A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series

Oliver Stegle[1], Katherine Denby[2], David L. Wild[2],
Zoubin Ghahramani[1], Karsten M. Borgwardt[1,3]

[1] University of Cambridge, UK
[2] University of Warwick, UK
[3] Max-Planck-Institutes for Developmental Biology and Biological Cybernetics, Tübingen, Germany

os252@cam.ac.uk,karsten.borgwardt@tuebingen.mpg.de

Abstract. Understanding the regulatory mechanisms that are responsible for an organism's response to environmental changes is an important question in molecular biology. A first and important step towards this goal is to detect genes whose expression levels are affected by altered external conditions. A range of methods to test for differential gene expression, both in static as well as in time-course experiments, have been proposed. While these tests answer the question *whether* a gene is differentially expressed, they do not explicitly address the question *when* a gene is differentially expressed, although this information may provide insights into the course and causal structure of regulatory programs. In this article, we propose a two-sample test for identifying *intervals* of differential gene expression in microarray time series. Our approach is based on Gaussian process regression, can deal with arbitrary numbers of replicates and is robust with respect to outliers. We apply our algorithm to study the response of *Arabidopsis thaliana* genes to an infection by a fungal pathogen using a microarray time series dataset covering 30,336 gene probes at 24 time points. In classification experiments our test compares favorably with existing methods and provides additional insights into time-dependent differential expression.

1 Introduction

Understanding regulatory mechanisms, in particular related to the response to changing external conditions, is of great interest in molecular biology. Changes in external conditions include environmental influences or treatments that an organism is exposed to, ranging from parasitic infections studied in plant biology to drug responses that are of interest in pharmacogenomics. A first step towards understanding these mechanisms is to identify genes that are involved in a particular response. This task can be reduced to a decision problem where we want to tell whether a gene is differentially expressed or not. In the past, most available datasets were static, such that this decision was based on measurements of

gene expression at a single time point (e.g. (1)). Increasingly, studies are being carried out that measure the expression profiles of large sets of genes over a time course rather than a single static snapshot.

The basic task however remains the same: given an observed time series in two conditions (treatment and control), the goal is to determine whether the observations originate from the same biological process or whether they are better described by means of independent processes specific to each condition. This is referred to as a two-sample problem in statistics. In the bioinformatics and statistics community, a wide range of methods have been proposed to test for differential gene expression, both from static microarray experiments (1; 2; 3; 4; 5) as well as from time series microarray data (6; 7; 8; 9). Among desirable properties of a useful test for time series data are the ability to handle multiple replicates of the same condition and robustness with respect to outliers in measured expression levels. While most tests can be applied to multiple replicates, only few of them are robust to outliers due to non-Gaussian errors (9). Furthermore, existing methods often make strong assumptions on the time series, for instance, gene expression levels being described by a linear model or over a finite basis (7; 8).

In this article, we propose a test for differential gene expression based on Gaussian processes (GP), a nonparametric prior over functions. The GP machinery allows attractive properties of existing two-sample tests to be combined: the capability to handle arbitrary numbers of replicates, robustness to outliers and a flexible model basis. In addition, our method is able to identify patterns of local differential expression, where gene expression levels are only differential in subintervals of the full time series. This feature can be used to understand *when* differential expression occurs. Such information is important in molecular biology, because it provides insights on the temporal order in which genes are activated or inhibited by environmental stimuli. For example, it allows to study whether there is a delay in response, whether the effect of the treatment is only temporary, or to identify a cascade of genes that trigger each other's activation during the response. The detection of *intervals* of differential expression can be considered the second central step towards uncovering gene regulatory mechanisms, which follows the first step of detecting differentially expressed genes. This main contribution is illustrated in Figure 1 (Top), where in addition to a score of differential expression, our test also allows to pinpoint the intervals in which a gene exhibits differential expression, as indicated by the Hinton diagrams in the top panel.

The remainder of this article is organized as follows. In Section 2 we describe our Gaussian process based two-sample test for microarray time series data. In Section 2.2 we show how a heavy-tailed noise model can be incorporated to gain additional robustness with respect to outliers. Section 3 concludes the methodological development by introducing a mixture model that can detect differential expression over parts of the time course. In our experimental evaluation, we compare our model to two state-of-the-art two-sample tests from the literature. On time series data for $30,336$ probes from *Arabidopsis thaliana*,

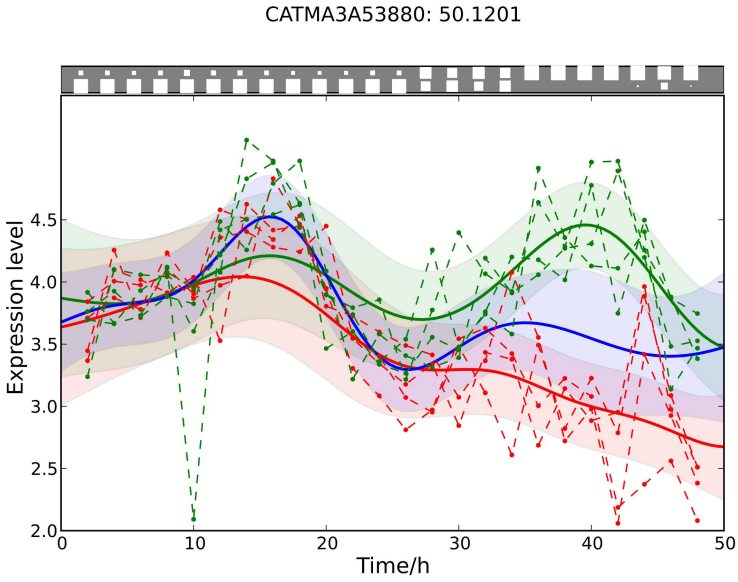


Fig. 1: An example result produced by the GPTwoSample temporal test. **Bottom:** Dashed lines represent replicates of gene expression measurements for control (green) and treatment (red). Thick solid lines represent Gaussian process mean predictions of the latent process traces; ± 2 standard deviation error bars are indicated by shaded areas. **Top:** Hinton diagrams illustrate the probability of differential expression for different time points. Size of upper bars indicates the probability of the genes being differentially expressed, size of lower bars that of being non-differentially expressed.

we assess the predictive performance (Section 2.4) and demonstrate that the detection of differential expression in intervals is useful to gain insights in the response of *Arabidopsis* to a fungal pathogen infection (Section 3.1).

