Genet. Sel. Evol. 33 (Suppl. 1) (2001) S319–S338

Original article

Experimental and theoretical evaluation of typing methods based upon random amplification of genomic restriction fragments (AFLP) for bacterial population genetics

Christophe Mougel[a], Sylvie Teyssier[a], Cathy d'Angelo[a], Karine Groud[a], Marc Neyra[a], Karim Sidi-Boumedine[b], Axel Cloeckaert[b], Michèle Peloille[b], Sylvie Baucheron[b], Élisabeth Chaslus-Dancla[b], Sophie Jarraud[c], Hélène Meugnier[c], Françoise Forey[c], François Vandenesch[c], Gérard Lina[c], Jérôme Étienne[c], Jean Thioulouse[d], Charles Manceau[e], Patrick Robbe[f], Renaud Nalin[f], Jérôme Briolay[f], Xavier Nesme[a,*]

[a] Écologie microbienne, Université Claude Bernard-Lyon 1, UMR CNRS 5557, Inra et IRD, 43 bd du 11 novembre 1918, 69622 Villeurbanne Cedex, France
[b] Station de pathologie aviaire, Institut national de la recherche agronomique, Centre de Tours, 37380 Nouzilly, France
[c] Département de microbiologie moléculaire et médicale, Université Claude Bernard-Lyon 1, EA1655, Faculté de Médecine Laënnec, rue Guillaume Paradin, 69372 Lyon Cedex 08, France
[d] Laboratoire de biométrie, génétique et biologie des populations, Université Claude Bernard-Lyon 1, UMR CNRS 5558, 43 bd du 11 novembre 1918, 69622 Villeurbanne Cedex, France
[e] UMR PaVé, Institut national de la recherche agronomique, INH, Université d'Angers, BP 57, 42 rue Georges Morel, 49071 Beaucouzé, France
[f] Développement technique pour l'analyse moléculaire et la biodiversité, et LIBRAGEN, Université Claude Bernard-Lyon 1, 43 bd du 11 novembre 1918, 69622 Villeurbanne Cedex, France

Abstract – The reliability and the level of taxonomic resolution of the amplified fragment length polymorphism (AFLP) method were evaluated with species of pathogenic bacteria involved in human, animal and plant diseases. The method was found to be very versatile as it can be adapted to the individual genome constraints of all

* Correspondence and reprints
E-mail: nesme@univ-lyon1.fr

tested species. The calculation of a genetic distance d corresponding to the average dissimilarity between actual overall genome sequences was proposed for comparing AFLP data. Bacterial models showed clearly different patterns between strains belonging to different genomic species, while patterns were clearly similar within a given species. The threshold which distinguishes between inter and infra-specific distances indicates a critical overall genome diversity of about 14% ($d = 0.14$). AFLP had more resolution power than serology, phage typing, PFGE and restriction analysis of ribosomal intergenic spacers. In the latter case, regression analysis showed that PCR-RFLP of ribosomal intergenic spacers can only be used to differentiate bacteria which have at least 3.4% ($d = 0.034$) nucleotide differences between their respective genomes. Finally, an improved procedure using newly developed software was also proposed in order to standardize the capture of reliable data and their numeric treatment for the future development of AFLP data bases.

AFLP / genomospecies / genetic distance / infraspecific diversity

Résumé – Évaluation expérimentale et théorique des méthodes de typages basées sur l'amplification aléatoire de fragments de restriction pour l'étude des populations bactériennes. La reproductibilité et le niveau de résolution taxonomique de l'amplification aléatoire de fragments de restriction (AFLP) ont été évalués avec divers modèles pathogènes de l'homme, des animaux et des plantes. La méthode est très adaptable et peut être modifiée en fonction des particularités génomiques de chaque espèce. La différence nucléotidique moyenne réelle entre les génomes est une mesure de la distance génomique entre bactéries qui peut être estimée à partir de l'AFLP. Cette mesure permet la comparaison de données AFLP obtenues de façons différentes. Avec la plupart des modèles, le seuil discriminant les distances inter- des distances infra-spécifiques correspond à des différences nucléotidiques de l'ordre de 14% ($d = 0,14$). L'AFLP s'est montrée plus résolutive que la sérologie, la lysotypie, l'électrophorèse en champs pulsés, et la PCR-RFLP de l'intergène ribosomique. L'analyse montre ainsi que la PCR-RFLP de l'intergène ribosomique permet de distinguer des bactéries présentant au moins 3,4% ($d = 0,034$) de différences entre leurs génomes respectifs. Une procédure standard d'aquisition et de traitement numérique des données incluant des logiciels adaptés est également proposée.

AFLP / espèce génomique / distance génétique / diversité infra-spécifique

1. INTRODUCTION

Epidemiologists and more generally microbial ecologists have a common interest in setting up efficient methods to accurately characterize bacterial genotypes involved in disease outbreaks or other biological activities. Methods are required to identify the reservoirs as well as to monitor bacterial behavior in complex biotopes such as water, soil, phytospheres and food chains. More generally, the management of complex ecosystems requires a better knowledge of gene flow among microbial populations.

Genes in Prokarya are generally clonally transmitted as a common heritage to heir cells. Nevertheless, genes are also parasexually transferred between

genomes through transduction, transformation or conjugation. Thus, bacterial populations are not always – or not entirely – clonal as can be tested by the linkage desequilibrium of multiple loci [15]. As a result, and contrary to the paradigm based on Eukarya, population genetics of Prokarya has to consider not only individuals from the same species, but also bacteria belonging to different species able to exchange genetic materials. Bacterial populations thus appear composed of individual cells with identical genomes (= clones), or closely related genomes derived from a common ancestor through binary fission (= clonal populations), or from different species and genera or even kingdoms, which have exchanged genes. The characterization of bacterial genotypes must thus be done at both species and infra-species levels.

