International Journal of Systematic and Evolutionary Microbiology (2002), **52**, 573–586 DOI: 10.1099/ijs.0.01956-0

A mathematical method for determining genome divergence and species delineation using AFLP

Christophe Mougel,[1] Jean Thioulouse,[2] Guy Perrière[2] and Xavier Nesme[1]

[1] Ecologie Microbienne UMR-CNRS 5557 and INRA, Bât. G. Mendel, Université Claude Bernard-Lyon 1, 43 bd du 11 novembre 1918, 69622 Villeurbanne cedex, France

[2] Biométrie, Génétique et Biologie des Populations UMR CNRS 5558, Université Claude Bernard-Lyon 1, Villeurbanne, France

Author for correspondence: Xavier Nesme. Tel: +33 4 72448289. Fax: +33 4 72431223. e-mail: nesme@univ-lyon1.fr

The delineation of bacterial species is presently achieved using direct DNA–DNA relatedness studies of whole genomes. It would be helpful to obtain the same genomically based delineation by indirect methods, provided that descriptions of individual genome composition of bacterial genomes are obtained and included in species descriptions. The amplified fragment length polymorphism (AFLP) technique could provide the necessary data if the nucleotides involved in restriction and amplification are fundamental to the description of genomic divergences. Firstly, in order to verify that AFLP analysis permits a realistic exploration of bacterial genome composition, the strong correspondence between predicted and experimental AFLP data was demonstrated using *Agrobacterium* strain C58 as a model system. Secondly, a method is proposed for determining current genome mispairing and evolutionary genome divergences between pairs of bacteria, based on arbitrary sampling of genomes by using AFLP. The measure of current genome mispairing was validated by comparison with DNA–DNA relatedness data, which itself correlates with base mispairing. The evolutionary genome divergence is the estimated rate of nucleotide substitution that has occurred since the strains diverged from a common ancestor. Current genome mispairing and evolutionary genome divergence were used to compare members of *Agrobacterium*, used as a model of closely related genomic species. A strong and highly significant correlation was found between calculated genome mispairing and DNA–DNA relatedness values within genomic species. The canonical 70% DNA–DNA hybridization value used to delineate genomic species was found to correspond to a range of current genome mispairing of 13–13·6%. These limits correspond to 0·097 and 0·104 nucleotide substitutions per site, respectively. In addition, experimental data showed that the large Ti and cryptic plasmids of *Agrobacterium* had little effect on the estimation of genome divergence. Evolutionary genome divergence was used for phylogenetic inferences. Data showed that members of the same genomic species clustered consistently, as supported by bootstrap resampling. On the basis of these results, it is proposed that the genomic delineation of bacterial species could be based, in future, on phylogenetic groups supported by bootstraps and genome descriptions of individual strains, obtained by AFLP analysis, recorded in accessible databases; this approach might eventually replace DNA–DNA hybridization studies.

Keywords: AFLP, genome divergence, evolutionary distance, phylogeny, *Agrobacterium* spp.

Abbreviations: AFLP, amplified fragment length polymorphism; RBR, relative binding ratio.

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

INTRODUCTION

The consensus definition for bacterial species is currently based on DNA–DNA relatedness studies (Wayne *et al.*, 1987). The genomic basis of bacterial species definition will still remain the 'gold standard' in bacterial taxonomy because it represents the underlying properties of bacteria. The zoological definition of a species is 'groups of potentially interbreeding populations that are reproductively isolated from other groups'. This definition stresses the requirement for reproductive isolation, which is also relevant to bacterial systematics. Sexuality exists in the *Prokarya* through conjugation, transformation and transduction. Even if the wide taxonomic range of bacterial recipients of some plasmids precludes the idea of complete reproductive isolation, the relatively small size of these autonomous genetic elements limits their impact on species delineation. In contrast, the entire genome is involved in integrative transformation. Integrative transformation is, however, limited to close relatives, because DNA recombination in the *Prokarya* is strictly controlled by both the SOS and mismatch repair systems (Matic *et al.*, 1995). Thus, the lack of homologous recombination plays the same role in prokaryotes as sexual incompatibility does in animals in maintaining the genome integrity of a species. As a result, speciation in bacteria is determined by a phenomenon that hinders recombination between distantly related bacteria (Matic *et al.*, 1995). Overall genome mispairing, currently measured by using DNA–DNA relatedness, is a fundamental concept in bacterial systematics.

DNA–DNA relatedness studies only establish direct genome comparisons, and the techniques involved are time-consuming and are difficult to apply to numerous strains (as is required for population genetics). Therefore, in order to find a method for comparing a large number of bacteria, genetic markers have been defined; this has aided the construction of appropriate databases suitable for indirect comparisons of bacteria (Perrière *et al.*, 2000). *rrs* (16S rRNA) gene sequences have received considerable attention for that reason (Maidak *et al.*, 2001). However, the resulting phylogeny is not sensitive enough to guarantee correct delineation of genomic species and, a fortiori, to study infraspecific diversity (Stackebrandt & Goebel, 1994; Fox *et al.*, 1992). Furthermore, as the method is used to consider a single locus, it is very sensitive to interspecific exchanges that introduce discrepancies between genomic and *rrs*-based phylogenies (Cohan, 1994; Yap *et al.*, 1999; Wang & Zhang, 2000). Alternative methods using multiple loci, such as multilocus sequencing (Spratt, 1999), or techniques using random or arbitrary amplification of genome fragments, such as randomly amplified polymorphic DNA analysis, the rep-PCR technique or amplified fragment length polymorphism (AFLP) analysis, have been proposed as methods complementary to DNA–DNA hybridization (Caetano-Anollés *et al.*, 1991; Versalovic *et al.*, 1991; Zabeau & Vos, 1993). Among those later methods, AFLP has received considerable support for taxonomic studies (Janssen *et al.*, 1996; Clerc *et al.*, 1998; Hauben *et al.*, 1999; Rademaker *et al.*, 2000) because it yields a clear delineation of bacteria belonging to the same genomic species (Clerc *et al.*, 1998). Validation of AFLP analysis as an alternative to DNA–DNA relatedness studies requires AFLP to be a descriptor of the genome structure and the demonstration of its relationship to DNA–DNA hybridization values.

The genus *Agrobacterium* is a good model for investigating the relatedness between DNA–DNA hybridization values and genome similarities obtained from AFLP. Conventional morpho-biochemical criteria recognize three biovars, as well as agrobacteria too distantly related to be included in these biovars (Keane *et al.*, 1970; Kersters *et al.*, 1973; Panagopoulos *et al.*, 1978; Bouzar *et al.*, 1995a; Bouzar & Jones, 2001). DNA–DNA hybridization studies supported by 16S rDNA analysis showed that the biovars correspond well to genomic species, corresponding (in two instances) to well-characterized nomenspecies: biovar 3, corresponding to *Agrobacterium vitis*, and biovar 2, corresponding to *Agrobacterium rhizogenes* (Ophel & Kerr, 1990; Sawada *et al.*, 1993; Yanagi & Yamasato, 1993). The case of biovar 1 is, however, much more complex. Proposals to define a single species for members of biovar 1 – *Agrobacterium radiobacter* (Sawada *et al.*, 1993), *Agrobacterium tumefaciens* (Bouzar, 1994) or *Rhizobium radiobacter* (Young *et al.*, 2001) – are not relevant in view of the genomic diversity of the biovar. De Ley *et al.* (1973), De Ley (1974) and Popoff *et al.* (1984) have already shown that agrobacteria belonging to biovar 1 group into at least nine genomic species that have not yet received accepted nomenspecies status. The nine genomic species that cluster in biovar 1 are separated by canonical values of relative DNA–DNA hybridization and ΔT_m. For this reason, biovar 1 as studied by Popoff *et al.* (1984) appeared to be a relevant model for the purpose of the present study because DNA–DNA hybridizations were performed with closely related bacteria belonging to the same biovar cluster along with the more distantly related species *Agrobacterium rubi* and *A. rhizogenes*.

In this work, experimental AFLP analyses were performed with agrobacteria in order to provide a dataset to be used to verify our AFLP predictions. AFLP predictions were made on the basis of constraining nucleotides. These constraining nucleotides are the basis of our mathematical method for determining genome base mispairing. This method was validated by comparing our prediction of current genome mispairing based on AFLP with that based on DNA–DNA hybridization (determined previously by Popoff *et al.*, 1984). A method for estimating evolutionary genome divergence was also proposed and then used for genome-based phylogenetic studies.

