Bioinformatics, 31(5), 2015, 728–735
doi: 10.1093/bioinformatics/btu699
Advance Access Publication Date: 28 October 2014
Original Paper

OXFORD

Systems biology

Detecting time periods of differential gene expression using Gaussian processes: an application to endothelial cells exposed to radiotherapy dose fraction

Markus Heinonen[1,2,*], Olivier Guipaud[3], Fabien Milliat[3], Valérie Buard[3], Béatrice Micheau[3], Georges Tarlet[3], Marc Benderitter[3], Farida Zehraoui[1] and Florence d'Alché-Buc[1,2*]

[1]IBISC, Universite d'Évry Val d'Essonne, 23 Boulevard de France, 91025 Évry, France, [2]AMIB, INRIA-Saclay, LRI UMR CNRS 8623, Université Paris Sud, Orsay, France, and [3]Institut de Radioprotection et de Sûreté Nucléaire, LRTE, 92262 Fontenay-aux-roses, France

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on March 20, 2014; revised on September 29, 2014; accepted on October 17, 2014

Abstract

Motivation: Identifying the set of genes differentially expressed along time is an important task in two-sample time course experiments. Furthermore, estimating at which time periods the differential expression is present can provide additional insight into temporal gene functions. The current differential detection methods are designed to detect difference along observation time intervals or on single measurement points, warranting dense measurements along time to characterize the full temporal differential expression patterns.

Results: We propose a novel Bayesian likelihood ratio test to estimate the differential expression time periods. Applying the ratio test to systems of genes provides the temporal response timings and durations of gene expression to a biological condition. We introduce a novel non-stationary Gaussian process as the underlying expression model, with major improvements on model fitness on perturbation and stress experiments. The method is robust to uneven or sparse measurements along time. We assess the performance of the method on realistically simulated dataset and compare against state-of-the-art methods. We additionally apply the method to the analysis of primary human endothelial cells under an ionizing radiation stress to study the transcriptional perturbations over 283 measured genes in an attempt to better understand the role of endothelium in both normal and cancer tissues during radiotherapy. As a result, using the cascade of differential expression periods, domain literature and gene enrichment analysis, we gain insights into the dynamic response of endothelial cells to irradiation.

Availability and implementation: R package 'nsgp' is available at www.ibisc.fr/en/logiciels_arobas

Contact: markus.heinonen@ibisc.fr or florence.dalche@ibisc.fr

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

1 Introduction

With the advent of high-throughput measurement technologies, large-scale systems biology experiments are now routinely performed. The first step towards understanding the system-level responses is determining the genes that are differentially expressed across samples obtained from two (Kerr *et al.*, 2000; Dudoit *et al.*, 2002) or multiple biological conditions (Kendziorski *et al.*, 2003), which is usually performed over static microarray measurements.

However, time-series measurements of the transcriptomic state of the cells are necessary to reveal additional information of the inherently dynamic regulation and function of the cells. Several differential expression methods have been designed for time series data, with majority of them testing whether the two-sample time-series are differential throughout the whole time course (Bar-Josephs *et al.*, 2003; Conesa *et al.*, 2006; Kalaitzis and Lawrence, 2011; Storey *et al.*, 2005; Tai *et al.*, 2006). In this article, we focus on the currently overlooked task of determining when genes are differentially expressed from time series experiments under two-sample setting (e.g. control and case).

Both Bayesian and frequentist statistical tests have been proposed for time series data (Stegle *et al.*, 2010; Storey *et al.*, 2005; Tai *et al.*, 2006). In the Bayesian approach, a Bayes factor between a null hypothesis—assuming no differential expression—and a differential hypothesis is often approximated by computing the likelihood ratios of the observed data against the competing hypotheses (Angelini *et al.*, 2007). In the differential hypothesis, separate time-series models are learned for both biological conditions, while in the null hypothesis a single model explains both samples. A difference is declared if the two time-series can be explained more confidently using separate differential models compared to a single null model. These approaches have been applied to testing whole time series to determine if a gene is differentially expressed or not.

Stegle *et al.* (2010) introduced the first test for estimating differential expression separately for individual observation times to produce time intervals of differential expression. A time period between two neighboring and differential measured time-points is assumed to be differential as well. This allows characterizing the starting and ending times of the differential expression, providing for a temporal characterisation of the underlying biological processes. The time periods are restricted to start and end at an observed time point. However, in the case of sparse or uneven observation times, it is highly desirable to be able to estimate differential expression accurately between measured timepoints, producing a continuous estimate of the differential expression time periods. In this article, we extend the approach of Stegle *et al.* (2010) by introducing a method for unconstrained detection of differentially expressed time periods, which need not to contain measured timepoints. We propose two likelihood ratio tests that measure the expected data likelihood instead of the observed data likelihood. These can be evaluated naturally over probabilistic underlying expression models.

We consider the Gaussian process regression (GPR) models, which have been commonly applied to model time course gene expression (Schliep *et al.*, 2005; Lawrence *et al.*, 2007), and are an apt model for likelihood ratio estimation (Stegle *et al.*, 2010). GPR models are a flexible class of non-parametric Bayesian models, which quantify the uncertainty of the underlying process estimates using Gaussian distributions (Rasmussen and Williams, 2006). GPR models of temporal gene expression have been extended with outlier detection (Cooke *et al.*, 2011), hierarchical replicate models and clustering (Hensman *et al.*, 2013), bootstrapping (Kirk and Stumpf, 2009) and with ordinary differential equation (ODE) model integrations (Äijö and Lähdesmäki, 2009; Gao *et al.*, 2008). GPR models naturally support replicate measurements (Stegle *et al.*, 2010), missing values and sparse observation times.

An important particularity of many gene expression measurements is that they are generally obtained from a perturbation of the basal system under study. As other non-parametric models, GPRs based on time-invariant parameters are not appropriate to model kinetics in response to a perturbation. To improve the accuracy of the modeling, we consider non-stationary Gaussian processes (Paciorek and Schervish, 2004), that is, GPR models that do not assume time-invariance. Hence, ideal to perturbation experiments, they can model the early and late perturbation responses by varying the model smoothness. We introduce a non-stationary Gaussian kernel, where the kernel variance is restricted to a logarithmic function of time, which is directly learned from the data.

