage automatisé	excellente	excellente	bon	bonne	bonne	variable	élevé
MLST (séquençage)	optimale	excellente	excellent	bonne	excellente	faible	élevé
SNPs (séquençage)	excellente	excellente	faible	bonne	excellente	faible	élevé
MLVA	excellente	excellente	excellent	excellente	excellente	bonne	faible

TAB. A.1 – Comparaison des différentes techniques de typage (adapté de (van Belkum, 2001))

Annexe B

Obtention du leader par alignement des séquences flanquantes

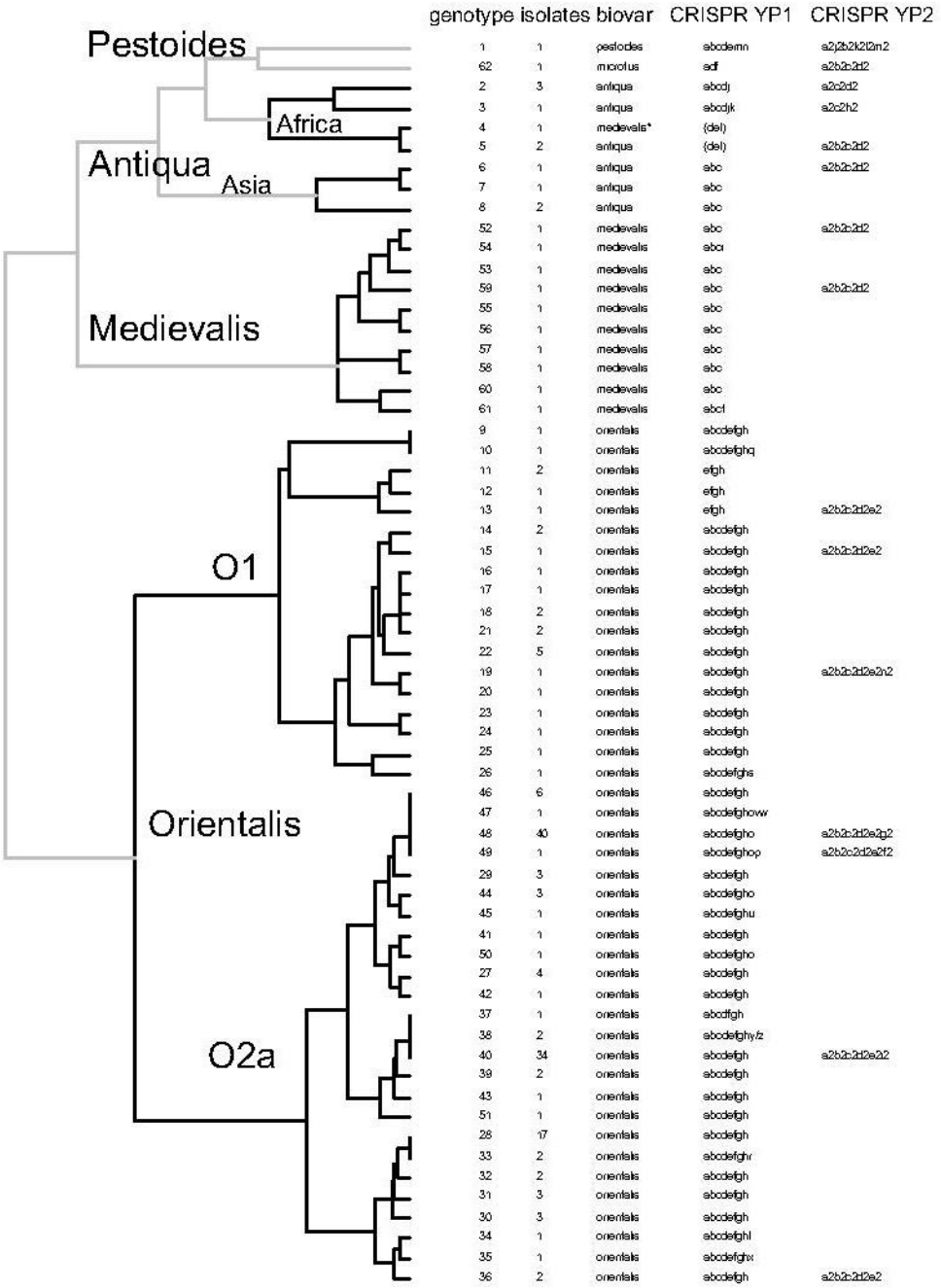
Utilisation de Flankalign pour identifier le leader des dix CRISPRs de *Aquifex aeolicus VF5*

```
NC_000918_1           -TAGCAATAAAAACTTAGA-CGAATTTATTATAATT-ATGT-TTAAGGGCTACAAAGCCC
NC_000918_3_RevComp   -ATTCCACATTAAATTAAA-AATATGTGATATACTT-TTAA-AAAGGGGCTAAAAAGCCC
NC_000918_9_RevComp   -GACACTTTAAACATTAAA-GATATGTGATATACTT-TTAA-AAAGGGGCTAAAAAGCCC
NC_000918_7           ---CAGCCTTTAAGTTTTC-CATTTAATGTATAATA-CTACCTGAAAGGGCTTCTAGCCC
NC_000918_8           ATATTGACCCGAAGTTTTT-ACTTTA--GTAAACTT-TTAATTGGAGGGCTACAGAGCCC
NC_000918_6           -AGGGCATAATAATTTATATGGTATAATTTTGAATAGCTCCTTAGAGGGCTACAATGCCT
NC_000918_4_RevComp   -GACAGACCATATATGTTT---TATGA-TTATAATGTTTAAATAAGGCCCTGCTGAGGGC
NC_000918_5_RevComp   -CATTTACGTAAAGAGAGAAAAGAAGTAGTATAATA---ATTTAAAGCCCTGCTGAGGGC
NC_000918_2_RevComp   --ATACTCCCGACCCGACCGATAATGTGATATAATC--TTAAAAGAGCTCCTTGGAGCAC
NC_000918_10          -AAAGCTTGGAAAAAACTTGAAAATGTTTTAAAATA-CTAAACAAAGCCCTTTTTGGGCT
                                 *                 *  * *                     *

NC_000918_1           CTCCTGGCTCCTTGGCAATTGAATA-TGGTTTATGGGTCTCAATGAGTGATGCTTTGATG
NC_000918_3_RevComp   CTCCTGGCTCCTTGGCAATTGAATA-TGGTTTATGGGTCTCAATGAGTGATGCTTTGATG
NC_000918_9_RevComp   CAGCCGGCTCCTTGGCAACTGAATA-GGGGTGGTGGGTCTTGATAAGTGATGTTTTGATG
NC_000918_7           CACCTGGCTCCTTGGCAATTGAATA-GGGTAGGTGGGTCCTAATAAGTGATGCTTTGATG
NC_000918_8           TACTTTGCTCCTTGGCAACTGAATA-GGGTAGGTGGGTCCTAATAAGTGATGTTTCGATG
NC_000918_6           C--TTGGCTCCTTGGCAATTGAATA-AAGCGGATTGGTC-TAATAAGTGATGCTTTGATT
NC_000918_4_RevComp   TTTCGGGCTCCTTGGCAATTGAATAAGGGTGGATAGGTCTTAGTGAGTGATGCTTTGATG
NC_000918_5_RevComp   TTTCGGGCTCCTTGGCAAGTGAATA-TGGTTTATGGGTCTTAACGAGTGATGTTTTGAGG
NC_000918_2_RevComp   TCTCAGGCTCCTTGGCAAGTGAATA-GGGTAGATTGGTCTTAATGAGTGATGCTTTGATA
NC_000918_10          ---CCGGCTCCTTGGCAATTGAATA-TGGTAAGTGAACCTTAATAAGTGATGCTTTGATG
                            ************ ******   *    *    *      ******* ** **
```

```
NC_000918_1            ACGTGATGAAATGAAATCATGATGCAATAGGGCTGAAGGGATTGAAAATCAAGGGTAAAA
NC_000918_3_RevComp    ACGTGATGAAATGAAATCATGATGCAATAGGGCTGAAGGGATTGAAAATCAAGGGTAAAA
NC_000918_9_RevComp    GTGTGATGAAATGAAATCATGATGCGATAAAGTTGAAAGGATTGAAAATCAAGGGTAAAA
NC_000918_7            ATGTGATGAAATGAAATCATGATGCAATAGGGTTAAAGGGATTGAAAATCAAGGATGGAA
NC_000918_8            GTGTGATGTATTGAAATCATGATGCAACATAGCTAAAAAGATTGAAAATCAAGGATGGAA
NC_000918_6            GTGTGATGGATTGATATCATGATGCAATAAACTTTAATAGACTGAAAATCAAAGATAAAA
NC_000918_4_RevComp    ATGTGATGAAATGAAATCATGATGCAATAGGGTTAAAGGGATTGAAAATCAAGGATGGAA
NC_000918_5_RevComp    GCGTGATGAGATGAAATCATGATGCAATAAGCTTTAAGGGATTGAAAATCAAGGATGGAA
NC_000918_2_RevComp    GCGTGATGAAATGAAATCGTGATACGACATAGTTAAATGGTTTGAAAATCAAGGGTGTAA
NC_000918_10           GTGTGATGTATTGAAATCATGATGCAACATAGCTAAAAAGATTGAAAATCAAGGGCATAA
                         ******   *** *** **** * * *    * **  *  ********** *    **

NC_000918_1            ATTTT-ATAAAGGTATACTCAAGTTTGCAGAACCTCAGGTTAGCGGAAATTAACGTTCGT
NC_000918_3_RevComp    ATTTT-ATAAAGGTATACTCAAGTTTGCAGAACCTCAGGTTAGCGGAAATTAACGTTCGT
NC_000918_9_RevComp    ATTTT-ATAAAGGTATACTCAAGCTTGCAGAACCTTGGGTTAGCGGAAATTAACATTCGT
NC_000918_7            ATTTT-ATAAAGGTATACTCAAGTTTGCAGAACTTCAGGTTAGCGGAAATTAACATTCAT
NC_000918_8            ATTTT-ATAAAGGTATACTCAAGTTTGCAGAACCTCGGGTTAGCGGAAATTAACATTCGT
NC_000918_6            ATTTT-ACAAATGTATGCTCAAGTTTGCAGAACTTGGGGTTAGCGGAAATTAACATTCGT
NC_000918_4_RevComp    ATTTT-ATAAAGGTATACTCAAGTTTGCAGAACTTCAGGTTAGCGGAAATTAACATTCAT
NC_000918_5_RevComp    ATTTT-ATAAAGGTATGCTCAAGTTTGCGGAACCTCGGGTTAGCGGAAATTAACATTCGT
NC_000918_2_RevComp    ATTTT-ATAAAGGCATACTCAACTTTGCAGAACCTCGGGTTAGCGGAAATTAACATTCAT
NC_000918_10           ATTTTTACAATGATATACTCAAGTTTGCAGAACCGCGGGTTAGCGGAAATTAACATTCGT
                       ***** * **    ** *****  **** ****    ********************* *

NC_000918_1            TTCGTGAGGGTCAGAAGTTTAATAAGTTATTGATAAAGCGTTATTTTTT-AGACACAGGC
NC_000918_3_RevComp    TTCGTGAGGGTCAGAAGTTTAATAAGTTATTGATAAAGCGTTATTTTTT-AGACACAGGC
NC_000918_9_RevComp    TTCGTGAGGGTCGGGTGTCTAATAAGTTATTGATAGTACGTTATTTTTT-AGACACAGGC
NC_000918_7            TTCGTGAAGGTATTGTGTCTAATAAGTTGTTGATAAAGCGTTATTTTTTTAGACACAGAC
NC_000918_8            TTCGTGAAGGTTAGGTGTCTAATAAGTTATTGATAAAGCGTTATTTTTT-GGACACAGGC
NC_000918_6            TTCGTGAAGGTTAGGTGTCTAATAAGTTATTGATAAAGCGTTATTTTTC-AGACATAGGC
NC_000918_4_RevComp    TTCGTGAAGGTTAGAAGTTTAATAAGTTGCTGATAAAACGTTATTTTTT-AGACACAGGC
NC_000918_5_RevComp    TTCGGGAGGATCGGGTGTCTAATAAGTTGTTGATAAAGCGTTATTTTTT-AGACACAGGC
NC_000918_2_RevComp    TTCGTGAAGGTCGGGTGTCTAATAAGTTGTTGATAAAGCGTTATTTTTT-AGACATAGGC
NC_000918_10           TTCGTGAAGGTTAGGTGTCTAATAAGTTATTGATAAAGCGTTATTTTTT-AGGCACAGGC
                       **** ** * *     ** *********  *****   **********   * ** ** *

NC_000918_1            GTGCAGG
NC_000918_3_RevComp    GTGCAGG
NC_000918_9_RevComp    ACGCAGG
NC_000918_7            ACGCAGG
NC_000918_8            ACACAGG
NC_000918_6            GCGCAGG
NC_000918_4_RevComp    ACGCAGG
NC_000918_5_RevComp    ACGCAGG
NC_000918_2_RevComp    ACGCAGG
NC_000918_10           GTGCAGG
                          ****
```