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

Table 1. *Agrobacterium* strains used in this study

Abbreviations: ATCC, American Type Culture Collection, Manassas, VA, USA; CIP, Collection de l'Institut Pasteur, Paris, France; CFBP, Collection Française de Bactéries Phytopathogènes, INRA, Angers, France; LMG, Bacteria Collection Laboratorium voor Microbiologie, Universiteit Gent, Gent, Belgium; SCRI, Scottish Crop Research Institute, Invergowrie, Dundee, UK. Genomic groups were determined by Popoff *et al.* (1984) or, for *A. vitis*, by Ophel & Kerr (1990). Strain TT111, strain C58 and its derivatives were assigned to genomic groups on the basis of other DNA–DNA.relatedness studies. *A. tumefaciens*, *A. rhizogenes* and *A. vitis* are designated according to Bouzar (1994), Sawada *et al.* (1993) and Ophel & Kerr (1990), respectively.

Strain	Received from/as:	Relevant properties	*rrs* accession no.*
Genomic group 1 (*Agrobacterium* sp.)			
TT111	CFBP 4716	Crown gall, USA	X67223
S56	LMG 321	Crown gall, USA	**AJ389895**
S377	LMG 326	Plant	(AJ389895)
S4	LMG 318	Plant	
ATCC 4720	LMG 182	Black raspberry, USA	
Genomic group 2 (*Agrobacterium* sp.)			
CIP 497-74	CFBP 2884	Human blood, France (Kiredjian, 1979)	**AJ389894**
CIP 28-75	M. Kiredjian, France	Human urine, France (Kiredjian, 1979)	(AJ389894)
CIP 127-76	M. Kiredjian, France	Human urine, Switzerland (Kiredjian, 1979)	(AJ389894)
CIP 43-76	M. Kiredjian, France	Human urine, France (Kiredjian, 1979)	(AJ389894)
Genomic group 3 (*Agrobacterium* sp.)			
CIP 111-78	M. Kiredjian, France	Human cephalo-rachidian liquid, France (Kiredjian, 1979)	**AJ389897**
Genomic group 4 (*Agrobacterium tumefaciens*)			
A. tumefaciens ATCC 23308^T	ATCC	Apple seedling, IA, USA (= Braun B6^T)	D12784
A. radiobacter ATCC 19358^T	ATCC	Soil	(AJ389897)
Genomic group 6 (*Agrobacterium* sp.)			
NCPPB 925	LMG 225	*Dahlia* sp., South Africa	(AJ012209)
Zutra F/1	LMG 296	*Dahlia* sp., Israel	**AJ389909**
Genomic group 7 (*Agrobacterium* sp.)			
RV3	LMG 317	No information available	**AJ389903**
Zutra 3/1	LMG 198	*Malus* sp., Israel	(AJ389903)
NCPPB 1641	LMG 228	*Flacourtia ramontchi*, UK	(AJ389903)
Genomic group 8 (*Agrobacterium* sp.)			
C58	CFBP 1903	Cherry tree gall, NY, USA	AJ012209
C58C1	CFBP 1902	C58 cured of Ti plasmid pTiC58	
C58pAt−	Y. Dessaux, France	C58 cured of cryptic plasmid pAtC58	
GMI 9023	T. Huguet, France	C58 cured of both pTiC58 and pAtC58 (Rosenberg & Huguet, 1984)	
T37	LMG 332	Walnut gall, CA, USA	(AJ012209)
TT9	LMG 64	No information available	
Mushin 6	LMG 201	Hop gall, Australia	(AJ012209)
Genomic group 9 (*Agrobacterium* sp.)			
Hayward 0363	LMG 27	John Innes potting soil, Australia	(AJ389894)
Hayward 0362	LMG 26	John Innes potting soil, South Australia	(AJ389894)
Undetermined genomic group (*Agrobacterium larrymoorei*)			
AF3.44	H. Bouzar, USA	*Ficus benjamina*, FL, USA (Bouzar *et al.*, 1995a)	Z30542
Fb33	A. Zoina, Italy	*Ficus benjamina*, Salerno, Italy	(Z30542)
Fb72	A. Zoina, Italy	*Ficus benjamina*, Salerno, Italy	(Z30542)
Genomic group 11 (*Agrobacterium rubi*)			
LMG 17935^T	LMG	*Rubus ursinus*, USA (= TR3^T)	D12787
***Agrobacterium vitis* (another genomic group)**			
CFBP 2617	CFBP	*Vitis vinifera*, France	**AJ389911**
CFBP 2678	CFBP	*Vitis vinifera*, France	**AJ389910**
CFBP 2736	CFBP	*Vitis vinifera*, Australia (= Kerr K305)	**AJ389912**

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

Table 1 (*cont.*)

Strain	Received from/as:	Relevant properties	*rrs* accession no.*
Genomic group 10 (*Agrobacterium rhizogenes*)			
CFBP 2408^T	CFBP	Apple, USA (= ATCC 11325^T)	D12788
Other strains			
SCRI 551	M. Perombelon, UK	Unassigned strain isolated from *Rubus* sp.	

* Accession numbers for sequences determined in this study are in bold. Numbers in parentheses represent sequences found to be identical to the accession shown.

METHODS

Bacterial strains and media. Bacteria used in the present study are members of the genus *Agrobacterium* (Table 1). Most of them belong to the genomic species (= genomic groups) determined by Popoff *et al.* (1984). This set of strains included type strains of *A. tumefaciens*, *A. radiobacter* (both of which are in genomic group 4), *A. rhizogenes* (group 10) and *A. rubi* (group 11). Other agrobacteria used in the study were as follows: (i) strains that could be attributed to the former genomic groups according to other DNA–DNA relatedness studies (De Ley *et al.*, 1973; De Ley, 1974); (ii) strains of *A. vitis*; and (iii) members of a novel species (isolated from fig trees) called *Agrobacterium larrymoorei* (Bouzar & Jones, 2001). The model also included derivatives of the standard strain C58 cured either of the Ti plasmid pTiC58 (C58C1) or of the cryptic plasmid pAtC58 (C58pAt−) or cured of both plasmids (GMI 9023).

Bacteria were grown routinely at 28 °C on mannitol/glutamate agar, as described by Bouzar *et al.* (1995b); LB broth (low salt; Gibco-BRL) was used to cultivate bacteria before DNA preparation.

Isolation and quantification of DNA. Total bacterial DNAs were extracted by using the Dneasy tissue kit (Qiagen) according to the manufacturer's instructions. DNAs were quantified by using MOLECULAR ANALYST software (Bio-Rad) according to the manufacturer's instructions.

Predictive AFLP. AFLP fragments were predicted as described by Arnold *et al.* (1999) by looking for DNA restriction sites and selective nucleotides, using a word-processor and spreadsheet. The non-redundant published sequences of strain C58 available in GenBank and used for the prediction represented approximately 120 kb (2·4 %) of the chromosomes (AF024659, AF033856/U38977/L24117, AF039940, AF044495, AF090987, AF111855, J03678, L07902/M36776, L18860, L38609, M24198, M38670, M58472, U32867, U59485/L63540, U91632, U95165, X68263, X69388, X87113, X95676) and 120 kb (50 %) of pTiC58 (J03320, AF034769, AF010180, AF057718, L22207, AJ237588/AF065243/AF126445) or the complete sequence of the related pTiSakura (AB016260). The fragment-size data were adjusted to allow for the addition of adaptors.

AFLP and data processing. The AFLP Microbial Fingerprinting kit of Applied Biosystems was used according to the recommendations of the manufacturer. AFLP analyses were performed using a selective primer containing two selective nucleotides at the 3′ end (-CA or -CC or -CG or -CT) of the fluorescent *Eco*RI-primer core (GACTGCGTACCAAT TC) and the non-fluorescent, non-selective *Mse*I primer (GATGAGTCCTGAGTAA). Primer sets and the resulting AFLPs are identified in this work as *Eco*RI+CA/*Mse*I+0, *Eco*RI+CC/*Mse*I+0, *Eco*RI+CG/*Mse*I+0 and *Eco*RI+CT/*Mse*I+0, respectively. The selective amplification of the two fragments released from *rrs* containing an *Eco*RI site was obtained as follows: (i) with a selective *Eco*RI primer with CA at the 3′ end used together with a selective *Mse*I primer with GTCA at the 3′ end (*Eco*RI+CA/*Mse*I+GTCA condition) for the fragment left of the *Eco*RI site; (ii) with a selective *Eco*RI primer containing CG at the 3′ end used together with a selective *Mse*I primer with TGCG at the 3′ end (*Eco*RI+CG/*Mse*I+TGCG condition) for the fragment right of the *Eco*RI site. DNA samples processed accordingly were loaded singly or in pools of three with different fluorescent dyes on a 6 % sequencing gel (Gel-Mix 6, containing 5·7 % acrylamide, 0·3 % *bis*-acrylamide and 7 M urea; Gibco-BRL) with an ABI PRISM 373 sequencer (Perkin-Elmer). The GENESCAN ANALYSIS software of Perkin-Elmer was used to extract data from electrophoregrams. At least two independent AFLP replicates were performed in order to retain, finally, only those peaks detected in all replicates. Assignment of fragments to discrete categories was done by using an iterative method called Lis, as follows: at each iteration, the program (i) finds the largest fragment; (ii) creates a category corresponding to the truncated value of the fragment size and (iii) assigns the same value to all fragment sizes between the maximum size and the maximum size minus 1 bp. The program continues with the next maximum. The program LecPCR was developed to transform fragment-size matrices into tabular binary matrices. The program DistAFLP calculates similarities (Jaccard or Dice index), current genome mispairing and evolutionary genome divergence with or without bootstrap resamplings. The programs LecPCR and DistAFLP are available 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 (Thioulouse *et al.*, 1997).