We consider a human endothelial cell irradiation response experiment to elucidate gene expression inhibitions and activations in the irradiated cells. Irradiation is an important class of stress to endothelial cells for both normal tissues and tumors during radiotherapy. Perturbation of the endothelial system by ionizing radiation, which remains mostly unknown, has important consequences on the radiosensitivity of both normal and cancer tissues. It is, therefore, of major importance to better understanding the biological consequences of exposure to ionizing radiation, and particularly to decipher the network perturbations that lead to a pathological phenotype of the cells. We assess the method against a realistical simulated gene expression time-series, and apply the novel ratios to the analysis of human umbilical vein endothelial cells (HUVEC) under a realistic radiotherapy dose fraction (2 Gy) against a control cell population. As a direct result of the method over this large-scale gene system, an insightful cascade of differential expression time periods of the observed genes emerges.

Below, we present a non-stationary Gaussian process model in Section 2, and introduce a non-stationary Gaussian kernel and two novel likelihood ratios in its subsections. In Section 3, we evaluate the method against a simulated dataset and apply it to study the large-scale dynamic genetic response to irradiation. We conclude in section 4.

2 Methods

We present a two-phase method for detection of differential time periods of two-sample time-series observations. We fit Gaussian process models on the biological conditions of each gene, and then proceed to compare the likelihood ratios of these GP-models along time domain. The method is demonstrated on gene expression time-series dataset with replicates and uneven observation times, but is readily applicable to any kind of quantitative biological time-series data, e.g. RNA or protein concentrations at even or uneven observation times, assuming Gaussian noise.

2.1 Overview of Gaussian process model

First, we construct smooth probabilistic models of the measured gene expression trajectories over time from point measurements using Gaussian processes. We model each gene expression time-series using an independent model. Let $\mathbf{y} = (y_{t_1}, \ldots, y_{t_N}) \in \mathbb{R}^N$ be the vector of N noisy gene expression measurements $y_t \in \mathbb{R}$ at input time points $T_{\text{obs}} = (t_1, \ldots, t_N) \in \mathbb{R}_+^N$ of a single gene. We assume R replicate measurements and denote the r'th replicate

measurement as $\mathbf{y}^r$. We assume that a true model $f(t)$ explains the observations through

$$y_t = f(t) + \varepsilon_t$$

under some Gaussian isotropic and time-dependent noise model $\varepsilon_t \sim \mathcal{N}(0, \omega_t^2)$. We collect the time-dependent noise variances $\omega_{t_1}^2, \ldots, \omega_{t_N}^2$ into a diagonal covariance matrix Ω.

GPR is a Bayesian non-parametric and non-linear method for regression. A Gaussian process is a generalization of distributions to functions, where any subset of function evaluations is jointly Gaussian (Rasmussen and Williams, 2006). A Gaussian process $\mathbf{f}_\star \sim \mathcal{GP}(\mu_\star, \Sigma_\star)$ represents a distribution over function samples $\mathbf{f}_\star = (f(t_1), \ldots, f(t_{N_\star}))$ at time points $T = (t_1, \ldots, t_{N_\star}) \in \mathbb{R}^{N_\star}$ through the mean vector $\mu_\star \in \mathbb{R}^{N_\star}$ and the covariance matrix $\Sigma \in \mathbb{R}^{N_\star \times N_\star}$.

According to the GPR modeling, we determine the function class by placing a Gaussian prior

$$\mathbf{f} \sim \mathcal{N}(0, K_{TT})$$

over the true model $f(t)$, where K_{TT} is a covariance, or more generally, a positive semi-definite kernel matrix between time points $T_{\text{obs}} \times T_{\text{obs}}$. We are interested in learning the Gaussian process given the data $\mathbf{y}$ and the function prior, which results in a 'posterior' distribution $\mathbf{f}_\star|\mathbf{y} \sim \mathcal{N}(\mu_\star, \Sigma_\star)$ defined by

$$\mu_\star = K_{\star T}(K_{TT} + \Omega)^{-1}\mathbf{y}$$
$$\Sigma_\star = K_{\star\star} - K_{\star T}(K_{TT} + \Omega)^{-1}K_{T\star},$$

where $K_{\star T} = K_{T\star}^T$ is the kernel K over $T_\star \times T_{\text{obs}}$.

The posterior of the true model can be visualized by the mean model $\mu_\star$ along with 95% confidence intervals $\pm 1.96\sqrt{\text{diag}\Sigma_\star}$. However, if we are interested in sampling from the estimated model with observational noise Ω, we use the distribution $\mathbf{y}_\star \equiv \mathbf{y}_\star|\mathbf{f}_\star \sim \mathcal{N}(\mu_\star, \Sigma_\star + \Omega)$ as the complete noisy kinetic model of the gene expression (Kirk and Stumpf, 2009) (Fig. 1).

2.2 Non-stationary Gaussian kernel

The kernel choice $K(t, t')$ plays an important role in determining the function space learned by the Gaussian process. The Gaussian kernel $K(t, t') = \exp(-||t - t'||^2/2\ell^2)$ is often used as a 'default' kernel due to its universality. It naturally gives high covariance for close time points, resulting in smooth regression models. However, the Gaussian kernel is a function of $t - t'$, and hence 'stationary'. In perturbation experiments, the cell's response is time-variant: the perturbation often subjects the cell's state to rapid changes, while the cells are reaching a more stable state. Perturbation experiments warrant 'non-stationary' covariance models. Non-stationary kernels for GPR were introduced by Gibbs (1997), while Paciorek and Schervish (2004) give a generalized construction for non-stationary version of kernels, where input values are associated with individual variances σ_t^2, with the drawback of high computational costs and non-analytical derivatives of model learning.

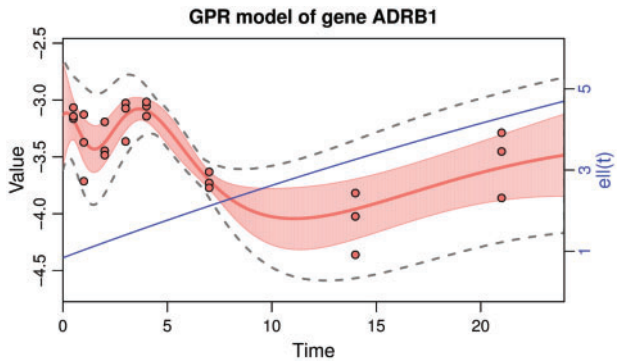

Fig. 1. Gaussian process posterior $\mathbf{f}_\star$ over the gene ADRB1 measurements (dots). The shaded region indicates the 95% confidence interval $\Sigma_\star$ of the underlying function, while the dashed line includes also the observational noise model with $\Sigma_\star + \Omega$. The black line indicates the learned length scale function $\ell(t)$ on the right *y*-axis