2 GPTwoSample - a robust two-sample test for time series using Gaussian processes

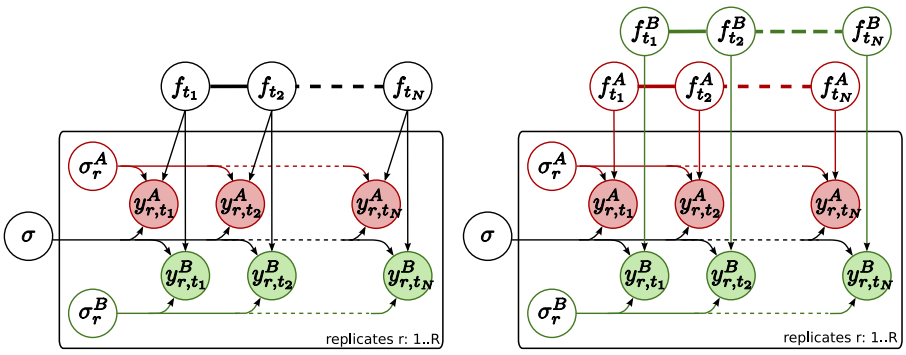
In line with previous approaches to test for differential expression, our test compares two alternative hypotheses: **Either** the time series measured in two conditions, A and B, can be described by a *shared* underlying process ($f(t)$), **or** they are better described by means of two *independent* processes, one for each condition ($f^A(t), f^B(t)$). Figure 2 shows a Bayesian network representation of both hypotheses. We assume that in both conditions, expression levels of R biological replicates are measured at discrete time points $t_1, \ldots, t_N$. For notational convenience, we assume that measurements from both conditions and for all replicates

are synchronized, i.e. share a common time discretization. However, this is not a requirement of the Gaussian process framework which can deal with arbitrary time representations and missing values.

The *Bayes factor* has been previously applied to test for differential expression (9; 10). Following this idea we score the two alternative hypotheses using the logarithm of the ratio of the corresponding model evidences (i.e. the logarithm of the *Bayes factor*)

$$\text{Score} = \log \frac{P(\mathcal{D}_A|\mathcal{H}_{\text{GP}}, \hat{\boldsymbol{\theta}}_{\text{I}})P(\mathcal{D}_B|\mathcal{H}_{\text{GP}}, \hat{\boldsymbol{\theta}}_{\text{I}})P(\hat{\boldsymbol{\theta}}_{\text{I}})}{P(\mathcal{D}_A, \mathcal{D}_B|\mathcal{H}_{\text{GP}}, \hat{\boldsymbol{\theta}}_{\text{S}})P(\hat{\boldsymbol{\theta}}_{\text{S}})}, \tag{1}$$

where $\mathcal{D}_{A,B}$ represent observed expression levels in both conditions A and B. The notation $\mathcal{H}_{\text{GP}}$ indicates that both models are Gaussian process models, where $P(\hat{\boldsymbol{\theta}}_{\text{I}})$ and $P(\hat{\boldsymbol{\theta}}_{\text{S}})$ are prior distributions over the model hyperparameters. The *Bayes factor* is computed conditioned on hyperparameters $\hat{\boldsymbol{\theta}}_{\text{I}}$ and $\hat{\boldsymbol{\theta}}_{\text{S}}$ of the *independent* (I) and *shared* (S) model respectively. Hyperparameters are set to their most probable value as described in the following.



(a) *Shared* model (b) *Independent* model

Fig. 2: Bayesian network for the two alternative models compared in the GPTwoSample test: **a)** *Shared* model where both conditions are explained by means of a single process $f(t)$, **b)** *Independent* model with processes $f^A(t)$ and $f^B(t)$ for each condition. Expression levels $y_{r,t}^{A,B}$ of a given gene are observed in two biological conditions A, B with $r : 1, \ldots, R$ biological replicates and at discrete time points $t : t_1, \ldots, t_N$. Observation noise is split into a global noise level σ and a per-replicate noise level $\sigma_r^{A,B}$. The smoothness induced by the Gaussian process priors is indicated by the thick band coupling the latent function values at different time points.

The accuracy of this score crucially depends on the model used to represent biological processes. A good model should provide sufficient flexibility to allow for noise and variation between biological replicates, but at the same time be able

to detect true differential expression. In addition, there are microarray specific requirements. Firstly, observations are likely to be sparse, i.e. only very few observations are made and potentially in irregular intervals. Secondly, expression data is highly susceptible to noise and prone to outliers which can obscure the test result, if not modeled accurately.

All these requirements can be accommodated by a Gaussian process (GP). This nonparametric prior over functions yields the required flexibility while still allowing specific beliefs, for instance about smoothness and length scales of the process, to be incorporated. Previously, Gaussian processes have been used to model gene expression time dynamics in the context of transcriptional regulation (11) and for biomarker discovery (12). In the context of hypothesis testing, Gaussian processes have been applied to gene expression profiles by Yuan (10).

