BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 20 2011, pages 2859–2865
doi:10.1093/bioinformatics/btr475

Gene expression

Advance Access publication August 16, 2011

The use of semiparametric mixed models to analyze PamChip® peptide array data: an application to an oncology experiment

Pushpike J. Thilakarathne[1,*], Lieven Clement[1,2], Dan Lin[1,2], Ziv Shkedy[2], Adetayo Kasim[3], Willem Talloen[4], Matthias Versele[4] and Geert Verbeke[1]

[1]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Katholieke Universiteit Leuven, B3000 Leuven, [2]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt, Agoralaan 1, B3590 Diepenbeek, Belgium, [3]Wolfson Research Institute, Durham University Queen's Campus, University Boulevard, Thornaby, Stockton-on-Tees, UK and [4]Janssen Pharmaceutica N.V., Beerse, Belgium

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Phosphorylation by protein kinases is a central theme in biological systems. Aberrant protein kinase activity has been implicated in a variety of human diseases (e.g. cancer). Therefore, modulation of kinase activity represents an attractive therapeutic approach for the treatment of human illnesses. Thus, identification of signature peptides is crucial for protein kinase targeting and can be achieved by using PamChip® microarray technology. We propose a flexible semiparametric mixed model for analyzing PamChip® data. This approach enables the estimation of the phosphorylation rate (Velocity) as a function of time together with pointwise confidence intervals.

Results: Using a publicly available dataset, we show that our model is capable of adequately fitting the kinase activity profiles and provides velocity estimates over time. Moreover, it allows to test for differences in the velocity of kinase inhibition between responding and non-responding cell lines. This can be done at individual time point as well as for the entire velocity profile.

Contact: pushpike@med.kuleuven.be

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

Received on February 24, 2011; revised on August 2, 2011; accepted on August 8, 2011

1 INTRODUCTION

Scientific advances in biotechnology bring us in a new era of targeted therapies and personalized medicine. Biomarkers are essential for selecting patients that benefit from certain treatments, and biomarker discovery is nowadays driven by novel platforms that measure more biological properties more precisely. Evidently, these technological advances often require the development of optimal analysis approaches for its generated data in order to increase its detection power.

Here, we propose a statistical model to efficiently analyze output from PamChip® microarrays. These high-content peptide arrays are capable of measuring phosphorylation changes over time for more than hundred peptides simultaneously (Hilhorst *et al.*, 2009; Nagaoka *et al.*, 2005; Perera *et al.*, 2006; Versele *et al.*, 2009). Compared to conventional microarrays, PamChips have two main advantages. First, proteins are closer to phenotypic expression than DNA or mRNA profiles. Second, PamChips are not measuring concentrations but biological activity, i.e. phosphorylation over time. An important application area of PamChip® microarrays is the study of protein kinase activity (Versele *et al.*, 2009). Protein kinase inhibitors can specifically block key signaling pathways in cancer, making kinases very popular targets in anticancer-targeted therapy discovery programs (Kondapalli *et al.*, 2005). PamChip® microarrays have a very distinct technical design (see Section 1 of the Supplementary Material for more details) and the data properties of the PamChip® read-out are atypical compared with more traditional microarrays. Hence, specific data analysis pipeline has to be developed for PamChip® microarrays.

In this article, we demonstrate the statistical methodology to model the kinase activity profiles and thereby estimate the phosphorylation rate, referred to as velocity. With the proposed approach, phosphorylation rates can be investigated and tested for each peptide. The article is organized as follows. Section 2 describes a real-life PamChip® microarray experiment that is used to demonstrate our method. In Section 3, we describe existing methods for analyzing PamChip® array data and motivate our approach. In Section 4, we give a brief review of penalized smoothing splines formulated as mixed models, specify the proposed model, specifically adapt it to the PamChip® array data at hand and focus on the inference problem of this study. The results are given in Section 5. The discussion and conclusions are presented in Sections 6 and 7, respectively.

2 REAL-LIFE PAMCHIP® ARRAY DATA

We use a real-life dataset described in Versele *et al.* (2009) as a motivating example for our method. It consists of an oncology experiment where phosphorylation of peptides is measured in 20 different cell line lysates in the absence (control) and presence of a kinase inhibitor (compound) using peptide arrays with a kinetic read-out. Ten out of twenty cell lines are known to be responsive and others are non-responsive to the specific kinase inhibitor under study. The cell lines were randomly assigned to six 96-well plates, of which each well contains a 144-peptide array. Each cell line was typically replicated 5–6 times. For quality control issues, some wells were excluded.