AFLP similarities between pairs of bacteria. Similarities of AFLP patterns were calculated in two ways. The pattern similarity used for the comparison of current genome traits was determined using the Jaccard index (S_{Jxy}). The similarity to the common ancestor of two strains required for phylogenetic studies was determined by using the Dice index (S_{Dxy}). The indexes are calculated as

$$S_{Jxy} = n_{xy}/(n_{xy}+\Delta_{xy}) \quad \text{(equation 1)}$$

$$S_{Dxy} = n_{xy}/[n_{xy}+(\Delta_{xy}/2)] \quad \text{(equation 2)}$$

in which n_{xy} is the number of fragments common to both strains x and y, and Δ_{xy} is the number of fragments found only in x or only in y.

Mathematical model for estimating genome divergence from AFLP data. Let us begin with the assumption that the occurrence of a common AFLP fragment in two strains requires the identity of r nucleotide sites (i.e. constraining

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

nucleotides) involved in both restriction and amplification (i.e. the sites recognized by endonucleases and selective nucleotides, respectively). As a consequence, we determined the proportion of common AFLP fragments, pF, as

$$pF = (1-d)^r \quad \text{(equation 3)}$$

where d is the rate of base mispairing between two genomes (i.e. genome mispairing). Conversely, we calculated the proportion of nucleotide differences as

$$d = 1-(pF)^{1/r} \quad \text{(equation 4)}$$

A measure of the current genome mispairing is given by d_{Jxy} after substitution of S_{Jxy} for pF in equation 4, because the likelihood for the current number of sites n_{xy} and Δ_{xy} is maximized by using the Jaccard index (Felsenstein, 2002).

For optimal phylogenetic studies, d_{Dxy} is calculated after substitution of S_{Dxy} for pF in equation 4, because the Dice index is suited for phylogenetic studies, since it assumes that all sites that are shared between two strains were present in a common ancestor halfway between them (Felsenstein, 2002). The evolutionary distance is then corrected to account for unobserved substitutions by using the standard Jukes–Cantor model (Swofford *et al.*, 1996), which assumes equal rates of substitution between all pairs of bases.

Thus, the evolutionary genome divergence, expressed as the number of nucleotide substitutions per site, was estimated as

$$\hat{t} = -(3/4)\ln[1-(4/3)d_{Dxy}] \quad \text{(equation 5)}$$

In equation 5, the maximum expected number of nucleotide differences per site is $d_{Dxy} = 0{\cdot}75$. If d_{Dxy} equals or exceeds this value, the distance becomes undefined. However, this is unlikely to occur with standard AFLP because $d_{Dxy} = 0{\cdot}75$ corresponds to $pF < 6\times10^{-8}$ for $r = 12$, which is obtained only if x and y have no fragment in common at all.

Correction for fragment dependence. Bands seen on electrophoregrams that differ between strains x and y (∂xy) are more numerous than AFLP fragments that really differ between strains (Δxy) because, with some point mutations, bands move but do not disappear. For genomes that have diverged substantially, this may not be a problem – fragments that differ between species may tend to have their presence or absence fixed within each species. For within-species inferences, we determined a correction for fragment dependence, K, that is given by:

$$\Delta_{xy} = K\partial_{xy} = (1/r)[E/(1+pEE.pE) + M/(1+pMM.pM) + N]\partial_{xy} \quad \text{(equation 6)}$$

in which E and M are the numbers of nucleotides in the *Eco*RI and *Mse*I sites, respectively, N is the number of selective nucleotides (note: $E+M+N = r$), pEE and pMM are the proportions of fragments that produce two different bands in AFLP patterns following mutations in the *Eco*RI or *Mse*I restriction site, respectively, and pE and pM are the respective probabilities of common *Eco*RI or *Mse*I sites in both genomes. pEE and pMM were estimated from the proportion of fragments in agrobacterial sequences with identical *Eco*RI extremities and identical *Mse*I extremities, respectively, that were short enough to be detected by the sequencing apparatus (i.e. below 500 bp with ABI PRISM 373). Note that '*Eco*RI and *Mse*I extremities' means not only nucleotides involved in restriction sites but also the selective nucleotides added to the AFLP primers. $pE = (1-d)^E$ and $pM = (1-d)^M$ are maximal for nearly identical strains and decrease with genome mispairing; thus, fragment dependence tends to be negligible between species. For simplicity, we use $pE = pM = 1$ in the present version of DistAFLP.

Phylogenetic analyses. Phylogenetic analyses were performed with the phylogenetic inference package PHYLIP (Felsenstein, 1993). Bootstrapped dendrograms were constructed by using the neighbour-joining method (Felsenstein, 1985; Saitou & Nei, 1987). In the case of bootstraps, distance data were not corrected for fragment dependence as explained above, but involved sampling K% of fragments at random among the 100%, according to the proposal of Felsenstein (1985), to cope with fragment dependence. Parsimony and maximum-likelihood analyses were done by using the DOLLO and RESTML programs of PHYLIP, respectively.

***rrs* sequencing and analysis of the ribosomal intergenic spacer.** Sequencing was performed by the company Genome Express with *rrs* amplicons obtained as described previously (Bouzar *et al.*, 1995a). The region lying between *rrs* and *rrl* (16S and 23S rRNA genes) was amplified and analysed as described previously (Ponsonnet & Nesme, 1994).

RESULTS

DNA–DNA relatedness study and inferred phylogeny of the model system

The present study was based mainly on the model system described by Popoff *et al.* (1984), consisting of nine closely related genomic species (groups 1–9) within the single biovar 1 cluster of *Agrobacterium*. A first inference of the phylogenetic structure of the model system was obtained with the two direct measurements of genome similarity or divergence [i.e. relative binding ratio (RBR) at 70 °C and at 80 °C, and ΔT_m at 70 °C] obtained by Popoff and co-workers. Only two clusters were obtained consistently by the three methods: one cluster containing strains RV3, Hayward 0363 and ATCC 23308 (respectively from groups 7, 9 and 4) and one cluster with strains T37 and NCPPB 925 (respectively groups 8 and 6). All other groupings were not supported by the three methods (data not shown). This phylogeny remained limited to a single strain per genomic group, because only one strain per genomic group was used to prepare the radioactively labelled (tracer) DNA. As a result, the consistency of the phylogeny obtained was unknown.

The bacterial collection of Popoff was then partly reconstructed for eight groups with the help of M. Kiredjian (Institut Pasteur, Paris, France) and M. Gillis (LMG, Gent, Belgium). PCR-RFLP analysis of the ribosomal intergenic spacer lying between *rrs* and *rrl* (16S and 23S rRNA genes) was used to verify that all the strains had readily differentiable genotypes. This was verified in all instances except for ICPB TT9 and T37 (from genomic group 8), which could not be differentiated using five different endonucleases (data not shown). On the other hand, sequencing revealed one *Eco*RI site in the *rrs* of all strains belonging to the biovar 1 cluster or to *A. rhizogenes*, although there is no *Eco*RI site in the *rrs* of *A. rubi*, *A. larrymoorei* and *A. vitis*. Thus, two *Eco*RI–*Mse*I fragments, of 82 and 194 bp, would be released from all agrobacteria be-

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

longing to biovar 1, but the 194 bp fragment alone was expected to occur in *A. rhizogenes*. These *rrs* fragments were used later to verify the quality of experimental AFLP analysis.

AFLP fingerprinting of *Agrobacterium* species

Digestion of genomic DNAs of *Agrobacterium* species with *Eco*RI and *Mse*I and amplifications with various primer sets characterized by two selective nucleotides consistently led to even distributions of AFLP fragments between 35 and 500 bp. By retaining only those fragments detected in two independent replicates (independent AFLP analyses performed with independently extracted DNA), the mean numbers of AFLP fragments per strain (±mean difference) were respectively 36·4±4·0, 26·6±3·4, 37·1±5·1 and 39·4±6·5 for *Eco*RI+CA/*Mse*I+0, *Eco*RI+CC/*Mse*I+0, *Eco*RI+CG/*Mse*I+0 and *Eco*RI+CT/*Mse*I+0 and 139·4±14·0 with combined AFLP data.

All data used in this study are available through anonymous FTP (ftp://pbil.univ-lyon1.fr/pub/datasets/IJSEM2001/).