We introduce a non-stationary Gaussian kernel. Adding non-stationarity allows varying smoothness along time, while retaining the favorable properties of the Gaussian kernel. A standard Gaussian kernel $K(t, t') = \langle \phi_\ell(t), \phi_\ell(t') \rangle$ admits an feature expansion, which implies a feature map over tuples $\phi(t, \ell)$:

$$\phi_\ell(t) = \exp\left(-\frac{t^2}{\ell^2}\right)\left(\sqrt{\frac{(2/\ell^2)^k}{k!}}t^k\right)_{k=0}^{\infty} \equiv \phi(t, \ell).$$

An inner product between the feature maps $\phi(t, \ell(t))$ and $\phi(t', \ell(t'))$ results in a non-stationary Gaussian kernel

$$K_\ell(t, t') = \sigma_f^2 \exp\left(-\left(\frac{t}{\ell(t)} - \frac{t'}{\ell(t')}\right)^2\right), \quad (1)$$

where the σ_f^2 is the kernel variance, and we, further, restrict the kernel by defining a length scale function following a logarithmic function

$$\ell(t) = \ell_t = \ell - (\ell - \ell_{\min})\mathrm{e}^{-ct},$$

controlled by three hyperparameters: maximum lengthscale ℓ, minimum lengthscale $\ell_{\min}$ (at time $t = 0$), and the curvature c controls how fast the function $\ell(t)$ approaches its maximum value. We assume that the data is normalized such that perturbation occurs at time 0. The kernel of Equation 1 remains analytically differentiable.

We note a simple alternative approach of achieving non-stationarity by applying a log transformation over the time domain. We compare the performance of the log-transformation to the non-stationary kernel in Section 3.2.

2.3 Model inference

The GPR framework provides a natural way to learn the hyperparameters $\theta = (\sigma_f, \ell, \ell_{\min}, c)$ of the kernel K_ℓ. In a Bayesian model inference, we would marginalize over the hyperparameters and the models implied by them. Due to computational tractability, we instead learn hyperparameters against the marginal log likelihood (MLL)

$$\log p(\mathbf{y}|T, \theta) = \log \int p(\mathbf{y}|\mathbf{f}, T)p(\mathbf{f}|\theta)\mathrm{d}\mathbf{f} \quad (2)$$

which follows $\mathbf{y} \sim \mathcal{N}(0, K_{TT} + \Omega)$ giving a log likelihood $-\frac{1}{2}\mathbf{y}^T(K_{TT} + \Omega)^{-1}\mathbf{y} - \frac{1}{2}\log|K_{TT} + \Omega| - \frac{N}{2}\log 2\pi$. We optimize the parameters θ by gradient descent over Equation 2 with L-BFGS. The noise model can be learned against the MLL (Rasmussen and Williams, 2006), which, however, leads to intractable inference if a varying noise model is considered. To avoid intractability, we predefine the time-dependent noise model $\Omega = \text{diag}(\omega_{t_1}^2, \ldots, \omega_{t_{N_\star}}^2)$ for each gene separately by interpolating the empirical replicate variances $\text{Var}((y_t^r)_{r=1}^R)$ using standard spline interpolant.

2.4 Framework for detection of differential time periods

Next, we are concerned with determining at which time periods two gene profiles $\mathbf{y}^A$ and $\mathbf{y}^B$ are exhibiting significant differential expression under different biological conditions (e.g. case and control).

The standard statistical test is to compare the two independent probabilistic models $\mathcal{H}^A$ and $\mathcal{H}^B$ fitted to time series $\mathbf{y}^A$ and $\mathbf{y}^B$ independently, against a joint probabilistic model $\mathcal{H}^S$ fitted to pooled data $\mathbf{y}^S = (\mathbf{y}^A; \mathbf{y}^B)$ (Storey *et al.*, 2005; Stegle *et al.*, 2010). These two cases correspond to a differential expression hypotheses (I) and to a null hypothesis (S), respectively.

In a Bayesian approach the Bayes factor score between the hypotheses is thresholded (Stegle *et al.*, 2010; Yuan, 2006; Angelini *et al.*, 2007). It can be approximated by evaluating the MLL ratio

$$S_{\mathrm{MLL}}(T_{\mathrm{obs}}|\mathbf{y}^A, \mathbf{y}^B) = \log \frac{p(\mathbf{y}^A|T_{\mathrm{obs}}, \hat{\theta}^A)p(\mathbf{y}^B|T_{\mathrm{obs}}, \hat{\theta}^B)}{p(\mathbf{y}^S|T_{\mathrm{obs}}, \hat{\theta}^S)} \quad (3)$$

where the three models are learned independently for the datasets $\mathbf{y}^A$, $\mathbf{y}^B$ and $\mathbf{y}^S$, respectively, resulting in optimal hyper parameters $\hat{\theta}^A$, $\hat{\theta}^B$ and $\hat{\theta}^S$. The evaluation of the MLL ratio of Equation 3 is done against the observed data $\mathbf{y}$ or its subsets over the corresponding time points. In particular, the MLL ratio can provide a single likelihood ratio for differential expression of the whole time course or ratios for each observed individual time point (Fig. 2a).

Given sparse observations, we would like to be able to evaluate the likelihood at arbitrary time points $\{t_\star\}$. We propose two novel likelihood ratio tests that can be evaluated at any time points: (i) the expected MLL ratio and (ii) the noisy posterior concentration (NPC) ratio. By evaluating them along time, we construct a smooth ratio curve indicating the precise time periods of differential expression according to a predefined threshold (Fig. 2).

2.4.1 Expected likelihood ratio test

A simple extension to the likelihood ratio is to evaluate the MLL against a sample $\{\hat{\mathbf{y}}_i\}$ drawn from the noisy posterior distribution $\mathbf{y}_\star$ (Kirk and Stumpf, 2009). The sample values $\hat{\mathbf{y}}_i$ are invariant of the measurement points and can be estimated to arbitrary temporal precision. The sample MLL converges into an 'expected' marginal likelihood $\mathbb{E}_{\hat{\mathbf{y}}\sim\mathbf{y}_\star} p(\hat{\mathbf{y}}|T_\star, \hat{\theta})$, which follows a Gaussian $\mathcal{N}(\mu_\star, \Sigma_\star + K_{\star\star} + 2\Omega)$ (See Supplementary data, Fig. 2b).