*To whom correspondence should be addressed.

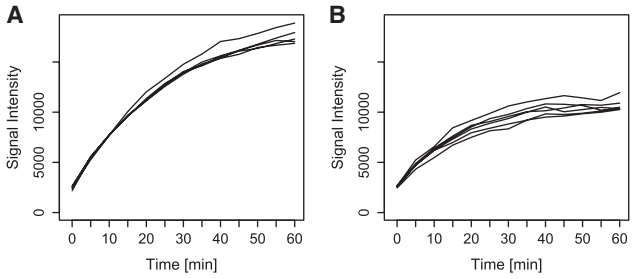

Fig. 1. PamChip® microarray kinetic output: observed kinase activity profiles for a responsive cell line. Profiles for a non-responsive cell line is given in Supplementary Figure S2. (**A**) Control responsive cell line; (**B**) Compound responsive cell line.

The intensities of 144 peptides were measured at 5 min time intervals over a period of 60 min using PamChip® technology. The first kinetic reading is done 11 min after the assay mix was pumped to the plate. The last reading is done at 71 min. For simplicity, we shift the time by subtracting 11 min from each time point leading to a time range $t = 0, ..., 60$ min. Therefore, $t = 0$ corresponds to the first kinetic read. Figure 1 shows kinase activity over time for a selected peptide of a responsive cell line under the treatment (Fig. 1B) and control (Fig. 1A). Profiles for a non-responsive cell line are given in Supplementary Figure S2. The graphical illustration of intensity measurements for a selected peptide across all the cell lines are given in Supplementary Figure S3.

3 MOTIVATION

Signal intensity profiles are the outcomes of the peptide microarray experiment and correspond to the degree of phosphorylation. For assessing the impact of the kinase inhibitor compound, however, the phosphorylation rate is required. Hilhorst *et al.* (2009) modeled peptide array data with an exponential model, $Y_{ij} = Y_o + Y_{\max}[1 - e^{-\eta t_j}]$, where the Y_{ij} values stand for the signal intensity of subject i at j-th time point, Y_o is the signal intensities at the beginning of the experiment and $Y_{\max}$ is the maximum observed intensity. η is the reaction rate constant and t_j is the time point when the image was recorded. Only the initial velocity of the peptide phosphorylation was used for subsequent data analysis. Versele *et al.* (2009) preprocessed each PamChip® profile separately using CurveFitHT software for deriving the initial velocity at the first kinetic time point read. They only used the first 30 min of the assay and based all subsequent data analysis on the estimated initial velocities. In particular, the estimated initial velocities were $\log_2$ transformed and are used for comparing responsive and non-responsive cell lines with a mixed model.

Both Versele *et al.* (2009) and Hilhorst *et al.* (2009) use a two stage approach where (i) exponential models are used in the first stage to summarize the individual time series into a single value i.e initial velocities and (ii) these estimated initial velocities are assessed in the downstream analysis without accounting for their associated uncertainty. Hence, they ignore the within replicate variability and only capture the technical variation between the replicates. This can have a severe effect on the outcome of the second stage analysis if the within variability of the replicates is large as compared to the between variability and in case that the uncertainty of the initial velocity estimates differs among replicates, cell lines or treatments. Hence, the Hilhorst method and the Versele method might suffer from a loss in power, efficiency and an incorrect control of the type I error (Verbeke and Molenberghs, 2000).

We propose a semiparametric mixed effects model that combines both the estimation and modeling stage into a single analysis. Moreover, the link between mixed models and smoothing provides a very flexible framework for estimating the kinase activity profiles in a data-driven way. Within the mixed model framework, it is very natural to account for the dependence structure of the intensity signals within and across activity profiles by modeling all data for a given peptide, simultaneously. The proposed semiparametric mixed model allows for testing whether the change in phosphorylation rate (velocity) induced by the treatment (compound) effect differs between responsive and non-responsive cell lines. The tests can be performed at any time point as well as for the entire profile, simultaneously. Further, velocity profiles can be derived as a function of time along with 95% pointwise confidence bands.

