BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 14 2011, pages 1979–1985
doi:10.1093/bioinformatics/btr289

Data and text mining

A statistical framework for biomarker discovery in metabolomic time course data

Maurice Berk[1], Timothy Ebbels[2] and Giovanni Montana[1,*]

[1]Statistics Section, Department of Mathematics, Imperial College London, Huxley Building and [2]Biomolecular Medicine, Department of Surgery and Cancer, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Metabolomics is the study of the complement of small molecule metabolites in cells, biofluids and tissues. Many metabolomic experiments are designed to compare changes observed over time under two experimental conditions or groups (e.g. a control and drug-treated group) with the goal of identifying discriminatory metabolites or *biomarkers* that characterize each condition. A common study design consists of repeated measurements taken on each experimental unit thus producing time courses of all metabolites. We describe a statistical framework for estimating time-varying metabolic profiles and their within-group variability and for detecting between-group differences. Specifically, we propose (i) a smoothing splines mixed effects (SME) model that treats each longitudinal measurement as a smooth function of time and (ii) an associated functional test statistic. Statistical significance is assessed by a non-parametric bootstrap procedure.

Results: The methodology has been extensively evaluated using simulated data and has been applied to real nuclear magnetic resonance spectroscopy data collected in a preclinical toxicology study as part of a larger project lead by the COMET (Consortium for Metabonomic Toxicology). Our findings are compatible with the previously published studies.

Availability: An R script is freely available for download at http://www2.imperial.ac.uk/~gmontana/sme.htm.

Contact: g.montana@imperial.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 14, 2011; revised on April 23, 2011; accepted on May 2, 2011

1 INTRODUCTION

Metabolomics, or metabolic profiling, experiments monitor the levels of the myriad of small molecules in biological systems and provide a snapshot of the metabolic state of the organism under study (Nicholson *et al.*, 1999; Raamsdonk *et al.*, 2001). Metabolomics is a key part of the systems biology approach to biological problems, reporting on a different layer of biomolecular organization to that assayed by proteomics and transcriptomics. To assay metabolite levels, metabolomics studies employ spectroscopic or spectrometric methods such as nuclear magnetic resonance (NMR) spectroscopy or mass spectrometry (MS), resulting in high quantities of complex yet information rich data. This complexity necessitates sophisticated statistical and bioinformatic approaches to data analysis both at the level of the raw data processing (e.g. peak detection, alignment, etc.) and also when investigating clustering, classification and other biological features in the data (Ebbels and Cavill, 2009; Goodacre, 2007; Steuer *et al.*, 2007).

One of the most common goals in metabolomic experiments is to discover biomarkers: metabolites whose concentrations are associated with metabolic status. For example, one may search for metabolites that respond to a biological stimulus or which change in concentration between healthy and diseased individuals. This goal extends to time series experiments, where a common objective is to find metabolites whose time profiles show significant differences between conditions. Such metabolites may reveal novel insights into the complex regulatory mechanisms underlying normal physiology and how they are altered in pathological conditions.

Figure 1 shows an example of selected time series for one spectral bin (or variable) taken from a real toxicology study on rats, which provides some insight into the difficulties encountered (see Section 2 for more details on this study). First, the time series are extremely short and it is rare to find metabolomics datasets with more than 3−10 time points. In contrast, classical approaches to time series analysis typically require several tens of data points in order to adequately model the time variation (Box *et al.*, 1994). Second, even within the same experimental groups, there can be a lot of biological variability; for instance, the temporal profiles for the three biological replicates (distinct biological units, in this case rats) shown in Figure 1 are different, although they were all observed under the same experimental conditions. Being able to accurately model this variability is therefore critical in order to detect meaningful temporal patterns. Third, there are missing observations, either by design or because they were deemed to be outliers and hence removed or due to errors in the measurement process. Many classical approaches would be unable to handle these missing observations without resorting to imputation procedures. Finally, the observations are clearly temporally correlated and the response is non-linear. To date, most analyses of time series data in metabolomics do not explicitly include the time ordering in the model; that is, the same results would be obtained if the time order of the data was permuted. An obvious consequence of ignoring the time ordering is that information is lost, resulting in less power to discover time-related structure. A notable exception is the batch modeling approach in which the time-ordered multivariate data are regressed against the time variable using Partial

*To whom correspondence should be addressed.

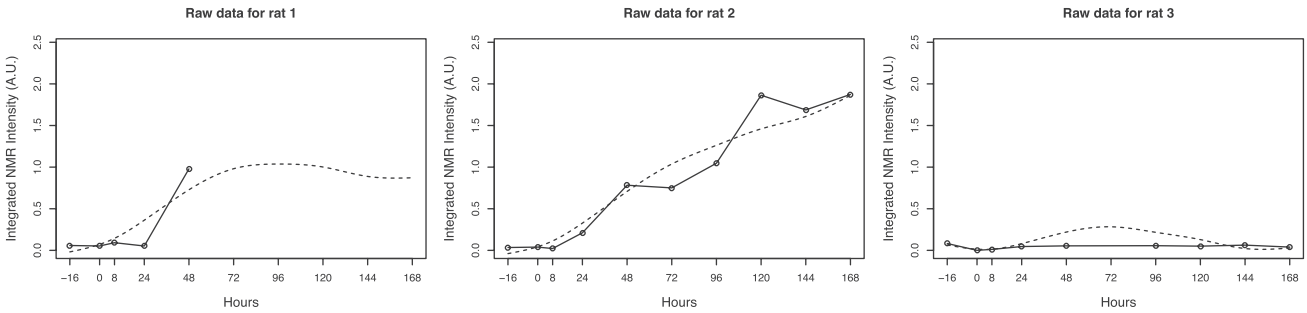

Fig. 1. Short individual time series observed for one metabolite (creatine, 3.0395 ppm) for three example rats in the high-dose group of our example dataset. Shown are the raw observations with fitted individual curves from our SME model (dashed lines).