We propose the ratio of expected marginal log likelihood (EMLL)

$$S_{\mathrm{EMLL}}(T_\star|\mathbf{y}^A, \mathbf{y}^B) = \log \frac{\sqrt{\mathbb{E}_{\mathbf{y}_\star} p(\hat{\mathbf{y}}|T_\star, \hat{\theta}^A)\mathbb{E}_{\mathbf{y}_\star} p(\hat{\mathbf{y}}|T_\star, \hat{\theta}^B)}}{\mathbb{E}_{\mathbf{y}_\star} p(\hat{\mathbf{y}}|T_\star, \hat{\theta}^S))} \quad (4)$$

analogously to Equation 3. The EMLL approach can be interpreted as replacing the measured data with the expected data $\mu_\star$ generated from our estimated noisy posterior model $\mathbf{y}_\star$. The model uncertainties include the covariance term $K_{\star\star} + \Omega$ from the prior, as well as, the noisy posterior covariance $\Sigma_\star + \Omega$.

2.4.2 NPC test

An alternative for the likelihood ratio of Equation 3 is to quantify the difference of concentrations of the noisy posterior distributions $(\mathbf{y}^A_\star, \mathbf{y}^B_\star)$ and $\mathbf{y}^S_\star$ under the independent (I) and shared (S) hypotheses. A concentration is proportional to the inverse of the variance and measures the certainty of the GPR model of the underlying gene expression. We expect that for differential genes, the two independent models attain smaller variances, than the joint model learned on shared data.

A natural measure of distribution concentration is the expected likelihood of its own distribution (Jebara *et al.*, 2004)

$$\mathbb{E}_{\hat{\mathbf{y}}\sim\mathbf{y}_\star} p(\hat{\mathbf{y}}|T_\star, \mathbf{f}_\star, \hat{\theta}) = \int \mathcal{N}(\hat{\mathbf{y}}|\mu_\star, \Sigma_\star + \Omega)^2 d\hat{\mathbf{y}},$$

which follows a Gaussian $\mathcal{N}(0, \Sigma_\star + \Omega)$ and gives a log likelihood of $-\frac{1}{2}\log|\Sigma_\star + \Omega|$ and a constant term. The log odds between the expected likelihoods of independent and shared hypotheses results in NPC score

$$S_{\mathrm{NPC}}(T_\star|\mathbf{y}^A, \mathbf{y}^B) = \log \frac{\sqrt{\mathbb{E}_{\mathbf{y}_\star} p(\hat{\mathbf{y}}|T_\star, \mathbf{f}_\star, \hat{\theta}^A)\mathbb{E}_{\mathbf{y}_\star} p(\hat{\mathbf{y}}|T_\star, \mathbf{f}_\star, \hat{\theta}^B)}}{\mathbb{E}_{\mathbf{y}_\star} p(\hat{\mathbf{y}}|T_\star, \mathbf{f}_\star, \hat{\theta}^S)}, \quad (5)$$

where each term is $-\frac{1}{2}\log|\Sigma^c_\star + \Omega^c|$, for $c \in \{A, B, S\}$.

As a measure of concentration, the NPC is invariant to the distribution means. The test takes into account both our confidences in the posterior variances σ^2_t of the function $f(t)$ and its estimated noise variances ω^2_t (Fig. 2b). To only compare the non-noisy posteriors, we drop the noise terms from the ratio test of Equation 5. In biological experiments, the measurement noise is a compound variance between biological variance and observation error (Kirk and Stumpf, 2009), and thus an inherent part of the model. We argue that the noisy test is necessary to capture the underlying process accurately.

3 Results

We evaluate the performance of the non-stationary Gaussian models and likelihood ratios against a realistically generated simulated dataset and conduct exploratory experiments on HUVEC under irradiation. We model the gene expression using the GPR models and estimate the time periods of differential gene expression under irradiation.

3.1 Materials and methods

We measured transcriptional profiles of 283 genes with real time qPCR under control and under a single irradiation dose of 2 Gy (case) at 0 h with measurements T_{obs} at 12 h, 1, 2, 3, 4, 7, 14 and 21

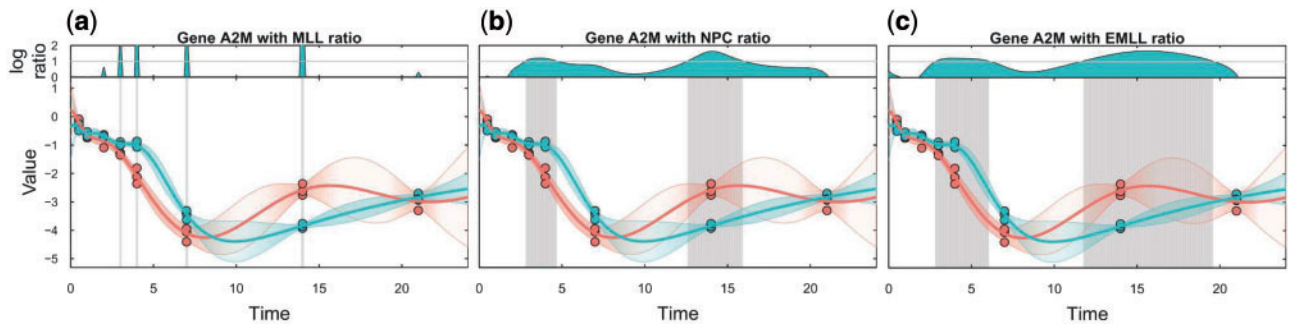

Fig. 2. The three differential expression tests for (a) MLL ratio, (b) noisy posterior concentration ratio and (c) for expected MLL ratio for the gene A2M (alpha-2-macroglobulin) under irradiation (teal) and control (red) with gray intervals indicating difference according to a log likelihood threshold of 1.0

days. Three biological replicate cell populations were separated from a single population just prior to experiments. GPR models are learned for each gene under both condition over prediction time points $T_\star$ that cover smoothly days 0 to 24. Learning a single GPR model takes approximately 1 min on a single core of a 2013 MacBook Pro with our R implementation.

Human umbilical vascular endothelial cells from Lonza (Verviers) were cultured in EGM-2-MV medium at 37°C with 5% CO_2. Confluent cells were irradiated at 2 Gy at passage 3 with a ^{137}Cesium source (IBL 637, CisBio; dose rate 1 Gy/min). For long term experiments (7–21 days post-irradiation), culture medium was changed every week.