4 METHODS

High-content PamChip® microarrays produce longitudinal time series profiles. For inference, it is necessary that we properly capture the form of the evolution of profiles over time. In longitudinal studies, the mean profile is often estimated by parametric functions. Examples of longitudinal studies and an extensive overview on longitudinal data analysis and inference using parametric models can be found in Fitzmaurice *et al.* (2008); Molenberghs and Verbeke (2005); Pinheiro and Bates (2000); Verbeke and Molenberghs (2000). However, many biomedical experiments typically generate non-linear data and imposing parametric functions might yield unsatisfactory results (Bowman and Azzalini, 1997; Wood, 2006). In the context of PamChip® experiments, for instance, the individual profiles are non-linear and parametric models may be too restrictive. Moreover, the experimental setup gives rise to a complex dependence structure, which can obscure the interpretation and inference on the non-linear phosphorylation profiles. Therefore, we propose a data-driven approach based on semiparametric regression models. Particularly, we use penalized thin plate regression splines to analyze PamChip® data. Penalized thin plate regression splines avoids problems of knot selection and knot placement, as they both emerge naturally from the mathematical formulation of the smoothing problem (Ramsay and Silverman, 2005; Wood, 2006). The smoothing parameters that control the degree of smoothing typically have to be tuned toward the specific application. Interestingly, these smoothing parameters can also be estimated within the linear mixed-model framework (Maringwa *et al.*, 2008; Ruppert *et al.*, 2003; Wood, 2006).

Let $Y_i(t)$ denote the $\log_2$ transformed and background corrected intensity measurement on cell line i ($i = 1, ..., n$), taken at time point t. The penalized spline model, with cell line-specific random effects $b_{0i}, ..., b_{pi}$ can be expressed as

$$Y_i(t) = \underbrace{\beta_0 + \sum_{l=1}^{v} \beta_l f_l(t)}_{S(t)} + b_{0i} + b_{1i}t + ... + b_{pi}t^p + \varepsilon_{i(t)}, \quad (1)$$

where $f_l(t)$s are a set of thin plate spline basis functions (Supplementary Figure S4), β_l are the coefficients of the basis functions (Ramsay and Silverman, 2005; Ruppert *et al.*, 2003; Wood, 2006), $[b_{oi}, ..., b_{pi}]^T \sim MVN(\mathbf{0}, \Sigma_b)$ and $\varepsilon_{i(t)}$ i.i.d. $N(0, \sigma_\varepsilon^2)$. The cell line-specific random effects b_{pi}'s accounts for the correlated nature of the intensities. Statistical significance of the cell line-specific random effects can be tested by using an approximate likelihood ratio test (Verbeke and Molenberghs, 2000). The spline model parameters can be

estimated by considering the goodness-of-fit and the degree of smoothness (Bowman and Azzalini, 1997; Ramsay and Silverman, 2005; Wood, 2006). The model (1) can be fitted by finding the function from an appropriate reproducing kernel Hilbert space which minimizes

$$\boldsymbol{Q}=||\mathbf{y}-\boldsymbol{S}||^2+\lambda\int[\boldsymbol{S}^m(t)]^2dt, \tag{2}$$

where $\int[\boldsymbol{S}^{(m)}(t)]^2dt$ is the roughness penalty which penalizes for the lack of smoothness of S. Roughness is quantified by the integral of squared m-th order derivatives. The most commonly used roughness penalty is $m=2$. Note that, for $m=2$ and a given λ, $\widehat{\boldsymbol{S}}$, the minimizer of $\boldsymbol{Q}$, is a cubic smoothing spline. In the literature, the smoothing parameter λ is often tuned by using an optimization criterion such as generalized cross-validation. Penalized thin plate regression splines (TPRS), however, can also be represented within the mixed model framework (Wood, 2006). It allows for the estimation of the smoothing parameter $\lambda=\sigma_\varepsilon^2/\sigma_s^2$, where σ_s^2 is the variance of the random effects that are involved in the mixed model representation of the spline.

In this study, our aim is to estimate the marginal phosphorylation rate over all cell lines, which is the first-order derivative of the smoother, $S(t)$, with respect to time t. Note that, the first-order derivative $S'(t)=\sum_{\iota=1}^{v}\beta_\iota f_\iota'(t)$ is a linear combination of the derivatives of the basis functions.