Bacterial species are presently defined through direct comparison of genome pairs by DNA/DNA hybridization studies leading to the concept of genomic species or genomospecies [32]. Direct comparison of genomes is not however easily applicable to the species identification of numerous isolates, and, thus, requires development of indirect methods. According to Woëse, the comparison of 16S ribosomal sequences is increasingly used to establish the bacterial phylogeny of newly isolated bacteria [14]. However, like other authors, we showed that 16S sequences are not precise enough to differentiate closely related genomic species, which have identical or nearly identical 16S rRNA [20, 21, 28]. Moreover, definition of species based upon the sequence of a single gene remains controversial in view of putative transfers including those encoding the ribosomal RNA [2, 34]. Thus, a multilocus approach, which is more in agreement with the conventional definition of the genomic species, should be preferred.

Analysis of multiple loci is also relevant for fingerprinting bacteria at the infraspecific level. This is currently achieved by random amplification of genome parts using degenerated or repeated sequences to design PCR primers, in RAPD and repetitive (Rep) PCR, respectively. These methods generally allow characterizing differences between bacterial strains, and are very useful for controlling the identity of a given strain. Vos *et al.* [31] proposed an alternative method for randomly amplifying restriction fragments of genomes. In our hands, this method was much more reliable than RAPD. We also showed that AFLP – but not RAPD – allowed a clear distinction of bacteria belonging to the same genomic species from others [6]. This finding suggested that AFLP could be a powerful method to characterize bacterial populations at both species and infra-species levels.

In the present work, we evaluated both versatility and resolution potential of the AFLP method on several bacterial species involved in human, animal and plant diseases. The calculation of genetic distances from AFLP data is proposed in order to compare AFLP data obtained in different experiments. Since fluorescent-AFLP allows gel electrophoresis and fragment visualization with automatic devices, a standardized procedure including the development of new programs was set up to gather reliable data in view of automated analysis of the genetic structure of bacterial populations.

2. MATERIALS AND METHODS

2.1. Bacterial strains

Bacteria considered in the present study were: *Agrobacterium* spp. (18 representative isolates from seven genomic species described by Popoff *et al.* [23] also analyzed by PCR-RFLP of the ribosomal intergenic region, ITS1-PCR-RFLP, as previously described [22]), *Staphylococcus aureus* (202 isolates belonging to four accessory gene regulatory groups (*agr*) also characterized by pulsed field gel electrophoresis, PFGE, as previously described [12,13]), *Xylophylus ampelinus* (16 isolates differentiated by conventional biochemical traits, but not by ITS1-PCR-RFLP), *Salmonella enterica* (35 Agona serotype, 40 Typhimurium serotype mostly from phage type DT104, also analyzed by PFGE), *Legionella pneumophila* (48 isolates also analyzed by PFGE [25]), *Pseudomonas syringae* (26 isolates gathering in genomovar III [9] also analyzed by RAPD [6]), *Xanthomonas populi* (12 isolates belonging to 5 physiologic races [19]), and *Frankia sp.* (isolates from different genomic species isolated from *Casuarinaceae* and *Betulaceae* [17]). Bacteria were grown and stored as described in related papers.

2.2. Isolation and quantification of DNA

Total DNAs were extracted by standard methods described by Sambrook *et al.* [26], or with rapid methods such as Dneasy^{TM} Tissue Kit (QIAGEN, Courtabœuf, France), or the CTAB technique [3]. A rapid method using colonies was also set up: bacterial suspensions (O.D.: 0.3 at 600 nm) were boiled for 10 min and immediately placed on ice for 10 min. Insoluble cellular residues were eliminated by centrifugation at 13000 rpm for 5 min. The supernatants were directly used for PCR or ligation reactions, or were stored at 4°C for several weeks. DNA quantifications, which are essential for reliable AFLP results, were performed spectrophotometrically or by comparison to a range of standard DNA quantities.

2.3. AFLP analysis

AFLP were generally performed according to instructions of the kit of fluorescent-AFLP (FAFLP), AFLP^{TM} Microbial Fingerprinting (Perkin-Elmer, Les Ulis, France), except for modifications indicated in the text. Adapters and primers were those provided by the manufacturer or synthesized by other companies using the same core sequences of adapters described by Zabeau and Vos [35]. Endonucleases used for genomic restriction and nucleotides added to PCR primers for selective amplifications were experimentally chosen according to individual responses of bacterial species as described in the result section. Samples processed accordingly were loaded singly or in pools of three with

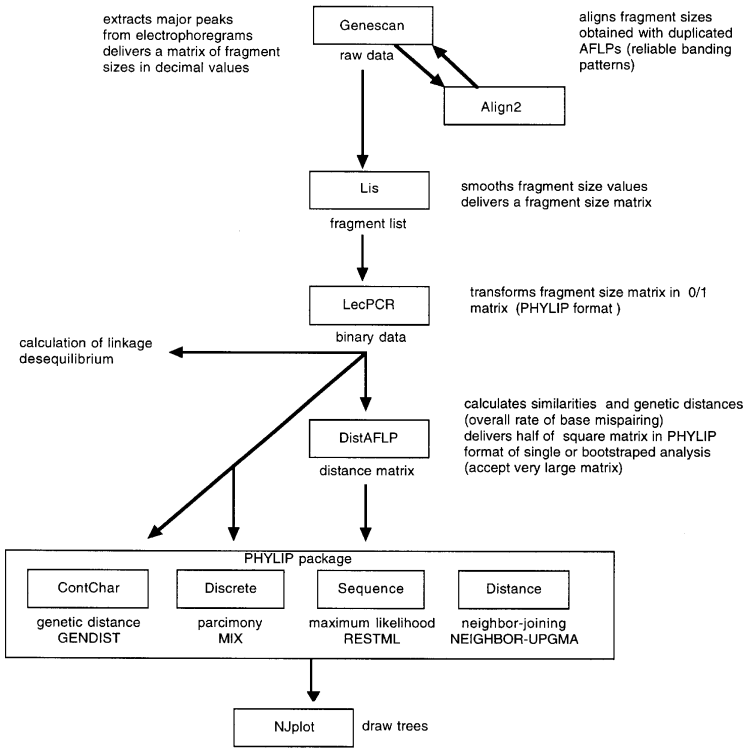

Figure 1. Standardized procedure for reliable data capture and processing of FAFLP analysis to the molecular phylogeny and population genetics of bacteria.

different fluorescent dyes in denaturing polyacrylamide gels for electrophoresis with automatic devices (ABI PRISM 373 or 377 DNA Sequencers, Perkin Elmer Corporation, Foster City, CA, USA).