Total RNA was prepared with the total RNA isolation kit (Rneasy Mini Kit, Qiagen). Total RNA integrity was analysed using Agilent 2100 and after quantification on a NanoDrop ND-1000 apparatus (NanoDrop Technologies). Reverse transcription was performed using the High Capacity Reverse Transcription Kit (Applied Biosystems) according to the manufacturers instructions. Gene expression assays were performed using a panel of premade TaqMan low density array gene Signature array (angiogenesis, inflammation, apoptosis, immune response and protein kinase) (Applied Biosystems). cDNA (400 ng) per sample was loaded onto the port of each gene signature array cards and PCR was performed with the ABI PRISM 7900 Sequence detection system (Applied Biosystems). Analyses were performed using RQ Manager and Data Assist software and relative mRNA quantification was performed using the comparative $\Delta\Delta CT$. Normalization was performed using a global normalization method (Mestdagh *et al.*, 2009), i.e. the software first finds the common assays among all samples and the median CT of those assays is used as the normalizer, on a per sample basis (Mestdagh *et al.*, 2009). Experiments were performed in triplicates for each time points of the kinetic.

3.2 Model and ratio evaluations

We performed simulation studies with 600 simulated gene expressions under two biological conditions. We reused the learned qPCR control time-course Gaussian processes by sampling new time-courses from them as realistic control time-series. Then, the case perturbation is modeled as a sample from another Gaussian process, which is only non-zero between a randomly chosen differential time period. A sample from the perturbation model was added to the control GP mean to obtain a case GP, from which the case time-series are replicated (See Supplementary data and Supplementary Figs. S2 and S3). We performed differential time period detection with the EMLL-ratio using the novel non-stationary kernel, as well as with a stationary kernel with a log-transform over the time domain to simulate the perturbation dynamics. Finally, we compared these two approaches to the method of Stegle *et al.* (2010) (GPTwoSample), which predicts the differential expression for observed time points only.

The method of Stegle *et al.* (2010) is the state-of-the-art Bayesian method for estimating differential expression for observed time-points. It uses the likelihood ratio test of Equation 3, instead of the expected likelihood ratios of Equations (4) and (5). They utilize a binary latent variable modeling the difference along observed time points with stationary Gaussian processes as the base model, which they learn using MCMC inference. Our method thresholds the likelihood ratios directly, and learns the non-stationary models by gradient descent.

Figure 3 indicates the AUROC curves (See Supplementary Fig. S4 for ROC curves) of the three methods on both continuous time

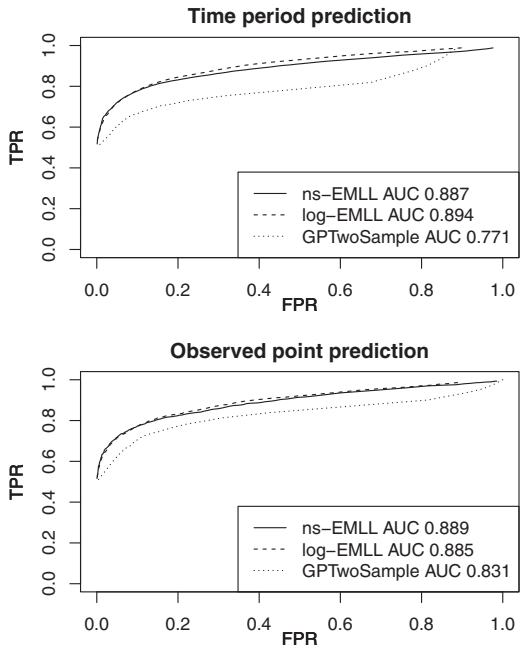

Fig. 3. True positive rate versus false positive rate over the likelihood threshold on simulated dataset. Top: prediction of time-periods with EMLL test using either non-stationary kernels or log-transform compared against the method of Stegle *et al.* (2010). Bottom: prediction of observed points only (13)

period detection (121 dense points) and when estimating over observed points (13 points). We extrapolated the results of the GPTwoSample method by declaring a time period between two observations as differential if the observations themselves were estimated to be differential. The EMLL achieves an AUC score of 0.89 against 0.77 of GPTwoSample on smooth prediction, and 0.89 against 0.83 when predicting observed points only. We note that the method of Stegle *et al.* (2010) was not designed for the former task.

We evaluated the performance of the GPR model with non-stationary Gaussian kernels against a standard Gaussian on the HUVEC dataset. A Supplementary Table S1 shows that utilising a non-stationary Gaussian kernel improves the model MLL (Equation 2) fits by—on average—7.3 on irradiated cells and 2.3 on control cells, on logarithmic scale. Furthermore, Supplementary Figure S1 indicates that non-stationary Gaussian kernel based model never decreases the model fit.

Additionally, we evaluate the EMLL ratio (Equation 4) and the concentration ratio (Equation 5), against the MLL ratio test (Equation 3) over the dataset. All of the tests use a comparable likelihood ratio, whose threshold acts as a precision-recall tradeoff, with a higher threshold giving more confident estimates.

Figure 4 shows a global view of irradiation by counting genes with a log ratio above a threshold 1.0 along time under the tests. The MLL ratio is the greediest and declares between 47 and 154 genes at any of the observed time point to have differential expression, a result which likely contains numerous false positives. The posterior concentration ratio fails due to not taking the noise model into account, while the noisy ratio fares noticeably better. Gene counts drop quickly around the measurement time points, implying weak generalizability. However, learning the GPR model using an expected MLL optimisation criteria (See Supplementary data) produces more informative models between observations, and shows little bias toward observed points (dashed green line). We did not

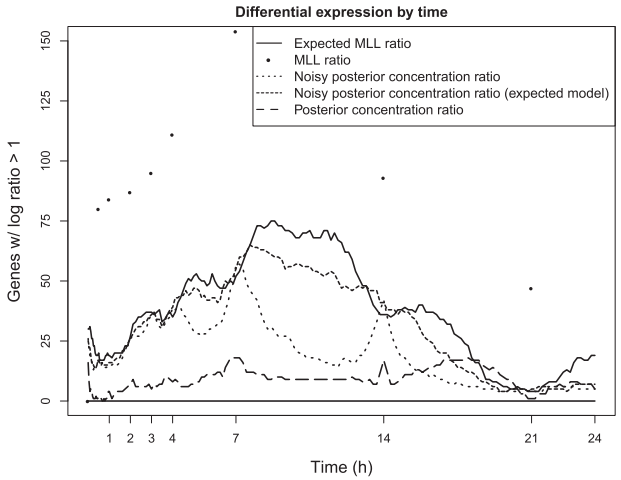

Fig. 4. The distribution of gene counts with a threshold ≥1.0 along time. The MLL ratio evaluations are constrained to observed time points