Annexe C

Figures



genotype	isolates	biovar	CRISPR YP1	CRISPR YP2
1	1	pestoides	abcdemn	a2j2b2k2l2n2
62	1	microtus	adf	a2b2c2d2
2	3	antiqua	abcdj	a2c2d2
3	1	antiqua	abcdjk	a2c2h2
4	1	medevalis*	(del)	
5	2	antiqua	(del)	a2b2c2d2
6	1	antiqua	abc	a2b2c2d2
7	1	antiqua	abc	
8	2	antiqua	abc	
52	1	medevalis	abc	a2b2c2d2
54	1	medevalis	abci	
53	1	medevalis	abc	
59	1	medevalis	abc	a2b2c2d2
55	1	medevalis	abc	
56	1	medevalis	abc	
57	1	medevalis	abc	
58	1	medevalis	abc	
60	1	medevalis	abc	
61	1	medevalis	abcf	
9	1	orientalis	abcdefgh	
10	1	orientalis	abcdefghq	
11	2	orientalis	efgh	
12	1	orientalis	efgh	
13	1	orientalis	efgh	a2b2c2d2e2
14	2	orientalis	abcdefgh	
15	1	orientalis	abcdefgh	a2b2c2d2e2
16	1	orientalis	abcdefgh	
17	1	orientalis	abcdefgh	
18	2	orientalis	abcdefgh	
21	2	orientalis	abcdefgh	
22	5	orientalis	abcdefgh	
19	1	orientalis	abcdefgh	a2b2c2d2e2n2
20	1	orientalis	abcdefgh	
23	1	orientalis	abcdefgh	
24	1	orientalis	abcdefgh	
25	1	orientalis	abcdefgh	
26	1	orientalis	abcdefghs	
46	6	orientalis	abcdefgh	
47	1	orientalis	abcdefghovw	
48	40	orientalis	abcdefgho	a2b2c2d2e2g2
49	1	orientalis	abcdefghop	a2b2c2d2e2f2
29	3	orientalis	abcdefgh	
44	3	orientalis	abcdefgho	
45	1	orientalis	abcdefghu	
41	1	orientalis	abcdefgh	
50	1	orientalis	abcdefgho	
27	4	orientalis	abcdefgh	
42	1	orientalis	abcdefgh	
37	1	orientalis	abcdfgh	
38	2	orientalis	abcdefghyfz	
40	34	orientalis	abcdefgh	a2b2c2d2e2a2
39	2	orientalis	abcdefgh	
43	1	orientalis	abcdefgh	
51	1	orientalis	abcdefgh	
28	17	orientalis	abcdefgh	
33	2	orientalis	abcdefghr	
32	2	orientalis	abcdefgh	
31	3	orientalis	abcdefgh	
30	3	orientalis	abcdefgh	
34	1	orientalis	abcdefghl	
35	1	orientalis	abcdefghx	
36	2	orientalis	abcdefgh	a2b2c2d2e2

Fig. C.2 – Dispersion du nombre de motifs par classe taxonomique.

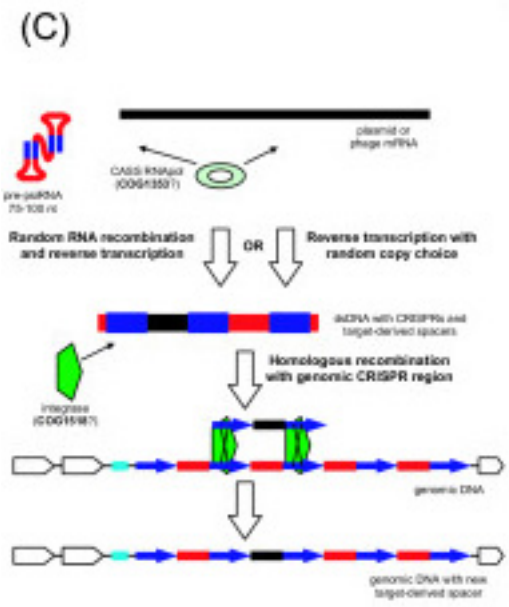


Fig. C.3 – Modèle d'acquisition des spacers selon Makarova *et col.* (Makarova, 2006)

Annexe D

Génomes contenant des CRISPRs

Liste des génomes contenant au moins un CRISPR, nombre de CRISPRs, de motifs et taille du CRISPR le plus long :

Génome	taxon	nbr CRISPRs	nbr motifs	CRISPR le plus long
269800,Thermobifida fusca YX,chromosome circular,NC_007333	Actinomycetales	14	253	65
326424,Frankia alni ACN14a,chromosome circular,NC_008278	Actinomycetales	7	155	93
340322,Corynebacterium glutamicum R,chromosome circular,NC_009342	Actinomycetales	1	111	111
227882,Streptomyces avermitilis MA-4680,chromosome circular,NC_003155	Actinomycetales	3	103	59
298653,Frankia sp. EAN1pec,chromosome circular,NC_009921	Actinomycetales	5	102	45
106370,Frankia sp. CcI3,chromosome circular,NC_007777	Actinomycetales	4	75	38
504474,Corynebacterium urealyticum DSM 7109,chromosome circular,NC_010545	Actinomycetales	2	71	69
306537,Corynebacterium jeikeium K411,chromosome circular,NC_007164	Actinomycetales	1	60	60
410289,Mycobacterium bovis BCG str. Pasteur 1173P2,chromosome circular,NC_008769	Actinomycetales	2	47	29
247156,Nocardia farcinica IFM 10152,chromosome circular,NC_006361	Actinomycetales	3	45	26
83332,Mycobacterium tuberculosis H37Rv,chromosome circular,NC_000962	Actinomycetales	2	40	23
233413,Mycobacterium bovis AF2122/97,chromosome circular,NC_002945	Actinomycetales	2	40	24
405948,Saccharopolyspora erythraea NRRL 2338,chromosome circular,NC_009142	Actinomycetales	4	40	19
419947,Mycobacterium tuberculosis H37Ra,chromosome circular,NC_009525	Actinomycetales	2	40	23
369723,Salinispora tropica CNB-440,chromosome circular,NC_009380	Actinomycetales	3	35	24
336982,Mycobacterium tuberculosis F11,chromosome circular,NC_009565	Actinomycetales	2	35	18
83331,Mycobacterium tuberculosis CDC1551,chromosome circular,NC_002755	Actinomycetales	2	33	18
257309,Corynebacterium diphtheriae NCTC 13129,chromosome circular,NC_002935	Actinomycetales	2	33	26
351607,Acidothermus cellulolyticus 11B,chromosome circular,NC_008578	Actinomycetales	1	23	23
196164,Corynebacterium efficiens YS-314,chromosome circular,NC_004369	Actinomycetales	1	21	21
243243,Mycobacterium avium 104,chromosome circular,NC_008595	Actinomycetales	1	12	12
100226,Streptomyces coelicolor A3(2),plasmid SCP1,NC_003903	Actinomycetales	1	4	4
402882,Shewanella baltica OS185,chromosome circular,NC_009665	Alteromonadales	1	83	83
319224,Shewanella putrefaciens CN-32,chromosome circular,NC_009438	Alteromonadales	1	81	81
357804,Psychromonas ingrahamii 37,chromosome circular,NC_008709	Alteromonadales	1	45	45

