3.

Chapter 3

Functional ANOVA with P-splines applied to the analysis of dissociation curves of nucleic acid duplexes

The organization of the chapters is as follows:

Chapter 1 emphasizes the characterization of microbial populations using molecular techniques based on phylogenetic markers.

Chapter 2 describes bacterial TNT transformation and shows that TNT denitration is promising for complete mineralization of the molecule.

Chapter 3 presents a functional ANOVA model to analyze dissociation curves from hybridizations to microarrays with thermal dissociation.

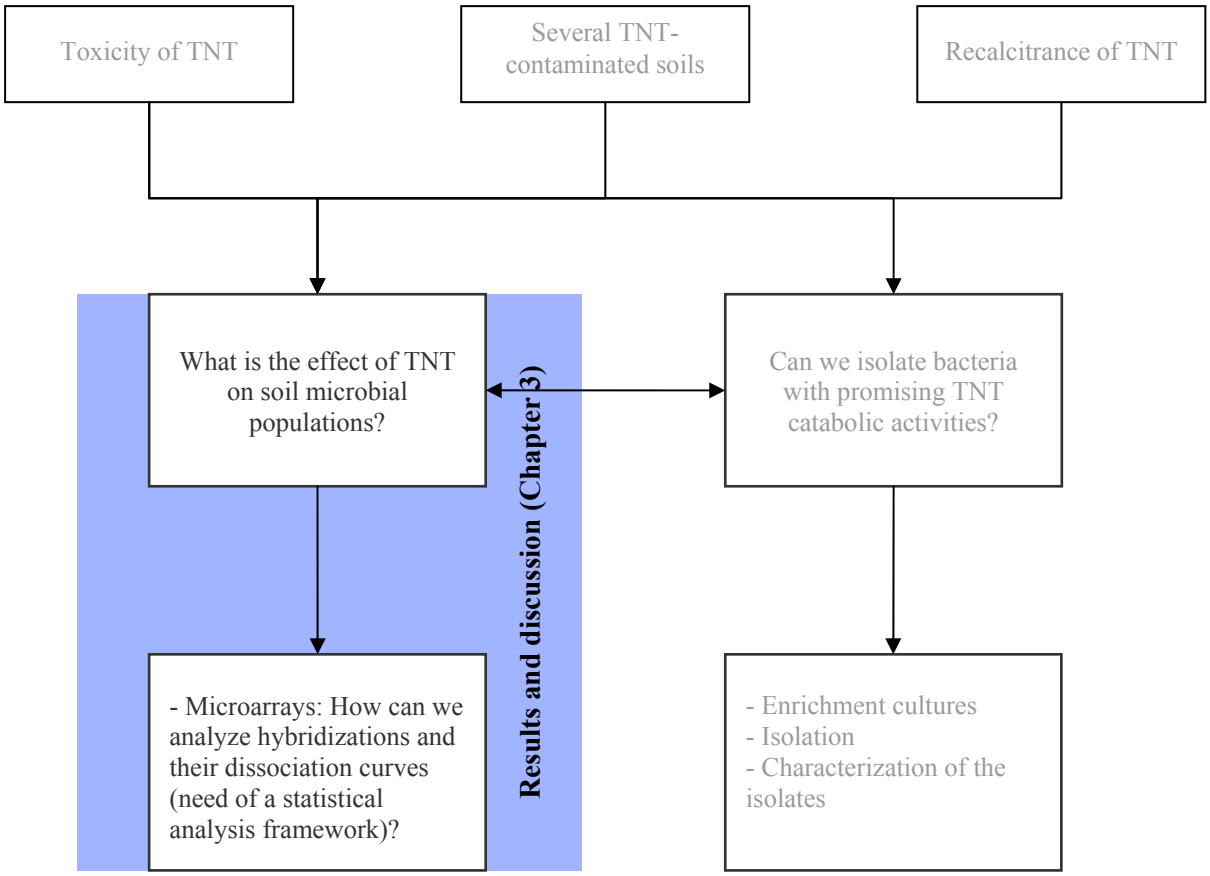
Chapter 4 compares hybridizations of in vitro transcribed 16S rRNA derived from an uncontaminated and a TNT-contaminated soil sample to a phylogenetic microarray with thermal dissociation analysis.

Chapter 5 compares extracted DNA and 16S PCR-DGGE fingerprints of uncontaminated and TNT-contaminated soil samples, and of a pristine soil artificially contaminated with TNT. Plate counts on agar plates inoculated with an uncontaminated and a TNT-contaminated soil sample are also evaluated.

Chapter 6 describes the isolation and characterization of a *Pseudomonas aeruginosa* strain that denitrates TNT in the absence of oxygen and presence of ferrihydrite.

Chapter 7 shows that a phenazine molecule produced by *Pseudomonas aeruginosa* denitrates TNT.

Chapter 8 discusses the results obtained from the characterization of bacterial populations of TNT-contaminated sites and the isolation of a TNT-degrading strain.


Toxicity of TNT
Several TNT-contaminated soils
Recalcitrance of TNT
What is the effect of TNT on soil microbial populations?
Can we isolate bacteria with promising TNT catabolic activities?
Results and discussion (Chapter 3)
- Microarrays: How can we analyze hybridizations and their dissociation curves (need of a statistical analysis framework)?
- Enrichment cultures
- Isolation
- Characterization of the isolates

Functional ANOVA with P-splines applied to the analysis of dissociation curves of nucleic acid duplexes

Céline Bugli[1,*], Philippe Lambert[1,2], James C. Smoot[3], Laura M. Smoot[3], Spiros N. Agathos[4], David A. Stahl[3] and Laurent Eyers[4,¥]

[1]Institut de Statistique, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium, [2]Unité d'épidémiologie, biostatistique et méthodes opérationnelles, Faculté de Médecine, Université catholique de Louvain, 1200 Bruxelles, Belgium, [3]Civil and Environmental Engineering, University of Washington, WA 98195, USA and [4]Unité de génie biologique, Faculté d'ingénierie biologique, agronomique et environnementale, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium

*To whom correspondence should be addressed.
Voie du Roman Pays, 20
B-1348 Louvain-la-Neuve, Belgium
E-mail : bugli@stat.ucl.ac.be
Tel: +32 10 47 4160
Fax: +32 10 47 3062

¥ L. Eyers developed the calculator under Matlab, analyzed the microarray data, and wrote the results and discussion section.

3.1. Abstract

Motivation: Several studies have examined the use of dissociation curves to provide information of probe-target nucleic acid duplexes. By comparing dissociation curves, the presence of specific target nucleic acid could be evaluated. However, these comparisons were generally done visually, as there have been no appropriate statistical tools available.

Results: We developed a functional ANOVA model to compare dissociation curves and determine if they significantly differed. To test and validate the model, we used dissociation curves obtained from in vitro transcribed nucleic acid amplified from two environmental samples hybridized to a microarray. Detection and rejection of outlier curves was important for appropriate discrimination between curves. The identification of significantly different curves was more efficient with the model than approaches relying on measurements at a single temperature.

Availability: A user-friendly interface in Matlab for analysis of curves with functional ANOVA is available at http://stahl.ce.washington.edu.

Contact: bugli@stat.ucl.ac.be

3.2. Introduction

Single nucleotide polymorphisms (SNPs) are the most fundamental genotypic variation that arises during natural selection and genetic drift. SNPs have profound implications in the genetic variation that occurs within and among species. For instance, mutations in the human *BRCA1* and *BRCA2* genes have been linked with the susceptibility to breast and ovarian cancer (Miki *et al.*, 1994; Wooster *et al.*, 1995), and mutations in the *APOE* gene are associated with an increased risk of developing Alzheimer's disease (Bullido *et al.*, 1998). In addition, the deep branching groups of bacteria, *Planctomycetes* and *Verrucomicrobia*, may be phylogenetically distinguished by SNPs in the small subunit ribosomal RNA gene.