Similar methods using standard visualization of amplified restriction fragments in agarose or acrylamide gels stained with ethidium bromide instead of fluorescence labeling were also considered: the infrequent-restriction-site amplification method (IRS-PCR) described by Mazureck *et al.* [16], and the simplified AFLP method for Prokarya described by Clerc *et al.* [6]. The latter method used only one endonuclease, a quadracutter, and one primer with generally three discriminating nucleotides. Thus each added discriminating nucleotide has a discriminating effect on both the upper and the lower strands.

2.4. Data capture and processing

A standardized procedure including newly designed sofwares was set up for capture and data processing (Fig. 1).

The software GeneScan® (PE Applied Biosystem) was used to visualize electrophoregrams, and to extract tabular data composed of lists of major peaks identified by their estimated size (in base pairs).

As a general rule, AFLPs were duplicated in order to keep only reliable data for further analysis. To facilitate the comparison of AFLP duplicates, we developed the program Align2, which ranks two sets of tabular data on a same scale, allowing an easy and rapid identification of orphan and duplicated peaks, as well as an easy visual verification of the correspondence between retained peaks and electrophoregrams. The retained data were the average sizes of duplicated peaks.

All polymorphic as well as monomorphic data were considered to calculate AFLP similarities between strains. The size values of all individual peaks were also stored for further analysis. This led to very large matrices of data composed of hundreds of strains and peaks, requiring the development of new software. Decimal peak size values obtained from GeneScan were then assigned in discrete categories. For this purpose, the program Lis was first used to optimally transform decimal peak size values into entire values, then the program LecPCR, was used to transform tabular lists of peak size into tabular binary data in a format suitable for numerical analysis.

2.5. Numerical analysis

The program distAFLP calculates similarity matrices and of genetic distances using equations defined in the present work. Outputs are simple or multiple matrices calculated after matrix bootstrapping.

Tabular binary data obtained from the LecPCR or distance matrix obtained from distAFLP were analyzed with the phylogenetic inference package PHYLIP [8], using the softwares MIX for parcimony or GENEDIST (binary matrix), as well as Neighbor-Joining or Maximum-Likelihood (distance matrix), and CONSENCE to determine the bootstrap values of dendrograms (multiple data).

Bacterial clonality was tested by calculation of the multilocus linkage disequilibrium of AFLP loci according to Maynard-Smith *et al.* [15].

Programs Align2, LecPCR, and DistAFLP are available for free on the ADE-4 web server http://pbil.univ-lyon1.fr/ADE-4/microb. In addition, DistAFLP can provide output files in the ADE-4 binary format suitable for multivariate analysis methods [30].

3. RESULTS

3.1. Experimental and simulated AFLP

The commercial combination of endonucleases and selective nucleotides were irrelevant with several species. AFLP fragments were too numerous, poorly separated, or on the contrary too few to provide the required level of discrimination. AFLP conditions thus had to be adapted to individual species characteristics. This could be obtained experimentally or by theoretical and simulated AFLP.

3.1.1. Experimental optimization of the AFLP technique for different genera

With the experimental approach, the most important step was to find a quadracutter endonuclease which led to a regular set of discrete fragments between 50 and 500 bp. High performance combinations of endonucleases were provided for several bacteria (Tab. I). Results of endonuclease screenings suggested that the GC content of genomes can modify AFLP results, as could be verified by a theoretical simulation of AFLP results.

3.1.2. Theoretical and predictive AFLP

The expected number of amplified fragments would vary according to the GC content of the three selective nucleotides. The simplified version of AFLP for Prokarya was used to test the hypothesis because this method allowed visualization of all amplified fragments whatever their size (contrary to FAFLP which is limited to fragments between 35 and 500 bp). In this case, if pC is the probability of the occurrence of a cytosine at a given nucleotide position in the genome, $p(\mathrm{Hpa}II) = (pC \times pC \times pG \times pG)$, and with three selective nucleotides X_1, X_2, X_3, $p(AFLP) = p(\mathrm{Hpa}II) \times p(X_1)^2 \times p(X_2)^2 \times p(X_3)^2$. For instance, with *X. populi* with a GC content of 64%, the expected numbers of fragments were, for instance 1.6 with AAA, 6 with CAA, 19 with CCA, 66 with CCC. This prediction was experimentally verified since reliable results were obtained with combinations like CAA or CCA, but not with AAA or CCC, presumably because too few or too many fragments were released, respectively (data not shown).

This theoretical approach facilitated the choice of endonucleases with FAFLP as well. However, the GC content alone was not sufficient to allow accurate predictions. For instance, with *Frankia sp.* with a GC content of 70%, *Hpa*II (CCGG) released too short fragments (below 200 bp), while *Hha*I recognizing a different sequence (GCGC) with the same GC content produced a large number of fragments, evenly distributed between 50 and 500 bp (data not shown). This showed that specific nucleotidic sequences of genomes were also involved

Table I. Combination of relevant endonucleases and selective nucleotides for FAFLP and related methods.

Bacteria	Hexacutter	Quadracutter	Selective nucleotides	*r*
Fluorescent-AFLP				
Staphylococcus aureus	*Eco*RI	*Taq*I	Eco+A/Taq+C,Eco+T/Taq+G	12
Salmonella enterica	*Eco*RI	*Mse*I	2 to 3 nucleotides	12/13
Xylophilus ampelinus	*Pst*I	*Mse*I	PstC+/Mse+TA,TC,TC,TT	13
Agrobacterium spp.	*Eco*RI	*Mse*I	Eco+CA,CC,CG,CT/Mse+0	12
Frankia sp.	*Pst*I	*Hha*I	PstI+0/HhaI+A,AG,G,GA	11/12
IRS-PCR				
Legionella pneumophila	PstI/XbaI	na	Pst+0/Xba+G	13
Simplified AFLP for Prokarya				
Chlamydia psittaci	na	*Hpa*II	HpaII+3 nucleotides	14
P. syringae genomovar III	na	*Hpa*II	HpaII+3 nucleotides	14
Xanthomonas populi	na	*Hpa*II	HpaII+3 nucleotides	14

The selective nucleotides are indicated after the endonuclease concerned by the primer.
0 indicates no added nucleotide. *r*: number of nucleotides constraining AFLP used to calculate AFLP genetic distances.
na: not applicable.

in AFLP results. This was verified by an experimental approach combining predictive and experimental AFLPs.