Predictive AFLP

To assess the accuracy of experimental AFLP analysis, the correspondence of predicted and experimental AFLPs was determined. Using all non-redundant chromosomal and Ti plasmid sequence data currently available in databases (representing about 5% of the whole genome of C58), 20 fragments were predicted by looking for restriction and selective nucleotide sites in sequences for *Eco*RI+CA/*Mse*I+0, *Eco*RI+CC/*Mse*I+0, *Eco*RI+CG/*Mse*I+0 and *Eco*RI+CT/*Mse*I+0 conditions. With C58 DNA, experimental AFLPs revealed a total of 152 fragments, of which 17 had the expected sizes of the 20 predicted by the simulation, corresponding to 85% prediction accuracy (Table 2). The correspondence between the expected and the experimental fragments was verified experimentally for both *rrs* and Ti plasmid fragments.

Detection of *rrs* fragments

As indicated above, an *Eco*RI site occurred in the *rrs* of most but not all agrobacteria. This gave us the opportunity to estimate the AFLP quality of individual strains by looking for the occurrence of *rrs* fragments in experimental AFLPs. The *rrs* sequencing showed that *Eco*RI+CA/*Mse*I+GTCA and *Eco*RI+CG/*Mse*I+TGCG conditions for AFLP analysis would allow amplification of two *rrs* restriction fragments in most strains of *Agrobacterium*. As predicted, a single fragment of approximately 109 bp (i.e. 82 bp from *rrs* plus 13 and 14 bp for *Eco*RI and *Mse*I adaptors, respectively) was obtained with *Eco*RI+CA/*Mse*I+GTCA conditions for all strains belonging to the biovar 1 cluster, but not for *A. rubi*, *A. larrymoorei* or *A. rhizogenes* (data not shown). A 109 bp fragment was also found among the numerous fragments obtained by using the less-selective *Eco*RI+CA/*Mse*I+0 conditions. As also predicted from *rrs* sequences, the *Eco*RI+CG/*Mse*I+TGCG condition allowed the amplification of a single fragment of approximately 223 bp in all strains belonging to biovar 1 and in *A. rhizogenes* (data not shown). A fragment of the same size was also found among the numerous fragments obtained with the less-selective *Eco*RI+CG/*Mse*I+0 conditions in the same strains. However, a 223 bp fragment was also found for *A. larrymoorei* strain AF3-44. This was not an *rrs* fragment, since there is no *Eco*RI site in the *rrs* of this species (Bouzar *et al.*, 1995a). Thus, this fragment is assumed to be one that co-migrated with the *rrs* fragment but which originated from an other genome region (i.e. homoplasy).

AFLP fragments originating from pTiC58

Comparisons of AFLPs performed with Ti plasmid-containing and plasmid-free strains were used to verify experimentally the correspondence between expected and observed fragments. Eleven fragments found in experimental AFLPs with C58 or C58pAt− were lacking from AFLPs performed with Ti plasmid-free derivatives C58C1 and GMI 9023 (see data followed by asterisks in Table 2). Nine of these fragments were effectively expected to originate from the Ti plasmid by predictive AFLP analysis performed with available sequences of pTiC58 or the complete sequence of the related Ti plasmid pTiSakura. Two fragments identified by experimental AFLP analysis to be specific for pTiC58 were not predicted. On the other hand, two fragments predicted to originate from the Ti plasmid were detected not only in C58 but also in C58 derivatives cured of the Ti plasmid (AFLP fragments of 187·9 and 261·7 bp with *Eco*RI+CA/*Mse*I+0 conditions). This suggests that the Ti-plasmid fragments co-migrated with chromosomal fragments and that the chromosomal fragments still remained in AFLP analysis performed with Ti plasmid-free strains.

AFLP fragments originating from pAtC58

Experiments also showed the occurrence of AFLP fragments specific to pAtC58 (Table 2). However, no sequences of pAtC58 were available for AFLP predictions.

Current genome mispairing and DNA–DNA reassociation values

With combined AFLP data, and after assignment of fragments to discrete categories, the number of common fragments per pair of strains was found to vary from 8 to 130 (data not shown). Current genome mispairings were calculated by using d_{1xy} with $r = 12$ and equations 1 and 4. With *Agrobacterium*, the *Eco*RI/*Mse*I conditions for AFLP analysis using two selective nucleotides added to the EcoRI primer required a correction factor ($K = 0{\cdot}89$) that was

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

Table 2. Comparison of predictive and experimental AFLP analyses performed with *A. tumefaciens* strain C58 and plasmid-free derivatives of C58 under four AFLP conditions

Sizes of fragments are indicated as Predicted (fragments theoretically released by double digestion with *Eco*RI and *Mse*I), Expected (tagged fragments theoretically amplified by AFLP under the conditions indicated) or Detected (mean sizes of fragments obtained in experimental AFLP analyses). Bold type indicates putative 16S rRNA gene fragments detected under the AFLP conditions indicated and confirmed as being true *rrs* fragments by using highly selective AFLP conditions (*Eco*RI+CA/*Mse*I+GTCA and *Eco*RI+CG/*Mse*I+TGCG; see text).

Fragment	Predicted	Expected	Detected
***Eco*RI+CA/*Mse*I+0**			
C58 chromosome			
1	**82**	**109**	**109·4**
2	192	219	220·7
3	270	297	ND
4	429	456	455·9
pTiC58			
1	10	37	ND
2	80	107	107·7*
3	161	188	187·9
4	234	261	261·7
pAtC58†			
1	–	–	72·3
2	–	–	78·7
***Eco*RI+CC/*Mse*I+0**			
C58 chromosome			
No fragments	–	–	–
pTiC58			
1	47	74	74·9*
2	82	109	109·3*
3	302	329	329·8*
4	361	388	389·9*
pAtC58†			
1	–	–	258·4
2	–	–	300·9
3	–	–	496·8
***Eco*RI+CG/*Mse*I+0**			
C58 chromosome			
1	168	195	196·2
2	**194**	**221**	**222·8**
3	424	451	453·5
pTiC58			
1	34	61	61·8*
2	NP	NP	111·74*
pAtC58†			
1	–	–	147·2
2	–	–	397·1
***Eco*RI+CT/*Mse*I+0**			
C58 chromosome			
1	430	457	458·0
pTiC58			
1	79	106	107·1*
2	81	108	ND
3	NP	NP	223·3*
4	261	288	289·4*
5	NP	NP	464·6*
pAtC58†			
1	–	–	260·7
2	–	–	303·2

Table 2 (*cont.*)

ND, Expected but not detected experimentally.

NP, Not predicted but found in C58 and absent from Ti plasmid-cured derivatives.

* Fragment found experimentally to be specific to pTiC58 (i.e. present in C58 or C58pAt– but absent from C58C1 or GMI 9023, C58 derivatives cured of pTiC58).

† No sequence was available for pAtC58 in order to perform predictive AFLP with this plasmid.

calculated from equation 6 after determination of p_{EE} and p_{MM} (0·014 and 0·439, respectively) with sequences of chromosomes of C58 and with pTiSakura.

Current genome mispairing – or current genome base similarity deduced from the former measure – was compared to DNA–DNA reassociation values obtained previously by Popoff *et al.* (1984). The RBRs at 70 and 80 °C were found to be highly correlated to current genome base similarity ($r^2 = 0{\cdot}81$ and 0·74, respectively), while ΔT_m was significantly correlated to current genome mispairing ($r^2 = 0{\cdot}52$). High correlation scores and significant linear regressions were obtained by considering only pairs of bacteria belonging to the same genomic group (Fig. 1). Significant correlations were also obtained when the linear model included inter-specific distances obtained with strains belonging to the most closely related genomic species revealed previously by DNA–DNA hybridization studies (i.e. groups 6 and 8 for one cluster and groups 4, 7 and 9 in another cluster) (data not shown). However, there was no significant correlation among distances for more-distantly related strains belonging to more-different genomic species ($r^2 = 0{\cdot}34$, 0·04 and 0·08, respectively; Fig. 1).

Current genome mispairing between closely and distantly related strains

The genome differences between C58 and its plasmid-cured derivatives are shown in Table 3. The maximum genome difference, 1·2%, was obtained between C58 and GMI 9023 (Table 3). For a comparison, the current genome mispairing was 1·9% between the most closely related strains of the present model – T37 and TT9, which could not be differentiated by PCR-RFLP analysis of the *rrs–rrl* intergene. This showed that large Ti and pAt plasmids had relatively little effect upon estimates of genome mispairing.