Least Squares (Antti *et al.*, 2002). This approach, however, implicitly assumes linear trends with respect to time and does not make use of any temporal autocorrelation structure present in the data. Other difficulties presented by the data include the high dimensionality, with thousands of spectral bins under study simultaneously, strong degree of collinearity between variables and the presence of noise.

Smilde *et al.* (2010) provides a comprehensive review of methods appropriate for the modelling of metabolomics time course datasets, ranging from differential equations to state-space models, while noting that the number of approaches is very limited and that they often fail to correctly take in to account the time ordering. Many of the methods discussed are motivated by datasets with many more time points than our example case study (145 versus 10) or where the focus is on modelling pre-selected metabolites rather than identifying potentially novel biomarkers. It is this established gap in the metabolomics data analysis toolbox that has motivated the development of our proposed statistical framework.

Our framework is designed for modelling short, high-dimensional and heterogenous metabolomic profiles and detecting significant differences between two biological groups. The first part of the framework is our proposed *smoothing spline mixed effects* (SME) model which sets out to address the challenges described above. The SME model has its roots in functional data analysis (FDA) (Ramsay and Silverman, 2006), a rapidly developing area of statistics that has been successfully applied to longitudinal data arising in genomics experiments that exhibit very similar characteristics to metabolic profiling data (Bar-Joseph *et al.*, 2003; Luan and Li, 2003; Ma *et al.*, 2006; Storey *et al.*, 2005) and other authors have noted the importance of taking inspiration from transcriptomics in developing metabolomics data analysis approaches (Smilde *et al.*, 2010). In the FDA paradigm, each observed time series is seen as the realization of an underlying stochastic process or smooth curve that needs to be estimated. These estimated curves are then treated as the basic observational unit in subsequent data analysis, such as the task of detecting variables that exhibit a significant difference between two groups that we focus on in this article. The second part of our framework, therefore, is our proposed moderated functional t-type statistic for quantifying the difference between two sets of curves that makes full use of the estimated within-group variability obtained from the SME model. Several solutions to this problem have been proposed in the literature such as the use of the functional L_2 distance (Cuevas *et al.*, 2004) or the heuristic procedure of Cox and Lee (2008) which discretizes the curves on a fine grid and then carries out point-wise t-tests. Our proposed test statistic builds on these approaches by *borrowing strength* across all variables in order to increase power.

This article is structured as follows. The suggested methodology, based on functional mixed-effect models, is described in Section 3. The experimental results are presented in Section 4. We summarize with some conclusions in Section 5.

2 BIOMARKER DISCOVERY AND THE COMET STUDY

Time series experiments are routine in metabolomics, and it is accepted that it is hard to understand a biological system without observing its behaviour over time. We note that one does not usually attempts to model metabolic reaction dynamics, which happen on very short-time scales, but instead wishes to characterize much slower processes such as disease progression or physiological response, which entail sampling over time scales of hours to months. For a given system, it is usually not possible to say *a priori* which time point best captures the desired effects of a stimulus and thus at least several time points must be sampled. These considerations, coupled with the expense of collecting and assaying many samples, lead to experimental designs with very small numbers of time points.

The Consortium for Metabonomic Toxicology (COMET) (Lindon *et al.*, 2005) illustrates these issues in the area of pre-clinical toxicology. The COMET project built a large database of metabolic profiles from laboratory animals treated with model toxins, with the aims of discovering biomarkers of key toxicological processes, thus aiding in the understanding of metabolic toxicology and improving predictions of the toxicological impact for unknown chemicals (e.g. a new drug). Here, we use one study from the COMET project, in which the temporal response of laboratory rats to the liver toxin hydrazine was assayed by ^{1}H NMR spectroscopy of their urine. A summary of the experimental set up is provided below and further details can be found in Bollard *et al.* (2005). In the study, 30 Sprague-Dawley rats were randomly assigned to three treatment groups (controls, low dose and high dose) in equal sample sizes of 10. Urine samples were collected at $0-8$ h and $8-24$ h on the day prior to treatment, and further samples were collected at 8, 24, 48, 72, 96, 120, 144 and 168 h (7 days) post-treatment. Five animals in each group were sacrificed at 48 h post-dose for histopathological and clinical chemistry evaluation. The ^{1}H NMR spectra were acquired at 600 MHz and automatically preprocessed

using an in-house MATLAB routine, including normalization to a constant total integrated intensity. The final dataset consisted of 8020 NMR intensities (bins) observed at 10 time points for all rats but those that were sacrificed at 48 h post-dose observed only 5 time points.

In this article, we use the SME approach to discover NMR spectral signals whose time profiles discriminate between the control and high-dose groups.

3 METHODS

3.1 Functional mixed-effects models

We postulate that the true levels of the metabolites, as representative of underlying biological processes, vary smoothly over time as part of normal biological cycles and in response to stimuli. We consider the two group setting and for clarity label the groups arbitrarily as control (C) and treatment (T). Within a given spectral bin, we let the integrated NMR intensity observed on individual (or equivalently, *biological* replicate) i belonging to group $k \in \{C,T\}$ at time t_{ij} be denoted by $y(t_{ij})$ where $i=1,2,\ldots,n_k, j=1,2,\ldots,m_i$, n_k is the sample size in group k and m_i is the number of observations for individual i for this spectral bin. We model the observations as

$$y(t_{ij}) = f(t_{ij}) + \epsilon_{ij} \tag{1}$$

where $f(\cdot)$ is the true smooth function we wish to estimate, that is the underlying metabolite volumes as a function of time. In order to account for the differing response between replicates, we take $f(\cdot)$ to be the additive sum of two components: a mean function $\mu(\cdot)$, the mean response across all individuals and an individual-specific deviation from this mean curve, $v_i(\cdot)$. That is, we assume that

$$f(t_{ij}) = \mu(t_{ij}) + v_i(t_{ij}) \tag{2}$$

Both $\mu(\cdot)$ and $v_i(\cdot)$ are treated as smooth functions of time. However, whereas $\mu(\cdot)$ is assumed to be a fixed, yet unknown, population curve, $v_i(\cdot)$ is treated as a random realization of an underlying Gaussian process with zero-mean and covariance function $\gamma(s,t)$. Finally, the additive error term ϵ_{ij} is assumed to have some group and spectral bin-specific variance σ^2.