2.1 Gaussian process model

Let us first consider the *shared* model (Figure 2a), where observations from both conditions are described by a single biological process $f(t)$. We split up the observation noise into a global noise component and a per-replicate noise component by introducing latent replicate observations $g^c_{r,t}$. The joint posterior distribution over unobserved function values $\mathbf{f}$ and the replicate observations $g^c_{r,t}$ for conditions $c \in \{A, B\}$ follows as

$$P(\mathbf{f}, \{g^c_{r,t}\} | \mathcal{D}_A, \mathcal{D}_B, \boldsymbol{\theta}_{\mathrm{S}}) \propto \mathcal{N}(\mathbf{f} \,|\, 0, K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_{\mathrm{K}})) \times \prod_{c=A,B} \prod_{r=1}^{R} \prod_{n=1}^{N} \mathcal{N}(g^c_{r,t_n} \,|\, f_{t_n}, \sigma^c_r) P_{\mathrm{L}}(y^c_{r,t_n} \,|\, g^c_{r,t_n}, \boldsymbol{\theta}_{\mathrm{L}}), \tag{2}$$

where $\boldsymbol{\theta}_{\mathrm{S}} = \{\boldsymbol{\theta}_{\mathrm{K}}, \boldsymbol{\theta}_{\mathrm{L}}, \{\sigma^c_r\}\}$ denotes the set of all hyperparameters for kernel, likelihood and the replicate noise levels respectively. The covariance matrix $K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_{\mathrm{K}})$ is derived from the covariance function $k(t, t' \,|\, \boldsymbol{\theta}_{\mathrm{K}})$ which specifies how function values at two time points t and t' covary. We use a covariance function that decays exponentially with squared time distance, $k_{SE}(t, t') = A \exp\{-\frac{1}{2}\frac{(t-t')^2}{L^2}\}$, which yields smooth functions with a typical squared amplitude A and a typical length-scale L. These kernel hyperparameters are summarized as $\boldsymbol{\theta}_{\mathrm{K}}$.

For simplicity, let us first assume Gaussian observation noise with variance σ, $P_{\mathrm{L}}(y^c_{r,t} \,|\, g^c_{r,t}, \boldsymbol{\theta}_{\mathrm{L}}) = \mathcal{N}(y^c_{r,t} \,|\, g^c_{r,t}, \sigma)$. Integrating out the latent replicate process observations $g^c_{r,t}$ results in a standard Gaussian process with an effective noise variance per replicate and condition

$$P(\mathbf{f} \,|\, \mathcal{D}_A, \mathcal{D}_B, \boldsymbol{\theta}_{\mathrm{S}}) \propto \mathcal{N}(\mathbf{f} \,|\, 0, K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_{\mathrm{K}})) \prod_{c=A,B} \prod_{r=1}^{R} \prod_{n=1}^{N} \mathcal{N}(y^c_{r,t_n} \,|\, f_{t_n}, \sigma^c_r), \tag{3}$$

where $\sigma^c_r = \sqrt{\sigma^{c2}_r + \sigma^2}$. Predictions from this model can be obtained by considering the joint distribution over training data and an unseen test input $t_\star$.

Completing the square leads to a Gaussian predictive distribution (see (13)) of the corresponding function value $f_\star \sim \mathcal{N}(\mu_\star, v_\star)$

$$\begin{aligned} \mu_\star &= K_{\star,\mathbf{T}} \left[K_{\mathbf{TT}} + \Sigma\right]^{-1} \mathbf{y} \\ v_\star &= K_{\star,\star} - K_{\star,\mathbf{T}} \left[K_{\mathbf{TT}} + \Sigma\right]^{-1} K_{\mathbf{T},\star}, \end{aligned} \tag{4}$$

where Σ is a diagonal matrix constructed from the noise levels $\{\sigma_r^c\}$ of the observed expression levels. Note that the dependence of the covariance matrices on hyperparameters $\boldsymbol{\theta}_\mathrm{K}$ is omitted for clarity. The *Bayes factor* in Eqn. 1 requires the evaluation of the log marginal likelihood. Again, this quantity can be calculated in closed form

$$\log P(\mathcal{D}_A, \mathcal{D}_B | \mathcal{H}_\mathrm{GP}, \boldsymbol{\theta}_\mathrm{S}) = -\frac{1}{2} \log \det K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_\mathrm{K}) - \frac{1}{2} \mathbf{y}^\mathsf{T} K_{\mathbf{T},\mathbf{T}}^{-1} \mathbf{y} - \frac{N}{2} \log 2\pi. \tag{5}$$

The most probable parameter settings $\hat{\boldsymbol{\theta}}_\mathrm{S}$ are determined by finding the maximum of the posterior probbility

$$\hat{\boldsymbol{\theta}}_\mathrm{S} = \arg\max_{\boldsymbol{\theta}_\mathrm{S}} \left[\log P(\mathcal{D}_A, \mathcal{D}_B | \mathcal{H}_\mathrm{GP}, \boldsymbol{\theta}_\mathrm{S}) + \log P(\boldsymbol{\theta}_\mathrm{S})\right], \tag{6}$$

where $P(\boldsymbol{\theta}_\mathrm{S})$ are priors on the hyperparameters. Prior distributions are set to incorporate *a priori* beliefs about parameter values. The prior on the amplitude A is uninformative and set to a broad gamma distribution $A \sim \Gamma(0.001, 1000)$. To ensure that noise is not explained by extremely short length scales, we set the prior on L such that the expectation value of the gamma prior corresponds to one fifth of the total length of the time series with a standard deviation of 50%. The noise hyperparameters are set to $\sigma \sim \Gamma(0.1, 10)$ and $\sigma_r^c \sim \Gamma(0.01, 10)$, which favors that noise variance is explained by the shared noise variance where possible.

The optimized marginal likelihood of the alternative hypothesis, assuming *independent* biological processes, $\log P(\mathcal{D}_A | \mathcal{H}_\mathrm{GP}, \hat{\boldsymbol{\theta}}_\mathrm{I}) + \log P(\mathcal{D}_B | \mathcal{H}_\mathrm{GP}, \hat{\boldsymbol{\theta}}_\mathrm{I}) + \log P(\hat{\boldsymbol{\theta}}_\mathrm{I})$, can be obtained analogously. Hyperparameters of the *independent* model are optimized jointly for both processes $f^A(t)$ and $f^B(t)$ where kernel parameters $\boldsymbol{\theta}_\mathrm{K}$ and the global noise variance σ are shared and hence the number of explicit hyperparameters is identical for both models.