High-throughput techniques are often used to identify SNPs present in gene sequences. The polymerase chain reaction (PCR) and high-density microarrays combined with fluorescence detection are two commonly used methods of identification. Microarrays contain thousands of elements (i.e. nucleotide probes) and simultaneously detect thousands of different genes and their variants. The presence and abundance of a specific gene sequence is generally determined by quantifying the emission of a fluorescent signal conferred by hybridization of a labeled target molecule to a complementary probe. This technique relies on the specificity of hybridization between a target nucleotide sequence and the complementary nucleotide probe on the microarray (formation of a perfect match (PM) duplex). However, false positive events may also occur when a sequence hybridizes to non-complementary probes (formation of a mismatch (MM) duplex). Hacia *et al.* (1996) reported that quantification of non-specific hybridization could be determined by subtracting the intensities of signals of MM duplexes from PM duplexes. However, this approach may not be suitable with samples containing a mixture of PM and MM sequences in unknown abundance. To discriminate between PM and MM duplexes, dissociation curves (also called “melting curves” or “melting profiles” in the literature) can be generated by linearly increasing the temperature of the microarray and recording signal intensities. The increase of temperature affects the binding of the target to the probe and leads to their dissociation. Plotting signal intensities versus temperature values results in a sigmoid-shaped curve. The exact shape of the curve depends on the sequence of the duplex; and, thus, PM duplexes may produce curves that are distinguishable from MM duplexes. For instance, Liu *et al.* (2001) demonstrated discrimination between target and non-target sequences differing by a single nucleotide.

PCR, which relies on amplification of a specific target sequence, is another technique to study nucleotide polymorphisms. The amplified product is typically analysed for its size and nucleotide sequence, but this process is often time-consuming and labour intensive. However, rapid characterisation of a PCR product for a SNP can be performed by incorporating fluorescent molecules during the thermocycling. For instance, Ririe *et al.* (1997) and Liew *et al.* (2004) showed that it is possible to differentiate amplified gene variants by incorporating a fluorescent molecule, increasing the temperature, and monitoring the evolution of the intensity of the molecule. Liew *et al.* (2004) were able to discriminate SNPs using the dissociation curves obtained from this analysis.

Data reduction techniques are often used to analyse dissociation curves. Calculation of T_d values (temperature at 50% probe-target dissociation) is one such method of data reduction. This approach was successfully used by Liu *et al.* (2001) and Liew *et al.* (2004) to discriminate sequences having a single nucleotide change with microarray and real-time PCR experiments, respectively. Urakawa *et al.* (2003) found that T_d did not effectively discriminate PM and MM duplexes in microarray experiments because it only accounts for signal intensities at 50% dissociation. Furthermore, in his study of SNPs with real-time PCR, Wittwer *et al.* (2003) showed that the entire curve should be included in the analysis. In an attempt to glean additional information beyond the T_d value, Urakawa *et al.* (2003) determined an optimum discrimination temperature defined as the maximum product of difference and ratio of the signal intensities between normalized dissociation curves of PM and MM duplexes, while El Fantroussi *et al.* (2003) evaluated the extent of the discrimination by visually comparing normalized dissociation curves. However, in these studies, interpretation of the comparisons between curves was limited, as no statistical analysis of the curves was used.

Although several statistical methods have been developed to compare signal intensities at a single temperature (an extensive review of the literature can be found at http://www.nslij-genetics.org/microarray), there is no available method to analyse curves generated from thermal dissociations. The objective of this study was to develop a statistical tool to compare dissociation curves and determine when they significantly differ.

3.3. Methods

Sample dataset. The dataset used to develop and test the calculator was obtained from hybridization of in vitro transcribed 16S rRNA of PCR amplified product of uncontaminated and 2,4,6-trinitrotoluene contaminated soil samples to a microarray targeting 16S rRNA genes (Eyers *et al.*, manuscript submitted). Hybridizations were done in triplicate for each sample. The microarray contained general (e.g. probe Eub338 targeting all *Bacteria*) and species-specific PM probes (e.g. probe Nse1472 targeting the species *Nitrosomonas europaea*). In addition, the microarray contained two MM probes for each PM probe (e.g. probe Eub338-t7g and Eub338-g11c, where G replaced T at the 7^{th} position and C replaced G at the 11^{th} position, respectively). A description of the microarray platform (i.e. description of the spotted probes, number of replicate probes, and location of the probes on the microarray) and the dataset (i.e. probe signal intensities before and during thermal dissociation) are available at the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under the accession numbers GPL2993 and GSE3499, respectively.

The methods used to hybridize the samples to the microarray and collect signal intensities are described elsewhere (Eyers *et al.*, manuscript submitted). In brief, dissociation curves were generated by increasing the temperature of the hybridized microarray (1°C/min) from 20°C to 68°C and recording signal intensities every 1°C. The hybridization signals were corrected by subtracting the local background signal (Fotin *et al.*, 1998). Comparisons of dissociation curves were both with curves from nucleic acid from the same sample hybridized to different probes (e.g. PM and MM probes) and with different nucleic acid samples hybridized to the same probe. Normalized curves were used for comparison, and normalization was done by setting the lowest intensity in a profile at a value of 0 and the highest intensity at 1 (i.e. signal intensities were subtracted by the minimum signal intensity and divided by the difference between the maximum and the minimum values). In the functional ANOVA calculator, a set of curves compared against another set was defined as a "group". Data from a PM and two respective MM probes hybridized to the same soil sample (3 groups) were used to develop the functional ANOVA model. Results of hybridizations of the two soil samples to PM and MM probes were used to validate the model (6 groups).

T_d computation. Calculation of T_d values was done using an adapted method presented by Urakawa *et al.* (2003). In brief, for each observed normalized curve, a linear regression was determined using 5 data points around the point with the observed normalized value closest to 0.5. The slope and intercept of the regression line were used to calculate the T_d at a normalized signal intensity of 0.5.

P-spline smoothing and Functional ANOVA with P-splines. P-splines are a combination of B-splines and difference penalties coefficients. B-splines are attractive for nonparametric modelling, but the choice of the optimal number and positions of knots can be difficult. A solution is to use a large number of equidistant knots and a penalty to control the smoothness of the fitted curve (O'Sullivan, 1986, 1988). Eilers and Marx (1996) proposed to use B-splines with a difference penalty on the spline coefficients.

Suppose that we observe a smooth function $y(t)$ at $t_1, t_2, \ldots, t_N$, yielding N pairs of observations $\{(t_n, y_n) : n = 1, \ldots, N\}$. In spline regression, we approximate *y(t)* by

$$\sum_{s=0}^{S} \beta_s \phi_s(t) \quad , \tag{1}$$

where $\phi_0, \phi_1, \ldots, \phi_S$ are known functions of t and $\beta = (\beta_0, \ldots, B_S)^T$ an appropriate coefficients vector. Examples of the ϕ_S might be B-splines, tensor products of B-splines, thin plate regression spline basis functions, the truncated power basis for cubic splines, or some radial basis functions (e.g. Ruppert *et al.*, 2003). In this application, we used B-spline basis functions.

A B-spline of degree q is a piecewise polynomial function with $q + 1$ pieces of degree q. At the q joining knots, derivatives up to order $q - 1$ are continuous (de Boor, 2001; Dierckx, 1995).

The coefficients β in Equation (1) are estimated by minimising:

$$\sum_{n=1}^{N} \left(y_n - \sum_{s=0}^{S} \beta_s \phi_s(t_n) \right)^2 + PEN(\lambda) \quad , \tag{2}$$

where $PEN(\lambda)$ is a roughness penalty depending on a smoothing parameter λ. Examples of this penalty are the thin plate spline penalty, the integrated square of second derivative

penalty for cubic spline, and the various difference penalties used in P-spline smoothing. Eilers and Marx (1996) introduced P-spline, a penalised B-spline regression model using equidistant knots associated with a convenient roughness penalty. P-splines are based on the difference of adjacent B-spline coefficients yielding the penalised likelihood:

$$\sum_{n=1}^{N}\left(y_n - \sum_{s=0}^{S}\beta_s\phi_s(t_n)\right)^2 + \lambda\sum_{s=d+1}^{S}\left(\Delta^d\beta_s\right)^2 \quad , \tag{3}$$

where Δ^d is the difference operator of order *d*.