Under these assumptions, model (1) is a functional mixed-effects model (Guo, 2002; Wu and Zhang, 2006). In practice, the infinite dimensional functions $\mu(\cdot)$ and $v_i(\cdot)$ must be projected onto some finite dimensional basis by choosing a suitable parameterization. In FDA, a common choice is regression splines, which are piecewise polynomials formed by placing *knots* that divide the time course up into distinct regions, within which a low degree polynomial is fit (de Boor, 1978). Judicious choice of the number and location of these knots is necessary for obtaining a good fit to the data while minimizing the number of model parameters. As an alternative to this approach, we propose to represent the fixed and random-effects as smoothing splines (Green and Silverman, 1994). With smoothing splines, a knot is placed at each distinct design time point, then the model is fit with a constraint on the curve's *roughness* (or conversely its smoothness), controlled through a single non-negative real valued smoothing parameter λ. As there can only be an integer number of knots in regression splines, smoothing splines offer a finer degree of control over the resulting smooth (Wu and Zhang, 2006). Furthermore, the task of model selection is reduced from determining both the number and location of the knots to a 1D optimization over the smoothing parameter. An example illustrating this process is given in Section A.2 of Supplementary Material. We call model (1)–(2) where $\mu(t)$ and $v_i(t)$ are represented as smoothing splines a *Smoothing splines Mixed-effects* or SME model. We contrast this with the widely used LME abbreviation standing for *Linear Mixed-effects* models.

A natural way of measuring the smoothness of a curve is by the means of its integrated squared second derivative, assuming that the curve is twice-differentiable. We call $\boldsymbol{\mu} = (\mu(\tau_1),\ldots,\mu(\tau_m))^T$ the vector containing the values of the mean curve estimated at all design time points and, analogously, the individual-specific deviations from the mean curve, for individual i, are collected in $\boldsymbol{v}_i = (v_i(\tau_1),\ldots,v_i(\tau_m))^T$. With this notation in place, the mean curve and individual curve are represented as, respectively, $\mu(t_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\mu}$ and $v_i(t_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{v}_i \quad i=1,2,\ldots,n$, where $\boldsymbol{x}_{ij} = (x_{ij1},\ldots,x_{ijm})^T$ and each element is

$$x_{ijr} = \begin{cases} 1 & \text{if } t_{ij} = \tau_r,\ r=1,\ldots,m \\ 0 & \text{else} \end{cases}$$

The fact that the individual curves are assumed to be realizations of a Gaussian process is captured by assuming that the individual deviations are random and follow a zero-centred Gaussian distribution with covariance $\boldsymbol{D}$, where $\boldsymbol{D}(r,s) = \gamma(\tau_s,\tau_r)$, $r,s=1,\ldots,m$. In matrix form, the suggested model can then be rewritten as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\mu} + \boldsymbol{X}_i\boldsymbol{v}_i + \boldsymbol{\epsilon}_i \tag{3}$$
$$\boldsymbol{v}_i \sim \text{MVN}(\boldsymbol{0},\boldsymbol{D}) \qquad \boldsymbol{\epsilon}_i \sim MVN(\boldsymbol{0},\boldsymbol{R}_i)$$

For simplicity, we assume that $\boldsymbol{R}_i = \sigma^2\boldsymbol{I}_{m_i}$. Clearly, the model accounts for the fact that, for a given variable, the observed repeated measurements for each biological replicate are correlated. Specifically, under the assumptions above, $\text{Cov}(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{D}\boldsymbol{X}_i^T + \sigma^2\boldsymbol{I}_{m_i}$ which separates out the between-individual variability around the mean curve and the variability due to measurement error. The model parameters to be estimated are the fixed-effects $\boldsymbol{\mu}$, the random-effects $\boldsymbol{v}_i$ and the variance components $\boldsymbol{D}$ and σ^2.

3.2 Parameter estimation and model selection

The proposed model (3) is in the form of a standard linear mixed-effects model. Common practice with parameter estimation in these models is to estimate the fixed- and random-effects by maximizing their joint likelihood (Robinson, 1991), equivalent to minimising the generalised log-likelihood (GLL). As we represent the mean and biological replicate curves with smoothing splines, we minimize the *penalised* GLL (PGLL),

$$\text{PGLL} = \sum_i^{n_k}\Big[\frac{1}{\sigma^2}||\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\mu} - \boldsymbol{X}_i\boldsymbol{v}_i||_2 + \log|\boldsymbol{D}| + \boldsymbol{v}_i^T\boldsymbol{D}^{-1}\boldsymbol{v}_i + m_i\log\sigma^2 + \lambda_v\boldsymbol{v}_i^T\boldsymbol{G}\boldsymbol{v}_i\Big] + \lambda_\mu\boldsymbol{\mu}^T\boldsymbol{G}\boldsymbol{\mu}$$