2.2 Robustness with respect to outliers

The presentation of the Gaussian process model so far makes a crucial simplification, namely that observation noise is Gaussian. However, for our full model we use a heavy-tailed noise model to acknowledge that a small fraction of the data points can be extremely noisy (outliers) while others are measured with considerably more precision. To reflect this belief we use a mixture model (14)

$$P_\mathrm{L}(y_{r,t}^c \,|\, g_{r,t}^c, \boldsymbol{\theta}_\mathrm{L}) = \pi_0 \, \mathcal{N}(y_{r,t}^c | g_{r,t}^c, \sigma) + (1 - \pi_0) \, \mathcal{N}(y_{r,t}^c | g_{r,t}^c, \sigma_\mathrm{inf}), \tag{7}$$

where π_0 represents the probability of the datum being a regular observation and $(1 - \pi_0)$ of being an outlier. The variance of the outlier component σ_{inf} is much larger than for regular observations and hence allows outliers to be discarded. Unfortunately when using this likelihood model the posterior in Eqn. 2 is no longer computable in closed form. To overcome this problem we use Expectation Propagation (EP) (15), a deterministic approximate inference algorithm. EP approximates the true posterior by a Gaussian process and is efficient enough to allow the algorithm to be applied on large scale datasets. EP for non-Gaussian likelihoods in Gaussian process models is discussed in (13); robust Gaussian process regression has been previously applied to biological data in (16). The derivation of EP for the robust likelihood and further references can be found in Appendix A.

2.3 Runtime

The computational complexity of a Gaussian process models scales with $(RN)^3$, where N is the number of observations per condition and R the number of replicates. Since microarray time series datasets are typically small in the sense that they cover few time points per gene this is not prohibitive. The robust Gaussian process method requires multiple cycles of EP updates which result in constant factor of additional computation. For the datasets studied below, including 24 time points with 4 replicates, the robust test takes approximately 10 seconds per gene on a standard desktop machine.

2.4 Differential gene expression in *Arabidopsis thaliana* after fungal infection

We applied GPTwoSample to study plant response to biotic stress on a dataset of microarray time series. Plant stress responses involve a significant degree of transcriptional change, with different stress stimuli activating common signalling components (17).

In this particular experiment, the stress response of interest is an infection of *Arabidopsis thaliana* by the fungal pathogen *Botrytis cinerea*. The ultimate goal is to elucidate the gene regulatory networks controlling plant defense against this pathogen. Finding differentially expressed genes and intervals of differential gene expression are important steps towards this goal.

Data were obtained from an experiment in which detached *Arabidopsis* leaves were inoculated with a *B. cinerea* spore suspension (or mock-inoculated) and harvested every 2h up to 48h post-inoculation (i.e. a total of 24 time points). *B. cinerea* spores (suspended in half-strength grape juice) germinate, penetrate the leaf and cause expanding necrotic lesions. Mock-inoculated leaves were treated with droplets of half-strength grape juice. At each time point and for both treatments one leaf was harvested from four plants in identical conditions (i.e. 4 biological replicates). Full genome expression profiles were generated from these whole leaves using CATMA arrays (18). Data preprocessing and normalization was carried out using a pipeline based on the MAANOVA package (19). The

experimental design is longitudinal in that subsequent time points should show related expression patterns, but also cross-sectional in that the biological replicates are all from independent plants. Due to this specific study design we expect particularly noisy observations and outliers within the time course of a single replicate plant. For each probe in the dataset, we applied our Gaussian process based test, including the robust noise model (GP robust) to the time courses measured in both conditions and all four replicates. As comparison we also applied two state-of-the-art methods from the literature, the *timecourse* method (TC) of Tai and Speed (8), and the F-Test (FT) as implemented in the MAANOVA package (19). For each of the three methods, we rank all probes based on their likelihood of being differentially expressed in descending order.

On a subset of 2000 randomly selected probes we asked a human expert to manually label each probe as either 'differentially expressed', 'not differentially expressed', or 'dubious case'. After removing the dubious cases, we used the remaining 1890 labeled probes as gold standard to benchmark the three methods. Figure 3 shows the area under the ROC curve for each method. To check the impact of our outlier-robust model, we also computed the area under the ROC curve for a variant of GPTwoSample that is not robust to outliers and instead uses a standard Gaussian noise model (GP standard). The area under the curve can be interpreted as the probability that a classifier ranks a randomly chosen positive instance (a differentially expressed gene) higher than a randomly chosen negative example (a non-differentially expressed gene). Hence a 'perfect' test would reach an AUC of 1, while a consistently failing test would yield an AUC of 0. On this randomly selected set, GPTwoSample with robust noise model (GP robust, AUC 0.986) and the simpler non-robust variant (GP standard, AUC 0.944) outperformed both benchmark models, F-Test (FT, AUC 0.859) and the *timecourse* method (TC, AUC 0.869). The model GP robust achieved an additional improvement over GP standard, showing the merits of a robust noise model.

As a second evaluation, we wanted to get an idea of how the different methods perform on reference datasets of '*controlled difficulty*' (rather than a *random* subset). For this purpose we created reference sets with 400 genes for every method. In each of these sets 100 genes were correctly classified as differentially expressed and 100 correctly as non-differentially expressed. The remaining two hundred genes were false positives and false negatives to equal proportions. Following this procedure, we obtained three labeled reference datasets of 400 genes for GP robust, FT and TC. On each of these datasets, we assessed the predictions of the remaining two methods by computing AUC scores (see Table 1). On the *timecourse* reference dataset, our GP robust achieved a higher area under the curve than the F-Test, and it outperformed *timecourse* on the F-Test reference dataset as well. Hence its ability to correctly detect differential gene expression on these reference datasets is again more than competitive with that of the state-of-the-art two-sample tests F-Test (7) and *timecourse* (8).