The highest current genome mispairing between members of the same species was 13% (between RV3 and

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

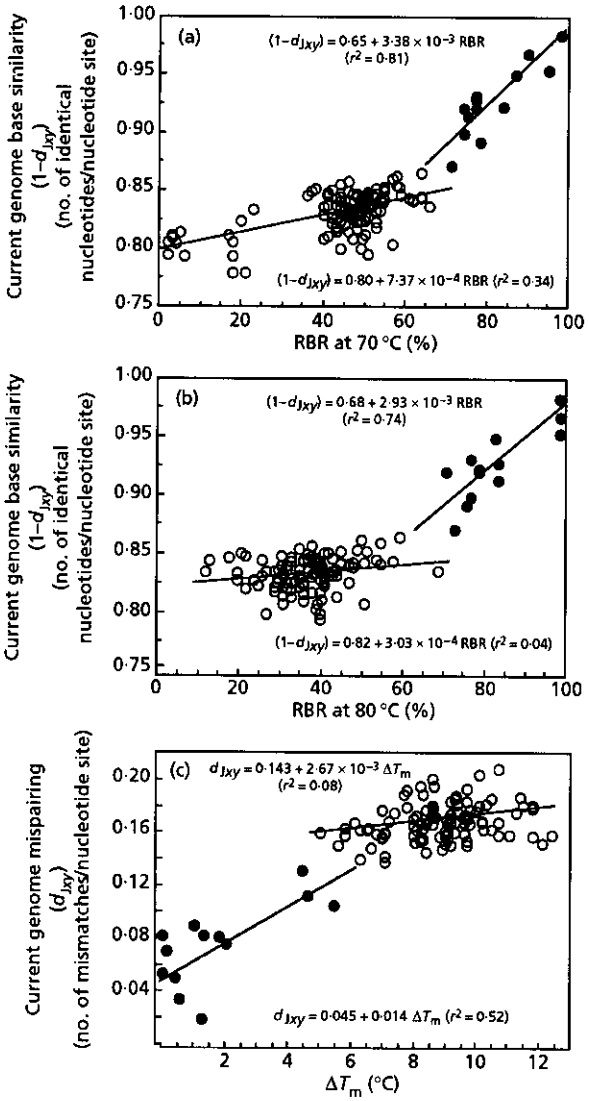

Fig. 1. Correlations between RBR at 70 °C (a), RBR at 80 °C (b) or ΔT_m (c) and current genome base similarity (or mispairing) determined by AFLP analysis. Values obtained between bacteria belonging to the same genomic species of *Agrobacterium* (●) and between different genomic species (○) are shown.

Zutra 3/1 within group 7). The shortest divergence found between members of different species was 13·6% (between RV3 and Hayward 0363, respectively from groups 7 and 9).

Phylogenetic analysis

The evolutionary genome divergence, which is an estimate of the rate of nucleotide substitution since the divergence from the common ancestor, was given by $\hat{t}$, obtained from equations 2, 4, 5 and 6. Using this metric, the phylogenetic analysis showed in all cases

Table 3. Genome differences between C58 and C58 derivatives cured of pTiC58 and/or pAtC58

Values above the diagonal are percentages of similarity of AFLP patterns (S_{Jxy}) and (in parentheses) current genome mispairing (d_{Jxy}). Values on the diagonal are combined numbers of fragments obtained between pairs of strains under *Eco*RI+CA/*Mse*I+0, *Eco*RI+CC/*Mse*I+0, *Eco*RI+CG/*Mse*I+0 and *Eco*RI+CT/*Mse*I+0 AFLP conditions. Values below the diagonal are combined numbers of common AFLP fragments between pairs of strains.

Strain	C58	C58C1	C58pAt−	GMI 9023
C58	152	93·5 (0·0056)	94·7 (0·0045)	87·1 (0·0115)
C58C1	141	141	88·1 (0·0105)	93·1 (0·0060)
C58pAt−	143	132	143	91·9 (0·0070)
GMI 9023	132	132	132	134*

* Two fragments not detected in C58, C58C1 or C58pAt− were obtained from GMI 9023.

that strains belonging to the same genomic species clustered significantly, as supported by high bootstrap values of dendrograms obtained with the neighbour-joining method (Fig. 2). This was verified for the targeted genomic species within biovar 1 and for other genomic species *A. vitis* and *A. larrymoorei*, but also for the type strain of *A. rubi*, which clustered significantly with another strain isolated from *Rubus*. Similar dendrograms were obtained with the maximum-likelihood method as well as with parsimony methods. The same significant branchings corresponding to genomic species were obtained with neighbour-joining and parsimony as well (data not shown). Other significant branchings were found at the infraspecific level, but no branching was significant at the interspecific level.

The highest evolutionary genome divergence ($\hat{t}$) found within a species was 0·097 nucleotide substitutions per site (for RV3 and Zutra 3/1) and the lowest between close, but different, species was 0·104 (for RV3 and Hayward 0363).

DISCUSSION

In this work, we propose a mathematical method, based on AFLP data, for determining genome divergence for microbial taxonomic and phylogenetic studies. For this purpose, we used the *Agrobacterium* strains studied by Popoff *et al.* (1984) as a model system, because the latter work provided data from DNA–DNA hybridizations performed with nine closely related genomic species belonging to the same biovar cluster. Moreover, the authors had used the nuclease S1 method considered most suited to that purpose by Grimont *et al.* (1980). In addition, complete tabular data for RBRs at 70 °C, at 80 °C and for ΔT_m were available; 70 and 80 °C are the optimal and stringent reassociation temperatures, respectively,

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

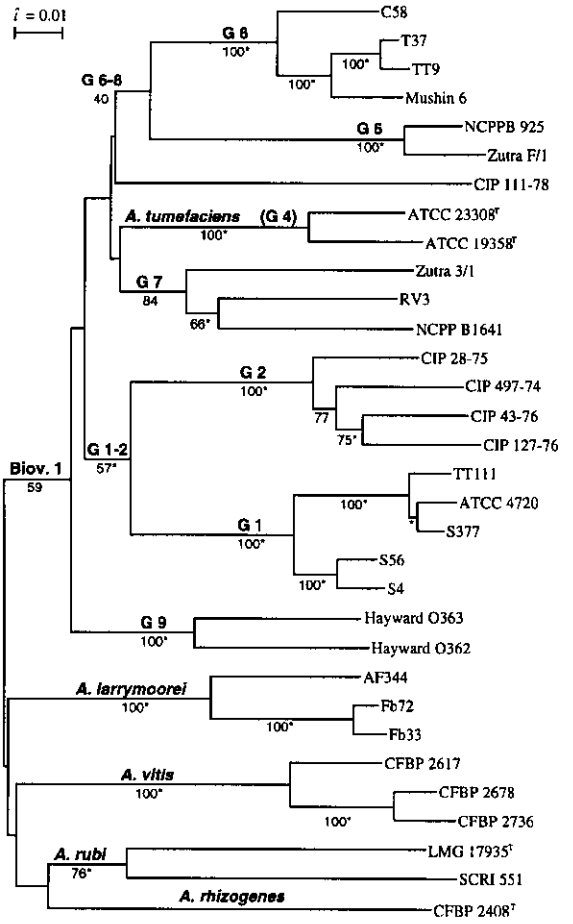

Fig. 2. Phylogenetic relationships amongst *Agrobacterium* strains inferred from evolutionary genome divergences determined by AFLP analysis. The dendrogram was obtained by using the neighbour-joining method with the estimated ratio of nucleotide substitution over whole genomes ($\hat{t}$). Branching robustness is expressed as the percentage reliability after 1000 bootstrap resamplings corrected for fragment dependence. Asterisks indicate branchings that were also supported by a bootstrapped parsimony method. The bar indicates the rate of base substitutions. G1–G9 are genomic species determined by Popoff *et al.* (1984) in biovar 1.

for bacteria with a G+C content between 58 and 60 mol% (Wetmur & Davidson, 1968). RBRs between the genomic species clustered in biovar 1 of *Agrobacterium* were in the critical range (between 40 and 69%) that required the ΔT_m to be clearly delineated. Thus, the model was found to be highly relevant for testing whether the AFLP method, correlated with the nuclease S1 method, is able clearly to delineate genomic species that are so closely related.

Predictive AFLP analysis was used to test the quality of experimental AFLP data by using partial sequences of the standard strain, C58. In view of the fact that AFLP prediction was not done for the complete genome sequence, but only for 5% of it, the present prediction accuracy of 85% (Table 2) is in agreement with the 92% accuracy achieved for predictions done for the complete genome sequence of *Escherichia coli* K-12 MG1655 by Arnold *et al.* (1999). Moreover, by using plasmid-containing and plasmid-free derivatives of the same strain, or by using AFLP conditions designed for specific amplification of *rrs* fragments known to be present in most (but not all) agrobacteria, we found an excellent correlation between expected and experimentally determined fragments. However, the rare discrepancies are caused by some limitations of the method. Arnold *et al.* (1999) reported that adjacent fragments can sometimes be confused because of the inability to discriminate single-base-pair differences in the lower part of the gel and that unpredicted fragments occurred more frequently with primers having two selective nucleotides. Conversely, the absence of a predicted fragment could be caused by the sensitivity of endonucleases to DNA methylation. To our knowledge, methylation of *Eco*RI and *Mse*I sites is not documented for *Agrobacterium*. However, DNA methylation is known to be a cell-cycle regulator in bacteria (Reisenauer *et al.*, 1999), and a recent report indicates that methylation of GA(N)TC sites by CcrM is an essential function that is cell-cycle regulated in *Agrobacterium* (Kahng & Shapiro, 2001). Thus, methylation could possibly affect experimental AFLPs and can cause pattern differences not related to real genome differences. This may explain the occurrence of two fragments in GMI 9023 that were not found in C58 (Table 2), i.e. if the absence of plasmids had modified the DNA methylation of the former. It could also explain why we never found environmental isolates of agrobacteria with completely identical AFLP patterns (our unpublished results). Alternatively, discrepancies could be caused by failures of the prediction itself, because it is known that sequencing errors are present in sequences deposited in databases. The success of predictive AFLPs indicates that AFLP data accurately reflect genome composition and that the method could be used to calculate genome mispairing.