Representative genome sequences of a given strain were used to predict FAFLP results and to determine the choice of endonucleases and discriminating nucleotides. A simulation performed with about 120 kb chromosomal sequences of strain C58 of *A. tumefaciens* (about 2.4% of the complete genome), showed a very low dinucleotide relative abundance of TA = 0.47, *i.e.* less than half the expected. This modified in turn the relative abundance of *Eco*RI (GAATTC) and *Mse*I (TTAA), that were multiplied by 1.95 and 0.45, respectively. As a consequence, the predictive AFLP performed with C58 chromosomal sequences showed that *Eco*RI can release ten fragments not cut by *Mse*I, with four out of ten between 93 and 477 bp detectable by FAFLP, and two out of four with the same first nucleotide after the two *Eco*RI sites (*i.e.* in *fla*D and *sig*A, the flagellin and sigma-70 factor genes, respectively). The inverted repeat configurations of the two latter fragments would allow their amplification and detection with the Eco-primer alone. Predictions of AFLP simulations were verified experimentally. Actually, repeated amplifications of the same putative *sig*A and *fla*D fragments were obtained in different AFLPs (data not shown).

3.2. Calculation of genetic distances from AFLP data

As demonstrated above, the banding pattern of AFLP strongly depends upon the genome composition. We considered this information in the calculation of a genetic distance between two strains. The genetic distance, d, is defined as the mean rate of nucleotide differences between the sequences of their two genomes (*i.e.* genome divergence), and $(1-d)$ is the probability that a given nucleotidic site is identical in two genomes.

AFLP similarities between couples of strains (x, y) were calculated using the Jaccard coefficient [27]:

$$\hat{S}_{Jxy} = n_{xy}/(n_x + n_y - n_{xy}), \tag{1}$$

in which n_x and n_y are the number of fragments found in strain x and strain y, respectively, and n_{xy} the number of common fragments. $\hat{S}_{Jxy}$ is an estimation of the probability for two strains to have AFLP fragments in common.

Occurrence of a common AFLP fragment between two strains requires at least the total identity of the r nucleotidic sites involved in both restriction and amplification (*i.e.* the sites recognized by endonucleases and discriminant nucleotides, respectively). As a consequence :

$$\hat{S}_{Jxy} = (1-d)^r. \tag{2}$$

Conversely, the genetic distance between two strains is given by:

$$d = 1 - \sqrt[r]{\hat{S}_{Jxy}}. \quad (3)$$

In most cases, with FAFLP, the constraint level is $r = 6 + 4 + 2 = 12$ for hexacutter, quadracutter and the two selective nucleotides, respectively. However $r = 13$ with *X. ampelinus*, and $r = 4 + 4 + (2 \times 3) = 14$ with the simplified AFLP method for Prokarya (Tab. I). This means that the information about genetic relatedness provided by $\hat{S}_{Jxy}$ values depends upon the constraint levels of AFLPs.

3.3. Application of AFLP to various bacterial models

FAFLP was performed on several bacterial species involved in human, animal and plant diseases. Average genetic distances were used to compare genetic diversity of bacteria belonging to closely related genomic species (Tab. II), the same species (Tab. III), or different infra-specific clusters (Tab. IV, V). As a result, average genetic distances within species or within clusters, found to exhibit significantly different calculated distances between species or between clusters (Tab. II, IV, V), were confirmed in all instances by significant bootstrap values in phylogenetic analysis and conversely (data not shown).

In addition, the genetic distance obtained by AFLP was plotted against the genetic distance obtained from PCR-RFLP analysis of the ribosomal region (Fig. 2). A strong correlation ($r^2 = 0.72$) was found between distances determined among bacteria belonging to the same genomic species. The linear regression crossed the ordinate at d = 0.034, showing that ribosomal intergene analysis allowed the discrimination of bacteria with more than 3.4% genome sequence divergence.

4. DISCUSSION

4.1. Standardization and evaluation of the AFLP method

4.1.1. Reliability, versatility and predictability

Methods like RAPD or repetitive PCR used for bacterial fingerprinting are severely limited by the occurrence of rather long DNA regions complementary to PCR primers in the explored genomes. Moreover, the regions must be present by pairs in inverted positions to allow the polymerase chain reaction. To overcome this constraint, primers are designed in repeated regions occurring frequently in genomes; they are also often degenerated to allow mismatches during annealing at low stringency. Resulting amplifications are thus only empirical and unpredictable. Results are also partly unreliable because PCR is

Table II. AFLP genetic distances between and within genomic species of *Agrobacterium* spp. clustered in the biovar 1.

Mean dist.	G1	G2	G4	G6	G7	G8	G9
G1 ($n = 4$)	0.065 ± 0.024	0.168 ± 0.008	0.183 ± 0.017	0.175 ± 0.009	0.176 ± 0.008	0.186 ± 0.011	0.198 ± 0.023
G2 ($n = 4$)		0.089 ± 0.006	0.193 ± 0.013	0.184 ± 0.007	0.195 ± 0.013	0.192 ± 0.008	0.205 ± 0.012
G4 ($n = 2$)			0.075	0.165 ± 0.001	0.161 ± 0.007	0.166 ± 0.006	0.205 ± 0.031
G6 ($n = 2$)				0.053	0.172 ± 0.008	0.159 ± 0.006	0.199 ± 0.014
G7 ($n = 3$)					0.128 ± 0.008	0.165 ± 0.007	0.190 ± 0.022
G8 ($n = 3$)						0.044 ± 0.016	0.183 ± 0.018
G9 ($n = 2$)							0.109

G1 to G9 are genomically different species delineated within the biovar 1 of *Agrobacterium* spp. by Popoff *et al.* [23]. n is the number of isolates tested per species.
Data are average genetic distances ± standard deviation.
Data on and above the diagonal are infra- and inter-specific distances, respectively.