This is the usual GLL with the addition of two terms accounting for the roughness of the fixed- and random-effects, with corresponding smoothing parameters λ_μ and λ_v and a *roughness matrix* $\boldsymbol{G}$ dependent upon the design time points and defined such that

$$\boldsymbol{\mu}^T\boldsymbol{G}\boldsymbol{\mu} = \int_{t_{min}}^{t_{max}}[\mu''(t)]^2\text{dt}, \quad \boldsymbol{v}_i^T\boldsymbol{G}\boldsymbol{v}_i = \int_{t_{min}}^{t_{max}}[v_i''(t)]^2\text{dt}$$

where $t_{\min}$ and $t_{\max}$ are the lower and upper bounds of the time course. A unique smoothing parameter λ_v is shared among all individuals as each curve is assumed to come from the same underlying Gaussian process. By collecting the terms involving $\boldsymbol{v}_i$, the PGLL can be rewritten in terms of the regularized covariance matrix $\boldsymbol{D}_v = (\boldsymbol{D}^{-1} + \lambda_v\boldsymbol{G})^{-1}$. This leads us to incorporate the roughness penalty on the random effects by refining our distributional assumptions so that $\boldsymbol{v}_i \sim MVN(\boldsymbol{0},\boldsymbol{D}_v)$ and hence $\boldsymbol{V}_i = \boldsymbol{X}_i\boldsymbol{D}_v\boldsymbol{X}_i^T + \sigma^2\boldsymbol{I}_{m_i}$.

Minimization of the PGLL with respect to $\boldsymbol{\mu}$ and $\boldsymbol{v}_i$ gives the best linear unbiased estimators/predictors of the fixed- and the random-effects as

$$\hat{\boldsymbol{\mu}} = (\textstyle\sum_i^{n_k}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i + \lambda_\mu\boldsymbol{G})^{-1}\sum_i^{n_k}\boldsymbol{X}_i\boldsymbol{V}_i^{-1}\boldsymbol{y}_i$$
$$\hat{\boldsymbol{v}}_i = \boldsymbol{D}_v\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\mu})$$

which rely on the unknown variance components $\boldsymbol{D}$ and σ^2. Had the random effects and errors been observed, the maximum likelihood (ML) estimators of these variance components would be $\boldsymbol{D} = 1/n_k\sum_{i=1}^{n_k}\boldsymbol{v}_i\boldsymbol{v}_i^T$ and $\sigma^2 = 1/N\sum_{i=1}^{n_k}\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i$. As they are generally unknown, however, we treat them as missing data and employ the Expectation–Maximization (EM) algorithm (Dempster *et al.*, 1977) for ML estimation. The sufficient statistics of $\boldsymbol{D}$ and

σ^2 are $\boldsymbol{v}_i\boldsymbol{v}_i^T$ and $\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i$, respectively, for all i. The conditional expectations required by the expectation step of the EM algorithm are $E[\boldsymbol{v}_i\boldsymbol{v}_i^T|\boldsymbol{y}_i]$ and $E[\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i|\boldsymbol{y}_i]$ which, under the assumptions posed, are given by

$$E[\boldsymbol{v}_i\boldsymbol{v}_i^T|\boldsymbol{y}_i]=\boldsymbol{v}_i\boldsymbol{v}_i^T+\boldsymbol{D}_v-\boldsymbol{D}_v\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\boldsymbol{D}_v$$
$$E[\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i|\boldsymbol{y}_i]=\boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i+m_i\sigma^2-\sigma^2tr(\boldsymbol{V}_i^{-1})$$

where $tr(\cdot)$ denotes the matrix trace. In the maximisation part of the algorithm, the ML estimators are used to obtain updated estimates of the variance components with the sufficient statistics replaced by their conditional expectations.

During the EM algorithm estimation procedure, it is assumed that the smoothing parameters λ_μ and λ_v are fixed. We search the 2D space $(\Lambda_\mu\times\Lambda_v)$ for values of λ_μ and λ_v that optimize the *corrected* Akaike Information Criterion (AICc) (Hurvich *et al.*, 1998), which includes a correction for small sample sizes, defined as

$$\text{AICc}=\text{AIC}+\frac{2k(k+1)}{n-k-1}$$

where $AIC=-2\mathcal{L}+2k$ is the standard AIC, $\mathcal{L}$ denotes the log-likelihood, k is the number of model parameters and n is the sample size. In our setting, n is the total number of observations for the given group and spectral bin across all individuals and k is the combined degrees of freedom of the smoothers of $\mu(\cdot)$ and $v_i(\cdot)$. The search is implemented using downhill simplex optimization (Nelder and Mead, 1965).

3.3 Moderated functional *t*-type statistic

The overarching goal of the analysis is to detect those spectral bins and hence metabolites which show a significant difference between the control and treatment groups. We, therefore, require a test statistic that quantifies the degree to which the two groups differ. After fitting model (3) to the data, independently for each spectral bin and each group, we obtain the estimated mean curves $\hat{\mu}_C(\cdot)$ and $\hat{\mu}_T(\cdot)$ describing the response profile in the control and treatment groups, respectively. In order to test the null hypothesis of no treatment effect over the entire period, we need to be able to assess whether the two curves are equal, that is whether the assumption that $\mu_C(t)=\mu_T(t)$ holds true for all t in $[t_{min},t_{max}]$. We propose a moderated functional t-type statistic of the form

$$Ft=\frac{l_2}{se+se_m} \tag{4}$$

where the numerator l_2 is the L_2 distance between the two mean curves $\mu_C(\cdot)$ and $\mu_T(\cdot)$, quantifying the difference between them and taking into account the entire time course. The l_2 term is the square root of $\int_{t_{min}}^{t_{max}}[\hat{\mu}_C(t)-\hat{\mu}_T(t)]^2dt$ with the integration carried out in practice numerically by discretizing the two curves on a fine grid of points and using the trapezoidal rule.