360107,Campylobacter hominis ATCC BAA-381,chromosome circular,NC_009714	Campylobacterales	1	71	71
360106,Campylobacter fetus subsp. fetus 82-40,chromosome circular,NC_008599	Campylobacterales	2	47	26
273121,Wolinella succinogenes DSM 1740,chromosome circular,NC_005090	Campylobacterales	2	30	21
360105,Campylobacter curvus 525.92,chromosome circular,NC_009715	Campylobacterales	2	11	8
407148,Campylobacter jejuni subsp. jejuni 81116,chromosome circular,NC_009839	Campylobacterales	1	7	7
360109,Campylobacter jejuni subsp. doylei 269.97,chromosome circular,NC_009707	Campylobacterales	1	5	5
192222,Campylobacter jejuni subsp. jejuni NCTC 11168,chromosome circular,NC_002163	Campylobacterales	1	4	4
195099,Campylobacter jejuni RM1221,chromosome circular,NC_003912	Campylobacterales	1	3	3
246195,Dichelobacter nodosus VCS1703A,chromosome circular,NC_009446	Cardiobacteriales	1	7	7
290317,Chlorobium phaeobacteroides DSM 266,chromosome circular,NC_008639	Chlorobiales	5	174	114
340177,Chlorobium chlorochromatii CaD3,chromosome circular,NC_007514	Chlorobiales	2	83	62
194439,Chlorobium tepidum TLS,chromosome circular,NC_002932	Chlorobiales	2	62	44
319225,Pelodictyon luteolum DSM 273,chromosome circular,NC_007512	Chlorobiales	1	39	39
290318,Prosthecochloris vibrioformis DSM 265,chromosome circular,NC_009337	Chlorobiales	2	15	12
324602,Chloroflexus aurantiacus J-10-fl,chromosome circular,NC_010175	Chloroflexaceae	3	539	276
357808,Roseiflexus sp. RS-1,chromosome circular,NC_009523	Chloroflexaceae	14	505	124
383372,Roseiflexus castenholzii DSM 13941,chromosome circular,NC_009767	Chloroflexaceae	7	490	213
187272,Alkalilimnicola ehrlichei MLHE-1,chromosome circular,NC_008340	Chromatiales	1	92	92
349124,Halorhodospira halophila SL1,chromosome circular,NC_008789	Chromatiales	2	13	10
323261,Nitrosococcus oceani ATCC 19707,chromosome circular,NC_007484	Chromatiales	1	6	6
203119,Clostridium thermocellum ATCC 27405,chromosome circular,NC_009012	Clostridiales	5	436	169
370438,Pelotomaculum thermopropionicum SI,chromosome circular,NC_009454	Clostridiales	5	423	240
477974,Candidatus Desulforudis audaxviator MP104C,chromosome circular,NC_010424	Clostridiales	4	207	90

138119,Desulfitobacterium hafniense Y51,chromosome circular,NC_007907	Clostridiales	5	150	47
246194,Carboxydothermus hydrogenoformans Z-2901,chromosome circular,NC_007503	Clostridiales	2	141	83
272563,Clostridium difficile 630,chromosome circular,NC_009089	Clostridiales	12	107	19
386415,Clostridium novyi NT,chromosome circular,NC_008593	Clostridiales	2	97	77
349161,Desulfotomaculum reducens MI-1,chromosome circular,NC_009253	Clostridiales	5	95	52
293826,Alkaliphilus metalliredigens QYMF,chromosome circular,NC_009633	Clostridiales	2	78	61
431943,Clostridium kluyveri DSM 555,chromosome circular,NC_009706	Clostridiales	5	71	24
212717,Clostridium tetani E88,chromosome circular,NC_004557	Clostridiales	8	66	33
289380,Clostridium perfringens SM101,chromosome circular,NC_008262	Clostridiales	1	60	60
498761,Heliobacterium modesticaldum Ice1,chromosome circular,NC_010337	Clostridiales	3	35	17
498214,Clostridium botulinum A3 str. Loch Maree,plasmid pCLK,NC_010418	Clostridiales	5	33	13
441772,Clostridium botulinum F str. Langeland,chromosome circular,NC_009699	Clostridiales	12	31	4
498213,Clostridium botulinum B1 str. Okra,plasmid pCLD,NC_010379	Clostridiales	4	29	13
441770,Clostridium botulinum A str. ATCC 19397 "Hall",chromosome circular,NC_009697	Clostridiales	8	22	5
413999,Clostridium botulinum A str. ATCC 3502,chromosome circular,NC_009495	Clostridiales	8	21	5
498213,Clostridium botulinum B1 str. Okra,chromosome circular,NC_010516	Clostridiales	5	19	6
334413,Finegoldia magna ATCC 29328,chromosome circular,NC_010376	Clostridiales	1	14	14
334413,Finegoldia magna ATCC 29328,plasmid pFMC,NC_010371	Clostridiales	3	13	6
195103,Clostridium perfringens ATCC 13124,chromosome circular,NC_008261	Clostridiales	2	6	3
43989,Cyanothece sp. ATCC 51142,chromosome circular,NC_010546	Cyanothece	5	20	10
255470,Dehalococcoides sp. CBDB1,chromosome circular,NC_007356	Dehalococcoides	1	18	18
216389,Dehalococcoides sp. BAV1,chromosome circular,NC_009455	Dehalococcoides	1	1	1
319795,Deinococcus geothermalis DSM 11300,plasmid 1,NC_008010	Deinococcales	1	66	66

319795,Deinococcus geothermalis DSM 11300,chromosome circular,NC_008025	Deinococcales	5	53	24
96561,Candidatus Desulfococcus oleovorans Hxd3,chromosome circular,NC_009943	Desulfobacterales	2	56	40
177439,Desulfotalea psychrophila LSv54,chromosome circular,NC_006138	Desulfobacterales	1	13	13
391774,Desulfovibrio vulgaris subsp. vulgaris DP4,plasmid pDVUL01,NC_008741	Desulfovibrionales	1	44	44
882,Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough,plasmid megaplasmid,NC_005863	Desulfovibrionales	1	27	27
207559,Desulfovibrio desulfuricans G20,chromosome circular,NC_007519	Desulfovibrionales	1	19	19
399550,Staphylothermus marinus F1,chromosome circular,NC_009033	Desulfurococcales	11	108	16
415426,Hyperthermus butylicus DSM 5456,chromosome circular,NC_008818	Desulfurococcales	2	94	48
272557,Aeropyrum pernix K1,chromosome circular,NC_000854	Desulfurococcales	3	85	41
453591,Ignicoccus hospitalis KIN4/I,chromosome circular,NC_009776	Desulfurococcales	12	85	23
243231,Geobacter sulfurreducens PCA,chromosome circular,NC_002939	Desulfuromonadales	2	179	142
338963,Pelobacter carbinolicus DSM 2380,chromosome circular,NC_007498	Desulfuromonadales	1	111	111
338966,Pelobacter propionicus DSM 2379,chromosome circular,NC_008609	Desulfuromonadales	1	43	43
351605,Geobacter uraniumreducens Rf4,chromosome circular,NC_009483	Desulfuromonadales	2	31	29
269799,Geobacter metallireducens GS-15,chromosome circular,NC_007517	Desulfuromonadales	1	12	12
338966,Pelobacter propionicus DSM 2379,plasmid pPRO1,NC_008607	Desulfuromonadales	1	3	3
502800,Yersinia pseudotuberculosis YPIII,chromosome circular,NC_010465	Enterobacteriales	3	77	46
290339,Enterobacter sakazakii ATCC BAA-894,chromosome circular,NC_009778	Enterobacteriales	4	69	30
243265,Photorhabdus luminescens subsp. laumondii TTO1,chromosome circular,NC_005126	Enterobacteriales	6	66	24
99287,Salmonella typhimurium LT2,chromosome circular,NC_003197	Enterobacteriales	3	55	32
349747,Yersinia pseudotuberculosis IP 31758,chromosome circular,NC_009708	Enterobacteriales	2	50	26
218491,Erwinia carotovora subsp. atroseptica SCRI1043,chromosome circular,NC_004547	Enterobacteriales	3	41	28
331111,Escherichia coli E24377A,chromosome circular,NC_009801	Enterobacteriales	3	39	25

399742,Enterobacter sp. 638,chromosome circular,NC_009436	Enterobacteriales	2	37	20
331112,Escherichia coli HS,chromosome circular,NC_009800	Enterobacteriales	3	32	19
273123,Yersinia pseudotuberculosis IP 32953,chromosome circular,NC_006155	Enterobacteriales	3	22	16
439855,Escherichia coli SECEC SMS-3-5,chromosome circular,NC_010498	Enterobacteriales	1	21	21
316407,Escherichia coli W3110,chromosome circular,AC_000091	Enterobacteriales	2	18	12
511145,Escherichia coli str. K-12 substr. MG1655,chromosome circular,NC_000913	Enterobacteriales	2	18	12
316385,Escherichia coli str. K-12 substr. DH10B,chromosome circular,NC_010473	Enterobacteriales	2	18	12
386656,Yersinia pestis Pestoides F,chromosome circular,NC_009381	Enterobacteriales	3	17	6
214092,Yersinia pestis CO92,chromosome circular,NC_003143	Enterobacteriales	3	16	8
364106,Escherichia coli UTI89,chromosome circular,NC_007946	Enterobacteriales	2	14	8
360102,Yersinia pestis Antiqua,chromosome circular,NC_008150	Enterobacteriales	3	12	6
321314,Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67,chromosome circular,NC_006905	Enterobacteriales	2	11	7
405955,Escherichia coli APEC O1,chromosome circular,NC_008563	Enterobacteriales	2	11	6
187410,Yersinia pestis KIM,chromosome circular,NC_004088	Enterobacteriales	3	10	4
295319,Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150,chromosome circular,NC_006511	Enterobacteriales	2	8	5
377628,Yersinia pestis Nepal516,chromosome circular,NC_008149	Enterobacteriales	3	8	4