Table III. Comparison of mean genetic distances within different bacterial species.

Bacteria	d max	d mean ± SE	Comments
Agrobacterium spp.	0.137	0.080 ± 0.032	several ribosomal IGS patterns
P. syr. genomovar III	0.064	0.034 ± 0.012	one ribosomal IGS pattern
Salmonella enterica	0.104	0.062 ± 0.008	two serotypes
Staphylococcus aureus	0.147	0.083 ± 0.018	4 agr groups, ≠ PFGE
Xylophilus ampelinus	0.035	0.022 ± 0.006	one ribosomal IGS pattern
Xanthomonas populi	0.074	0.032 ± 0.011	one ribosomal IGS pattern

Table IV. AFLP genetic distances between and within serotypes of *Salmonella enterica*.

Mean distance	Typhimurium DT104	Agona
Typhimurium DT104 ($n = 31$)	0.023 ± 0.019	0.062 ± 0.008
Agona ($n = 36$)	0.062 ± 0.008	0.026 ± 0.015

Typhimurium and Agona are *S. enterica* serotypes.
DT104 is a phage type of serotype Typhimurium.
Data are average genetic distances ± standard error.

Table V. AFLP genetic distances between and within agr groups of *Staphylococcus aureus*.

Mean dist.	I	II	III	IV
I ($n = 56$)	0.075 ± 0.02	0.080 ± 0.015	0.086 ± 0.015	0.086 ± 0.013
II ($n = 43$)		0.074 ± 0.018	0.092 ± 0.016	0.089 ± 0.016
III ($n = 37$)			0.067 ± 0.021	0.091 ± 0.012
IV ($n = 42$)				0.060 ± 0.017

Groups *agr*I to *agr*IV are accessory gene regulator groups as defined by sequencing of the gene *agr*.
Data are average genetic distances ± standard error.
Data on and above the diagonal are infra- and inter-*agr* group distances, respectively.

performed at low stringency. Finally, the number of amplified fragments can be difficultly modified to particular genome constraints of individual species. Conversely, the exploration of the genome diversity by amplification of restriction fragments is a versatile method that can be adjusted to individual specie's characteristics. Fragments are selected by using PCR in stringent conditions with primers containing additional nucleotides in 3' end. Most AFLP patterns are thus reliable [33]. Rare pattern inconsistencies are however problematic when the purpose of the study is the precise characterization of bacterial populations as requested for population genetics and epidemiology. Reliable results

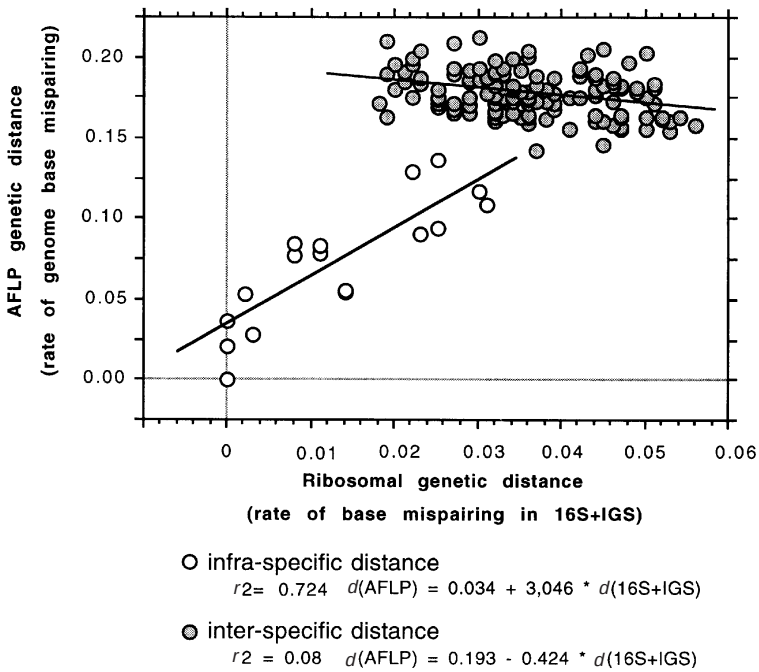

Figure 2. Relationships between the ribosomal genetic distance (16S + IGS) and the genetic distance obtained by AFLP amongst bacteria of the biovar 1 cluster of genomic species of *Agrobacterium* spp.

were obtained by using a standardized AFLP procedure including at least two independent AFLP analyses per strain (Fig. 1).

As AFLP is only based upon the occurrence of restriction sites rather than repetitive sequences, the method can be adjusted according to individual species characteristics particularly their guanine + cytosine contents as well as their response to endonucleases. This could be obtained either experimentally (Tab. I) or after simulations with available genome sequences. The predictive AFLP based upon complete genome sequences is an efficient approach to molecular epidemiology [1]. We have shown, however, that incomplete but representative genome sequences of a given strain could also be used to predict AFLP results and to determine the choice of endonucleases and discriminating nucleotides. Each genome has a specific "signature" defined as the ratios between the observed dinucleotide frequencies and the frequencies expected if neighbours were chosen at random (dinucleotide relative abundances) [5]. We have shown that the occurrences of restriction sites and thus AFLP probabilities in turn also depend upon the genome "signature". In addition, the predictive AFLP demonstrated that dependent fragments can be obtained in different FAFLP. To overcome this difficulty, two discriminating nucleotides were added

to the primer designed for the hexacutter, but not to that for the quadracutter (see *Agrobacterium*, Tab. I), ledding to the amplification of independent fragments. In our opinion, this is more suitable for the accurate estimation of bacterial genetic relatedness.

4.1.2. Genetic distances

We assumed that pattern similarities could give rise to a direct measure of genome similarities. Briefly, as already assumed by Nei and Li [18] for RFLP, AFLP is constrained by the number of nucleotidic sites involved in the reaction. Eq. 3, $(1 - \sqrt[r]{pF})$ is equivalent to the approximated formula $(-Ln(pF)/r)$ given by Nei and Li [18] for studying genetic variations in terms of restriction endonucleases in which r is the number of nucleotides in the recognition sequence of the endonuclease.