In this work, mathematical formulae are proposed for calculating genome divergence for comparison with DNA–DNA hybridization values and for determining phylogenic relationships. The current degree of genome mispairing was determined by using the number of nucleotide sites constraining the technique. Our method is identical to that of Li & Graur (1991) and gives results similar to the log formula of Upholt (1977); however, both of these were proposed for the analysis of restriction-site data rather than AFLP. Our proposition is in agreement with that of Felsenstein (2002), who indicated that, from a mathematical point of view and in spite of their detection as fragments, the statistical behaviour of AFLP bands is similar to that of restriction sites (i.e. unlike RFLP bands) and that restriction-site distance and parsimony methods can be used with AFLP data. A major feature of interest in

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

the mathematical methods presently proposed is that they clearly show the importance of the constraining nucleotides in calculations of evolutionary genome divergence. Contrary to what has been done by most authors, including us in a previous study (Clerc *et al.*, 1998), fingerprint similarity should not be used directly for phylogenetic purposes because the relationship between genome divergence and fingerprint similarity is not linear.

A correction of AFLP-pattern similarity was done to take into account the probability of a given fragment mutating into two detectable bands in electrophoregrams. To our knowledge, this is the first time that such a correction has been proposed, because AFLPs are generally assumed to give independent fragments that originate from different genome regions (Zabeau & Vos, 1993; Felsenstein, 2002). This assumption is reasonable for AFLP analyses performed with large eukaryotic genomes that require a large number of constraining nucleotides ($r = 14$ or more). However, the shorter, prokaryotic genomes require fewer constraining nucleotides ($r = 12$ in the present study). The estimation of the correction factor (K) requires the availability of significantly long sequences of targeted genomes to predict the mean length of a restriction fragment. Sequence analysis showed that *Eco*RI sites and *Mse*I sites are respectively 1·95- and 0·45-fold more frequent than expected from the G+C content (data not shown) because of the characteristic genome 'signature' (*sensu* Karlin *et al.*, 1998) of *Agrobacterium*. This resulted in the occurrence of a detectable number of *Eco*RI–*Eco*RI fragments containing no *Mse*I site, leading us to add the selective nucleotides to the *Eco*RI primer to avoid the amplification of *Eco*RI–*Eco*RI fragments. Nevertheless, the prediction also showed that K depends upon the number and the positions of the selective nucleotides determined by the characteristic genomic signature of the target bacteria, and adding selective nucleotides to both *Eco*RI and *Mse*I primers would lead to a larger correction factor. In our opinion, this correction factor must be applied in order to achieve the most accurate estimate of genome divergence at the infraspecific level, but also for correcting bootstraps, as proposed by Felsenstein (1985). The determination of the correction factor will become more and more feasible with the increasing availability of complete genome sequences for numerous phyla.

The current genome mispairing calculated from AFLP data in the present work seems to agree with intrinsic genome parameters. The reduction in the thermal stability of reassociated hybrid DNA is directly proportional to the sequence difference (as a percentage base-pair mismatch) between reannealed single strands (Werman *et al.*, 1996). The conversion between ΔT_m values and the percentage mismatch compiled from the literature by Werman *et al.* (1996) varied from 0·7 to 2·0% base-pair mismatch per degree Celsius depression in ΔT_m, and the rate of divergence, calculated more rigorously by Springer *et al.* (1992), was 1·18% sequence divergence per degree Celsius. In the present work, the regression calculated between ΔT_m and d_{Jxy} gave a slope of 1·4% mismatch per degree Celsius depression (between 0·5 and 2·2% at the 95% confidence level) (Fig. 1), which fits very well with the data reported by Werman *et al.* (1996). This shows that the proposed mathematical method gave a correct determination of the current genome mispairing.

A highly significant linear correlation was found between RBRs and current genome mispairing calculated from AFLP data. A linear correlation between sequence identities and hybridization values was, however, not expected, since the latter give only relative similarity values between genomes (Rossello-Mora & Amann, 2001). Single-stranded DNA from two different strains will reassociate to a measurable extent and form a DNA hybrid if the strands contain more than 85% base similarity (Ullmann & McCarthy, 1973). Thus, for a value of 50% DNA–DNA hybridization, half of the genomes have less than 85% base similarity and do not hybridize, while the half that hybridizes will have more than 85% base similarity, resulting in a median value of 85%. In the present work, an extrapolated value of 82% current genome mispairing was obtained for 50% DNA–DNA hybridization with the linear regression (RBR at 70 °C, Fig. 1), agreeing well with expectations and showing the accuracy of the current genome mispairing determination.

The relationship found between AFLP similarities and DNA–DNA hybridizations was not linear for all data (Fig. 1). This result, obtained in the present work for members of several genomic species of *Agrobacterium*, was similar to those obtained with several genomic species of *Xanthomonas* by Rademaker *et al.* (2000), who used a second-degree regression curve to fit the relationship. We found, however, that the correlation was linear when limited to related bacteria. We chose to perform separate statistics for infraspecific and interspecific distances because AFLP analysis has the unique property of clearly separating distributions of infraspecific and interspecific similarities. This was found in the present work with *Agrobacterium*, as well as with *Pseudomonas tomato* versus *Pseudomonas syringae* (Clerc *et al.*, 1998), and was evidenced by the scatter plots presented by Rademaker *et al.* (2000) for *Xanthomonas* species. This clear-cut result is given by AFLP – or by related methods involving both restriction and selective amplification, such as the simplified AFLP method for *Prokarya* (Clerc *et al.*, 1998) and presumably the infrequent restriction site/PCR method (Mazurek *et al.*, 1996) – but not by randomly amplified polymorphic DNA analysis, BOX or other rep-PCR techniques (Clerc *et al.*, 1998; Rademaker *et al.*, 2000). This shows that AFLP analysis and related methods are particularly suited to the delineation of genomic species. As the relationship between AFLP similarities and DNA–DNA hybridizations is not the same for related and non-related strains, it is likely that estimates of genome similarity based on AFLP analysis

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

(and, in turn, current genome mispairing and evolutionary genome divergence) are correct at the infraspecific level but not the interspecific level. The reason for the overestimation of interspecific similarities is likely to be related to homoplasy. AFLP fragments of the same size were assumed to originate from the same genome region, but the chance that they have originated from different genome regions increases with distantly related bacteria. As a consequence, AFLP fragments were falsely considered to be identical, and the genome divergences were underestimated in the case of interspecific comparisons. This could explain the lack of significance of the deepest branchings in phylogenetic studies.

The proposed method for determining genome divergence assumes that most differences between AFLP patterns are caused by point mutations. However, differences can also be caused by indels (i.e. insertions and deletions) or the presence of large plasmids. The effect of the latter source of variation was tested experimentally. Results indicated a small effect of plasmids upon the estimated genome mispairing. The entire C58 genome is 5750 kb in size, containing two chromosomes of 3000 and 2100 kb plus 450 kb for pAtC58 and 200 kb for pTiC58 (Allardet-Servent *et al.*, 1993). The two large plasmids represent 13·3 % of the entire C58 genome. The dissimilarity of AFLP patterns between C58 and its plasmid-free derivative, GMI 9023, was 12·9 % (Table 3), which is in good agreement with the expected value. This dissimilarity led to a genome mispairing equivalent to 1·1 %. This value is smaller than the smallest genome mispairing value found between pairs of different strains in the present study. This very small genome divergence, 1·8 % genome mispairing, occurred between strains T37 and TT9. These two strains could not be distinguished by RFLP-PCR analysis of the 16S–23S intergene, a method that is suited to the distinction of closely related agrobacteria (Ponsonnet & Nesme, 1994). Thus, almost identical agrobacteria, included, until now, in the same 'strain' (*sensu* Tenover *et al.*, 1995; i.e. isolates undistinguishable by methods such as ribotyping), showed higher current genome mispairing than strains having (or not having) two large plasmids. This is also true for larger plasmids or indels, since a calculation with equation 3 showed that, for $d_{Jxy} = 13{\cdot}6\,\%$ current genome mispairing (which is the threshold between genomic species determined in the present study), the proportion of AFLP fragments differing between strains belonging to the same genomic species could be as high as 83 % (for pF with $r =$ 12 constraining nucleotides). Thus, large horizontally transferred indels, such as the 32 % sequence difference recorded between the genomes of *Escherichia coli* strains O157:H7 and K-12 MG1655 (Hayashi *et al.*, 2001), would be equivalent to only 3 % current genome mispairing and would therefore have only a limited effect upon strain assignation to the same genomic species.