229193,Yersinia pestis biovar Microtus str. 91001,chromosome circular,NC_005810	Enterobacteriales	2	7	4
209261,Salmonella enterica subsp. enterica serovar Typhi Ty2,chromosome circular,NC_004631	Enterobacteriales	1	6	6
220341,Salmonella enterica subsp. enterica serovar Typhi str. CT18,chromosome circular,NC_003198	Enterobacteriales	1	5	5
155864,Escherichia coli O157 :H7,chromosome circular,NC_002655	Enterobacteriales	2	4	3
386585,Escherichia coli O157 :H7 str. Sakai,chromosome circular,NC_002695	Enterobacteriales	2	4	3
300269,Shigella sonnei Ss046,chromosome circular,NC_007384	Enterobacteriales	2	4	2
290338,Citrobacter koseri ATCC BAA-895,chromosome circular,NC_009792	Enterobacteriales	1	3	3
198214,Shigella flexneri 2a str. 301,chromosome circular,NC_004337	Enterobacteriales	1	2	2
198215,Shigella flexneri 2a str. 2457T,chromosome circular,NC_004741	Enterobacteriales	1	2	2
362663,Escherichia coli 536,chromosome circular,NC_008253	Enterobacteriales	1	2	2
373384,Shigella flexneri 5 str. 8401,chromosome circular,NC_008258	Enterobacteriales	1	2	2
300268,Shigella boydii Sb227,chromosome circular,NC_007613	Enterobacteriales	1	1	1
402612,Flavobacterium psychrophilum JIP02/86,chromosome circular,NC_009613	Flavobacteriales	1	20	20
190304,Fusobacterium nucleatum subsp. nucleatum ATCC 25586,chromosome circular,NC_003454	Fusobacteriaceae	1	15	15
272569,Haloarcula marismortui ATCC 43049,plasmid pNG400,NC_006392	Halobacteriales	3	80	51
272569,Haloarcula marismortui ATCC 43049,plasmid pNG300,NC_006391	Halobacteriales	1	47	47
348780,Natronomonas pharaonis DSM 2160,chromosome circular,NC_007426	Halobacteriales	2	12	8
348780,Natronomonas pharaonis DSM 2160,plasmid PL23,NC_007428	Halobacteriales	1	7	7
362976,Haloquadratum walsbyi DSM 16790,chromosome circular,NC_008212	Halobacteriales	2	5	3
348780,Natronomonas pharaonis DSM 2160,plasmid PL131,NC_007427	Halobacteriales	1	2	2
390333,Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842,chromosome circular,NC_008054	Lactobacillaceae	1	40	40
272621,Lactobacillus acidophilus NCFM,chromosome circular,NC_006814	Lactobacillaceae	1	32	32
362948,Lactobacillus salivarius subsp. salivarius,chromosome circular,NC_007929	Lactobacillaceae	1	28	28
321967,Lactobacillus casei ATCC 334,chromosome circular,NC_008526	Lactobacillaceae	1	20	20

321956,Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365,chromosome circular,NC_008529	Lactobacillaceae	1	20	20
387344,Lactobacillus brevis ATCC 367,chromosome circular,NC_008497	Lactobacillaceae	2	9	5
297245,Legionella pneumophila str. Lens,chromosome circular,NC_006369	Legionellales	2	64	52
297245,Legionella pneumophila str. Lens,plasmid pLPL,NC_006366	Legionellales	1	53	53
297246,Legionella pneumophila str. Paris,chromosome circular,NC_006368	Legionellales	1	33	33
267671,Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130,chromosome I,NC_005823	Leptospiraceae	1	3	3
272626,Listeria innocua Clip11262,chromosome circular,NC_003212	Listeriaceae	1	10	10
169963,Listeria monocytogenes EGD-e,chromosome circular,NC_003210	Listeriaceae	1	4	4
187420,Methanothermobacter thermautotrophicus str. Delta H,chromosome circular,NC_000916	Methanobacteriales	2	169	123
339860,Methanosphaera stadtmanae DSM 3091,chromosome circular,NC_007681	Methanobacteriales	3	121	61
420247,Methanobrevibacter smithii ATCC 35061,chromosome circular,NC_009515	Methanobacteriales	1	43	43
243232,Methanocaldococcus jannaschii DSM 2661,chromosome circular,NC_000909	Methanococcales	20	178	25
406327,Methanococcus vannielii SB,chromosome circular,NC_009634	Methanococcales	2	103	100
402880,Methanococcus maripaludis C5,chromosome circular,NC_009135	Methanococcales	1	27	27
419665,Methanococcus aeolicus Nankai-3,chromosome circular,NC_009635	Methanococcales	1	18	18
323259,Methanospirillum hungatei JF-1,chromosome circular,NC_007796	Methanomicrobiales	7	264	79
410358,Methanocorpusculum labreanum Z,chromosome circular,NC_008942	Methanomicrobiales	1	18	18
368407,Methanoculleus marisnigri JR1,chromosome circular,NC_009051	Methanomicrobiales	1	6	6
190192,Methanopyrus kandleri AV19,chromosome circular,NC_003551	Methanopyrales	5	36	11
349307,Methanosaeta thermophila PT,chromosome circular,NC_008553	Methanosarcinales	2	155	96
192952,Methanosarcina mazei Go1,chromosome circular,NC_003901	Methanosarcinales	4	128	80
269797,Methanosarcina barkeri str. fusaro,chromosome 1,NC_007355	Methanosarcinales	5	100	50

259564,Methanococcoides burtonii DSM 6242,chromosome circular,NC_007955	Methanosarcinales	2	84	64
188937,Methanosarcina acetivorans C2A,chromosome circular,NC_003552	Methanosarcinales	5	69	30
243233,Methylococcus capsulatus str. Bath,chromosome circular,NC_002977	Methylococcales	2	69	62
265072,Methylobacillus flagellatus KT,chromosome circular,NC_007947	Methylophilales	2	94	92
449447,Microcystis aeruginosa NIES-843,chromosome circular,NC_010296	Microcystis	3	172	112
233150,Mycoplasma gallisepticum R,chromosome circular,NC_004829	Mycoplasmataceae	2	72	71
267748,Mycoplasma mobile 163K,chromosome circular,NC_006908	Mycoplasmataceae	1	62	62
262723,Mycoplasma synoviae 53,chromosome circular,NC_007294	Mycoplasmataceae	1	11	11
246197,Myxococcus xanthus DK 1622,chromosome circular,NC_008095	Myxococcales	4	139	57
243365,Chromobacterium violaceum ATCC 12472,chromosome circular,NC_005085	Neisseriales	4	92	26
122587,Neisseria meningitidis Z2491,chromosome circular,NC_003116	Neisseriales	2	17	16
242231,Neisseria gonorrhoeae FA 1090,chromosome circular,NC_002946	Neisseriales	3	3	1
122586,Neisseria meningitidis MC58,chromosome circular,NC_003112	Neisseriales	2	3	2
272831,Neisseria meningitidis FAM18,chromosome circular,NC_008767	Neisseriales	2	2	1
335283,Nitrosomonas eutropha C91,chromosome circular,NC_008344	Nitrosomonadales	3	23	11
228410,Nitrosomonas europaea ATCC 19718,chromosome circular,NC_004757	Nitrosomonadales	2	16	8
240292,Anabaena variabilis ATCC 29413,chromosome circular,NC_007413	Nostocaceae	11	184	42
103690,Nostoc sp. PCC 7120,chromosome circular,NC_003272	Nostocaceae	13	106	28
400668,Marinomonas sp. MWYL1,chromosome circular,NC_009654	Oceanospirillales	2	148	108
349521,Hahella chejuensis KCTC 2396,chromosome circular,NC_007645	Oceanospirillales	2	87	55
290398,Chromohalobacter salexigens DSM 3043,chromosome circular,NC_007963	Oceanospirillales	2	52	31
272843,Pasteurella multocida subsp. multocida str. Pm70,chromosome circular,NC_002663	Pasteurellales	4	73	35
221988,Mannheimia succiniciproducens MBEL55E,chromosome circular,NC_006300	Pasteurellales	3	59	32
434271,Actinobacillus pleuropneumoniae serovar 3 str. JL03,chromosome circular,NC_010278	Pasteurellales	1	34	34

416269,Actinobacillus pleuropneumoniae L20,chromosome circular,NC_009053	Pasteurellales	1	25	25
339671,Actinobacillus succinogenes 130Z,chromosome circular,NC_009655	Pasteurellales	2	17	11
62977,Acinetobacter sp. ADP1,chromosome circular,NC_005966	Pseudomonadales	3	117	90
509173,Acinetobacter baumannii AYE,chromosome circular,NC_010410	Pseudomonadales	1	59	59
399739,Pseudomonas mendocina ymp,chromosome circular,NC_009439	Pseudomonadales	2	54	38
379731,Pseudomonas stutzeri A1501,chromosome circular,NC_009434	Pseudomonadales	1	42	42
208963,Pseudomonas aeruginosa UCBPP-PA14,chromosome circular,NC_008463	Pseudomonadales	2	35	21
381754,Pseudomonas aeruginosa PA7,chromosome circular,NC_009656	Pseudomonadales	2	9	5
349106,Psychrobacter sp. PRwf-1,chromosome circular,NC_009524	Pseudomonadales	1	4	4
316057,Rhodopseudomonas palustris BisB5,chromosome circular,NC_007958	Rhizobiales	1	48	48
402881,Parvibaculum lavamentivorans DS-1,chromosome circular,NC_009719	Rhizobiales	1	48	48
426117,Methylobacterium sp. 4-46,chromosome circular,NC_010511	Rhizobiales	3	30	16
316056,Rhodopseudomonas palustris BisB18,chromosome circular,NC_007925	Rhizobiales	1	19	19
323098,Nitrobacter winogradskyi Nb-255,chromosome circular,NC_007406	Rhizobiales	1	17	17
318586,Paracoccus denitrificans PD1222,chromosome 1,NC_008686	Rhodobacterales	1	6	6
76114,Azoarcus sp. EbN1,chromosome circular,NC_006513	Rhodocyclales	3	115	75
76114,Azoarcus sp. EbN1,plasmid 1,NC_006823	Rhodocyclales	2	22	14
269796,Rhodospirillum rubrum ATCC 11170,chromosome circular,NC_007643	Rhodospirillales	12	204	40
391165,Granulibacter bethesdensis CGD-NIH1,chromosome circular,NC_008343	Rhodospirillales	4	117	46
349163,Acidiphilium cryptum JF-5,plasmid pA-CRY02,NC_009468	Rhodospirillales	1	39	39
349163,Acidiphilium cryptum JF-5,plasmid pA-CRY04,NC_009470	Rhodospirillales	1	3	3
266117,Rubrobacter xylanophilus DSM 9941,chromosome circular,NC_008148	Rubrobacterales	4	49	31
264203,Zymomonas mobilis subsp. mobilis ZM4,chromosome circular,NC_006526	Sphingomonadales	3	15	8
243275,Treponema denticola ATCC 35405,chromosome circular,NC_002967	Spirochaetaceae	1	57	57