The mathematical model given in equation (3) requires a realistic estimate of the relative number of sites involved in AFLP. Rather different values of similarities are provided by the coefficients of Dice and Jaccard that could be used for this purpose [18, 27]. The choice for one or the other index is directed by the fact that fragments are or are not independent characters. As discussed later, the question of fragment independence is questionable in the case of asymmetric AFLP. However, in the symetric AFLP, like the simplified AFLP for Prokarya, any mutation in the constraining nucleotides anihilates amplification. As a result, all detected fragments are assumed to be independent characters. In this case, we proposed to calculate similarities with the Jaccard coefficient, instead of the Dice coefficient.

The present mathematical model showed that the information about genetic relatedness described by S_{xy} values strongly depends upon r, the constraint level of AFLP. For this reason, comparison of AFLPs performed with different constraint levels should be done after transformation in genetic distances with equation (3), and not simply by the usual value $(1 - S_{xy})$, even though this simple transformation has been used to date by most authors [6].

4.2. Genetics diversity of bacterial populations

4.2.1. Species delineation and speciation

Biovar 1 of *Agrobacterium* is a cluster of closely related genomic species gathering in the so-called biovar 1 [23]. In all experiments, the model has shown that the genetic distances calculated between strains belonging to different species were significantly greater than distances calculated between members of the same species (Tab. II). As a result, the differentiation of genomic species should be easily done by AFLP, which thus appeared as a truly potential alternative to DNA/DNA hybridization studies. This was verified for *Agrobacterium* spp. as well as with various other bacterial species like *Pseudomonas syringae* genomovars, *Burkholderia* spp., *Aeromonas* spp., *Xanthomonas* spp.,

Stenotrophomonas spp., etc. [6, 7, 10, 11, 24]. However, thanks to the calculation of genetic distances, our approach also provides an estimate of the genome divergence between species.

Actually, the difference found between and within genetic distances indicates a gap which could be related to the intrinsic base of the speciation itself. The shortest genetic distances between *Agrobacterium* species were 0.143 and the largest within species was 0.137 (Tab. III). This could indicate that, in the biovar 1 cluster of *Agrobacterium* at least, speciation occurred when the genome divergences reached a critical threshold of 14%. This estimation needs to be confirmed with other bacterial systems and more precisely by direct measures of the divergences of complete genome sequences. Nevertheless, this value – 14% genome divergence meaning an average of one difference per 7.1 nucleotides – should be realistic. Due to the codon degeneration, this ratio should lead to a very high level of protein similarity and even identity as it is expected to occur within species. It would also lead to very high DNA/DNA hybridization levels.

4.2.2. Genetic diversity within bacterial species and populations

The high genotype resolution obtained by AFLP will be very useful for precise epidemiological studies as previously suggested with *Chlamydia psittaci* [4]. Providing standardized data capture and processing are used, we could also expect epidemiological lineages could be discriminated among very closely related bacteria belonging for example to the same phage type.

The genetic diversity of bacterial species and bacterial populations might be estimated by the average genetic distance. Within the various species, the average genetic distances were in agreement with the level of diversity obtained with other methods. Bacteria like *Agrobacterium* spp. or *S. aureus* already found to be largely diverse by PFGE or ribosomal intergene analysis, had the largest average genetic distances. Conversely, other bacteria like *P. syringae, X. ampelinus, X. populi* had a smaller range of genetic distances (Tab. III). Models providing the largest infra-specific diversity showed an average genetic distance within species of 0.08 ± 0.02 with a maximum of about 0.14.

Lower taxonomic levels corresponding to infra-specific clusters were also analyzed. Strains belonging to the Agona serotype of *S. enterica* or to the phage type DT104 of the serotype Typhimurium, showed average genetic distances of 0.026 and 0.023, respectively (Tab. IV). This was significantly lower than previously found within entire species, or between the two serotypes (0.06). *S. enterica* serotypes could be thus readily discriminated by AFLP. This was confirmed by complete phylogenetic analysis which showed that isolates of the two serotypes fell repeatedly into two clearly separated clusters after data bootstrappings (data not shown).

Within *S. aureus, agr* groups were defined on the basis of sequence differences of a single gene, the *agr* locus. AFLP as well as PFGE both showed that groups *agr*III and *agr*IV are likely to be monoclonal forming single clusters,

contrary to the polyphyletic groups *agr*I and *agr*II (data not shown). However, *agr*III and *agr*IV clusters were not confirmed by significant bootstrap values (allowed with AFLP but not PFGE data). Moreover, genetic distances calculated between and within *agr* groups were not significantly different (Tab. V). This could indicate an insufficient number of AFLP data (*i.e.* AFLP fragments) as required for bootstrap. However, another hypothesis was that since agr groups are defined upon a single locus, this single locus could have been parasexually exchanged.

The question was asked whether AFLP could also be used for population genetics. The important point for that purpose is the independence of characters. From a theroretical point of view, the character independence could be ascertained only for the simplified AFLP for Prokarya, but not in all instances with FAFLP or IRS-PCR. The simplified AFLP for Prokarya is a symetric AFLP which is based upon the occurrence of inverted regions flanking the amplified fragments. As a consequence, any mutation in the flanking regions prevents their amplification, ascertaining thus that two different fragments obtained with two different strains are independent characters. On the other hand, FAFLP and IRS-PCR are asymmetric AFLPs using two endonucleases. In both cases, the detection (FAFLP) or the amplification (IRS-PCR) of a fragment is determined by only one of its extremities, generally those corresponding to the hexacutter. As a result, only mutations in the constraining region related to the hexacutter would prevent the detection (or the amplification) of the fragment. Conversely, mutations in the quadracutter region could lead to shorter or larger fragments. In the latter case, two different fragments in two different strains could originate from the same genome region, and would, thus, be two alleles of the same character. Increasing the constraint to the primer designed for the hexacutter extremity would likely increase the fragment independence. This is the reason why two discriminating nucleotides with the *Eco*RI primer were used with *Agrobacterium* spp. We verified experimentally that this location of the discriminating nucleotides produced less dependent fragments especially with combined FAFLP (data not shown). Nevertheless, the extent AFLP fragments are truly independent characters with FAFLP or IRS-PCR is difficult to determine. Thus, in absence of experimental evidences of fragment independence, the use of FAFLP or IRS-PCR data for the genetical analysis of populations must be done with care.