The term *se* is the functional standard error. Identical to the standard t-test, its role is to scale the L_2 distance so that the statistic Ft depends only on the relative difference between the two mean curves and is comparable across different spectral bins observed on different scales. *se* is computed as the square root of $(\hat{s}_C^2/n_C)+(\hat{s}_T^2/n_T)$, with $\hat{s}_C^2$ and $\hat{s}_T^2$ being the sample functional variance estimates in the control and treatment groups, respectively. These estimates are calculated from the fitted $\hat{v}_{C_i}(\cdot)$ and $\hat{v}_{T_i}(\cdot)$ individual curves and quantify the between-replicate variation as the mean of the squared size of the replicate-level effects. For instance, for the control group we have

$$\hat{s}_C^2=1/n_C\sum_{i=1}^{n_C}\int_{t_{\min}}^{t_{\max}}[\hat{v}_{C_i}(t)]^2\mathrm{d}t.$$

Estimation of the variance terms $\hat{s}_C^2$ and $\hat{s}_T^2$ is made possible with the SME model via the fitted individual-level curves $\hat{v}_{C_i}(\cdot)$ and $\hat{v}_{T_i}(\cdot)$, unlike simpler models that do not attempt to fully model the biological variation. When the temporal profiles in the treated group are suspected to vary widely while untreated individuals tend to have more homogeneous temporal profiles, heteroscedasticity can be assumed, as above. When this assumption is not deemed necessary, a *pooled* variance estimate could be obtained instead by lumping together the individual curves from both groups.

We add a small positive value, se_m, calculated from *all* spectral bins, to the functional standard error in order to *regularise* or moderate the Ft statistic. This is necessary because when the functional variances are small, as may be the case for spectral bins observed on a small scale, even small estimation errors can lead to inflated values of the test statistic when carrying out the division. This is further compounded by the fact that good estimates of the variances are difficult to obtain given the small number of biological replicates available (Tusher *et al.*, 2001). With the addition of se_m, the relative impact of small estimation errors in *se* to the size of the denominator can be greatly reduced. In practice, we have adapted the *ad hoc* correction of Tusher *et al.* (2001) in order to find an optimal se_m that minimizes the coefficient of variation of the Ft statistic across all spectral bins. This approach is preferred for its simplicity in comparison to other methods which tend to rely on computationally expensive hierarchical Bayesian models (Smyth, 2004) with complex distributional assumptions (Opgen-Rhein and Strimmer, 2007). The process of calculating se_m on a global basis is similar to the idea of *shrinkage* whereby the individual estimates of *se* for each spectral bin are improved by shrinking them towards the mean. In this sense, we say that the moderated Ft statistic *borrows strength* across all spectral bins.

The distribution of the moderated Ft statistic under the null hypothesis of no difference between the two mean curves is unknown. It is necessary, therefore, to resort to a resampling procedure in order to approximate its sampling distribution. Such resampling procedures generate data under the null hypothesis of equal mean curves for the two groups while preserving the observed between-individual variation and level of noise. When the sample size is large, a non-parametric bootstrap approach can be used, which simulates null variables by resampling from the model fits to the real data. To generate a null variable from a given spectral bin, first a mean curve is chosen at random from either of the fitted mean curves, $\mu_C(\cdot)$ or $\mu_T(\cdot)$. This is then shared among both simulated groups in order to maintain the null hypothesis. Between individual-variation is preserved by resampling, with replacement, the individual effects $v_i(\cdot)$ and adding them to the chosen mean curve. Similarly, the level of noise is preserved by resampling the error terms ϵ_{ij} with replacement and adding them to the mean curve and individual effects to produce the final simulated observations. Alternatively, when the sample size is small and there are few individual curves to sample from, a parametric bootstrap approach or permutation procedure may be more appropriate.

After empirically determining raw P-values for each spectral bin, the large number of hypothesis tests being carried out simultaneously requires that we correct them for multiple testing to avoid incorrect inference. This correction is carried out by controlling the false discovery rate (FDR) using the approach of Benjamini and Hochberg (1995).

A flow chart giving an overview of our entire framework is available in Section A.1 in the Supplementary Material, .

3.4 A hierarchical model for time course simulation

In order to assess the ability of our proposed SME model to recover the true unobserved mean and replicate curves, as well as the the power of the moderated functional t-statistic in detecting differences between mean curves, we propose a data simulation procedure. The procedure has been designed to be flexible and produce observations that resemble real experimental data.

We generate artificial longitudinal data compatible with model (1)–(2), parameterizing the fixed- and random-effects as natural cubic splines. For a given simulated dataset, we generate S spectral bins and, for each simulated spectral bin s, the observations for a given individual i belonging to the control group C are generated according to the following model

$$\boldsymbol{y}_i^{(s)}=\boldsymbol{X}_i\boldsymbol{\mu}_C^{(s)}+\boldsymbol{X}_i\boldsymbol{v}_i^{(s)}+\boldsymbol{\epsilon}_i^{(s)}$$

where $\boldsymbol{X}_i$ is the natural cubic spline basis matrix, of dimension $m_i\times q$, and q is the dimension of the basis given by $q=K+2$, with K being the number of knots chosen for the spline. The mean curve spline coefficients $\boldsymbol{\mu}_C$ are

taken to be multivariate normally distributed with zero mean and covariance $\boldsymbol{D}_{\mu_C}$, which is chosen to be a first-order autoregression covariance matrix of dimension $q \times q$ with each $\boldsymbol{D}_{\mu_C}(a,b)$ entry being $\xi \times \rho^{|a-b|}$. The parameter ξ controls the overall size of the spline coefficients while ρ controls their correlation, and hence the roughness of the simulated curves. The individual-level spline coefficients $\boldsymbol{v}_i$ are taken to be multivariate normally distributed with zero mean and covariance $\tau_C^{(s)}\boldsymbol{D}_{\mu_C}$. The scalar value τ_C serves the purpose of scaling the individual-level curves to be of an order smaller than the mean curve, and is drawn from a uniform distribution. The noise term $\boldsymbol{\epsilon}_i$ is also normally distributed with zero mean and covariance matrix $\sigma_C^{(s)^2}\boldsymbol{I}_{m_i}$ where $\sigma_C^{(s)^2}$ is log normally distributed.