$S'(t)$ is a smooth function of time when using cubic splines. However, they give rise to biased velocity estimates at the boundaries (Supplementary Section S9) and are not suitable in our situation. Therefore, we will penalize on the third-order derivative, which provides a more smooth first derivative that does not suffer from boundary artifacts (Ramsay and Silverman, 2005; Wood, 2006). Because the TPRS are adopted within the mixed model framework, standard errors and confidence bands for the velocity follow from standard theory of mixed models (e.g. Verbeke and Molenberghs, 2000; Ruppert, Wand and Carroll, 2003, Section 6.8.1; and Wood, 2006).

Fitting penalized splines models within the linear mixed model framework has some appealing advantages, such as the estimation of the smoothing parameter λ, a unified framework for inference and straightforward extension of the model. Moreover, the mixed model representation also provides a natural way for handling clustered data, missing data and measurement error (Durbán *et al.*, 2004; Maringwa *et al.*, 2008; Ruppert *et al.*, 2003; Wood, 2006). In the following section, we describe how we draw the inference from the proposed approach.

4.1 Inference on first-order derivative

In this particular study, we are interested in marginal velocity. Hypothesis tests on the velocity can be formulated by using general linear hypotheses.

$$H_0:\boldsymbol{LV}=0 \quad \text{versus} \quad H_a:\boldsymbol{LV}\neq 0,$$

where $\boldsymbol{L}$ is the contrast matrix and $\boldsymbol{V}$ is a vector which contains coefficients of the first derivatives of the basis functions ($f_\iota(t)$) discussed in Section 4.

4.2 Formulation of the model for PamChip ® array data

This section provides the formulation of the spline model (1) for the PamChip® array data at hand. We propose a full factorial model structure for treatment and responsive status. For a desired flexibility, the penalized thin plate regression smoothing splines with penalization on the third-order derivative ($m=3$) are considered (Ramsay and Silverman, 2005; Wood, 2006).

For the present study, we have two treatments (control and compound) and two responsive statuses (responsive: R and non-responsive: NR) for cell lines. This data structure formulates four distinct groups namely, $g=1$: Compound - R, $g=2$: Compound - NR, $g=3$: Control - NR and $g=4$: Control - R.

Let Y_{gijl} denote the $\log_2$ transformed and background corrected intensity measurement of cell line i, replicate j, measured at time t, on the l-th plate for group g, with $i=1,...,20; j=1,...,6; t=0,5,10,...,55,60$ min, $l=1,...,6$ and $g=1,...,4$. We propose the following semiparametric mixed model with quadratic random effects structures ($p=2$) for replicates as well as for cell lines

$$\begin{aligned} Y_{gijl} &= S_g(t)+b_{0gi}+b_{1gi}t+b_{2gi}t^2 \\ &+ c_{0j(gi)}+c_{1j(gi)}t+c_{2j(gi)}t^2 \\ &+ a_l+\varepsilon_{gijl(t)}. \end{aligned} \tag{3}$$

In this proposed model, the cell line-specific random intercept is considered to capture correlation of the intensity measurement over time within the cell line. We assumed cell line-specific random slopes for linear as well as for quadratic time effects to capture different evolution of kinase activity over time. Moreover, we allow these cell lines-specific random structures to be different for each group. We assumed group-specific random structure for cell lines for which $[b_{0gi},b_{1gi},b_{2gi}]^T\sim MVN(\mathbf{0},\Sigma_b)$ where g is the group indicator and Σ_b denote the variance covariance matrix of cell line-specific random effects. There are technical replicates for each cell line under each treatment and therefore, the replicate-specific random effect within the cell line, $c_{0j(gi)}$, is specified. Replicate-specific random slopes for linear and quadratic time effects nested within the cell line, capture the different evolutions of replicate-specific profiles (Durbán *et al.*, 2004). We assumed that the replicate-specific random structure is $[c_{0j(gi)},c_{1j(gi)},c_{2j(gi)}]^T\sim MVN(\mathbf{0},\Sigma_c)$ and Σ_c denote the variance covariance matrix of replicate-specific random effects. A random 96-well plate-specific effect, $a_l\sim N(0,\sigma_a^2)$, is assumed for the correlation of the cell lines on the same 96-well plate. Ten out of twenty cell lines are known to be responsive and others are non-responsive. Therefore, we assumed heterogeneous error variability across the four different groups. Hence, the residuals $\varepsilon_{gijl(t)}$ are assumed to be i.i.d. $N(0,\sigma_{\epsilon g}^2)$.