In terms of population genetics, the abundance of genomic traits involved in AFLP could be used to evaluate how clonal bacteria are. However, in our experience we found great linkage disequilibrium with all bacterial models, even when horizontal gene transfers were ascertained. Actually, FAFLP was done with four isogenic strains of *Agrobacterium* containing zero, one or the two large plasmids (the Ti and the cryptic plasmids), which yielded eleven and nine specific fragments, respectively (data not shown). In spite of the fact that the two plasmids could be sexually transmitted independently in our model, a strong multilocus linkage disequilibrium was found by using the

model of Maynard-Smith *et al.* [15] with an association index $I_a = [V_o/V_e] - 1 = 9.32$. This discrepancy, between the expected and the calculated results suggested that using AFLP data with the model of Maynard-Smith was not relevant to demonstrate that sexuality must occur between bacteria. The cause for the discrepancy is perhaps related to the the fact that not all FAFLP fragments were independent as indicated above. On the other hand, several characters originating from the same plasmid were detected by AFLP. These linked fragments led to the estimation of a strong linkage disequilibrium. The clonality of bacterial populations still requires further research, but it is likely that AFLP could be used for this purpose if fragment independence is verified.

4.2.3. Comparison of different pathogenic models

The present work gave the opportunity to compare different pathogenic species by using the mean genome divergence of populations as a common scale to measure population diversities. One interesting finding is that in spite of a quest for a maximal strain diversity, the present study showed that bacteria isolated from epiphytic diseases (*P. syringae, X. ampelinus, X. populi*) were much less diverse than bacteria isolated from other biotopes. The whole species *X. ampelinus* for instance was no more diverse than the single phage type DT104 of *S. enterica* Typhimurium. Like DT104, the epiphytic bacteria appeared poorly diverse, perhaps as a consequence of the particular constraints of this habitat.

On the other hand, the soil pathogen *Agrobacterium* is very diverse, and includes several species with several genotypes identified by ribosomal intergene PCR-RFLP [22]. Several genotypes differentiated by the ribosomal intergene of *Agrobacterium* generally coexist in tumors. We found that different ribosomal intergenes occurred only for a minimal 3.4% genome divergence (Fig. 2), thus much more than the average genetic diversity found in epiphytic bacteria (Tab. III). This indicates that agrobacterial populations found in a single biotope are probably much more diverse than the entire species of epiphytic bacteria. This is probably because the agrobacterial populations – but not the epiphytic populations – are frequently connected to the large soil reservoir of biodiversity, and more precisely because the pathogenicity of this bacterium is determined by a conjugative Ti plasmid which can diffuse in very diverse bacterial genotypes [29].

ACKNOWLEDGEMENTS

The authors wish to thank Marie-Andrée Poirier for skillful technical assistance. C.M. is a research fellow of the Institut national de la recherche agronomique. This work was made possible by a grant from the *Bureau des*

ressources génétiques to X.N., C.M., E.C.D., J.E., J.B., and by the INCO-DC European project ERB1C18CT970198 to X.N.

REFERENCES

[1] Arnold C., Metherell L., Clewley P., Stanley J., Predictive modelling of fluorescent AFLP: a new approach to the molecular epidemiology of *E. coli*, Res. Microbiol. 150 (1999) 33–44.

[2] Asai T., Zaporojets D., Squires C., Squires C.L., An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria, Proc. Natl. Acad. Sci. USA 96 (1999) 1971–1976.

[3] Ausubel F.M., Brent R., Kingston R.E., Moore D.D., Smith J.A., Seidman J.G., Struhl K., Current protocols in molecular biology, Greene Publishing Associates, Wiley Interscience, New-York, 1992.

[4] Boumedine K.S., Rodolakis A., AFLP allows the identification of genomic markers of ruminant *Chlamydia psittaci* strains useful for typing and epidemiological studies, Res. Microbiol. 149 (1998) 735–744.

[5] Campbell A., Mrazek J., Karlin S., Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA, Proc. Natl. Acad. Sci. USA 96 (1999) 9184–9189.

[6] Clerc A., Manceau C., Nesme X., Comparison of Random Amplified Polymorphism DNA (RAPD) with Amplified Fragment Lenght Polymorphism (AFLP) to assess the genetic diversity and genetic relatednesses within the genospecies III of *Pseudomonas syringae*, Appl. Environ. Microbiol. 64 (1998) 1180–1187.

[7] Coenye T., Schouls L.M., Govan J.R.W., Kersters K., Vandamme P., Identification of *Burkholderia* species and genomovars from cystic fibrosis patients by AFLP fingerprinting, Int. J. Syst. Bacteriol. 49 (1999) 1657–1666.

[8] Felsenstein J., PHYLIP (Phylogeny Inference Package) version 3.5c, Distributed by the author, Department of genetics, University of Whashington, Seattle, USA, 1993.

[9] Gardan L., Shafik H., Belouin S., Broch R., Grimont F., Grimont P.A., DNA relatedness among the pathovars of *Pseudomonas syringae* and description of *Pseudomonas tremae* sp. nov. and *Pseudomonas cannabina* sp. nov. (ex Sutic and Dowson 1959), Int. J. Syst. Bacteriol. 49 (1999) 469–478.

[10] Hauben L., Vauterin L., Moore E.R., Hoste B., Swings J., Genomic diversity of the genus *Stenotrophomonas*, Int. J. Syst. Bacteriol. 49 (1999) 1749–1760.

[11] Huys G., Coopman R., Janssen P., Kersters K., High-resolution genotypic analysis of the genus *Aeromonas* by AFLP fingerprinting, Int. J. Syst. Bacteriol. 46 (1996) 572–580.