Each spectral bin s is then paired with an indicator variable which flags the variable as being a *significant* variable with fixed probability p. If the bin s is not significant, observations in the treatment group are generated under the assumed model with the same model parameters $\boldsymbol{\mu}_C(s), \boldsymbol{D}_{\mu_C}, \tau_C^{(s)}$ and $\sigma_C^{(s)}$ used to generate data in the control group. On the other hand, if s is flagged as significant, the observations are generated by a model with parameters $\boldsymbol{\mu}_T^{(s)} = \boldsymbol{\mu}_C^{(s)} + \boldsymbol{\mu}_\delta^{(s)}$ where $\boldsymbol{\mu}_\delta \sim MVN(0, \boldsymbol{D}_{\mu_\delta})$ and $\boldsymbol{D}_{\mu_\delta}$ differs from $\boldsymbol{D}_{\mu_C}$ as it uses a different ρ value. The vector $\boldsymbol{\mu}_\delta$ is normalized so that $||\boldsymbol{\mu}_\delta||_2$ is equal to some pre-specified value, in order to control the effect size. Parameters $\tau_T^{(s)}$ and $\sigma_T^{(s)}$ are simulated from their distributions and such that $\tau_T^{(s)} \neq \tau_C^{(s)}$ and $\sigma_T^{(s)} \neq \sigma_C^{(s)}$. The number of observations flagged as missing and removed are drawn from a Poisson distribution.

Based on extensive simulations, we have found that this hierarchical model for data generation is flexible enough as it gives control over many aspects of the simulated curves, thus reproducing the characteristics observed in real studies. First, the complexity of the curves can be controlled by adjusting the number of knots, K, of the natural cubic spline. Second, the AR(1) covariance structure for the fixed- and random-effects ensures that successive spline coefficients are correlated and hence enforces temporal dependence. Third, by allowing the scalar τ_C to differ from τ_T for significant spectral bins, the simulated effect of the treatment manifests itself both in terms of the mean and the variance of the two sets of curves. This maps closely with the example given in Figure 1 where the treatment induces a range of responses in different rats. Finally, the size of the difference in mean curves between treatment and control groups for significant variables can also be controlled. Example of simulated observations illustrating the output from our procedure can be found in Supplementary Figure S3A.

4 RESULTS

4.1 Performance assessment using simulated data

Using simulated data, we set out to assess the performance of our proposed SME model in estimating the true mean and biological replicate curves and in identifying significant biomarkers and compared it to the closely related functional mixed-effects 'EDGE' model of Storey *et al.* (2005). Their approach differs from ours in two key respects. First, they treat the individual curves as simple scalar shifts from the mean, in order to minimize the number of model parameters. Second, they use an F-type test statistic to test for differences between the two groups.

In our experiments, the number of biological replicates was varied between 3, 7 or 10 and the number of time points between 4, 7 or 10. For each combination of the number of replicates and number of time points, 50000 spectral bins were simulated with 5% of them set to have a significant difference between the control and treatment groups. After fitting both our proposed SME model and EDGE to each dataset, we discretized the estimated and true mean curves on a fine grid in order to calculate the mean squared error (MSE), which provides a measure of goodness of fit. This process was averaged over the 50000 variables to give a single MSE for each condition, shown in the top half of Supplementary Table S1A. The MSE of the replicate level curves was calculated in exactly the same way, as given in the bottom half of Supplementary Table S1A.

Table 1. Statistical power comparison based on 50000 simulated spectral bins

Time points	EDGE			SME		
	4	7	10	4	7	10
Replicates						
3	0.36	0.43	0.46	0.56	0.59	0.60
7	0.63	0.66	0.73	0.90	0.89	0.90
10	0.69	0.78	0.81	0.95	0.96	0.96

Our SME model uses the moderated Ft-statistic and EDGE uses the F-statistic.

In terms of goodness of fit, our SME model outperforms EDGE in every case except for when there are only four time points. With less few time points, this is a particularly challenging condition for the model selection process and we found that our proposed approach tended to oversmooth in this case. Unsurprisingly, the MSE of the replicate level curves was higher for EDGE in all cases as its scalar shifts failed to adequately model the simulated data. The scalar shift replicate-effects yield individual temporal profiles that all have the same shape, specifically that of the mean curve $\mu(\cdot)$, which does not accurately reflect the range of profiles present in the simulated data.

To determine the power of each approach, we calculated the Ft-statistic and the F-type statistic for the SME model and EDGE, respectively, and then ranked the variables by significance. Stepping through the ranked list, for each variable we determined the number of true/false positives and true/false negatives if that variable was used as a significance cut-off, then calculated the corresponding sensitivity and specificity. The sensitivity at a fixed specificity of 10% of each approach can be found in Table 1. Interestingly, our SME model outperforms EDGE in—including the scenarios with only four time points where it suffered from a higher MSE for the mean curves. This is likely due to the benefits of the Ft-statistic, specifically the incorporation of the —which in contrast to the mean curve did have a lower MSE for the SME model at four time points.