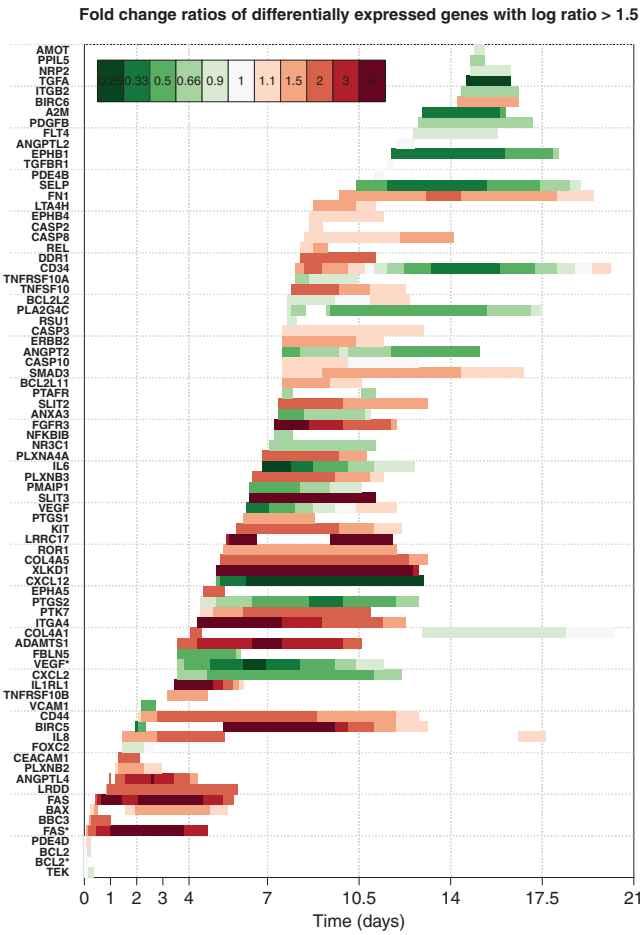

Fig. 5. The cascade of 77 differentially expressed genes over irradiation at threshold 1.5. Asterisks indicate a measurement of a secondary gene probe

record similar improvements on the other ratios, which implies that pairing a more complex optimization criteria with a simple likelihood ratio test provides a good separability. Finally, the expected MLL ratio achieves the highest separability between the case and control, and shows little bias around observed time points. In contrast, the EMLL ratio differentiates more genes at non-observed time points between days 7 and 14, than at days 7 and 14. However, this might also indicate a slight bias toward intervals between observations.

3.3 Differentially expressed genes in endothelial cells

To get a broad picture of the behavior of differentially expressed genes in response to irradiation, we displayed their fold change ratios throughout the 3 weeks of study post-irradiation. Figure 5 highlights the temporal cascade of the 80 gene probes corresponding to 77 genes with significant gene expression difference with threshold 1.5 (See Supplementary Fig. S7 for threshold 1.0 with 174 genes). The maximum of differentially expressed genes occurs between 8 and 12 days post-irradiation. Interestingly, this response is transitory since at 3 weeks post-irradiation, almost no gene displays differential expression anymore. Conversely, the immediate response to irradiation is fast as there is 10 genes with differential expression starting during the first day, and four of them are only active for less than 12 h, immediately after irradiation. These are TEK (TIE2) and PDE4D, both involved in signal transduction through MAP kinase signaling, and BCL2 and FAS, both involved in apoptosis. The TEK receptor tyrosine kinase is expressed almost exclusively in endothelial cells. This angiopoietin 1 and 2 receptor has been involved in increasing survival in presence of angiopoietin-1 after irradiation of endothelial cells (Kwak *et al.*, 2000). PDE4D degrades cAMP, which acts as a signal transduction molecule in multiple cell types, including vascular cells stimulated by the proinflammatory cytokine TNF (Miro *et al.*, 2000). Interestingly, PDE4D has never been described to play a role in the response to irradiation.

Apoptosis has been extensively studied under irradiation. In particular, endothelial apoptosis could be the primary lesion initiating intestinal radiation damage (Paris *et al.*, 2001). Supporting this result, 30 other genes related to regulation of apoptosis were also differential during the 3 weeks. However, the main regulators of apoptosis (ANGPTL4, BAX, BBC3, BCL2, BIRC5, CD44, FAS, LRDD, TNFRSF10B) return to the level of the control cells within 6 days, suggesting that apoptosis occurs primarily within few days after exposure and then returns to normal in most cells. This is remarkably illustrated by the expression profile of BIRC5 (survivin), the inhibitor of apoptosis protein (Tamm *et al.*, 1998), which is first repressed around day 2 and then highly overexpressed from day 5.5 until around 2 weeks after irradiation before returning to control cell levels. Interestingly, many genes involved in positive or negative regulation of apoptosis remain differentially expressed throughout the experiment, suggesting that irradiated cells may still express apoptotic signals, but to a lesser extent than in early times.

3.4 Gene Ontology analysis

To further analyse the set of differential genes, we performed GO enrichment analysis corrected for non-genome-wide analysis of the 77 differential genes with the tool DAVID (Huang *et al.*, 2009), using the observed gene set as a background set (See Supplementary Table S2). We found out GO enrichments related to kidney development, cell adhesion and migration, morphogenesis, and steroid stimulus. These results suggest that endothelial cells subjected to irradiation could initiate a transcriptional program related to the renewal of vasculature (development, morphogenesis and response to a stimulus), and are in accordance with previous studies showing they acquire a pro-inflammatory phenotype associated with an