176279,Staphylococcus epidermidis RP62A,chromosome circular,NC_002976	Staphylococcus	1	3	3
299768,Streptococcus thermophilus CNRZ1066,chromosome circular,NC_006449	Streptococcaceae	1	41	41
388919,Streptococcus sanguinis SK36,chromosome circular,NC_009009	Streptococcaceae	3	39	20
264199,Streptococcus thermophilus LMG 18311,chromosome circular,NC_006448	Streptococcaceae	2	37	33
322159,Streptococcus thermophilus LMD-9,chromosome circular,NC_008532	Streptococcaceae	3	27	16
467705,Streptococcus gordonii str. Challis substr. CH1,chromosome circular,NC_009785	Streptococcaceae	1	26	26
208435,Streptococcus agalactiae 2603V/R,chromosome circular,NC_004116	Streptococcaceae	1	24	24
205921,Streptococcus agalactiae A909,chromosome circular,NC_007432	Streptococcaceae	1	14	14
211110,Streptococcus agalactiae NEM316,chromosome circular,NC_004368	Streptococcaceae	1	13	13
160490,Streptococcus pyogenes M1 GAS,chromosome circular,NC_002737	Streptococcaceae	2	9	6
293653,Streptococcus pyogenes MGAS5005,chromosome circular,NC_007297	Streptococcaceae	2	7	4
370551,Streptococcus pyogenes MGAS9429,chromosome circular,NC_008021	Streptococcaceae	1	7	7
370553,Streptococcus pyogenes MGAS2096,chromosome circular,NC_008023	Streptococcaceae	1	6	6
210007,Streptococcus mutans UA159,chromosome circular,NC_004350	Streptococcaceae	1	5	5
319701,Streptococcus pyogenes MGAS6180,chromosome circular,NC_007296	Streptococcaceae	2	5	4
370552,Streptococcus pyogenes MGAS10270,chromosome circular,NC_008022	Streptococcaceae	2	5	3
370554,Streptococcus pyogenes MGAS10750,chromosome circular,NC_008024	Streptococcaceae	1	5	5
273063,Sulfolobus tokodaii str. 7,chromosome circular,NC_003106	Sulfolobales	5	457	121
273057,Sulfolobus solfataricus P2,chromosome circular,NC_002754	Sulfolobales	7	416	102
399549,Metallosphaera sedula DSM 5348,chromosome circular,NC_009440	Sulfolobales	4	373	161
330779,Sulfolobus acidocaldarius DSM 639,chromosome circular,NC_007181	Sulfolobales	4	223	132
292459,Symbiobacterium thermophilum IAM 14863,chromosome circular,NC_006177	Symbiobacterium	3	113	57
321332,Synechococcus sp. JA-2-3B'a(2-13),chromosome circular,NC_007776	Synechococcus	6	119	34
321327,Synechococcus sp. JA-3-3Ab,chromosome circular,NC_007775	Synechococcus	9	90	41
1148,Synechocystis sp. PCC 6803,plasmid pSYSA,NC_005230	Synechocystis	3	143	56

56780,Syntrophus aciditrophicus SB,chromosome circular,NC_007759	Syntrophobacterales	2	163	84
335543,Syntrophobacter fumaroxidans MPOB,chromosome circular,NC_008554	Syntrophobacterales	2	148	79
335541,Syntrophomonas wolfei subsp. wolfei str. Goettingen,chromosome circular,NC_008346	Syntrophomonas	5	300	185
300852,Thermus thermophilus HB8,plasmid pTT27,NC_006462	Thermales	9	94	23
262724,Thermus thermophilus HB27,plasmid pTT27,NC_005838	Thermales	8	66	15
300852,Thermus thermophilus HB8,chromosome circular,NC_006461	Thermales	2	18	14
262724,Thermus thermophilus HB27,chromosome circular,NC_005835	Thermales	2	8	6
273068,Thermoanaerobacter tengcongensis MB4,chromosome circular,NC_003869	Thermoanaerobacterales	3	306	216
399726,Thermoanaerobacter sp. X514,chromosome circular,NC_010320	Thermoanaerobacterales	4	296	219
340099,Thermoanaerobacter pseudethanolicus ATCC 33223,chromosome circular,NC_010321	Thermoanaerobacterales	7	186	48
264732,Moorella thermoacetica ATCC 39073,chromosome circular,NC_007644	Thermoanaerobacterales	2	69	46
186497,Pyrococcus furiosus DSM 3638,chromosome circular,NC_003413	Thermococcales	7	200	51
70601,Pyrococcus horikoshii OT3,chromosome circular,NC_000961	Thermococcales	6	149	66
69014,Thermococcus kodakarensis KOD1,chromosome circular,NC_006624	Thermococcales	3	74	36
272844,Pyrococcus abyssi GE5,chromosome circular,NC_000868	Thermococcales	4	57	26
263820,Picrophilus torridus DSM 9790,chromosome circular,NC_005877	Thermoplasmatales	4	116	82
273075,Thermoplasma acidophilum DSM 1728,chromosome circular,NC_002578	Thermoplasmatales	1	46	46
273116,Thermoplasma volcanium GSS1,chromosome circular,NC_002689	Thermoplasmatales	2	34	18
444157,Thermoproteus neutrophilus V24Sta,chromosome circular,NC_010525	Thermoproteales	10	224	39
368408,Thermofilum pendens Hrk 5,chromosome circular,NC_008698	Thermoproteales	10	147	34
178306,Pyrobaculum aerophilum str. IM2,chromosome circular,NC_003364	Thermoproteales	6	130	80
340102,Pyrobaculum arsenaticum DSM 13514,chromosome circular,NC_009376	Thermoproteales	4	127	89
410359,Pyrobaculum calidifontis JCM 11548,chromosome circular,NC_009073	Thermoproteales	7	90	37

384616,Pyrobaculum islandicum DSM 4184,chromosome circular,NC_008701	Thermoproteales	5	72	36
381764,Fervidobacterium nodosum Rt17-B1,chromosome circular,NC_009718	Thermotogaceae	2	190	178
243274,Thermotoga maritima MSB8,chromosome circular,NC_000853	Thermotogaceae	8	106	40
390874,Thermotoga petrophila,chromosome circular,NC_009486	Thermotogaceae	8	100	39
391009,Thermosipho melanesiensis BI429,chromosome circular,NC_009616	Thermotogaceae	5	94	51
401614,Francisella tularensis subsp. novicida U112,chromosome circular,NC_008601	Thiotrichales	2	22	13
484022,Francisella philomiragia subsp. philomiragia ATCC 25017,chromosome circular,NC_010336	Thiotrichales	1	4	4
203124,Trichodesmium erythraeum IMS101,chromosome circular,NC_008312	Trichodesmium	1	3	3
298386,Photobacterium profundum SS9,plasmid pPBPR1,NC_005871	Vibrionales	1	64	64
345073,Vibrio cholerae O395,chromosome 2,NC_009457	Vibrionales	1	39	39
298386,Photobacterium profundum SS9,chromosome 2,NC_006371	Vibrionales	1	26	26
196600,Vibrio vulnificus YJ016,chromosome II,NC_005140	Vibrionales	2	11	9
338187,Vibrio harveyi ATCC BAA-1116,chromosome II,NC_009784	Vibrionales	1	3	3
223926,Vibrio parahaemolyticus RIMD 2210633,chromosome II,NC_004605	Vibrionales	1	2	2
291331,Xanthomonas oryzae pv. oryzae KACC10331,chromosome circular,NC_006834	Xanthomonadales	1	59	59
342109,Xanthomonas oryzae pv. oryzae MAFF 311018,chromosome circular,NC_007705	Xanthomonadales	1	48	48
190486,Xanthomonas axonopodis pv. citri str. 306,chromosome circular,NC_003919	Xanthomonadales	1	18	18
156889,Magnetococcus sp. MC-1,chromosome circular,NC_008576	unknown	1	137	137
374847,Candidatus Korarchaeum cryptofilum OPF8,chromosome circular,NC_010482	unknown	2	121	119
351160,Uncultured methanogenic archaeon RC-I,chromosome circular,NC_009464	unknown	1	114	114
228908,Nanoarchaeum equitans Kin4-M,chromosome circular,NC_005213	unknown	2	41	28
329726,Acaryochloris marina MBIC11017,chromosome circular,NC_009925	unknown	1	3	3

Bibliographie

Abouelhoda, M., Kurtz, S., and Ohlebusch, E. (2004). Replacing suffix trees with enhanced suffix arrays. *Journal of Discrete Algorithms*, **2**, 53–86.

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., and Carniel, E. (1999). *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc Natl Acad Sci U S A*, **96**, 14043–14048.