4.2 COMET longitudinal study

The hydrazine toxicity data exemplify a widely encountered problem in metabolomics where we wish to distinguish between two groups, on the basis of the time profiles, and discover the variables responsible for such discrimination. Here, we focus on the task of identifying the spectral bins showing a significant difference over time between the controls and the high-dose group.

Application of the SME model to the hydrazine data yielded a P-value and associated FDR for each of the 8020 spectral intensity bins. Of the 303 bins with FDR lower than 2.5%, all but 7 could be associated with peaks from metabolites previously known to be differentially regulated in hydrazine toxicity (Bollard *et al.*, 2005; Nicholls *et al.*, 2001). These include the endogenous metabolites hippurate, argininosuccinate, creatine, 2-aminoadipate, citrulline, N-acetyl-citrulline, creatinine, 2-oxoglutarate, succinate, citrate, beta-alanine and also metabolites of hydrazine itself.

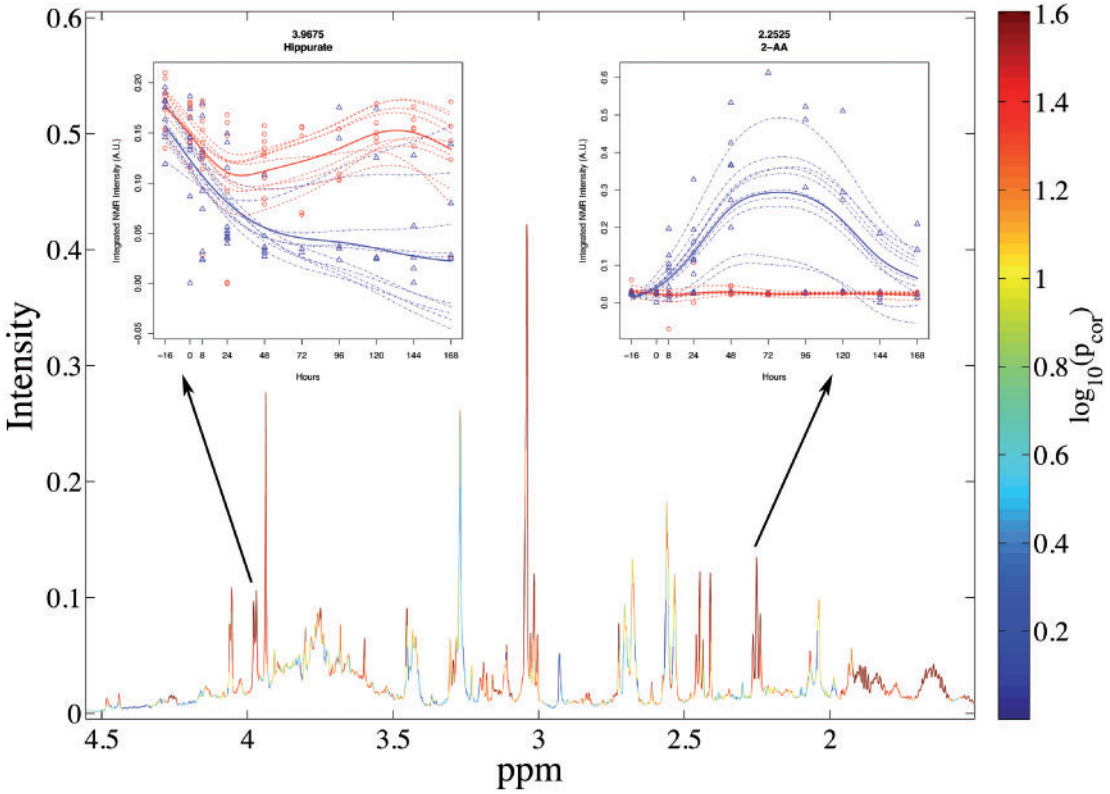

Fig. 2. Biomarker selection using the SME model. The negative $\log_{10}$ FDR corrected *P*-values from the functional *t*-test are mapped on to the mean high-dose NMR spectrum using a colour scale. The insets show the model fits for two significant resonances, hippurate and 2-aminoadiapate (2-AA). High-dose group observations are represented by triangles and controls by circles. The thick solid lines correspond to the fitted mean curves for each group. The dashed lines are individual curves for the control group, and the dot dash lines are individual curves for the high-dose group. Both the illustrated bins have an SME FDR corrected *P*-value of 0.0248.

Figure 2 illustrates the results of applying the SME model to the hydrazine data. NMR signals significant at $FDR < 2.5\%$ correspond to the deep red colour (1.60 on the $-\log$ scale). It can be seen that several regions of the spectrum significantly differentiate the groups according to the functional *Ft*-test. The model fit and raw data for two significant metabolites, hippurate and 2-aminoadipate, are shown in the insets. While changes in urinary hippurate are commonly seen in rats undergoing physiological stress, increased levels of 2-aminoadipate are highly specific to the mode of action of hydrazine in the rat. Thus, the high significance assigned to this bin by the SME model and moderated *Ft*-test statistic builds confidence that the results of this method can have meaningful biological interpretations. For each of the metabolites, the fitted curves are seen to faithfully trace the time course data, following the onset and recovery from the toxic episode. While for 2-AA, only the high-dose animals exhibit a time profile deviating from constant excretion, for hippurate both groups show changes over time, but to differing extents. Both types of group behaviour are successfully captured by the SME model and are reflected in the highly significant *P*-value associated with each variable.

Although most animals in each group follow the mean curve closely, in each case there are a small number which react differently to the stimulus. This is a common effect seen in metabolomics studies where deviating animals can be differentiated into strong/weak and fast/slow responders depending on their individual dynamics (Nicholson *et al.*, 2002). The ability of the proposed approach to model individuals separately is thus a great advantage for these and similar studies.