To further validate the quality of the gene list produced by GPTwoSample we clustered genes considered to be differentially expressed using the SplineCluster

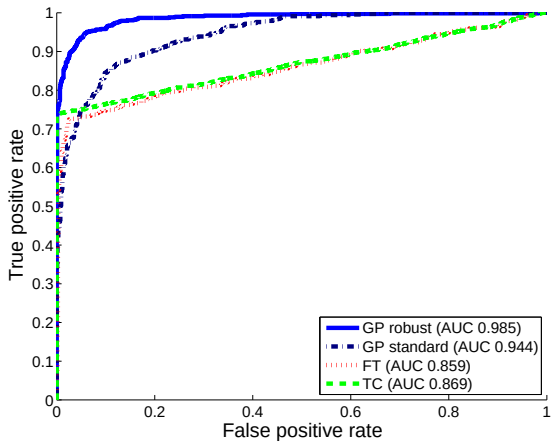


Fig. 3: Predictive accuracy of four different methods measured by the area under the ROC curve. Each method has been evaluated on the random benchmark dataset of 1890 genes as described in the text.

Dataset / Method	GP robust	TC	FT
GP robust dataset	———	0.937	0.929
TC dataset	0.959	———	0.805
FT dataset	0.986	0.956	———

Table 1: AUC scores of three methods on reference datasets created from genome-wide results on GP robust, TC, and FT.

method of Heard et al. (20; 21). We analyzed the resulting clusters for statistically significantly over-represented Gene Ontology(GO) annotations related to a given cluster of genes. The probability that this over-representation is not found by chance can be calculated by the use of a hypergeometric test, implemented in the R/Bioconductor package *GOstats* (22). Because of the effects of multiple testing, a subsequent correction of the p-values is necessary. We apply a Bonferroni correction, which gives a conservative (and easily calculated) correction for multiple testing. In the supplementary material (23) we show the GO annotations for the clusters which are significant at Bonferroni-corrected p-values of 0.01 and 0.05. These GO groupings of the clusters derived from GPTwoSample are intuitively meaningful in the context of plant-pathogen interactions.

3 Detecting intervals of differential gene expression

On knowing that a particular gene is differentially expressed, it is interesting to ask in which time intervals this difference in expression is present and in which time intervals the time series are similar. To tackle this questions we use a mixture model, switching between the two hypotheses, corresponding either to the *shared* model (Figure 2a) or the *independent* model (Figure 2b) as a function of

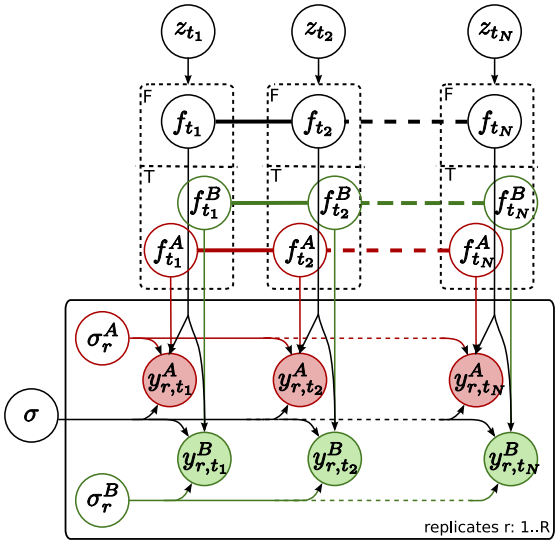


Fig. 4: Bayesian network for the time local mixture model. At each observed time point t_n binary indicator variables z_{t_n} determine whether the observation is explained by the single Gaussian process expert ($f(t)$) or the expert corresponding to the *independent* model ($f^A(t)$ and $f^B(t)$). This switch is graphically represented as dotted boxes around the processes $f(t)$ and $f^A(t)$, $f^B(t)$ respectively. If the switch is true (T) the *independent* expert is used, if the switch is false (F) the *shared* expert.

time. Figure 4 shows the Bayesian network representation of this temporal two-sample test. This model is related to mixtures of Gaussian process experts, which have been studied previously (24; 25). In our setting, we have a fixed number of two experts, where one expert is a single Gaussian process describing both conditions, while the second expert models each condition with a separate process. In order to retain the computational speed required to apply this algorithm on large scale, performing thousands of tests, we use a simplistic gating network. Binary switches z_{t_n} at every observed time point determine which expert describes the expression level at this particular time point. *A priori* the indicator variables are independent Bernoulli distributed, $P(z_{t_n}) = \text{Bernoulli}(z_{t_n} \,|\, 0.5)$, assigning both experts equal probability.

The joint probability of both experts and all model parameters, conditioned on the observed data from both conditions, can be written as

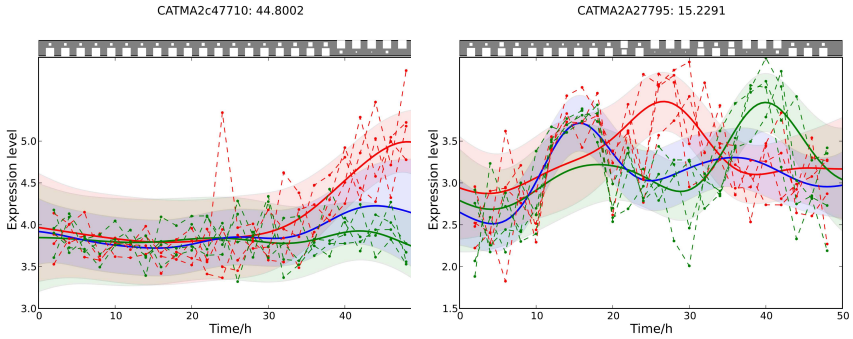
$$P(\mathbf{f}, \mathbf{f}^A, \mathbf{f}^B, \mathbf{Z} \,|\, \mathcal{D}_A, \mathcal{D}_B, \boldsymbol{\theta}_\mathrm{S}, \boldsymbol{\theta}_\mathrm{I}) \propto P(\mathbf{f} \,|\, \boldsymbol{\theta}_\mathrm{K}) P(\mathbf{f}^A \,|\, \boldsymbol{\theta}_\mathrm{K}) P(\mathbf{f}^B \,|\, \boldsymbol{\theta}_\mathrm{K}) \times$$
$$\prod_{r=1}^{R} \prod_{n=1}^{N} \left[\mathcal{N}(f_{t_n} \,|\, y^A_{r,t_n}, \sigma^A_r) \mathcal{N}(f_{t_n} \,|\, y^B_{r,t}, \sigma^B_r) \right]^{(z_{t_n}=0)} \times$$
$$\left[\mathcal{N}(f^A_{t_n} \,|\, y^A_{r,t_n}, \sigma^A_r) \mathcal{N}(f^B_{t_n} \,|\, y^B_{r,t}, \sigma^B_r) \right]^{(z_{t_n}=1)}, \tag{8}$$