Aklujkar, M. A., Boles, A. R., Haveman, S. A., DiDonato, R. J., Postier, B. L., and Lovley, D. R. (2007). Comparative genomic evidence of histidyl-trna synthetase inhibition by crispr as a factor in the evolution of *Pelobacter carbinolicus*. Poster in American Society for Microbiology conference.

Andersson, A. F. and Banfield, J. F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science*, **320**, 1047–1050.

Aranaz, A., Liébana, E., Mateos, A., Dominguez, L., Vidal, D., Domingo, M., Gonzolez, O., Rodriguez-Ferri, E. F., Bunschoten, A. E., Van Embden, J. D., and Cousins, D. (1996). Spacer oligonucleotide typing of *Mycobacterium bovis* strains from cattle and other animals : a tool for studying epidemiology of tuberculosis. *J Clin Microbiol*, **34**, 2734–2740.

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

Barrow, P. A. and Soothill, J. S. (1997). Bacteriophage therapy and prophylaxis : Rediscovery and renewed assessment of potential. *Trends Microbiol*, **5**, 268–271.

Beloglazova, N., Brown, G., Zimmerman, M. D., Proudfoot, M., Makarova, K. S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W., Koonin, E. V., Edwards, A. M., Savchenko, A., and Yakunin, A. F. (2008). A novel family of sequence-specific endoribonucleases associated with the Clustered Regularly Interspaced Short Palindromic Repeats. *J Biol Chem*.

Benson, G. (1999). Tandem repeats finder : a program to analyze DNA sequences. *Nucleic Acids Res*, **27**, 573–580.

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., and Hugenholtz, P. (2007). CRISPR recognition tool (CRT) : a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209–209.

Blount, Z. D. and Grogan, D. W. (2005). New insertion sequences of *Sulfolobus :* functional properties and implications for genome evolution in hyperthermophilic archaea. *Mol Microbiol*, **55**, 312–325.

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

Boysen, A., Ellehauge, E., Julien, B., and Søgaard-Andersen, L. (2002). The DevT protein stimulates synthesis of FruA, a signal transduction protein required for fruiting body morphogenesis in *Myxococcus xanthus*. *J Bacteriol*, **184**, 1540–1546.

Brookes, A. J. (1999). The essence of SNPs. *Gene*, **234**, 177–186.

Brudey, K., Driscoll, J. R., Rigouts, L., Prodinger, W. M., Gori, A., Al-Hajoj, S. A., Allix, C., Aristimuño, L., Arora, J., Baumanis, V., Binder, L., Cafrune, P., Cataldi, A., Cheong, S., Diel, R., Ellermeier, C., Evans, J. T., Fauville-Dufaux, M., Ferdinand, S., Garcia de Viedma, D., Garzelli, C., Gazzola, L., Gomes, H. M., Guttierez, M. C., Hawkey, P. M., van Helden, P. D., Kadival, G. V., Kreiswirth, B. N., Kremer, K., Kubin, M., Kulkarni, S. P., Liens, B., Lillebaek, T., Ho, M. L., Martin, C., Martin, C., Mokrousov, I., Narvskaïa, O., Ngeow, Y. F., Naumann, L., Niemann, S., Parwati, I., Rahim, Z., Rasolofo-Razanamparany, V., Rasolonavalona, T., Rossetti, M. L., Rüsch-Gerdes, S., Sajduda, A., Samper, S., Shemyakin, I. G., Singh, U. B., Somoskovi, A., Skuce, R. A., van Soolingen, D., Streicher, E. M., Suffys, P. N., Tortoli, E., Tracevska, T., Vincent, V., Victor, T. C., Warren, R. M., Yap, S. F., Zaman, K., Portaels, F., Rastogi, N., and Sola, C. (2006). *Mycobacterium tuberculosis* complex genetic diversity : Mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol*, **6**, 23–23.

Brügger, K., Torarinsson, E., Redder, P., Chen, L., and Garrett, R. A. (2004). Shuffling of *Sulfolobus* genomes by autonomous and non-autonomous mobile elements. *Biochem Soc Trans*, **32**, 179–183.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S., and Venter, J. C. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058–1073.

Callen, J. (1999). *Biologie Cellulaire.* Dunod.

Chan, M. S., Maiden, M. C., and Spratt, B. G. (2001). Database-driven multi locus sequence typing (MLST) of bacterial pathogens. *Bioinformatics*, **17**, 1077–1083.

DeBoy, R. T., Mongodin, E. F., Emerson, J. B., and Nelson, K. E. (2006). Chromosome evolution in the Thermotogales : large-scale inversions and strain diversification of CRISPR sequences. *J Bacteriol*, **188**, 2364–2374.

Denoeud, F. and Vergnaud, G. (2004). Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a web-based resource. *BMC Bioinformatics*, **5**, 4–4.

Desiere, F., Lucchini, S., Canchaya, C., Ventura, M., and Brüssow, H. (2002). Comparative genomics of phages and prophages in lactic acid bacteria. *Antonie Van Leeuwenhoek*, **82**, 73–91.

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D. A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *J Bacteriol*, **190**, 1390–1400.

D'Herelle, F. (1917). Sur un microbe invisible antagoniste des bacilles dysentériques. *C. R. Acad. Sci. Paris*, pages 165–373.

Diancourt, L., Passet, V., Chervaux, C., Garault, P., Smokvina, T., and Brisse, S. (2007). Multilocus sequence typing of *Lactobacillus casei* reveals a clonal population structure with low levels of homologous recombination. *Appl Environ Microbiol*, **73**, 6601–6611.

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

Durand, P., Mahé, F., Valin, A. S., and Nicolas, J. (2006). Browsing repeats in genomes : Pygram and an application to non-coding region analysis. *BMC Bioinformatics*, **7**, 477–477.

Edgar, R. C. (2007). PILER-CR : fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, **8**, 18–18.

Edgar, R. C. and Myers, E. W. (2005). PILER : identification and classification of genomic repeats. *Bioinformatics*, **21 Suppl 1**, i152–i158.

Edwards, R. A. and Rohwer, F. (2005). Viral metagenomics. *Nat Rev Microbiol*, **3**, 504–510.

Fabre, M., Koeck, J. L., Le Flèche, P., Simon, F., Hervé, V., Vergnaud, G., and Pourcel, C. (2004). High genetic diversity revealed by variable-number tandem repeat genotyping and analysis of hsp65 gene polymorphism in a large collection of mycobacterium canettii strains indicates that the *M. tuberculosis* complex is a recently emerged clone of *M. canettii*. *J Clin Microbiol*, **42**, 3248–3255.

Filliol, I., Driscoll, J. R., Van Soolingen, D., Kreiswirth, B. N., Kremer, K., Valétudie, G., Anh, D. D., Barlow, R., Banerjee, D., Bifani, P. J., Brudey, K., Cataldi, A., Cooksey, R. C., Cousins, D. V., Dale, J. W., Dellagostin, O. A., Drobniewski, F., Engelmann, G., Ferdinand, S., Gascoyne-Binzi, D., Gordon, M., Gutierrez, M. C., Haas, W. H., Heersma, H., Källenius, G., Kassa-Kelembho, E., Koivula, T., Ly, H. M., Makristathis, A., Mammina, C., Martin, G., Moström, P., Mokrousov, I., Narbonne, V., Narvskaya, O., Nastasi, A., Niobe-Eyangoh, S. N., Pape, J. W., Rasolofo-Razanamparany, V., Ridell, M., Rossetti, M. L., Stauffer, F., Suffys, P. N., Takiff, H., Texier-Maugein, J., Vincent, V., De Waard, J. H., Sola, C., and Rastogi, N. (2002). Global distribution of *Mycobacterium tuberculosis* spoligotypes. *Emerg Infect Dis*, **8**, 1347–1349.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.

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., and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science*, **269**, 496–512.

Forterre, P. (1995). Thermoreduction, a hypothesis for the origin of prokaryotes. *C R Acad Sci III*, **318**, 415–422.

Fouts, D. E., Mongodin, E. F., Mandrell, R. E., Miller, W. G., Rasko, D. A., Ravel, J., Brinkac, L. M., DeBoy, R. T., Parker, C. T., Daugherty, S. C., Dodson, R. J., Durkin, A. S., Madupu, R., Sullivan, S. A., Shetty, J. U., Ayodeji, M. A., Shvartsbeyn, A., Schatz, M. C., Badger, J. H., Fraser, C. M., and Nelson, K. E. (2005). Major structural differences and novel potential virulence mechanisms from the genomes of multiple *Campylobacter* species. *PLoS Biol*, **3**, e15–e15.

Frothingham, R. and Meeker-O'Connell, W. A. (1998). Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. *Microbiology*, **144 (Pt 5)**, 1189–1196.

Gilson, E., Clément, J. M., Brutlag, D., and Hofnung, M. (1984). A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli*. *EMBO J*, **3**, 1417–1421.

Godde, J. S. and Bickerton, A. (2006). The repetitive DNA elements called CRISPRs and their associated genes : Evidence of horizontal transfer among prokaryotes. *J Mol Evol*, **62**, 718–729.

Goyal, M., Saunders, N. A., van Embden, J. D., Young, D. B., and Shaw, R. J. (1997). Differentiation of *Mycobacterium tuberculosis* isolates by spoligotyping and IS6110 restriction fragment length polymorphism. *J Clin Microbiol*, **35**, 647–651.

Greve, B., Jensen, S., Brügger, K., Zillig, W., and Garrett, R. A. (2004). Genomic comparison of archaeal conjugative plasmids from *Sulfolobus*. *Archaea*, **1**, 231–239.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007a). CRISPRFinder : a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res*, **35**, W52–W57.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007b). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, **8**, 172–172.

Grissa, I., Vergnaud, G., and Pourcel, C. (2008a). CRISPRcompar : a website to compare clustered regularly interspaced short palindromic repeats. in press.

Grissa, I., Bouchon, P., Pourcel, C., and Vergnaud, G. (2008b). On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. *Biochimie*, **90**, 660–668.