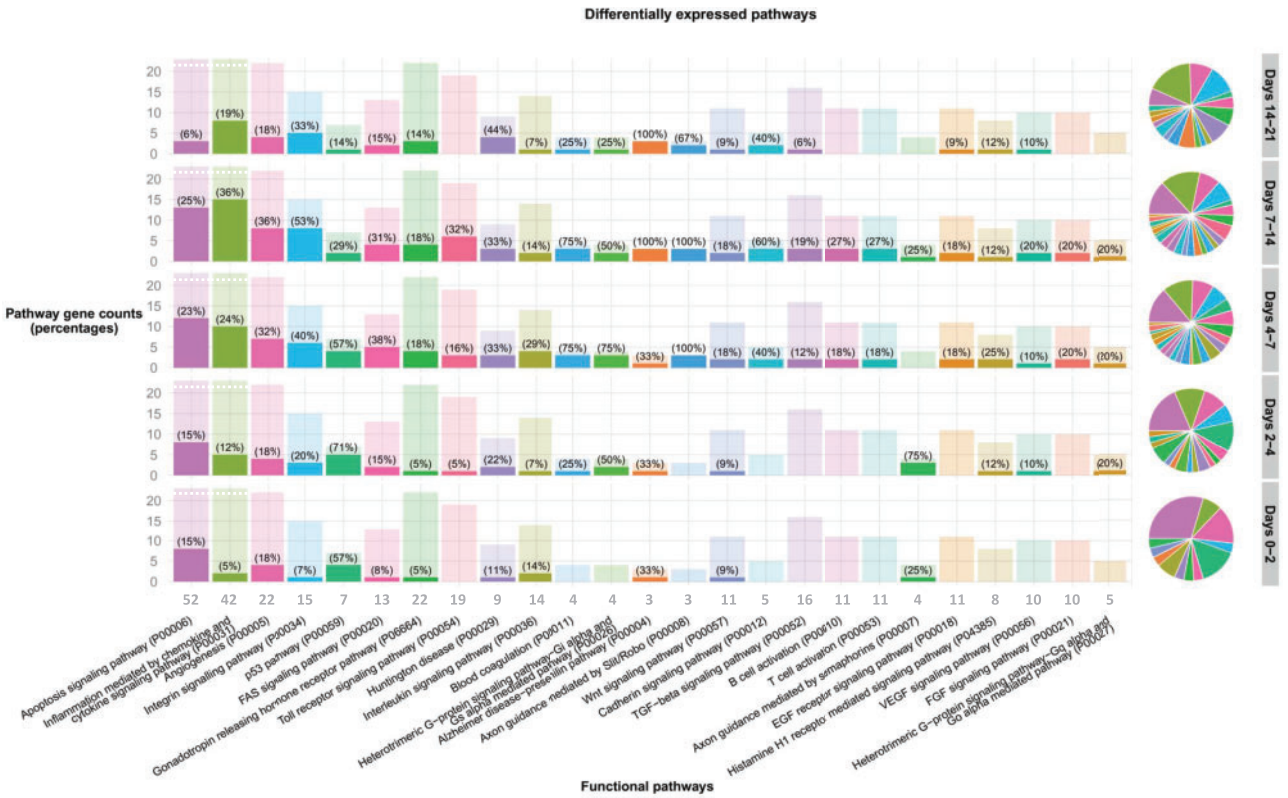

Fig. 6. Division of the 174 differentially expressed genes (threshold 1.0) into PANTHER pathway classification over five time periods. The solid color bars indicate the number of differential genes found per pathway, with background bars and numbers below indicating the total number of genes measured per pathway. The ratio between them is denoted as percentage. The pie charts represent proportions of genes belonging to various pathways during the time period. Parentheses indicate the PANTHER pathway identifiers. See Supplementary Figure S6 for a smooth visualisation of differential pathways along time

increase of cell adhesion and leukocyte migration (Panes *et al.*, 1995; Vereycken-Holler *et al.*, 2002).

3.5 Functional pathway analysis

We studied the pathways of the differential genes also using PANTHER (Mi *et al.*, 2013), which is a classification system combining gene functions, ontologies and pathways (Fig. 6 and Supplementary Fig. S6). This allowed us to highlight pathways of genes related to p53, FAS, integrin, interleukin signaling and others. For example, it has been established that alteration of the plasma membrane can generate apoptosis through the FAS signaling pathway (Corre *et al.*, 2010). Also, it is well-known that ionizing radiation induces DNA damage that triggers the stabilization of p53 and the phosphorylation at different amino acid sites, leading to the transcription of many genes controlled by this factor (Fei and El-Deiry, 2003). We found most differential genes related to p53 pathway were found between 0.5 and 6 days post-irradiation, whereas no or at most 1 gene was differentially expressed between 7 and 21 days post-irradiation (See Supplementary Fig. S6). In the same way, p53 pathway takes a greater proportion of all differential genes in early times (0–2 and 2–4 time intervals) than in late times, as emphasized by Figure 6 (pie charts). We verified the p53 expression patterns with additional bead-based experiments (Supplementary Fig. S5). These protein expression patterns are in accordance with results of the pathway analyses since irradiation of primary endothelial cells induces early, but not late, changes in total and phosphorylated p53.

The Figure 6 shows the amount of genes related to the 25 main pathways (with at least 2 differential genes in one of the intervals) of a total 63 pathways identified with PANTHER for the 174 differential genes (threshold 1.0). Although apoptosis pathway genes were over-represented in our experimental design (background bars), it is noticeable that apoptosis related genes take an even greater part considering all differential genes within the early time intervals 0–2 and 2–4 days (pie charts). Similarly, inflammation, angiogenesis, integrin, p53 and FAS pathways are all over-represented comparatively to the measured genes composition.

3.6 Biological perspectives

Finally, we are able to build a dynamic picture of gene expression changes after irradiation. In the early times, apoptosis, interleukin and p53 signaling pathways are over-represented. Then, they decrease progressively post-irradiation. In contrast, integrin and inflammation pathways become increasingly more differentiated over time. This may reflect that cells exposed to a relatively small dose of ionizing radiation express genes related to death by apoptosis first, and then for those who survived, modify the expression of genes related to long-lasting activation of pathways. Interestingly, inflammation mediated by chemokine and cytokine signaling pathways are early and continuously activated after irradiation at this dose, suggesting that endothelial cells may present an inflamed phenotype all along a radiotherapy course, with possible consequences on the vasculature of both normal tissues and tumors.

These results are highly consistent with domain literature. Moreover, the temporal cascade allows us to propose for the first time a temporal view on the response to irradiation primary endothelial cells exposed at a radiotherapy dose fraction of 2 Gy. We were able to determine the apoptotic signal triggers timings and shed light on the continuous activation of inflammation pathways,

suggesting that endothelial cells may present an inflamed phenotype all along a radiotherapy course, with possible consequences on the vasculature of irradiated tissues, in accordance with clinical observation in patients treated by radiotherapy for head and neck cancers where a sustained inflammation due to NF-κB activation occurred in human arteries and veins (Halle *et al.*, 2010). These results have implications for a better understanding of the molecular networks involved in the dynamic response of endothelial cells to irradiation. High dose fractionated radiotherapy is commonly used for the treatment of solid tumors but the optimization of the response of cancer and normal tissues to radiation remains an important challenge (Moding *et al.*, 2013).