where $P(\mathbf{f} \mid \boldsymbol{\theta}_{\mathrm{K}})P(\mathbf{f}^A \mid \boldsymbol{\theta}_{\mathrm{K}})P(\mathbf{f}^B \mid \boldsymbol{\theta}_{\mathrm{K}})$ denotes the independent Gaussian process priors on all three processes. Again we simplify the presentation by considering a Gaussian noise model.

Inference in this model is achieved using a variational approximation (26). The joint posterior distribution (Eqn. 8) is approximated by a separable distribution of the form $Q(\mathbf{f})Q(\mathbf{f}^A)Q(\mathbf{f}^B)\prod_{n=1}^{N} Q(z_{t_n})$.
Iterative variational inference updates the approximate posteriors over the latent processes $Q(\mathbf{f}), Q(\mathbf{f}^A), Q(\mathbf{f}^B)$ given the current state of $Q(\mathbf{Z})$ and vice versa, until convergence is reached. A variational approximation per se is not suited to perform inference in a mixture of Gaussian process model, due to the coupling of target values induced by the GP priors. However, in this specific application, the approximate posteriors over the indicator variables are sufficiently accurate. Finally, to decide whether a time point is differentially expressed, we use the inferred mixing state $Q(z_{t_n})$ with a threshold value of 0.5.

3.1 Detecting transition points in the Arabidopsis time series data

We applied the temporal GPTwoSample model to detect intervals of differential expression of genes from the same *Arabidopsis* time series dataset as in Section 2.4. Figure 5 shows raw data and the inference results for two selected example genes.



(a) Delayed differential expression (b) Periodic differential expression

Fig. 5: Two example results of the temporal GPTwoSample model on the *Arabidopsis* data. The bottom panel in each plot illustrates the inferred posterior distributions from the Gaussian processes (blue: the process describing the *shared* biological behavior; red and green: the two separate processes modelling differential gene expression). The Hinton diagrams in the top panel indicate whether at a given point in time the gene is likely to be differentially expressed or not. The size of the dots in each row is proportional to the probability of differential expression (top row) and of no differential expression (bottom row).

Delayed differential expression Having the inferred time intervals of differential and non-differential expression at hand, it is possible to analyze the time information and its distribution over genes.

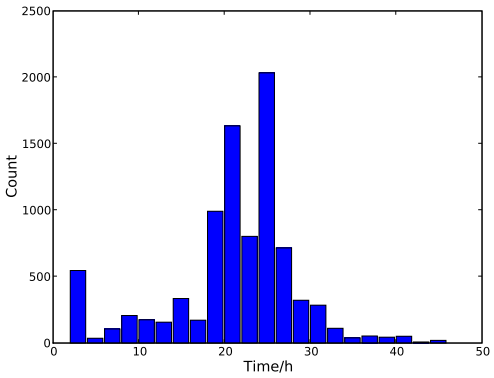


Fig. 6: Histogram of the most likely start of differential expression for the top 9000 differentially expressed genes.

Here we took the top 9000 genes which have a score suggesting significant differential expression and determined the first time point where the probability of differential expression exceeded 0.5. Figure 6 shows the histogram of this start time. Identification of transition points for individual gene expression profiles shows that a significant change in the transcriptional program begins around 20h post inoculation. This program of gene expression change appears to have two waves peaking around 22h and 26h after inoculation. We expect transcription factors (if regulated by differential expression) to be expressed at earlier time points than the downstream genes whose expression they control. Hence transcription factor genes whose expression first changes in the 22h wave (or earlier) would be of particular interest when designing further experiments to elucidate transcriptional networks mediating the defense response against *B. cinerea*.

4 Conclusion

Detecting differential gene expression and patterns of its temporal dynamics are important first steps towards understanding regulatory programs on a molecular level. In this paper, we proposed a Gaussian process framework which provides answers to these problems. Our test not only determines which genes are differentially expressed, but also infers subintervals of differential expression over time. The analysis carried out on the *Arabidopsis thaliana* expression datasets

demonstrates that this additional knowledge can be used to gain an understanding of pathways and the timing in which, as in this example, the effect of a fungus infection spreads. Source code and additional information about the used dataset is available online (23).

The natural next question to ask is in which manner these genes interact as part of a regulatory program. The algorithmic task is here to infer a network of regulatory interactions from gene expressions measurements and prior knowledge. In future work, we will study how the detection of differential expression can be combined with regulatory network inference.

Acknowledgments The authors would like to thank Andrew Mead and Stuart McHattie for data preprocessing. We acknowledge support from the Cambridge Gates Trust (OS), grants BBSRC BB/F005806/1 (KD and DW), EU Marie Curie IRG 46444 (DW) and NIH GM63208 (KB).