For phylogenetic analysis, we used the evolutionary genome divergence instead of the current genome mispairing to take into account multiple base substitutions that had occurred since divergence. In agreement with the clear differences found previously between infraspecific and interspecific distances, genomic species were clearly delineated by phylogenetic analysis. Clusters corresponding to genomic species were supported in all instances by bootstrap resampling (Fig. 2), and strains that do not belong to the model of Popoff *et al.* (1984), such as C58, TT11 or SCRI 551, appeared in the expected clusters. Moreover, a significant clustering was also obtained with members of *A. larrymoorei*, which was expected to be a genomic species on the basis of *rrs* studies (Oger *et al.*, 1998). Even though experimental DNA–DNA reassociation studies are still required to fulfil the recommendations of Wayne *et al.* (1987), the proposal that *A. larrymoorei* is a genomic species is supported by the AFLP results given in the present work. The recent description of *A. larrymoorei* by Bouzar & Jones (2001) confirms our evaluation and validates the ability of the AFLP method to delineate new genomic species.

Species assignment based on AFLP data requires a measure of the genome divergence that must be in agreement with former taxonomic recommendations. Wayne *et al.* (1987) indicate that species include strains with approximately $\geqslant 70\,\%$ DNA–DNA relatedness and with a $\Delta T_m \leqslant 5$ °C. The present study establishes that genomic species of *Agrobacterium* would include strains with approximately $\leqslant 13{\cdot}6\,\%$ whole-genome mispairing. A slightly higher value was calculated for *Streptococcus aureus* (our unpublished results), showing that the maximum genome mispairing of genomic species must be determined in several other taxa in order to determine a cut-off point equivalent to the 70 % RBR. Conversely, the mispairing level found between *Agrobacterium* species determined in the present study is probably underestimated, but agrees with the mean sequence divergences found between closely related, but different, species reported in the literature for other taxa, such as that found between *E. coli* and *Salmonella typhimurium* (16 % sequence mismatch; Matic *et al.*, 1995).

Both current genome mispairing and evolutionary genome divergence are interesting in the context of bacterial taxonomy. Work based upon partial sequencing has shown that sequence mispairing is directly responsible for the sexual isolation that leads to speciation in various taxa (Shen & Huang, 1986; Vulic *et al.*, 1997; Majewski & Cohan, 1998; Majewski *et al.*, 2000). Whole-genome mispairing estimated, for example, by AFLP (as described in this work) appears to be an intrinsic parameter specifically suited to bacterial phylogeny or population genetics, since it indicates directly how bacteria are potentially isolated sexually. For this reason, current genome mispairing could be used in future as an alternative to RBRs or ΔT_m to describe genomic species. On the other hand, the evolutionary genomic distance is more suited to phylogenetic purposes. Interestingly, the threshold

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

values of evolutionary distance obtained within and between genomic species (respectively 0·097 and 0·104 nucleotide substitutions) were determined objectively from the robust tree topology obtained after bootstrap resamplings of AFLP data.

Bootstrap resampling appears to be a method of paramount importance for the utilization of AFLP in taxonomy, since it supports clear and objective delineations of genomic species. The bootstrap method assumes that characters evolve independently (Felsenstein, 1985), and this assumption is generally recognized in AFLP data (Felsenstein, 2001). We found a small degree of fragment dependence between closely related bacteria (less than 11%), which is assumed to be negligible between species. The bootstrap method can be used to delineate genomic species with AFLP data because a correction factor is used to cope with fragment dependence, as proposed by Felsenstein (1985). It is remarkable that the modified bootstrap method supports the inclusion of strain RV3 in genomic group 7 together with Zutra 3/1 (Fig. 2) in spite of the 13% genome mispairing found between the two strains, since this value is large and very close to the 13·6% genome mispairing with Hayward 0363 from genomic group 9. Very interestingly, the RBR values between RV3 and Zutra 3/1 and between RV3 and Hayward 0363 were respectively 72 and 68%, and Popoff *et al.* (1984) needed to use ΔT_m (respectively 4·5 and 8·2 °C) to delineate genomic groups 7 and 9 clearly. Thus, it is likely that the bootstrap method could be used as an alternative to ΔT_m to delineate very closely related genomic groups, providing the species contain several strains to form a cluster.

Since the proposal of Zuckerkandl & Pauling (1965), macromolecules have been used increasingly for phylogeny. Genomes are probably the ultimate macromolecule for use for this purpose, and AFLP analysis could be used along with complete genome sequence analysis to estimate genome divergence. Whole genomes probably diverge more rapidly than is reported for genes, in agreement with the neutral mutation model of Kimura (1983), which is the basic hypothesis behind the concept of a molecular 'clock'. The calibration of the molecular clock proposed by Ochman & Wilson (1987) indicated a sequence divergence of 0·02% per million years (My) for *rrs* and 0·7–0·8% My^{-1} for genes encoding proteins. Taking the latter conversion as a maximum, if the shortest evolutionary genome divergence between species is 10·4% nucleotide substitution, this suggests that the agrobacterial genomic species probably diverged 13 My ago at the earliest. At the strain level, agrobacteria that are not differentiated by means of ribosomal intergenic spacers (like T37 and ICPB TT9 in the present study) show an evolutionary genome divergence of 1%, resulting, perhaps, in a divergence that occurred 1·3 My ago. Dating of divergence, however, needs both a better calibration of the molecular clock for whole-genome divergence and an accurate estimate of genome divergence through AFLP.

The AFLP method as described in this study is a very useful and powerful tool for exploring bacterial genomes. The further delineation of bacterial species by a genomic approach could be based, in future, upon genome descriptions of individual strains obtained by AFLP analysis or equivalent methods and clustering supported by bootstrap resampling, in order eventually to replace DNA–DNA reassociation studies. In addition, this method could provide interesting estimates of genomic substitution rates suitable for reconstructing the evolutionary history of prokaryotes.

ACKNOWLEDGEMENTS

The authors wish to thank M. A. Poirier and K. Groud for skillful technical assistance, M. Gouy and J. Felsenstein for scientific counsel and T. Vogel for discussion and manuscript revision. The research was made possible with the kind help of M. Kiredjian and M. Gillis, who sent agrobacteria from the collection studied by Popoff. C.M. is a Research Fellow of the Institut National de la Recherche Agronomique. This work was part of INCO-DC European project ERB1C18CT970198 ('Integrated control of crown gall in Mediterranean countries') and was performed using the sequencing facilities of the DTAMB at Université Claude Bernard-Lyon 1.

REFERENCES

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

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

Bouzar, H. (1994). Request for a judicial opinion concerning the type species of *Agrobacterium*. *Int J Syst Bacteriol* **44**, 373–374.

Bouzar, H. & Jones, J. B. (2001). *Agrobacterium larrymoorei* sp. nov., a pathogen isolated from aerial tumours of *Ficus benjamina*. *Int J Syst Evol Microbiol* **51**, 1023–1026.

Bouzar, H., Chilton, W. S., Nesme, X., Dessaux, Y., Vaudequin, V., Petit, A., Jones, J. B. & Hodge, N. C. (1995a). A new *Agrobacterium* strain isolated from aerial tumours on *Ficus benjamina* L. *Appl Environ Microbiol* **61**, 65–73.

Bouzar, H., Jones, J. B. & Bishop, A. L. (1995b). Simple cultural tests for identification of Agrobacterium biovars. *Methods Mol Biol* **44**, 9–13.

Caetano-Anollés, G., Bassam, B. J. & Gresshoff, P. M. (1991). DNA amplification fingerprinting using very short arbitrary oligonucleotide primers. *Biotechnology* **9**, 553–557.

Clerc, A., Manceau, C. & Nesme, X. (1998). Comparison of randomly amplified polymorphic DNA with amplified fragment length polymorphism to assess genetic diversity and genetic relatedness within genospecies III of *Pseudomonas syringae*. *Appl Environ Microbiol* **64**, 1180–1187.

Cohan, F. M. (1994). Genetic exchange and evolutionary divergence in prokaryotes. *Trends Ecol Evol* **9**, 175–180.

De Ley, J. (1974). Phylogeny of procaryotes. *Taxon* **23**, 291–300.

De Ley, J., Tijtgat, R., De Smedt, J. & Michiels, M. (1973). Thermal stability of DNA:DNA hybrids within the genus *Agrobacterium*. *J Gen Microbiol* **78**, 241–252.