In the case of longitudinal data (i.e. repeated observations over time), we have several curves $y_i(t)\ i = 1,\ldots, I$ observed at multiple occasions, say $t = (t_1, t_2,\ldots, t_N)$. Multiple regressions were used to model these curves and their association to factors (i.e. explanatory variables). For example, suppose that *P* binary covariates $(x_{i1},\ldots, x_{iP})$ are associated with the *i*th curve (the model can be extended to non-binary variables). Each curve was decomposed as the sum of several curves: one reference curve *f*(*t*) that we named the "mean curve" and P corrections curves $g_p(t)\ (p = 1,\ldots, P)$ (named "effect curves") that summarise the effect of each factor:

$$E(Y_i(t_n)) = \mu_i^n = f(t_n) + \sum_{p=1}^{P} b_{ip}g_p(t_n) \quad . \tag{4}$$

These curves can be modelled using B-splines (Hastie and Tibshirani, 1991).

As with spline smoothing, the evolution of the curvature of a B-spline fit is strongly dependent on the number and the positions of the knots. Rather than optimising the number and position of the knots (which is a difficult numerical problem), we used a relatively large number of equidistant knots and avoided over-fitting by adding a difference penalty on the coefficients of adjacent B-splines. This is the P-spline approach recommended by Eilers and Marx (1996).

The full model for a B-splines basis $\{\phi_s(t) : s = 1,\ldots, S\}$ is, for curve $y_i(t)$:

$$E(y_i t)) = f(t) + \sum_{p=1}^{P} b_p g_p(t) = \sum_{s=1}^{S}\beta_{0s}\phi_s(t) + \sum_{p=1}^{P}\left(b_{ip}\sum_{s=1}^{S}\beta_{ps}\phi_s(t)\right) \quad . \tag{5}$$

If Y is the matrix such that: $Y_{in} = y_i(t_n)$, then Equation (5) can be written as:

$$E(Y) = XB\Phi^T \quad , \tag{6}$$

where X is the matrix indicating presence or absence of factors of interest $(p = 1,\ldots, P+1)$ for each curve $(i = 1,\ldots, I)$, Φ is the matrix associated to the B-spline basis (such that $\Phi_{ns} = \phi_s(t_n), s = 1,\ldots, S; n = 1,\ldots, N$), and B is the matrix of the B-spline coefficients (such that $B_{ks} = \beta_{k-1,s}, k = 1,\ldots, P+1; s = 1,\ldots, S$). Subsequently, the notations of the above problem were transformed to obtain an index notation, which conveys the same information but is intended for mathematical convenience rather than machine calculation. We used the vectorial form of this model with the vec(.) operator defined by: $x = \text{vec}(X)$ such that $x_{(j-1)N+i} = X_{ij} (i = 1,\ldots, I; j = 1,\ldots, J)$.

For the above spline regression model, we have:

$$\text{vec}\left(E(Y)^T\right) = \text{vec}\left((XB\Phi^T)^T\right) = \text{vec}\left(\Phi B^T X^T\right) \quad . \tag{7}$$

Henderson and Searle (1981) showed that

$$\text{vec}\left(\Phi B^T X^T\right) = (X \otimes \Phi)\text{vec}\left(B^T\right) \quad , \tag{8}$$

where $(X \otimes \Phi)$ is the Kroenecker product of the matrices X and Φ. Therefore, Equation (5) can be rewritten as

$$E(y) = \tilde{X}b \quad , \tag{9}$$

with $(X \otimes \Phi) = \tilde{X}$ and $\text{vec}\left(B^T\right) = b$.

If D is the difference matrix associated to the difference operator of order d such that $\sum_{s=d+1}^{S} \left(\Delta^d \beta_{ps}\right)^2 = \beta_p^T D^T D \beta_p$, then the estimates of the spline parameters will be obtained by minimising:

$$(y - \tilde{X}b)^T (y - \tilde{X}b) + \sum_{p=0}^{P} \lambda_p \beta_p^T D^T D \beta_p \quad , \tag{10}$$

where λ_p is the penalty parameter associated to p^{th} curve. The penalty term in that penalised likelihood can be rewritten as $b^T \left(\Lambda_f \otimes D^T D\right) b$ where $\Lambda_f = \text{diag}\{\lambda_0, \lambda_1, \ldots, \lambda_p\}$.

Conditionally based on the penalty parameters, the following spline parameter estimators were obtained:

$$\hat{b} = \left(\tilde{X}^T \tilde{X} + \Lambda_f \otimes D^T D\right)^{-1} \tilde{X}^T y \quad , \tag{11}$$

and the fitted values:

$$\hat{y} = \tilde{X}\hat{b} = \tilde{X}\left(\tilde{X}^T \tilde{X} + \Lambda_f \otimes D^T D\right)^{-1} \tilde{X}^T y = Hy \quad . \tag{12}$$

The model was presented in a least square setting to keep the explanation simple. Because of its regression structure, it can be extended to a generalised additive model (Eilers and Marx, 1996). A more detailed discussion of functional ANOVA with P-spline is provided by Bugli and Lambert (2004).

Application of functional ANOVA to discriminate PM and MM probes. As discussed previously, the model in matrix notation is:

$$E(Y) = XB\Phi^T \quad , \tag{13}$$

where X is the matrix indicating presence or absence of factors of interest $(p = 1,\ldots, P)$ for each curve $(i = 1,\ldots, I)$,

$$X = \begin{bmatrix} 1 & b_{11} & b_{12} & \ldots & b_{1P} \\ 1 & b_{21} & b_{22} & \ldots & b_{2P} \\ & & \vdots & & \\ 1 & b_{I1} & b_{I2} & \ldots & b_{IP} \end{bmatrix} \quad , \tag{14}$$

Φ is the matrix of the B-spline basis:

$$\Phi = \begin{bmatrix} \phi_1(t_1) & \phi_2(t_1) & \ldots & \phi_S(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \ldots & \phi_S(t_2) \\ & \vdots & & \\ \phi_1(t_N) & \phi_2(t_N) & \ldots & \phi_S(t_N) \end{bmatrix} \quad , \tag{15}$$

and B is the matrix of the B-spline coefficients (Bugli and Lambert, 2004):

$$B = \begin{bmatrix} \beta_{01} & \beta_{02} & \ldots & \beta_{0S} \\ \beta_{11} & \beta_{12} & \ldots & \beta_{1S} \\ & \vdots & & \\ \beta_{P1} & \beta_{P2} & \ldots & \beta_{PS} \end{bmatrix} \quad . \tag{16}$$

In the following equations, we transformed the notations of the above problem to obtain an index notation, which conveys the same information but is intended for human manipulation rather than machine calculation. We used the vectorial form of this model with the vec(.) operator defined by: $\underline{x} = \mathrm{vec}(X)$ such that $\underline{x}_{(j-1)N+i} = X_{ij}$ (Bugli and Lambert, 2004). We obtained:

$$\mathrm{vec}\left(Y^T\right) = \underline{y} = \begin{bmatrix} Y_{1,1} \\ Y_{1,2} \\ \vdots \\ Y_{1,N} \\ Y_{2,1} \\ \vdots \\ Y_{I,N} \end{bmatrix} = \begin{bmatrix} y_1(t_1) \\ y_1(t_2) \\ \vdots \\ y_1(t_N) \\ y_2(t_1) \\ \vdots \\ y_I(t_N) \end{bmatrix} \quad . \tag{17}$$