5 DISCUSSION

In this article, we have presented our complete framework for estimation and testing in metabolomics time course datasets with the goal of identifying significant biomarkers that discriminate between two treatment groups. We have focused on the development of a flexible functional model that can accurately describe longitudinal metabolomic data and a functional *t*-test that fully exploits the model parameters to increase power to detect significant differences between the two groups. We note that while we have demonstrated our approach with NMR, the framework can be applied to any kind of longitudinal data including widely used techniques such as mass spectrometry where its ability to handle inter-subject and inter-metabolite heterogeneity can be considered a useful advantage.

The general functional mixed-effects model (1) has appeared in a number of different specific forms in proposed approaches for the analysis of longitudinal data arising from genomics experiments. These approaches vary in their ability to handle biological replicates, analysis goals—typically either clustering (Luan and Li, 2003; Ma

et al., 2006) or detecting differentially expressed genes (Bar-Joseph *et al.*, 2003; Storey *et al.*, 2005)—model selection procedures and the representations used for the fixed- and random-effects functions. Despite the wide range of resulting models, the message is clear: functional mixed-effects models are well suited to handling time series data that is very short, noisy and replicated.

In this work, we have set out to improve upon the existing methodology and adapt it to better suit our case study in three key ways. First, we have modelled the individual-level effects as full curves, yielding a flexible model that can more accurately describe the often extremely heterogeneous responses we observe among different biological replicates than simpler models such as EDGE. Second, we perform model selection on a per-variable basis, avoiding the possibility of overfitting that the more flexible model presents. This procedure is made possible by the smoothing spline representation that simplifies the optimization for each variable to a 2D search of real-valued smoothing parameters and avoids having to select knots. Third, we have introduced a moderated functional *t*-statistic that makes full use of the fitted model parameters by incorporating the replicate-level effects and borrows strength across all spectral bins to increase power.

The results from our simulation study demonstrate the tangible benefits that the combined use of the SME model and moderated functional *t*-statistic present compared with the most closely related existing approach. Furthermore, the ability of the model and test statistic to give biologically meaningful results in the context of the COMET study suggests these benefits are equally applicable to real datasets.

ACKNOWLEDGEMENTS

The authors wish to thank the members of the Consortium for Metabonomic Toxicology (COMET) for access to data.

Funding: This work was supported by the Wellcome Trust [080715/Z/06/Z to M.B.].

Conflict of Interest: none declared.

REFERENCES

Antti,H. *et al.* (2002) Batch statistical processing of 1H NMR-derived urinary spectral data. *J. Chemometrics*, **16**, 461–468.

Bar-Joseph,Z. *et al.* (2003) Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc. Natl Acad. Sci. USA*, **100**, 10146–10151.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.

Bollard,M.E. *et al.* (2005) Comparative metabonomics of differential hydrazine toxicity in the rat and mouse. *Toxicol. Appl. Pharmacol.*, **204**, 135–151.

Box,G. *et al.* (1994) *Time Series Analysis, Forcasting and Control*. Prentice-Hall, Englewood Cliffs, New Jersey.

Cox,D.D. and Lee,J.S. (2008) Pointwise testing with functional data using the Westfall–Young randomization method. *Biometrika*, **95**, 621–634.

Cuevas,A. *et al.* (2004) An ANOVA test for functional data. *Comput. Stat. Data Anal.*, **47**, 111–122.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* , **39**, 1–38.

de Boor,C. (1978) *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer, New York, 1978.

Ebbels,T. and Cavill,R. (2009) Bioinformatic methods in NMR-based metabolic profiling. *Progr. Nucl. Magn. Reson. Spectrosc.*, **55**, 361–374.

Goodacre,R. *et al.* (2007) Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, **3**, 231–241.

Green,P.J. and Silverman,B.W. (1994) *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London.

Guo,W. (2002) Functional mixed effects models. *Biometrics*, **58**, 121–128.

Hurvich,C.M. *et al.* (1998) Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *J. R. Stat. Soc. Ser. B* , **60**, 271–293.

Lindon,J.C. *et al.* (2005) The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. *Pharmacogenomics*, **6**, 691–699.

Luan,Y. and Li,H. (2003) Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, **19**, 474–482.

Ma,P. *et al.* (2006) A data-driven clustering method for time course gene expression data. *Nucleic Acids Res.*, **34**, 1261–1269.

Nelder,J.A. and Mead,R. (1965) A simplex method for function minimization. *Comput. J.*, **7**, 308–313.

Nicholls,A.W. *et al.* (2001) Metabonomic investigations into hydrazine toxicity in the rat. *Chem. Res. Toxicol.*, **14**, 975–987.

Nicholson,J.K. *et al.* (2002) Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.*, **1**, 153–161.

Nicholson,J. *et al.* (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, **29**, 1181.

Opgen-Rhein,R. and Strimmer,K. (2007) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol.*, **6**, 9.

Raamsdonk,L.M. *et al.* (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.*, **19**, 45–50.

Ramsay,J. and Silverman,B. (2006) *Functional Data Analysis*. Springer, New York.

Robinson,G.K. (1991) That BLUP is a good thing: the estimation of random effects. *Stat. Sci.*, **6**, 15–32.

Smilde,A.K. *et al.* (2010) Dynamic metabolomic data analysis: a tutorial review. *Metabolomics*, **6**, 3–17.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 3.

Steuer,R. *et al.* (2007) A gentle guide to the analysis of metabolomic data. *Methods Mol. Biol.*, **358**, 105–126.

Storey,J.D. *et al.* (2005) Significance analysis of time course microarray experiments. *Proc. Natl Acad. Sci. USA*, **102**, 12837–12842.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wu,H. and Zhang,J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*. Wiley, Hoboken, New Jersey.