Further, we assume a different smoother, $S_g(t)$, for the four different groups, which is defined as

$$S_g(t)=\beta_{0g}+\sum_{\iota=1}^{v}\beta_{\iota g}f_\iota(t),$$

with group-specific smoothing parameter λ_g. Note, that the design matrix, corresponding to the smoother, is a block diagonal with each diagonal entry corresponding to a particular group.

4.3 Comparing group-specific velocities

The data structure for the present study formulates four distinct groups as described in Section 4.2. We define $S_g'(t)$ as the velocity for each group g, $g=1,...,4$, and we are interested in investigating whether kinase inhibitor induces a similar velocity change in responsive and non-responsive cell lines. To test whether the treatment effect is the same for responsive and non-responsive cell lines, we formulate the general linear hypothesis.

$$\begin{aligned} H_o:&S_1'(t)-S_4'(t) = S_2'(t)-S_3'(t) \text{ versus} \\ H_a:&S_1'(t)-S_4'(t) \neq S_2'(t)-S_3'(t). \end{aligned} \tag{4}$$

This hypothesis can be tested at any time point t as well as for the entire time profile $\forall t\in[0,...,60]$. The null hypothesis obviously implies that the velocity contrast between the treatment and control is the same for responsive and non-responsive cell lines. Supplementary Figure S5 illustrates a hypothetical example for a possible scenario related to the evolution of velocity over time under the alternative hypothesis. The left panel of Supplementary Figure S5 shows a velocity at $t=0$ under the alternative hypothesis and the right panel shows the evolution of velocity over time for each group. Four groups have different velocities over time and the groups differ in both linear and smooth parts.

First we consider testing at a particular time point $t=t^*$. Let $\boldsymbol{L}=[S_1'(t^*) -S_2'(t^*)\ S_3'(t^*) -S_4'(t^*)]$ be the contrast for which $S_g'(t^*)=[f_1'(t^*),...,f_v'(t^*)]$ is the vector of first derivatives of basis functions. $\boldsymbol{V}=[\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\beta}_3,\boldsymbol{\beta}_4]^T$ is the vector of group-specific coefficients of the basis functions for which $\boldsymbol{\beta}_g=[\beta_{1g},...,\beta_{vg}]$. Therefore, the null hypothesis is $H_0:\boldsymbol{LV}=0$ and the

test statistic $\frac{\boldsymbol{L}\widehat{\boldsymbol{V}}}{\sqrt{\boldsymbol{L}\widehat{\Sigma}_V\boldsymbol{L}^T}}$, asymptotically follows a $N(0,1)$ under H_0. $\widehat{\Sigma}_V$ is the estimated variance covariance matrix of the coefficients of the basis functions. Second, we illustrate how we can perform the test (4) for entire time profile i.e. $\forall t \in [0,60]$. The contrast matrix $\boldsymbol{L}$ is now simplified to

$$\boldsymbol{L}=\begin{bmatrix} 1 & -1 & 1 & -1 & & & & & & & & \\ & & & & 1 & -1 & 1 & -1 & & & & \\ & & & & & & & & \cdot & & & \\ & & & & & & & & & \cdot & & \\ & & & & & & & & & & 1 & -1 & 1 & -1 \end{bmatrix}.$$

The ι-th row represents the contrast on the coefficients of the ι-th basis function $f_\iota(t)$ for four different groups, with $\iota = 1,\ldots,v$. In this case, hypothesis (4) can then be tested using the test statistic $\boldsymbol{L}\widehat{\boldsymbol{V}}(\boldsymbol{L}\widehat{\Sigma}_V\boldsymbol{L}^T)^{-1}\boldsymbol{L}\widehat{\boldsymbol{V}}^T$, which follows a chi-squared distribution with degrees of freedom equal to the rank of $\boldsymbol{L}$. The test can be performed for each peptide and illustrations are given in Section 5. We provided a proof for the equivalence between testing on the entire profiles and testing on the coefficients of the basis functions in the Supplementary Section S11.