In the application of nucleic acid duplex discrimination, we used the following model:
Suppose that there were 10 curves measured at 50 temperatures, in 2 groups: 5 are PM and 5 are MM:

$$\mu = \left(\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \otimes \Phi \right) \mathrm{vec}\left(B^T\right) \tag{18}$$

$$= \begin{bmatrix} \Phi & 0 \\ \Phi & 0 \\ \Phi & 0 \\ \Phi & 0 \\ \Phi & 0 \\ 0 & \Phi \\ 0 & \Phi \\ 0 & \Phi \\ 0 & \Phi \\ 0 & \Phi \end{bmatrix} \begin{bmatrix} \beta^{(PM)} \\ \beta^{(MM)} \end{bmatrix} \tag{19}$$

$$= \tilde{X}\underline{b} \quad ,$$

where each block Φ denotes the matrix of B-spline bases with (N=50 rows and S=20 columns), and thus:

$$\Phi = \begin{bmatrix} \phi_1(t_1) & \phi_2(t_1) & \cdots & \phi_{20}(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \cdots & \phi_{20}(t_2) \\ & \vdots & & \\ \phi_1(t_{50}) & \phi_2(t_{50}) & \cdots & \phi_{20}(t_{50}) \end{bmatrix} \tag{20}$$

and

$$B = \begin{bmatrix} \beta^{(PM)^T} \\ \beta^{(MM)^T} \end{bmatrix} = \begin{bmatrix} \beta_1^{(PM)} & \cdots & \beta_{20}^{(PM)} \\ \beta_1^{(MM)} & \cdots & \beta_{20}^{(MM)} \end{bmatrix} \quad . \tag{21}$$

Matlab interface. A user-friendly interface of the functional ANOVA calculator was made in Matlab for automated statistical analysis of comparison of dissociation curves. In addition, the software was compiled to compare initial signal intensities and T_d values. The calculator is available at http://stahl.ce.washington.edu. It is an open-source. Users can view and alter the algorithms and add or substitute modules to perform different functions. A stand-alone executable version of the calculator was also created and is available at the same webpage. A user manual is provided and gives additional information for installation, requirements for the datasets, use, and performance of the calculator.

After loading the data, the functional ANOVA calculator computes and normalizes the individual curves. If specified by the user, detection and removal of outlier curves can be done. In this case, outliers are not included in the functional ANOVA model, calculation and comparison of T_d values. Then, an average curve and confidence interval with an α value of

0.05 are calculated for each group. The normality of initial signal intensities and T_d values are tested. Signal intensities and T_d values are compared between groups using Wilcoxon and Student's *t*-tests with an α value of 0.05. Figures are automatically created at each step and can be saved. A summary window shows the removed outliers (if specified) (Figure 1). The summary also shows pair of groups with significantly different dissociation curves and provides for these groups the area between the intervals of confidence of the pair of groups, the range of temperature for which curves are significantly different, the maximum difference in normalized signal intensities (MAX_{DCSD}, where DCSD stands for "dissociation curves significantly different") between the lower limit of the interval of confidence of the upper dissociation curve and the upper limit of the lower dissociation curve, and the temperature corresponding to the MAX_{DCSD}. The area, range of temperature and MAX_{DCSD} values indicate the extent of the discrimination between the 95% confidence intervals. The summary gives values of the T_d for each individual curve, as well as the mean and standard deviation for each group. Finally, the summary provides the results of the tests of normality, Wilcoxon and Student's *t*-tests for initial signal intensities and T_d values (Figure 1).

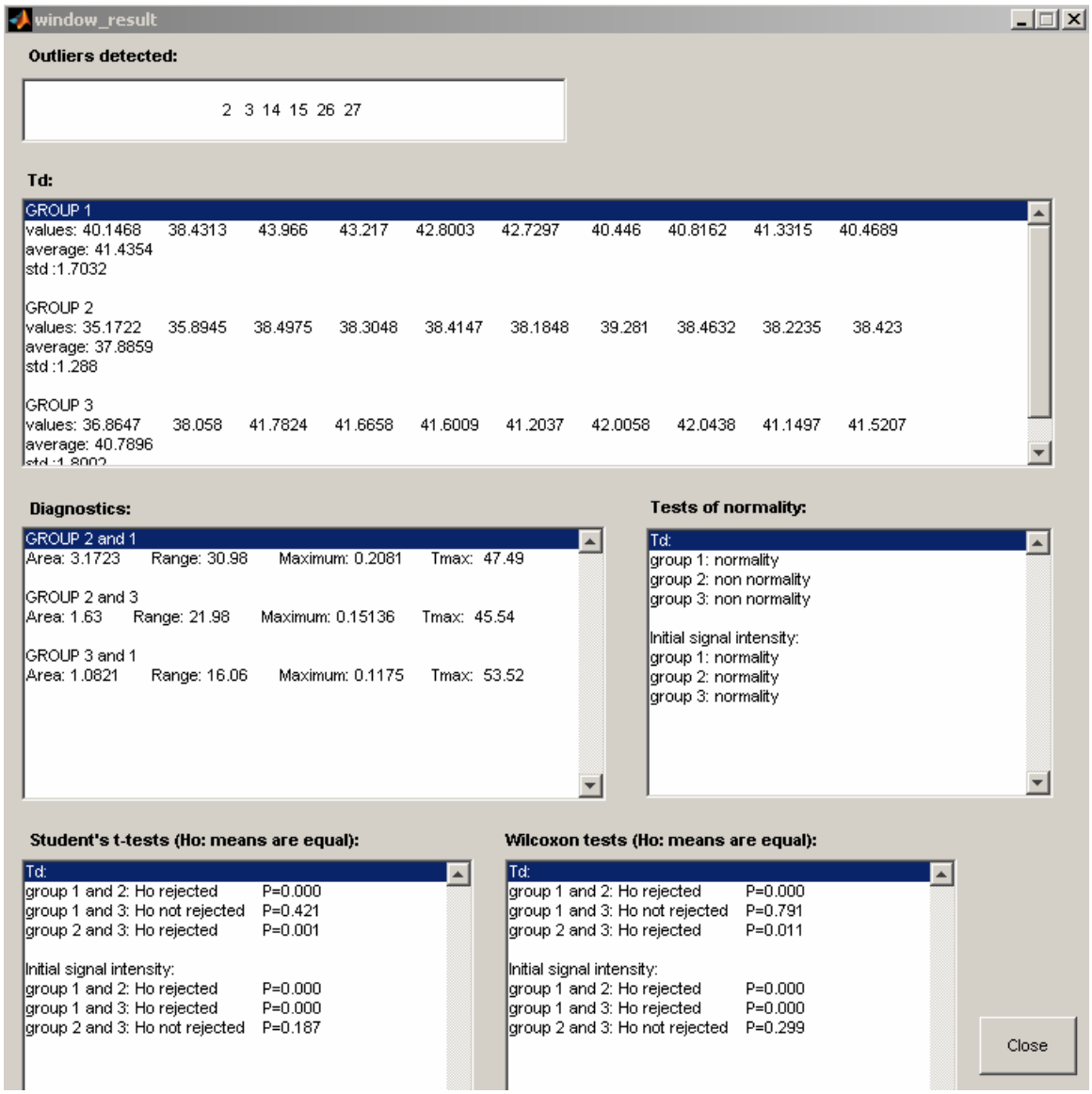


Figure 1. Screen snapshot of the output summary obtained using the functional ANOVA calculator with data of in vitro transcribed 16S rRNA of contaminated soil hybridized to probes Eub338 (group 1), Eub338-g11c (group 2) and Eub338-t7g (group 3).

3.4. Results and discussion

There are several approaches to nonparametric modelling: serial-based smoothers including wavelets (Tarter and Lock, 1993), kernel methods (Härdle, 1990; Silverman, 1986) including local regression (Fan and Gijbels, 1996; Wand and Jones, 1995), and splines methods including smoothing splines (Eubank, 1988; Green and Silverman, 1994; Wahba, 1990), regression splines (Friedman, 1991; Stone, *et al.*, 1997), and B-splines (de Boor, 2001; Dierckx, 1995). All these methods can be used efficiently, but the nature of the data is important in the choice among them. There is a growing interest in modelling longitudinal data with curves, in the style of ANOVA. Specifically, Ramsay and Silverman (1997) discussed the use of functional ANOVA models, and their use in several applications was recently presented (Brumback and Rice, 1998; Rice and Wu, 2001). The approach used in our paper is based on functional ANOVA with P-splines. Although the smoothing parameter, lambda, has drawbacks, the smoothness of the original data minimises the influence of lambda.