[12] Jarraud S., Lyon G.J., Figueiredo A.M., Lina G., Vandenesch F, Etienne J, Muir TW, Novick RP.Exfoliatin-producing strains define a fourth *agr* specificity group in *Staphylococcus aureus*. J. Bacteriol. 22 (2000) 6517–6522.

[13] Lina G., Vandenesch F., Etienne J., Kreiswirth B., Fleurette J., Comparison of coagulase-negative staphylococci by pulsed-field gel electrophoresis, FEMS Microbiol. Lett. 92 (1992) 133–138.

[14] Maidak B.L., Cole J.R., Parker C.T.Jr., Garrity G.M., Larsen N., Li B., Lilburn T.G., McCaughey M.J., Olsen G.J., Overbeek R., Pramanik S., Schmidt

T.M., Tiedje J.M., Woese C.R., A new version of the RDP (Ribosomal Database Project), Nucleic Acids Res. 27 (1999) 171–173.

[15] Maynard-Smith J., Smith N.H., O'Rourke M., Spratt B.G., How clonal are bacteria?, Proc. Natl. Acad. Sci. USA 90 (1993) 4384–4388.

[16] Mazurek G., Reddy V., Marston B.J., Haas W.H., Crawford J.T., DNA fingerprinting by infrequent-restriction-site amplification, 1996. J. Clin. Microbiol. 34 (1996) 2386–2390.

[17] Navarro E., Rouvier C., Normand P., Domenach A.M., Simonet P., Prin Y., Evolution of *Frankia-Casuarinaceae* interactions, Genet. Select. Evol. 30 (Suppl. 1) (1998) S357–S372.

[18] Nei M., Li. W.H., Mathematical model for studying genetic variation in terms of restriction endonucleases, Proc. Natl. Acad. Sci. USA 76 (1979) 5269–5273.

[19] Nesme X., Steenackers M., Steenackers V., Picard C., Ménard M., Ridé S., Ridé M., Differential host-pathogen interactions among clones of poplar and strains of *Xanthomonas populi* pv. *populi*, Phytopathology 84 (1994) 101–107.

[20] Nesme X., Vaneechoutte M., Orso S., Hoste B., Swings J., Diversity and genetic relatedness within genera *Xanthomonas* and *Stenotrophomonas* using restriction endonuclease site differences of PCR-amplified 16SrRNA gene, System. Appl. Microbiol. 18 (1995) 127–135.

[21] Oger P., Dessaux Y., Petit A., Gardan L., Manceau C., Chomel C., Nesme X., Validity sensitivity, and resolution limit of PCR-RFLP analysis of the *rrs* (16S rRNA gene) as a tool to identify soil-borne and plant-associated bacterial populations, Genet. Select. Evol. 30 (Suppl. 1) (1998) S311–S332.

[22] Ponsonnet C., Nesme X., Identification of *Agrobacterium* strains by PCR-RFLP analysis of pTi and chromosomal regions, Arch. Microbiol. 161 (1994) 300–309.

[23] Popoff M.Y., Kersters K., Kiredjian M., Miras I., Coynault C., Position taxonomique des souches de *Agrobacterium* d'origine hospitalière, Ann. Microbiol. (Inst. Pasteur) 135A (1984) 427–442.

[24] Rademaker J.L., Hoste B., Louws F.J., Kersters K., Swings J., Vauterin L., Vauterin P., de Bruijn F.J., Comparison of AFLP and rep-PCR genomic fingerprinting with DNA-DNA homology studies: *Xanthomonas* as a model system, Int. J. Syst. Evol. Microbiol. 2 (2000) 665–677.

[25] Riffard S., Lo Presti F., Vandenesch F., Forey F., Reyrolle M., Etienne J., Comparative analysis of infrequent-restriction-site PCR and pulsed-field gel electrophoresis for epidemiological typing of *Legionella pneumophila* serogroup 1 strains, J. Clin. Microbiol. 36 (1998) 161–167.

[26] Sambrook J.E., Fristsch F., Maniatis T., Molecular cloning: a laboratory manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989.

[27] Sokal R.R., Sneath P.H.A., The construction of a taxonomic system, in: Principles of numerical taxonomy, chap. 7, Freeman, San Fransisco, USA, 1989, pp. 169–210.

[28] Stackebrandt E., Goebel B.M., Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology, Int. J. Syst. Bacteriol. 44 (1994) 846–849.

[29] Teyssier-Cuvelle S., Mougel C., Nesme X., Direct conjugal transfers of Ti plasmid to soil microflora. Mol. Ecol. 8 (1999) 1273–1284.

[30] Thioulouse J., Chessel D., Dolédec S., Olivier J.M. ADE-4: a multivariate analysis and graphical display software. Stat. Comput. 7 (1997) 75–83.

[31] Vos P.R., Hogers R., Bleeker M., Reijans M., van der Lee T., Hornes M., Fridjers A., Pot J., Peleman J., Kuiper M., Zabeau M., AFLP: a new technique for DNA fingerprinting, Nucleic Acids Res. 23 (1995) 4407–4414.

[32] Wayne L.G., Brenner D.J., Colwell R.R., Grimont P.A.D., Kandler O., Krichevsky M.I., Moore L.H., Moore W.E.C., Murray R.G.E., Stackebrandt E., Starr M.P., Trüper H.G., Report of the ad hoc comittee on reconciliation of approaches to bacterial systematics, Int. J. Syst. Bacteriol. 37 (1987) 463–464.

[33] Willems A.F., Doignon-Bourcier F., Coopman R., Hoste B., de Lajudie P., Gillis M., AFLP fingerprint analysis of *Bradyrhizobium* strains isolated from *Faidherbia albida* and *Aeschynomenes* species, Syst. Appl. Microbiol. 23 (2000) 137–147.

[34] Yap W.H., Zhang Z., Wang Y., Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon, J. Bacteriol. 181 (1999) 5201–5209.

[35] Zabeau M., Vos P., Selective restriction fragment amplification: a general method for DNA fingerprinting, Publication 0 534 858 A1, European Patent Office, Münich, 1993.