4 Conclusion

In this article, we have proposed a novel Bayesian likelihood ratio test for detecting time-periods of differential gene expression in time course data. We record major improvements on the perturbation model learning using non-stationary GPR models as the underlying model class. For systems of genes, the method estimates a temporal cascade of differentially expressed genes providing a large-scale view on the genetic progression of the irradiation response. The next step entails combining the GPR modeling with the inference of molecular networks. Recent works have explored how to exploit GPR to facilitate parameter estimation in ODEs (Calderhead *et al.*, 2009; Dondelinger *et al.*, 2013), which is undoubtedly a promising research direction.

Funding

This work was supported by Electricité de France (Groupe Gestion Projet-Radioprotection) and Institut de Radioprotec-tion et de Sûreté nucléaire (programme ROSIRIS).

Conflict of Interest: none declared.

References

Äijö,T. and Lähdesmäki,H. (2009) Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. *Bioinformatics,* **25**, 2937–2944.

Angelini,C. *et al.* (2007) A bayesian approach to estimation and testing in time-course microarray experiments. *Statist. Appl. Genet. Mol. Biol.*, **6**, 1–13.

Bar-Josephs,Z. *et al.* (2003) Continuous representations of time-series gene expression data. *J. Comp. Biol.,* **10**, 341–356.

Calderhead,B., Girolami,M. and Lawrence,N. (2009) Accelerating bayesian inference over nonlinear differential equations with gaussian processes. *NIPS,* **21**, 217–224.

Conesa,A. *et al.* (2006) masigpro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics,* **22**, 1096–1102.

Cooke,E.J. *et al.* (2011) Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. *BMC Bioinformatics,* **12**, 399.

Corre,I. *et al.* (2010) Plasma membrane signaling induced by ionizing radiation. *Mutat. Res.,* **704**, 61–67.

Dondelinger,F. *et al.* (2013) Ode parameter inference using adaptive gradient matching with gaussian processes. *JMLR,* **31**, 216–228.

Dudoit,S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica,* **12**, 111–140.

Fei,P. and El-Deiry,W. (2003) P53 and radiation responses. *Oncogene,* **22**, 5774–5783.

Gao,P. *et al.* (2008) Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics,* **24**, i70–i75.

Gibbs,M.N. (1997) Bayesian Gaussian Processes for Regression and Classification. Ph.D. thesis, Department of Physics, University of Cambridge.

Halle,M. *et al.* (2010) Endothelial activation with prothrombotic response in irradiated microvascular recipient veins. *J. Plastic Reconstr. Aesthetic Surg.,* **63**, 1910–1916.

Hensman,J. *et al.* (2013) Hierchical bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics,* **14**, 252.

Huang,D. *et al.* (2009) Systematic and integrative analysis of large gene lists using David bioinformatics resources. *Nat. Protocols,* **4**, 44–57.

Jebara,T. *et al.* (2004) Probability product kernels. *J. Mach. Learn. Res.,* **5**, 819–844.

Kalaitzis,A. and Lawrence,N. (2011) A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC Bioinformatics,* **12**, 180.

Kendziorski,C. *et al.* (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.,* **22**, 3899–3914.

Kerr,M.K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.,* **7**, 819–837.

Kirk,P.D.W. and Stumpf,M.P.H. (2009) Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. *Bioinformatics,* **25**, 1300–1306.

Kwak,H. *et al.* (2000) Angiopoietin-1 inhibits irradiation- and mannitol-induced apoptosis in endothelial cells. *Circulation,* **101**, 2317–2324.

Lawrence,N. *et al.* (2007) Modelling transcriptional regulation using gaussian processes. In: Schölkopf,B., Platt,J.C. and Hofmann,T. (eds). *Advances in Neural Information Processing Systems,* **vol. 19**. MIT Press Cambridge, MA, pp.785–792.

Mestdagh,P. *et al.* (2009) A novel and universal method for microrna rt-qpcr data normalization. *Genome Biol.,* **10**, R64.

Mi,H. *et al.* (2013) Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.,* **41**, D377–386.

Miro,X. *et al.* (2000) Phosphodiesterases 4d and 7a splice variants in the response of huvec cells to tnf-alpha(1). *Biochem. Biophys. Res. Commun.,* **274**, 415–421.

Moding,E. *et al.* (2013) Strategies for optimizing the response of cancer and normal tissues to radiation. *Nat. Rev.,* **12**, 526–542.

Paciorek,C. and Schervish,M.J. (2004) Nonstationary covariance functions for gaussian process regression. In: Saul,L.K., Weiss,Y. and Bottou,L. (eds), *Advances in Neural Information Processing Systems,* **vol. 16**. MIT Press, Cambridge MA, USA.

Panes,J. *et al.* (1995) Role of leukocyteendothelial cell adhesion in radiation-induced microvascular dysfunction in rats. *Gastroenterology,* **108**, 17z61–1769.

Paris,F. *et al.* (2001) Endothelial apoptosis as the primary lesion initiating intestinal radiation damage in mice. *Science,* **293**, 293–297.

Rasmussen,C. and Williams,K. (2006) *Gaussian processes for machine learning*. MIT Press, Cambridge MA, USA.

Schliep,A. *et al.* (2005) Analyzing gene expression time-courses. *IEEE/ACM Trans. Comput. Biol. Bioinf.,* **2**, 179–193.

Stegle,O. *et al.* (2010) A robust bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *J. Comput. Biol.,* **17**, 355–367.

Storey,J.D. *et al.* (2005) Significance analysis of time course microarray experiments. *PNAS,* **102**, 12837–12842.

Tai,Y.C. *et al.* (2006) A multivariate empirical bayes statistic for replicated microarray time course data. *Ann. Stat.,* **34**, 2387–2412.

Tamm,I. *et al.* (1998) Iap-family protein survivin inhibits caspase activity and apoptosis induced by fas (cd95), bax, caspases, and anticancer drugs. *Cancer Res.,* **58**, 5315–5320.

Vereycken-Holler,V. *et al.* (2002) Radiation effects on circulating and endothelial cell interactions studied by quantitative real-time videomicroscopy. *Int. J. Radiat. Biol.,* **78**, 923–930.

Yuan,M. (2006) Flexible temporal expression profile modelling using the gaussian process. *Comput. Stat. Data Anal.*, **51**, 1754–1764.