A Expectation Propagation for robust Gaussian process regression

Predictions (Eqn. 4) and the log marginal likelihood (Eqn. 5) are only available in closed form for a Gaussian likelihood model P_{L}. When using a complicated likelihood function, such as the mixture model in Eqn. 7, Expectation Propagation (EP) (15) can be used to obtain a tractable approximation.

In our application the exact posterior distribution over latent functions $f(t)$ for a given dataset $\mathcal{D} = \{t_n, y_n\}_{n=1}^N$ is

$$\begin{aligned} P\left(\mathbf{f}|\mathcal{D},\boldsymbol{\theta}\right) &\propto \mathcal{N}(\mathbf{f}\,|\,0, K_{\mathbf{T}\,\mathbf{T}}(\boldsymbol{\theta}_{\mathrm{K}})) \prod_{n=1}^{N} P_{\mathrm{L}}\left(y_n \,|\, f_n, \boldsymbol{\theta}_{\mathrm{L}}\right) \\ &= \mathcal{N}(\mathbf{f}\,|\,0, K_{\mathbf{T}\,\mathbf{T}}(\boldsymbol{\theta}_{\mathrm{K}})) \prod_{n=1}^{N} \left[\pi_0\, \mathcal{N}(y_n|f_n,\sigma) + (1-\pi_0)\, \mathcal{N}(y_n \,|\, f_n, \sigma_{\mathrm{inf}})\right], \end{aligned} \tag{9}$$

where again we define $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\mathrm{K}}, \boldsymbol{\theta}_{\mathrm{L}}\}$. The goal of EP is to approximate this exact posterior with a tractable alternative

$$Q(\mathbf{f}\,|\,\mathcal{D},\boldsymbol{\theta}) \propto \mathcal{N}(\mathbf{f}\,|\,0, K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_{\mathrm{K}})) \prod_{n=1}^{N} g_n\left(f_n\right), \tag{10}$$

where $g_n\left(f_n\right)$ denote approximate factors. Following (14) we choose unnormalized Gaussians as approximate factors

$$g_n\left(f_n \,|\, C_n, \tilde{\mu}_n, \tilde{\nu}_n\right) = C_n \exp\left(-\frac{1}{2\tilde{\nu}_n}(f_n - \tilde{\mu}_n)^2\right), \tag{11}$$

which leads to an approximate posterior distribution of $f(t)$ that is a Gaussian process again. Evaluated at the training inputs the distribution over function values is a multivariate Gaussian

$$Q(\mathbf{f} \mid \mathcal{D}, \boldsymbol{\theta}_\mathrm{K}, \boldsymbol{\theta}_\mathrm{L}) \propto \mathcal{N}(\mathbf{f} \mid 0, K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_\mathrm{K})) \prod_{n=1}^{N} g_n\left(f_n \mid C_n, \tilde{\nu}_n, \tilde{\nu}_n\right) \tag{12}$$

$$= \mathcal{N}(\mathbf{f} \mid 0, K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_\mathrm{K}))\mathcal{N}(\mathbf{f} \mid \tilde{\boldsymbol{\mu}}, \tilde{\Sigma}), \tag{13}$$

where we define $\tilde{\boldsymbol{\mu}} = \{\nu_1, \ldots, \nu_N\}$ and $\tilde{\Sigma} = \operatorname{diag}\left(\{\nu_1^2, \ldots, \nu_N^2\}\right)$.

The idea of EP is to iteratively update one approximate factor leaving all other factors fixed. This is achieved by minimizing the Kullback–Leibler (KL) divergence, a distance measure for distributions (27). Updates for a single approximate factor i can be derived by minimizing

$$\mathrm{KL}\Bigg[\mathcal{N}(\mathbf{f} \mid 0, K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_\mathrm{K})) \prod_{n \neq i} q_n\left(f_n \mid C_n, \tilde{\mu}_n, \tilde{\nu}_n\right) \overbrace{P_\mathrm{L}(y_i \mid f_i, \boldsymbol{\theta}_\mathrm{L})}^{\text{exact factor}} \,\Big|\Big|$$

$$\mathcal{N}(\mathbf{f} \mid 0, K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_\mathrm{K})) \prod_{n \neq i} q_n\left(f_n \mid C_n, \tilde{\mu}_n, \tilde{\nu}_n\right) \underbrace{g_i\left(f_i \mid C_i, \tilde{\mu}_i, \tilde{\nu}_i\right)}_{\text{approximation}}\Bigg] \tag{14}$$

with respect to the ith factor's parameters $\tilde{\mu}_i, \tilde{\nu}_i$ and C_i. This is done by matching the moments between the two arguments of the KL divergence which can then be translated back into an update for factor parameters. It is convenient to work in the natural parameter representation of the distributions where multiplication and division of factors are equivalent to addition and subtraction of the parameters.

There is no convergence guarantee for EP, but in practice it is found to converge for the likelihood model we consider (14). The fact that the mixture of Gaussians likelihood is not log-concave is problematic, as it may cause invalid EP updates, leading to a covariance matrix that is not positive definite. We avoid this problem by damping the updates (14; 28).

After EP converged, we obtain a Gaussian process as approximate posterior distribution again and hence can evaluate a predicted mean and variance as for the Gaussian noise model (Eqn. 4).