During the testing of the functional ANOVA model, we found that the shape of some individual dissociation curves was atypical. In other words, they presented a shape that deviated from the general sigmoid shape of the majority of the curves. Three distinct "classes" of outlier curves were identified in the dataset. The first class corresponded to curves with signal intensities sharply decreasing at temperatures lower than 35°C (Figure 2A and 2B). The shape of the second class of outliers was relatively linear instead of sigmoidal (Figure 3A and 3B). The third class of outliers consisted of sigmoid-shaped curves with unusual variation of the signal intensities at high temperatures (Figure 4A and 4B). The observed outliers may result from different experimental artefacts such as movement of the microarray during thermal dissociation analysis which results in an incorrect placement of the grid used to determine signal intensities of each microarray element, sporadic occurence of fluorescent particles, and formation of bubbles within the hybridization chamber at high temperatures.

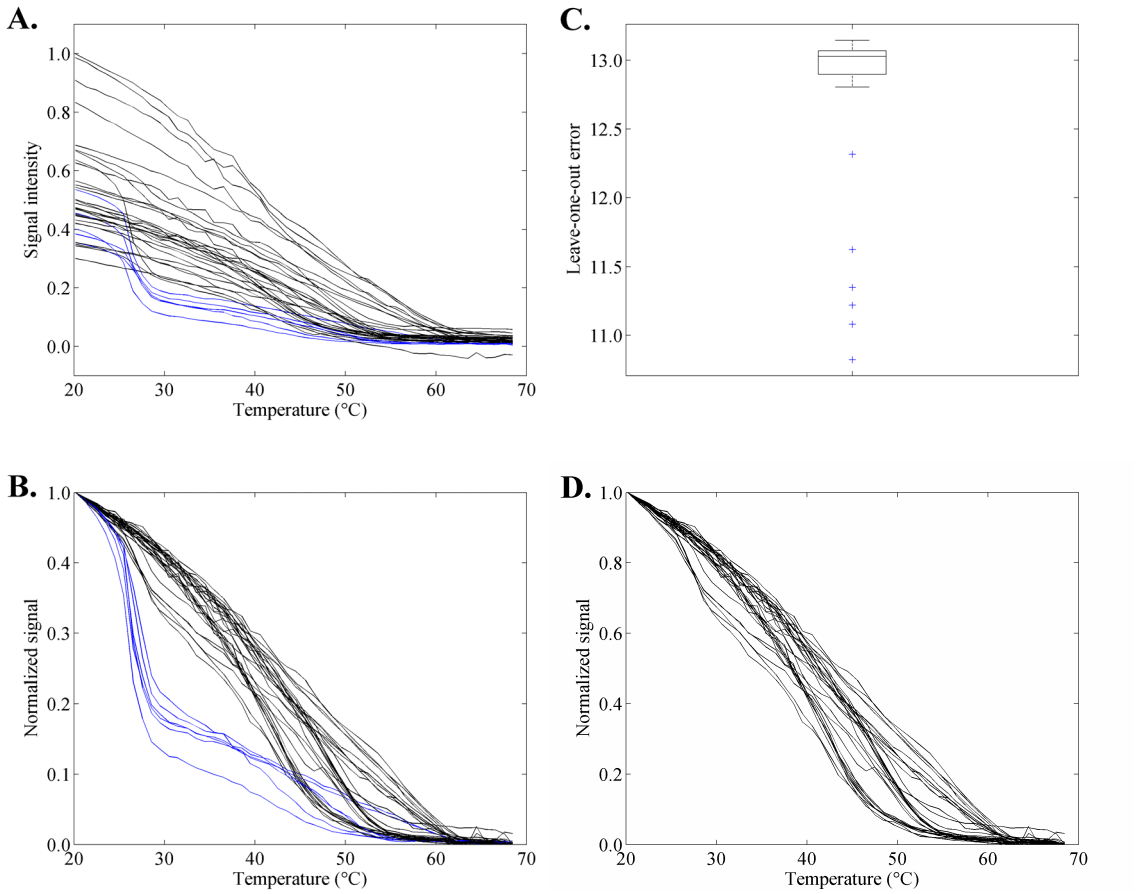


Figure 2. Detection of outlier dissociation curves of the first class with in vitro transcribed 16S rRNA of contaminated soil. Dissociation curves of probe Eub338, Eub338-g11c, and Eub338-t7g (**A**). Normalized dissociation curves of probe Eub338, Eub338-g11c, and Eub338-t7g (**B**). Boxplot[1] of the leave-one-out errors of normalized dissociation curves of probe Eub338, Eub338-g11c, and Eub338-t7g (**C**). Normalized dissociation curves of probe Eub338, Eub338-g11c, and Eub338-t7g after rejection of outliers (**D**). Outlier dissociation curves are indicated in blue.

[1] The upper edge of the box indicates the 75th percentile of the data set and the lower edge the 25th percentile. The median is shown by a line in the box. Vertical lines represent the maximum and minimum values, unless outliers (shown by a + cross) are present in which case the lines extend to a maximum of 1.5 times the inter-quartile range

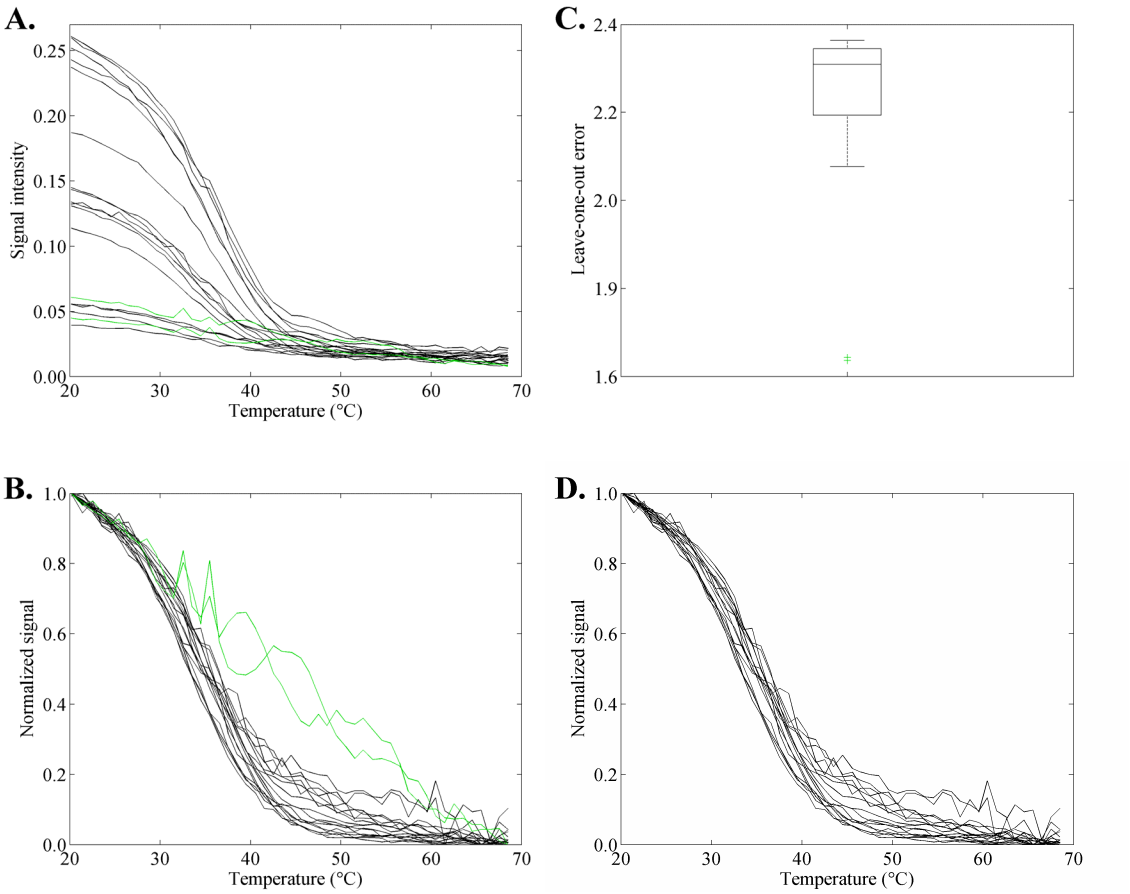


Figure 3. Detection of outlier dissociation curves of the second class with in vitro transcribed 16S rRNA of uncontaminated soil. Dissociation curves of probe Lowgc353, Lowgc353-a6c, and Lowgc353-c9g (**A**). Normalized dissociation curves of probe Lowgc353, Lowgc353-a6c, and Lowgc353-c9g (**B**). Boxplot of the leave-one-out errors of normalized dissociation curves of probe Lowgc353, Lowgc353-a6c, and Lowgc353-c9g (**C**). Normalized dissociation curves of probe Lowgc353, Lowgc353-a6c, and Lowgc353-c9g after rejection of outliers (**D**). Outlier dissociation curves are indicated in green.