Groenen, P. M., Bunschoten, A. E., van Soolingen, D., and van Embden, J. D. (1993). Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis* ; application for strain differentiation by a novel typing method. *Mol Microbiol*, **10**, 1057–1065.

Gusfield, D. (1997). Algorithms on strings, trees and sequences. *Cambridge University Press, NY*.

Haft, D. H., Selengut, J., Mongodin, E. F., and Nelson, K. E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol*, **1**, e60–e60.

Hermans, P. W., van Soolingen, D., Bik, E. M., de Haas, P. E., Dale, J. W., and van Embden, J. D. (1991). Insertion element IS987 from *Mycobacterium bovis* bcg is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. *Infect Immun*, **59**, 2695–2705.

Hoe, N., Nakashima, K., Grigsby, D., Pan, X., Dou, S. J., Naidich, S., Garcia, M., Kahn, E., Bergmire-Sweat, D., and Musser, J. M. (1999). Rapid molecular genetic subtyping of serotype M1 group A *Streptococcus* strains. *Emerg Infect Dis*, **5**, 254–263.

Horvath, P., Barrangou, R., F. C., Boyaval, P., and Romero, D. (2007). International patent application. 2007025097.

Horvath, P., Coûté-Monvoisin, A. C., Romero, D. A., Boyaval, P., Fremaux, C., and Barrangou, R. (2008a). Comparative analysis of crispr loci in lactic acid bacteria genomes. in press.

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

Hulton, C. S., Higgins, C. F., and Sharp, P. M. (1991). ERIC sequences : a novel family of repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium* and other enterobacteria. *Mol Microbiol*, **5**, 825–834.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J Bacteriol*, **169**, 5429–5433.

Jansen, R., van Embden, J. D. A., Gaastra, W., and Schouls, L. M. (2002a). Identification of a novel family of sequence repeats among prokaryotes. *OMICS*, **6**, 23–33.

Jansen, R., Embden, J. D. A., Gaastra, W., and Schouls, L. M. (2002b). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*, **43**, 1565–1575.

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., and van Embden, J. (1997). Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J Clin Microbiol*, **35**, 907–914.

Kunin, V., Sorek, R., and Hugenholtz, P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol*, **8**, R61–R61.

Kunin, V., He, S., Warnecke, F., Peterson, S. B., Garcia Martin, H., Haynes, M., Ivanova, N., Blackall, L. L., Breitbart, M., Rohwer, F., McMahon, K. D., and Hugenholtz, P. (2008). A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res*, **18**, 293–297.

Kurland, C. G. (2000). Something for everyone. Horizontal gene transfer in evolution. *EMBO Rep*, **1**, 92–95.

Kurtz, S. and Schleiermacher, C. (1999). REPuter : fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15**, 426–427.

Kurtz, S., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2000). Computation and visualization of degenerate repeats in complete genomes. *Proc Int Conf Intell Syst Mol Biol*, **8**, 228–238.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter : the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res*, **29**, 4633–4642.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Le Flèche, P., Hauck, Y., Onteniente, L., Prieur, A., Denoeud, F., Ramisse, V., Sylvestre, P., Benson, G., Ramisse, F., and Vergnaud, G. (2001). A tandem repeats database for bacterial genomes : Application to the genotyping of *Yersinia pestis* and *Bacillus Anthracis*. *BMC Microbiol*, **1**, 2–2.

Le Flèche, P., Jacques, I., Grayon, M., Al Dahouk, S., Bouchon, P., Denoeud, F., Nöckler, K., Neubauer, H., Guilloteau, L. A., and Vergnaud, G. (2006). Evaluation and selection of tandem repeat loci for a *Brucella* MLVA typing assay. *BMC Microbiol*, **6**, 9–9.

Lillestøl, R. K., Redder, P., Garrett, R. A., and Brügger, K. (2006). A putative viral defence mechanism in archaeal cells. *Archaea*, **2**, 59–72.

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

Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B., and Koonin, E. V. (2002). A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res*, **30**, 482–496.

Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I., and Koonin, E. V. (2006). A putative RNA-interference-based immune system in prokaryotes : Computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct*, **1**, 7–7.

Meunier, J. R. and Grimont, P. A. (1993). Factors affecting reproducibility of random amplified polymorphic DNA fingerprinting. *Res Microbiol*, **144**, 373–379.

Mojica, F. J., Juez, G., and Rodríguez-Valera, F. (1993). Transcription at different salinities of *Haloferax mediterranei* sequences adjacent to partially modified PstI sites. *Mol Microbiol*, **9**, 613–621.

Mojica, F. J., Ferrer, C., Juez, G., and Rodríguez-Valera, F. (1995). Long stretches of short tandem repeats are present in the largest replicons of the archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. *Mol Microbiol*, **17**, 85–93.

Mojica, F. J., Díez-Villaseñor, C., Soria, E., and Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol*, **36**, 244–246.

Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol*, **60**, 174–182.

Mokrousov, I., Narvskaya, O., Limeschenko, E., and Vyazovaya, A. (2005). Efficient discrimination within a *Corynebacterium diphtheriae* epidemic clonal group by a novel macroarray-based method. *J Clin Microbiol*, **43**, 1662–1668.

Mokrousov, I., Limeschenko, E., Vyazovaya, A., and Narvskaya, O. (2007). *Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci. *Biotechnol J*, **2**, 901–906.

Mongodin, E. F., Hance, I. R., Deboy, R. T., Gill, S. R., Daugherty, S., Huber, R., Fraser, C. M., Stetter, K., and Nelson, K. E. (2005). Gene transfer and genome plasticity in *Thermotoga maritima*, a model hyperthermophilic species. *J Bacteriol*, **187**, 4935–4944.

Mrázek, J., Xie, S., Guo, X., and Srivastava, A. (2008). AIMIE : a web-based environment for detection and interpretation of significant sequence motifs in prokaryotic genomes. *Bioinformatics*, **24**, 1041–1048.

Nakata, A., Amemura, M., and Makino, K. (1989). Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli K-12* chromosome. *J Bacteriol*, **171**, 3553–3556.

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O., Salzberg, S. L., Smith, H. O., Venter, J. C., and Fraser, C. M. (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, **399**, 323–329.

Onteniente, L. (2004). *Etude du polymorphisme associé aux répétitions en tandem pour le typage de bactéries pathogènes : Pseudomonas aeruginosa et Staphylococcus aureus*. Biologie cellulaire et moléculaire, Université Evry Val d'Essonne.

Peng, X., Blum, H., She, Q., Mallok, S., Brügger, K., Garrett, R. A., Zillig, W., and Prangishvili, D. (2001). Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2 : relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. *Virology*, **291**, 226–234.

Peng, X., Brügger, K., Shen, B., Chen, L., She, Q., and Garrett, R. A. (2003). Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in *Sulfolobus* genomes. *J Bacteriol*, **185**, 2410–2417.

Perry, J., Staley, J., and Lory, S. (2004). *Microbiologie*. Éditions Dunod.

Philippe, H. and Douady, C. J. (2003). Horizontal gene transfer and phylogenetics. *Curr Opin Microbiol*, **6**, 498–505.

Pourcel, C., André-Mazeaud, F., Neubauer, H., Ramisse, F., and Vergnaud, G. (2004). Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*. *BMC Microbiol*, **4**, 22–22.

Pourcel, C., Salvignol, G., and Vergnaud, G. (2005). CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, **151**, 653–663.

Pourcel, C., Visca, P., Afshar, B., D'Arezzo, S., Vergnaud, G., and Fry, N. K. (2007). Identification of variable-number tandem-repeat (VNTR) sequences in *Legionella pneumophila* and development of an optimized multiple-locus VNTR analysis typing scheme. *J Clin Microbiol*, **45**, 1190–1199.

Prangishvili, D., Forterre, P., and Garrett, R. A. (2006). Viruses of the Archaea : a unifying view. *Nat Rev Microbiol*, **4**, 837–848.

Price, E. P., Smith, H., Huygens, F., and Giffard, P. M. (2007). High-resolution DNA melt curve analysis of the clustered, regularly interspaced short-palindromic-repeat locus of *Campylobacter jejuni*. *Appl Environ Microbiol*, **73**, 3431–3436.

Rice, George .and Tang, L., Stedman, K., Roberto, F., Spuhler, J., Gillitzer, E., Johnson, J. E., Douglas, T., and Young, M. (2004). The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. *Proc Natl Acad Sci U S A*, **101**, 7716–7720.

Russell, W. M., Barrangou, R., and Horvath, P. (2006). Detection and typing of bacterial strains. US Patent Application. 20060199190.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, **230**, 1350–1354.

Salcedo, C., Arreaza, L., Alcalá, B., de la Fuente, L., and Vázquez, J. A. (2003). Development of a multilocus sequence typing method for analysis of *Listeria monocytogenes* clones. *J Clin Microbiol*, **41**, 757–762.

Sanders, M. E. (1988). Phage resistance in lactic acid bacteria. *Biochimie*, **70**, 411–422.

Schouls, L. M., Reulen, S., Duim, B., Wagenaar, J. A., Willems, R. J. L., Dingle, K. E., Colles, F. M., and Van Embden, J. D. A. (2003). Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing : Strain diversity, host range, and recombination. *J Clin Microbiol*, **41**, 15–26.

Sebaihia, M., Wren, B. W., Mullany, P., Fairweather, N. F., Minton, N., Stabler, R., Thomson, N. R., Roberts, A. P., Cerdeño-Tárraga, A. M., Wang, H., Holden, M. T. G., Wright, A., Churcher, C., Quail, M. A., Baker, S., Bason, N., Brooks, K., Chillingworth, T., Cronin, A., Davis, P., Dowd, L., Fraser, A., Feltwell, T., Hance, Z., Holroyd, S., Jagels, K., Moule, S., Mungall, K., Price, C., Rabbinowitsch, E., Sharp, S., Simmonds, M., Stevens, K., Unwin, L., Whithead, S., Dupuy, B., Dougan, G., Barrell, B., and Parkhill, J. (2006). The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat Genet*, **38**, 779–786.