By capturing the zeroth moment of the exact distribution with the explicit normalization constant C_n, we obtain an approximation to the log marginal

likelihood

$$\log P(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{K}}, \boldsymbol{\theta}_{\mathrm{L}}) = \ln \int df\, \mathcal{N}(\mathbf{f} \,|\, 0, K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_{\mathrm{K}})) \prod_{n=1}^{N} P_{\mathrm{L}}\left(f_n \,|\, y_n, \boldsymbol{\theta}_{\mathrm{L}}\right)$$

$$\approx \ln \int df\, \mathcal{N}(\mathbf{f} \,|\, 0, K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_{\mathrm{K}})) \prod_{n=1}^{N} g_n\left(f_n \,|\, C_n, \tilde{\mu}_n, \tilde{\nu}_n\right) \tag{15}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\left(\ln \tilde{\nu}_n^2 + \ln C_n\right) - \frac{1}{2}\ln\left|K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_{\mathrm{K}}) + \tilde{\Sigma}\right|$$

$$- \frac{1}{2}\tilde{\boldsymbol{\mu}}^{\mathsf{T}}\left(K_{\mathbf{T},\mathbf{T}}(\boldsymbol{\theta}_{\mathrm{K}}) + \tilde{\Sigma}\right)\tilde{\boldsymbol{\mu}}. \tag{16}$$

This log marginal likelihood approximation enables us to optimize hyperparameters of the kernel $\boldsymbol{\theta}_{\mathrm{K}}$, as well as the from likelihood $\boldsymbol{\theta}_{\mathrm{L}}$ and serves as approximation when evaluating the *Bayes factor* in Eqn. 1.

Bibliography

[1] Kerr, M., Martin, M., Churchill, G.: Analysis of Variance for Gene Expression Microarray Data. Journal of Computational Biology **7**(6) (2000) 819–837

[2] Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P.: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica **12** (2002) 111–140

[3] Efron, B., Tibshirani, R., Storey, J.D., Tusher, V.: Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association **96** (2001) 1151–1160

[4] Ishwaran, H., Rao, J.: Detecting differentially expressed genes in microarrays using Bayesian model selection. Journal of the American Statistical Association **98** (2003) 438–455

[5] Lonnstedt, I., Speed, T.: Replicated microarray data. Statistica Sinica **12** (2002) 31–46

[6] Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D.K., Jaakkola, T.S.: Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. Proceedings of the National Academy of Sciences of the United States of America **100** (September 2003) 10146–51

[7] Storey, J.D., Xiao, W., Leek, J.T., Tompkins, R.G., Davis, R.W.: Significance analysis of time course microarray experiments. Proceedings of the National Academy of Sciences of the United States of America **102** (September 2005) 12837–42

[8] Tai, Y.C., Speed, T.P.: A multivariate empirical Bayes statistic for replicated microarray time course data. Annals of Statistics **34** (2006) 2387–2412

[9] Angelini, C., De Canditiis, D., Mutarelli, M., Pensky, M.: A Bayesian Approach to Estimation and Testing in Time-course Microarray Experiments. Statistical Applications in Genetics and Molecular Biology **6** (September 2007)

[10] Yuan, M.: Flexible temporal expression profile modelling using the Gaussian process. Computational Statistics and Data Analysis **51** (2006) 1754–1764

[11] Lawrence, N.D., Sanguinetti, G., Rattray, M.: Modelling transcriptional regulation using Gaussian Processes. In: Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA (2007) 785–792

[12] Chu, W., Ghahramani, Z., Falciani, F., Wild, D.: Biomarker discovery in microarray gene expression data with Gaussian processes. Bioinformatics **21**(16) (2005) 3385–3393

[13] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press (December 2006)

[14] Kuss, M., Pfingsten, T., Csato, L., Rasmussen, C.: Approximate Inference for Robust Gaussian Process Regression. Technical report, Max Planck Institute for Biological Cybernetics, Tubingen, 2005

[15] Minka, T.: Expectation propagation for approximate Bayesian inference. In: Uncertainty in Artificial Intelligence. Volume 17. (2001) 362–369
[16] Stegle, O., Fallert, S.V., MacKay, D.J., Brage, S.: Gaussian process robust regression for noisy heart rate data. IEEE Trans Biomed Eng **55** (2008) 2143–51
[17] Fujita, M., Fujita, Y., Noutoshi, Y., Takahashi, F., Narusaka, Y., Yamaguchi-Shinozaki, K., Shinozaki, K.: Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks. Current Opinion in Plant Biology **9** (August 2006) 436–442
[18] Allemeersch, J., Durinck, S., Vanderhaeghen, R., Alard, P., Maes, R., Seeuws, K., Bogaert, T., Coddens, K., Deschouwer, K., Hummelen, P.V., Vuylsteke, M., Moreau, Y., Kwekkeboom, J., Wijfjes, A.H., May, S., Beynon, J., Hilson, P., Kuiper, M.T.: Benchmarking the catma microarray. a novel tool forarabidopsis transcriptome analysis. Plant Physiol. **137** (February 2005) 588–601
[19] Wu, H., Kerr, M., Cui, X., Churchill, G.: MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. The Analysis of Gene Expression Data: Methods and Software 313–341
[20] Heard, N., Holmes, C., Stephens, D., Hand, D., Dimopoulos, G.: Bayesian coclustering of Anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. Proceedings of the National Academy of Sciences **102**(47) (2005) 16939–16944
[21] Heard, N., Holmes, C., Stephens, D.: A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. Journal of the American Statistical Association **101**(473) (2006) 18
[22] Falcon, S., Gentleman, R.: Using GOstats to test gene lists for GO term association. Bioinformatics **23**(2) (2007) 257
[23] Stegle, O., Denby, K., Wild, D.L., Ghahramani, Z., Borgwardt, K.: Supplementary material: A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series (2009) http://www.inference.phy.cam.ac.uk/os252/projects/GPTwoSample.
[24] Yuan, C., Neubauer, C.: Variational Mixture of Gaussian Process Experts. In: Advances in Neural Information Processing Systems 19, Cambridge, MA, MIT Press (2008)
[25] Rasmussen, C.E., Ghahramani, Z.: Infinite Mixtures of Gaussian Process Experts. In: Advances in Neural Information Processing Systems 19, Cambridge, MA, MIT Press (2001) 881–888
[26] Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L.: An introduction to variational methods for graphical models. Machine Learning **37** (1999) 183–233
[27] Kullback, S., Leibler, R.: On Information and Sufficiency. The Annals of Mathematical Statistics **22**(1) (1951) 79–86
[28] Seeger, M.: Expectation Propagation for Exponential Families. Technical report, University of California at Berkeley, 2005.