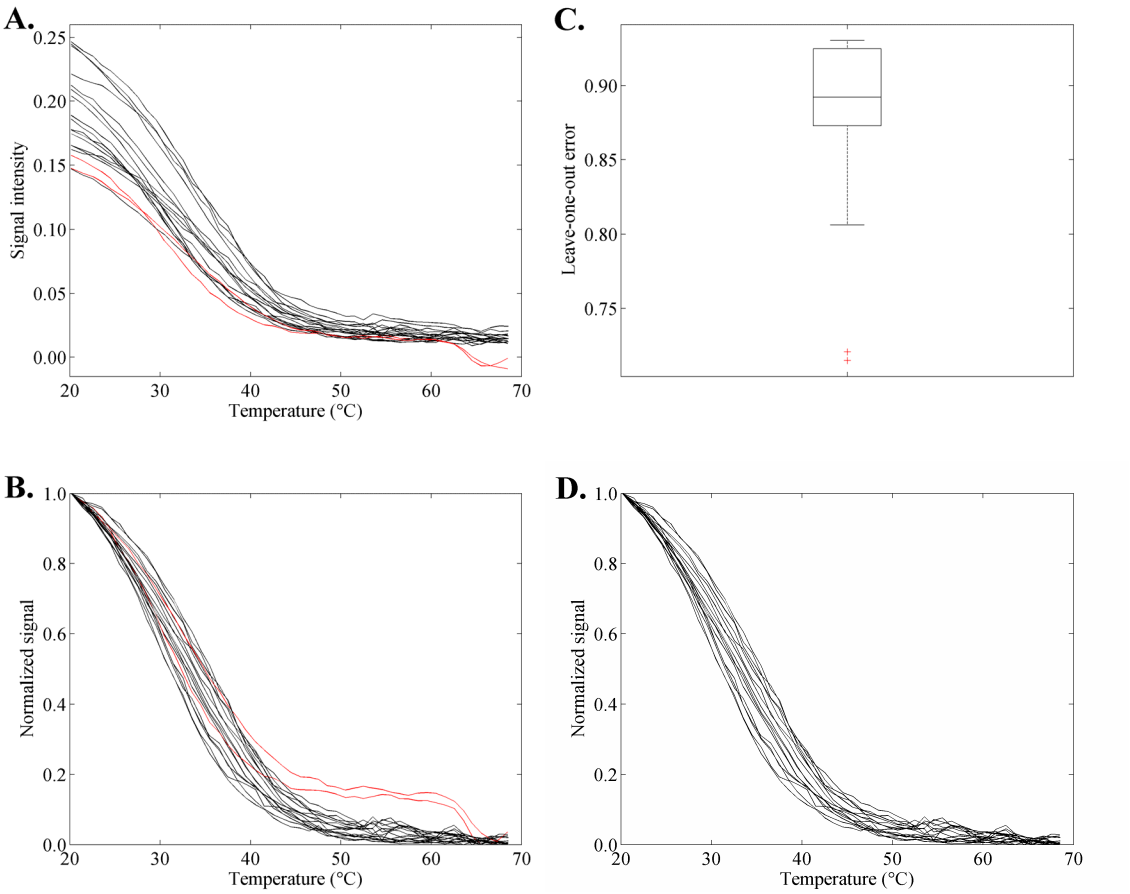


Figure 4. Detection of outlier dissociation curves of the third class with in vitro transcribed 16S rRNA of uncontaminated soil. Dissociation curves of probe Nse1472, Nse1472-g7c, and Nse1472-c15g (**A**). Normalized dissociation curves of probe Nse1472, Nse1472-g7c, and Nse1472-c15g (**B**). Boxplot of the leave-one-out errors of normalized dissociation curves of probe Nse1472, Nse1472-g7c, and Nse1472-c15g (**C**). Normalized dissociation curves of probe Nse1472, Nse1472-g7c, and Nse1472-c15g after rejection of outliers (**D**). Outlier dissociation curves are indicated in red.

The outlier curves affected the functional ANOVA model and the interpretation of dissociation curves. Figure 5A shows the average curves for each group obtained with the functional ANOVA calculator from the individual curves of Figure 2B.

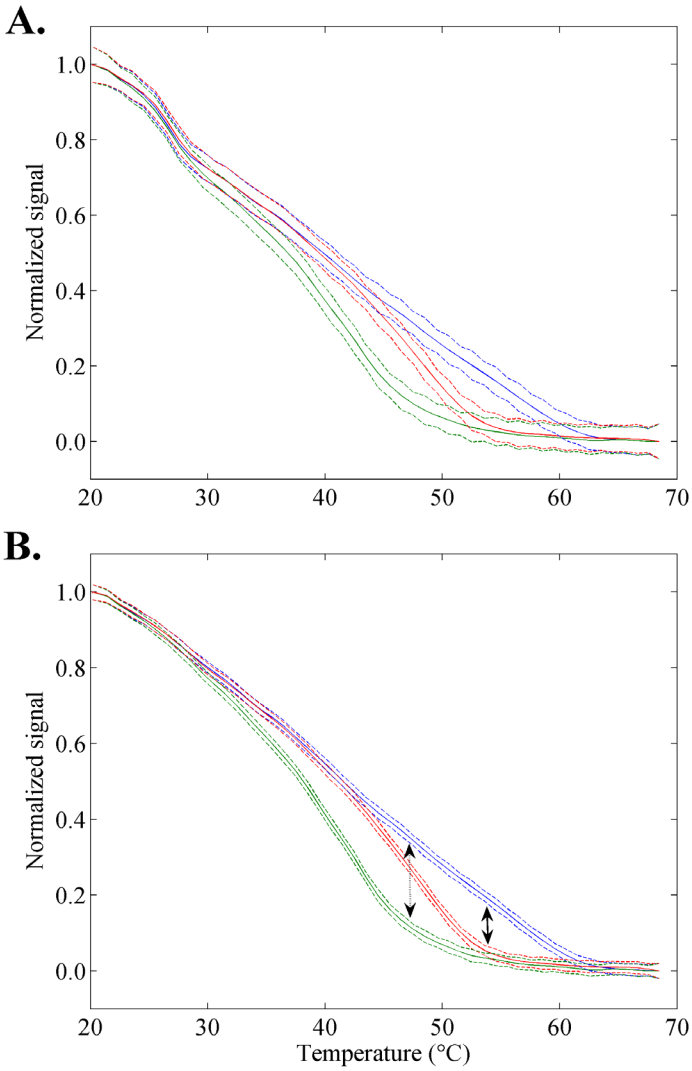


Figure 5. Dissociation curves analysed with the functional ANOVA calculator. Average normalized dissociation curves of probe Eub338 (blue line, *n*=12), Eub338-g11c (green line, *n*=12), and Eub338-t7g (red line, *n*=12) from hybridizations with in vitro transcribed 16S rRNA of contaminated soil (**A**). Average normalized dissociation curves of probe Eub338 (blue line, *n*=10), Eub338-g11c (green line, *n*=10), and Eub338-t7g (red line, *n*=10) from hybridizations with in vitro transcribed 16S rRNA of contaminated soil after rejection of outliers (**B**). Confidence bands are indicated by non-continuous lines. The dashed and solid arrows represent, respectively, the MAX_{DCSD} of Eub338/Eub338-g11c and Eub338/Eub338-t7g probe sets.