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

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

Felsenstein, J. (1993). PHYLIP: Phylogenetic Inference Package, version 3.5c. Seattle: University of Washington.

Felsenstein, J. (2002). *Inferring Phylogenies.* Sunderland, MA: Sinauer.

Fox, G. E., Wisotzkey, J. D. & Jurtshuk, P., Jr (1992). How close is close: 16 rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol* **42**, 166–170.

Grimont, P. A. D., Popoff, M. Y., Grimont, F., Coynault, C. & Lemelin, L. (1980). Reproducibility and correlation study of three deoxyribonucleic acid hybridization procedures. *Curr Microbiol* **4**, 325–330.

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

Hayashi, T., Makino, K., Ohnishi, M. & 19 other authors (2001). Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* **8**, 11–22.

Janssen, P., Coopman, R., Huys, G., Swings, J., Bleeker, M., Vos, P., Zabeau, M. & Kersters, K. (1996). Evaluation of the DNA fingerprinting method AFLP as a new tool in bacterial taxonomy. *Microbiology* **142**, 1881–1893.

Kahng, L. S. & Shapiro, L. (2001). The CcrM DNA methyltransferase of *Agrobacterium tumefaciens* is essential, and its activity is cell cycle regulated. *J Bacteriol* **183**, 3065–3075.

Karlin, S., Campbell, A. M. & Mrazek, J. (1998). Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32**, 185–225.

Keane, P. J., Kerr, A. & New, P. B. (1970). Crown gall of stone fruit. II: Identification and nomenclature of *Agrobacterium* isolates. *Aust J Biol Sci* **23**, 585–595.

Kersters, K., De Ley, J., Sneath, P. H. A. & Sackin, M. (1973). Numerical taxonomic analysis of *Agrobacterium*. *J Gen Microbiol* **78**, 227–239.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge: Cambridge University Press.

Kiredjian, M. (1979). Le genre *Agrobacterium* peut-il être pathogène pour l'homme? *Med Malad Infect* **9**, 223–235.

Li, W. H. & Graur, D. (1991). *Fundamentals of Molecular Evolution.* Sunderland, MA: Sinauer.

Maidak, B. L., Cole, J. R., Lilburn, T. G. & 7 other authors (2001). The RDP-II (Ribosomal Database Project). *Nucleic Acids Res* **29**, 173–174.

Majewski, J. & Cohan, F. M. (1998). The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. *Genetics* **148**, 13–18.

Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M. & Dowson, C. G. (2000). Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J Bacteriol* **182**, 1016–1023.

Matic, I., Rayssiguier, C. & Radman, M. (1995). Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. *Cell* **80**, 507–515.

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

Ochman, H. & Wilson, A. C. (1987). Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J Mol Evol* **26**, 74–86.

Oger, P., Dessaux, Y., Petit, A., Gardan, L., Manceau, C., Chomel, C. & Nesme, X. (1998). Validity, sensitivity and resolution limit of the PCR-RFLP analysis of the *rrs* (16S rRNA gene) as a tool to identify soil-borne and plant-associated bacterial populations. *Genet Sel Evol* **30**, S311–S321.

Ophel, K. & Kerr, A. (1990). *Agrobacterium vitis* sp. nov. for strains of *Agrobacterium* biovar 3 from grapevines. *Int J Syst Bacteriol* **40**, 236–241.

Panagopoulos, C., Psallidas, P. G. & Alivizatos, A. S. (1978). Studies on biotype 3 of *Agrobacterium radiobacter* var. *tumefaciens*. In *Proceedings of the IVth International Conference on Plant Pathogenic Bacteria*, vol. I, pp. 221–228. Angers: Station de Pathologie Végétale et Phytobactériologie, INRA.

Perrière, G., Duret, L. & Gouy, M. (2000). HOBACGEN: database system for comparative genomics in bacteria. *Genome Res* **10**, 379–385.

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

Popoff, M. Y., Kersters, K., Kiredjian, M., Miras, I. & Coynault, C. (1984). Position taxonomique de souches de *Agrobacterium* d'origine hospitalière. *Ann Microbiol* **135**, 427–442.

Rademaker, J. L. W., Hoste, B., Louws, F. J., Kersters, K., Swings, J., Vauterin, L., Vauterin, P. & de Bruijn, F. J. (2000). Comparison of AFLP and rep-PCR genomic fingerprinting with DNA–DNA homology studies: *Xanthomonas* as a model system. *Int J Syst Evol Microbiol* **50**, 665–677.

Reisenauer, A., Kahng, L. S., McCollum, S. & Shapiro, L. (1999). Bacterial DNA methylation: a cell cycle regulator? *J Bacteriol* **181**, 5135–5139.

Rosenberg, C. & Huguet, T. (1984). The pAtC58 plasmid of *Agrobacterium tumefaciens* is not essential for tumour induction. *Mol Gen Genet* **150**, 53–61.

Rossello-Mora, R. & Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiol Rev* **25**, 39–67.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.

Sawada, H., Ieki, H., Oyaizu, H. & Matsumoto, S. (1993). Proposal for rejection of *Agrobacterium tumefaciens* and revised descriptions for the genus *Agrobacterium* and for *Agrobacterium radiobacter* and *Agrobacterium rhizogenes*. *Int J Syst Bacteriol* **43**, 694–702.

Shen, P. & Huang, H. V. (1986). Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics* **112**, 441–457.

Spratt, B. G. (1999). Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet. *Curr Opin Microbiol* **2**, 312–316.

Springer, M. S., Davidson, E. H. & Britten, R. J. (1992). Calculation of sequence divergence from the thermal stability of DNA heteroduplexes. *J Mol Evol* **34**, 379–382.

Stackebrandt, E. & Goebel, B. M. (1994). 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**, 846–849.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996). Phylogenic inference. In *Molecular Systematics*, 2nd edn, pp. 407–514. Edited by D. M. Hillis, C. Moritz & B. K. Mable. Sunderland, MA: Sinauer.

Tenover, F. C., Arbeit, R. D., Goering, R. V., Mickelsen, P. A., Murray, B. E., Persing, D. H. & Swaminathan, B. (1995). Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *J Clin Microbiol* **33**, 2233–2239.

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

Ullmann, J. S. & McCarthy, B. J. (1973). The relationship between mismatched base pairs and the thermal stability of DNA duplexes. *Biochim Biophys Acta* **249**, 416–424.

Upholt, W. B. (1977). Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. *Nucleic Acids Res* **4**, 1257–1265.

Versalovic, J., Koeuth, T. & Lupski, J. R. (1991). Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Res* **19**, 6823–6831.

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.

Vulic, M., Dionisio, F., Taddei, F. & Radman, M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci U S A* **94**, 9763–9767.

Wang, Y. & Zhang, Z. (2000). Comparative sequence analyses reveal frequent occurrence of short segments containing an abnormally high number of non-random base variations in bacterial rRNA genes. *Microbiology* **146**, 2845–2854.

Wayne, L. G., Brenner, D. J., Colwell, R. R. & 9 other authors (1987). International Committee on Systematic Bacteriology. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Bacteriol* **37**, 463–464.

Werman, S. D., Springer, M. S. & Britten, R. J. (1996). Nucleic acids I: DNA-DNA hybridization. In *Molecular Systematics*, 2nd edn, pp. 169–201. Edited by D. M. Hillis, C. Moritz & B. K. Mable. Sunderland, MA: Sinauer.

Wetmur, J. G. & Davidson, N. (1968). Kinetics of renaturation of DNA. *J Mol Biol* **31**, 349–370.

Yanagi, M. & Yamasato, K. (1993). Phylogenetic analysis of the family *Rhizobiaceae* and related bacteria by sequencing of 16S rRNA gene using PCR and DNA sequencer. *FEMS Microbiol Lett* **107**, 115–120.

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

Young, J. M., Kuykendall, L. D., Martínez-Romero, E., Kerr, A. & Sawada, H. (2001). A revision of *Rhizobium* Frank 1889, with an emended description of the genus, and the inclusion of all species of *Agrobacterium* Conn 1942 and *Allorhizobium undicola* de Lajudie *et al.* 1998 as new combinations: *Rhizobium radiobacter*, *R. rhizogenes*, *R. rubi*, *R. undicola* and *R. vitis*. *Int J Syst Evol Microbiol* **51**, 89–103.

Zabeau, M. & Vos, P. (1993). *Selective Restriction Fragment Amplification: a General Method for DNA Fingerprinting.* Publication 0534858A1. Munich: European Patent Office.

Zuckerkandl, E. & Pauling, L. (1965). Molecules as documents of evolutionary history. *J Theor Biol* **8**, 357–366.

Mougel, C., J. Thioulouse, G. Perrière, and X. Nesme. 2002. A mathematical method for determining genome divergence and species delineation using AFLP. International Journal of Systematic and Evolutionary Microbiology **57**:573-586.