Selander, R. K., Caugant, D. A., Ochman, H., Musser, J. M., Gilmour, M. N., and Whittam, T. S. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl Environ Microbiol*, **51**, 873–884.

Sharples, G. J. and Lloyd, R. G. (1990). A novel repeated DNA sequence located in the intergenic regions of bacterial chromosomes. *Nucleic Acids Res*, **18**, 6503–6508.

She, Q., Phan, H., Garrett, R. A., Albers, S. V., Stedman, K. M., and Zillig, W. (1998). Genetic profile of pNOB8 from *Sulfolobus* : the first conjugative plasmid from an archaeon. *Extremophiles*, **2**, 417–425.

She, Q., Singh, R. K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M. J., Chan-Weiher, C. C., Clausen, I. G., Curtis, B. A., De Moors, A., Erauso, G., Fletcher, C., Gordon, P. M., Heikamp-de Jong, I., Jeffries, A. C., Kozera, C. J., Medina, N., Peng, X., Thi-Ngoc, H. P., Redder, P., Schenk, M. E., Theriault, C., Tolstrup, N., Charlebois, R. L., Doolittle, W. F., Duguet, M., Gaasterland, T., Garrett, R. A., Ragan, M. A., Sensen, C. W., and Van der Oost, J. (2001). The complete genome of the crenarchaeon *Sulfolobus solfataricus P2*. *Proc Natl Acad Sci U S A*, **98**, 7835–7840.

Singh, U. B., Suresh, N., Bhanu, N. V., Arora, J., Pant, H., Sinha, S., Aggarwal, R. C., Singh, S., Pande, J. N., Sola, C., Rastogi, N., and Seth, P. (2004). Predominant tuberculosis spoligotypes, Delhi, India. *Emerg Infect Dis*, **10**, 1138–1142.

Singleton, P. (2005). *Bactériologie : Pour la médecine, la biologie et les biotechnologies.* Dunod, cours, 6e edition.

Smith, G. R. (1988). Homologous recombination in procaryotes. *Microbiol Rev*, **52**, 1–28.

Sola, C., Horgen, L., Maïsetti, J., Devallois, A., Goh, K. S., and Rastogi, N. (1998). Spoligotyping followed by double-repetitive-element PCR as rapid alternative to IS6110 fingerprinting for epidemiological studies of tuberculosis. *J Clin Microbiol*, **36**, 1122–1124.

Sontheimer, E. J. (2005). Assembly and function of RNA silencing complexes. *Nat Rev Mol Cell Biol*, **6**, 127–138.

Sorek, R., Kunin, V., and Hugenholtz, P. (2008). CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol*, **6**, 181–186.

Stern, M. J., Ames, G. F., Smith, N. H., Robinson, E. C., and Higgins, C. F. (1984). Repetitive extragenic palindromic sequences : a major component of the bacterial genome. *Cell*, **37**, 1015–1026.

Sturino, J. M. and Klaenhammer, T. R. (2006). Engineered bacteriophage-defence systems in bioprocessing. *Nat Rev Microbiol*, **4**, 395–404.

Summers, W. C. (1993a). Cholera and plague in India : the bacteriophage inquiry of 1927-1936. *J Hist Med Allied Sci*, **48**, 275–301.

Summers, W. C. (1993b). How bacteriophage came to be used by the Phage Group. *J Hist Biol*, **26**, 255–267.

Sutter, V. L., Hurst, V., and Fennell, J. (1965). A Standardized System For Phage Typing Pseudomonas aeruginosa. *Health Lab Sci*, **2**, 7–16.

Tang, G. (2005). siRNA and miRNA : an insight into RISCs. *Trends Biochem Sci*, **30**, 106–114.

Tang, T. H., Bachellerie, J. P., Rozhdestvensky, T., Bortolin, M. L., Huber, H., Drungowski, M., Elge, T., Brosius, J., and Hüttenhofer, A. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc Natl Acad Sci U S A*, **99**, 7536–7541.

Thöny-Meyer, L. and Kaiser, D. (1993). devRS, an autoregulated and essential genetic locus for fruiting body development in *Myxococcus xanthus*. *J Bacteriol*, **175**, 7450–7462.

Twort, F. (1915). *An investigation on the nature of ultra-microscopic viruses*, volume 186, chapter Part II, pages 1241–2143. Lancet.

Tyson, G. W. and Banfield, J. F. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol*, **10**, 200–207.

Valjevac, S., Hilaire, V., Lisanti, O., Ramisse, F., Hernandez, E., Cavallo, J. D., Pourcel, C., and Vergnaud, G. (2005). Comparison of minisatellite polymorphisms in the *Bacillus cereus* complex : a simple assay for large-scale screening and identification of strains most closely related to *Bacillus anthracis*. *Appl Environ Microbiol*, **71**, 6613–6623.

van Belkum, A., Scherer, S., van Alphen, L., and Verbrugh, H. (1998). Short-sequence DNA repeats in prokaryotic genomes. *Microbiol Mol Biol Rev*, **62**, 275–293.

van Belkum, A., Struelens, M., de Visser, A., Verbrugh, H., and Tibayrenc, M. (2001). Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. *Clin Microbiol Rev*, **14**, 547–560.

van Embden, J. D., van Gorkom, T., Kremer, K., Jansen, R., van Der Zeijst, B. A., and Schouls, L. M. (2000). Genetic variation and evolutionary origin of the direct repeat locus of *Mycobacterium tuberculosis* complex bacteria. *J Bacteriol*, **182**, 2393–2401.

van Soolingen, D., Hermans, P. W., de Haas, P. E., Soll, D. R., and van Embden, J. D. (1991). Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains : Evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *J Clin Microbiol*, **29**, 2578–2586.

van Soolingen, D., de Haas, P. E., Hermans, P. W., Groenen, P. M., and van Embden, J. D. (1993). Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of *Mycobacterium tuberculosis*. *J Clin Microbiol*, **31**, 1987–1995.

van Soolingen, D., van der Zanden, A. G., de Haas, P. E., Noordhoek, G. T., Kiers, A., Foudraine, N. A., Portaels, F., Kolk, A. H., Kremer, K., and van Embden, J. D. (1998). Diagnosis of *Mycobacterium microti* infections among humans by using novel genetic markers. *J Clin Microbiol*, **36**, 1840–1845.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H., and Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.

Vergnaud, G. and Denoeud, F. (2000). Minisatellites : mutability and genome architecture. *Genome Res*, **10**, 899–907.

Vergnaud, G., Li, Y., Gorgé, O., Cui, Y., Song, Y., Zhou, D., Grissa, I., Dentovskaya, S. V., Platonov, M. E., Rakin, A., Balakhonov, S. V., Neubauer, H., Pourcel, C., Anisimov, A. P., and Yang, R. (2007). Analysis of the three *Yersinia pestis* CRISPR loci provides new tools for phylogenetic studies and possibly for the investigation of ancient DNA. *Adv Exp Med Biol*, **603**, 327–338.

Viswanathan, P., Murphy, K., Julien, B., Garza, A. G., and Kroos, L. (2007). Regulation of dev, an operon that includes genes essential for *Myxococcus xanthus* development and CRISPR-associated genes and repeats. *J Bacteriol*, **189**, 3738–3750.

Vu-Thien, H., Corbineau, G., Hormigos, K., Fauroux, B., Corvol, H., Clément, A., Vergnaud, G., and Pourcel, C. (2007). Multiple-locus variable-number tandem-repeat analysis for longitudinal survey of sources of *Pseudomonas aeruginosa* infection in cystic fibrosis patients. *J Clin Microbiol*, **45**, 3175–3183.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain : the primary kingdoms. *Proc Natl Acad Sci U S A*, **74**, 5088–5090.

Yeramian, E. and Buc, H. (1999). Tandem repeats in complete bacterial genome sequences : Sequence and structural analyses for comparative studies. *Res Microbiol*, **150**, 745–754.

Résumé

Les CRISPRs constituent une famille particulière d'éléments génétiques, retrouvés dans de nombreux génomes de procaryotes (la moitié des bactéries et presque toutes les archées). Des études récentes suggèrent que cette structure représente un système de défense contre les ADNs étrangers fonctionnant grâce à un mécanisme d'interférence ARN. Les CRISPRs consistent en la succession de régions très bien conservées (DR) dont la taille varie de 23 à 47 pb, séparées par des séquences uniques d'une taille similaire et ayant en général une origine virale. Le polymorphisme observé entre différentes souches de la même espèce fait du CRISPR un marqueur génétique intéressant pour des analyses comparatives de souches bactériennes très proches et pour des études phylogénétiques. Cette thèse décrira trois outils bioinformatiques accessibles à l'adresse (`http://crispr.u-psud.fr`) et leurs applications dans l'investigation et le typage des CRISPRs. Le premier outil est CRISPRFinder qui est un programme d'identification des CRISPRs à partir des séquences génomiques. Le deuxième est la base de données CRISPRdb et ses utilitaires qui fournissent un accès aux CRISPRs de tous les génomes procaryotes séquencés. le troisième outil est CRISPRcompar qui sert à identifier et comparer les CRISPRs dans des génomes proches pour faciliter la procédure de typage.

Abstract

Clustered, Regularly Interspaced Short Palindromic Repeat (CRISPR) system is a putative RNA-interference-based immune system conferring an acquired resistance against foreign DNA agressions in prokaryotes. It consists of a succession of highly conserved regions (DR) varying in size from 23 to 47 bp, separated by similarly sized unique sequences (spacer) of usually viral origin. Polymorphism can be observed in different strains of a species and may be used for genotyping. This thesis describes three bioinformatics tools available freely on the web (`http://crispr.u-psud.fr`) and their applications for the CRISPR investigations and typing. The first one is CRISPRFinder, a tool that allows the identification of CRISPRs from published genomes. The second one is CRISPRdb, a database of CRISPRs in the sequenced prokaryotes and several associated tools to facilitate its query and the CRISPRs investigations. The third tool is CRISPRcompar for the identification and comparison of CRISPRs loci and their component determination.