Removal of the outliers generally resulted in an improvement of the quality of the average curves, in terms of shape and confidence interval (i.e. smoothed curves with less variation along the sigmoid shape and smaller confidence intervals); and, subsequently, rejection of outliers provided a greater discrimination between average curves. Figure 5B shows the

average curves for each group from the individual curves of Figure 2D, i.e. after rejection of outlier curves. The outlier curves were detected with the functional ANOVA calculator using leave-one-out error, as it shows extreme values for influential data. The "leave-one-out error" is the error of the computed functional ANOVA model determined by successive single exclusion of each observed curve. The error was calculated by summing for each point the square difference between observed points and estimated points determined with the model. A boxplot of the leave-one-out errors for each curve allowed the detection of extreme values of the errors (Figure 2C, 3C, and 4C). Removal of the curves associated with these extreme values resulted in the rejection of outliers (Figure 2D, 3D, and 4D). Of note, a minimum of 4 curves for each group is needed to detect potential outlier curves in appropriate way as with less than 4 curves, a valid boxplot can not be drawn.

One hundred fifty-one individual curves from a total of 1548 were detected as outliers. Of these 151 curves, 91 were of the first class, 24 of the second class and 36 of the third class. The microarray elements producing these outliers were identified for each replicate hybridization (Figure 6). Outliers of the second and third class appeared at different locations on the microarray among the replicates. Most of the outliers of the first class were detected with the first replicate hybridization carried out with the contaminated soil (Figure 6D). In addition, these outliers were mainly located at the lower right region of the microarray. Together, these findings may indicate that the microarray of this replicate effectively moved during thermal dissociation in such way that it affected preferentially the curves collected at the lower right region of the microarray.

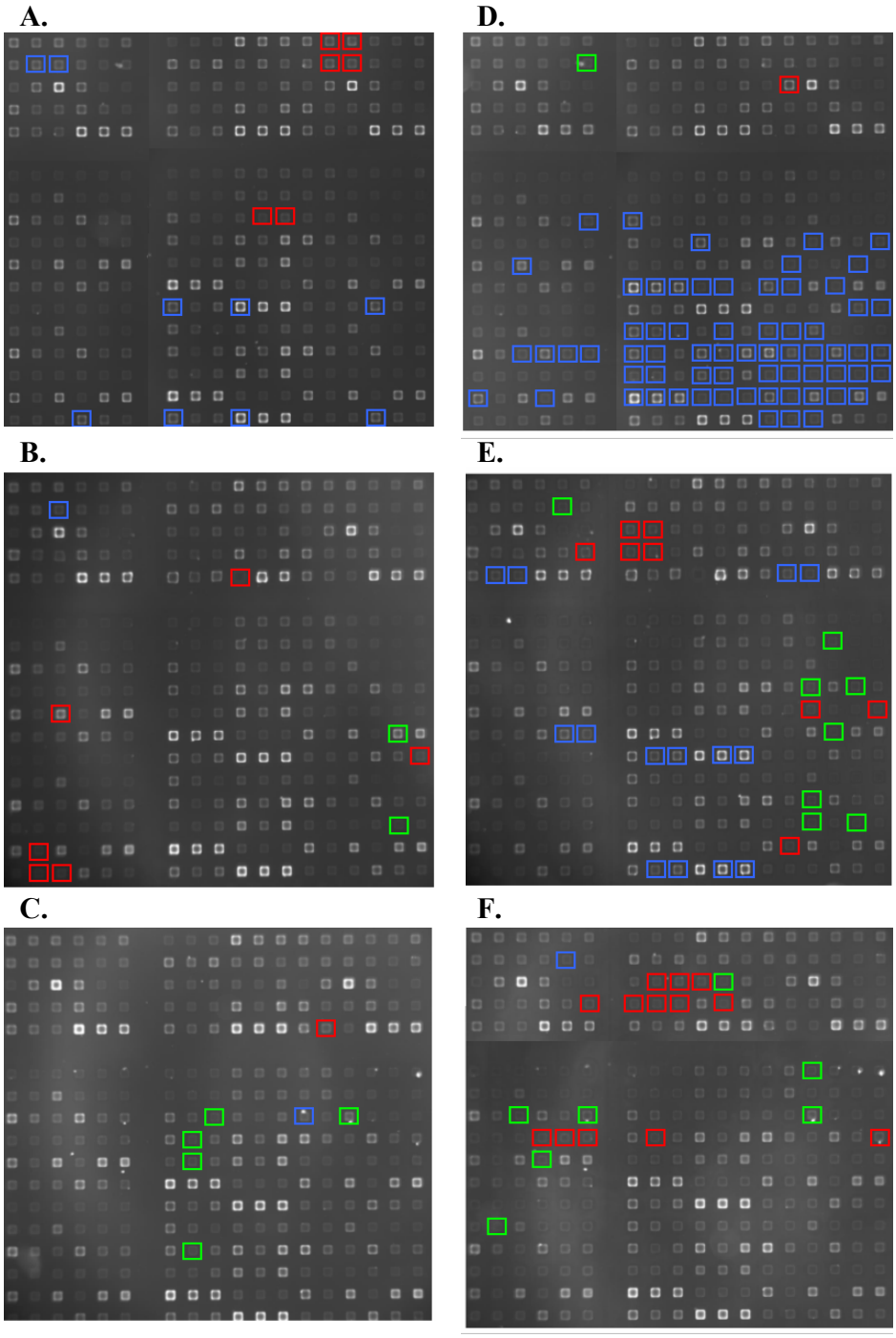


Figure 6. Images at 20°C of the microarray after hybridization with in vitro transcribed 16S rRNA of soil samples. Uncontaminated soil sample, first replicate (**A**). Uncontaminated soil sample, second replicate (**B**). Uncontaminated soil sample, third replicate (**C**). Contaminated soil sample, first replicate (**D**). Contaminated soil sample, second replicate (**E**). Contaminated soil sample, third replicate (**F**). Microarray elements with dissociation curves identified as outliers are indicated by colored squares (blue: outliers of the first class; green: outliers of the second class; red: outliers of the third class).

Validation of the functional ANOVA calculator was realised with data from hybridizations of nucleic acids from uncontaminated and 2,4,6-trinitrotoluene contaminated soil samples. The functional ANOVA calculator revealed significant differences (Student's *t*-tests) in initial signal intensities for most of the probes between the two soil samples and between the PM-MM probe sets (Eyers *et al.*, manuscript submitted). For instance, a significant difference in initial signal intensities of a contaminated soil sample hybridized to the PM probe Eub338 and the MM probe Eub338-t7g was detected (Figure 7A). As previously mentioned, Hacia *et al.* (1996) subtracted signal intensities of MM duplexes to PM duplexes to take into account the presence of non-specific hybridizations to the PM probe. However, the sample used in this study may contain PM and MM targets in unknown abundance. For this reason, the approach used by Hacia *et al.* is not suitable. To analyse the specificity of each probe, it was necessary to compare the dissociation curves. Comparison of T_d values calculated from the dissociation curves of probes Eub338 and Eub338-t7g hybridized to contaminated soil showed no significant differences (Student's *t*-test) between these probes (Figure 7B). However, functional ANOVA effectively discriminated the dissociation curves for these probes. The confidence bands of the dissociation curves of probe Eub338 and Eub338-t7g did not overlap between 44 and 60°C (Figure 5B); and, thus, between this range of temperature, the curves were significantly different. The T_d values were not significant because discrimination did not occur until after 50% probe-target dissociation (Figure 5B).

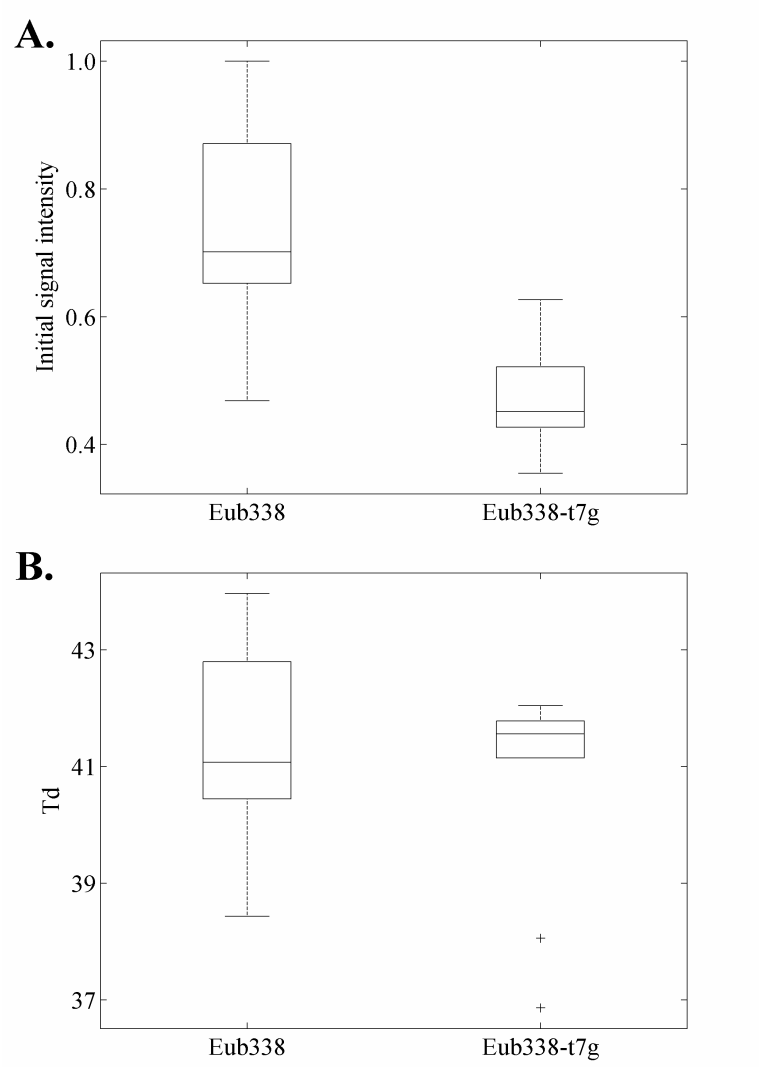


Figure 7. Comparison of hybridizations with in vitro transcribed 16s rRNA of the contaminated soil sample. Boxplot of initial signal intensities of probe Eub338 and Eub338-t7g (**A**). Boxplot of T_d values of probe Eub338 and Eub338-t7g (**B**).

Among the 126 comparisons made between groups (Eyers *et al.*, manuscript submitted), 55 were significantly different (Student's *t*-tests) with the T_d approach. An additional 31 comparisons were significantly different with the functional ANOVA approach. For these 31 comparisons, discrimination generally occurred at higher temperatures than the respective T_d values. Therefore, functional ANOVA had a higher performance for detecting significantly different curves than the T_d approach. Moreover, functional ANOVA offered the possibility to determine the temperature at which discrimination between curves was maximal, as well as the range of temperatures over which the curves were discriminated.

3.5. Conclusions

In this paper, we demonstrated the application of functional ANOVA to discriminate among dissociation curves of different probe-target nucleic acid duplexes. Functional ANOVA with P-splines was shown to be a very useful tool to model curves. Indeed, the model was easily and clearly stated. The model allowed a straightforward comparison of the shapes of the curves, and influent outlier curves were easily detected and removed. It also had an increased sensitivity to detect significantly different curves than other approaches based on comparisons at a single temperature. Finally, the model was tested and validated with microarray data, but it is also applicable to other molecular techniques that generate dissociation curves, such as those derived from the melting of PCR amplicons.

Acknowledgements

Philippe Lambert thanks the IAP network nr P5/24 of the Belgian state (Federal Office for Scientific, Technical and Cultural Affairs). The “Fonds National pour la Recherche Scientifique (FNRS)”, Belgium is gratefully acknowledged for financial support through a FRIA research grant. Financial support through a patronage of Eli Lilly and Company is also gratefully acknowledged. Additional support was received from NIH/NIDCR (U01 DE14955) and US DARPA (DABT63-99-1-0009) to D.A.S., and NASA (MSMT-2004-0045-0066) to D.A.S and J.C.S.

3.6. References

Brumback,B. and Rice,J. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Am. Stat. Ass.*, **93**, 961-976.

Bugli,C. and Lambert,P. (2004) Functional ANOVA with random functional effects: an application to event-related potentials modelling for electroencephalograms analysis. *Stat. Med.*, in press.

Bullido,M.J. *et al.* (1998) A polymorphism in the regulatory region of *APOE* associated with risk for Alzheimer's dementia. *Nat. Genet.*, **18**, 69-71.

de Boor,C. (2001) *A practical guide to splines*. Springer, NY.

Dierckx,P. (1995) *Curve and surface fitting with splines*. Clarendon, Oxford.

Eilers,P. and Marx,B. (1996) Flexible smoothing with B-splines and penalties (with discussion). *Stat. Science*, **11**, 89-121.

El Fantroussi,S. *et al.* (2003) Direct profiling of environmental microbial populations by thermal dissociation analysis of native rRNAs hybridized to oligonucleotide microarrays. *Appl. Environ. Microbiol.*, **69**, 2377-2382.

Eubank,R.L. (1988) *Spline smoothing and nonparametric regression*. Marcel Dekker, NY.

Fan,J. and Gijbels,I. (1996) *Local polynomial modelling and its applications*. Chapman and Hall, London.

Fotin,A.V. *et al.* (1998) Parallel thermodynamic analysis of duplexes on oligodeoxyribonucleotide microchips. *Nucleic Acids Res.*, **26**, 1515-1521.

Friedman,J.H. (1991) Multivariate adaptive regression splines. *Ann. Stat.*, **19**, 1-141.

Green,P. and Silverman,B.W. (1994) *Nonparametric regression and generalized linear models*. Chapman and Hall, London.

Hacia,J.G. *et al.* (1996) Detection of heterozygous mutations in *BRCA1* using high density oligonucleotides arrays and two-colour fluorescent analysis. *Nat. Genet.*, **14**, 441-447.

Härdle,W. (1990) *Smoothing techniques with implementation in S*. Springer-Verlag, NY.

Hastie,T. and Tibshirani,R. (1991) *General additive models for medical research*. Chapman and Hall, London.

Henderson,H. and Searle,S. (1981) The vec-permutation matrix, the vec operator and Kroenecker product: a review. *Lin. Multilin. Algebra*, **9**, 271-288.

Liew,M. *et al.* (2004) Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. *Clin. Chem.*, **50**, 1156-1164.

Liu,W.T. *et al.* (2001) Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. *Environ. Microbiol.*, **3**, 619-629.

Miki,Y. *et al.* (1994) A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science*, **266**, 66-71.

O'Sullivan,F. (1986) A statistical perspective on ill-posed inverse problems. *Stat. Science*, **1**, 505-527.

O'Sullivan,F. (1988) Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Comput.*, **9**, 363-379.

Ramsay,J.O. and Silverman,B.W. (1997) *Functional data analysis*. Springer, NY.

Rice,J. and Wu,C. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253-259.

Ririe,K.M. *et al.* (1997) Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. *Anal. Biochem.*, **245**, 154-160.

Ruppert,D. *et al.* (2003) *Semiparametric regression*. Cambridge University Press, NY.

Silverman,B.W. (1986) *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Stone,C. *et al.* (1997) Polynomial splines and their tensor products in extended linear modelling. *Ann. Stat.*, **25**, 1371-1470.

Tarter,M.E. and Lock,M.D. (1993) *Model-free estimation*. Chapman and Hall, NY.

Urakawa,H. *et al.* (2003) Optimization of single-base-pair mismatch discrimination in oligonucleotide microarrays. *Appl. Environ. Microbiol.*, **69**, 2848-2856.

Wahba,G. (1990) Spline models for observation data. *Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia.

Wand,M.P. and Jones,M.C. (1995) *Kernel smoothing*. Chapman and Hall, NY.

Wittwer,C. *et al.* (2003) High-resolution genotyping by amplicon melting analysis using LCGreen, *Clin. Chem.*, **49**, 853-860.

Wooster,R. *et al.* (1995) Identification of the breast cancer susceptibility gene *BRCA2*. *Nature*, **378**, 789-792.