In what follows, we apply the proposed semiparametric mixed model (3) to the PamChip® array data and test the scenarios illustrated in Supplementary Figure S5.

5 RESULTS

We apply the proposed method to a publicly available dataset from Versele *et al.* (2009), which we described in Section 2. Similar to microarray technology, PamChip® data are also subject to noise and we use background corrected intensity measurements for the data analysis. Prior to fitting the model, we examine the data for unusual observations and extreme profiles. Outlier removal has been done in three steps. First, we notice that some of the observations are negative for a given kinase activity profile. The PamChip® software reports background corrected intensities. Hence, negative observations might occur when the measured intensity for a particular peptide in a certain well does not exceed the background signal. Negative intensity measurements are replaced by half of the positive minimum intensity measurement of that profile. Second, we compute the area under the curve (AUC) for each profile of a given cell line. Analysis of variance has been carried out based on those AUC values while taking the treatment as covariate with two levels. Extreme profiles whose residuals are beyond $\pm 2\times$ the square root of the mean squared error are removed. Third, we look at the observations within a profile and remove the extreme observations if they differ more than $\pm 2\times$ standard deviations of the profile-specific mean. In this step, we are able to remove extremely high intensity measurements (spikes). All these steps are repeated for each peptide and each cell line. Supplementary Figures S6 and S7 illustrate these preprocessing steps for peptide `AMPE_5_17_Y12` and cell line `N87`. The filtered data are then $\log_2$ transformed to stabilize the variability over time.

5.1 Results on a single peptide

We begin with analyzing peptide, `TYRO3_679_691_Y686`, which is known to be biologically significant (Versele *et al.*, 2009). The proposed penalized thin plate regression splines model is fitted for the particular peptide by using the *gamm* procedure available in R package *mgcv*. The smoothing parameters can be estimated via generalized cross-validation (Bowman and Azzalini, 1997; Ramsay and Silverman, 2005; Wood, 2006). In this study, we use the link between mixed model and splines. This is an advantage of fitting the thin plate regression splines within the mixed model framework.

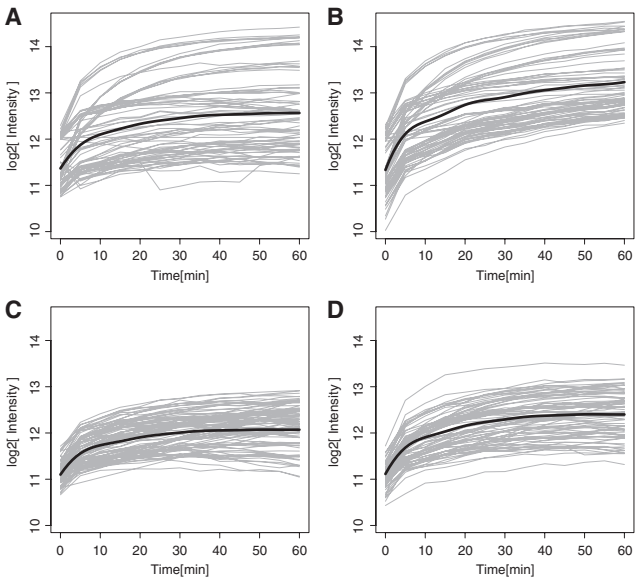

Fig. 2. Fitted smoothing spline for peptide 'TYRO3_679_691_Y686'. Note, that roughness penalty is set to $m=3$. The thick solid curve shows the group-specific mean smoothing spline and the gray lines shows observed $\log_2$ transformed time profiles. The cell line-specific fit and replicate-specific fits are given in Supplementary Figures S8 and S9, respectively. Replicate-specific fits for all the cell lines are shown in Supplementary Figure S10. (**A**) NR–Control; (**B**) R–Control; (**C**) NR–Compound; (**D**) R–Compound.

We used penalized thin plate regression splines with a roughness penalty on the third-order derivative $m=3$ to obtain a smooth first-order derivative.

For this particular peptide, we test whether the group-specific error variances are the same. The null hypotheses H_o: $\sigma^2_{R,\text{Compound}} = \sigma^2_{R,\text{Control}}$ and $H_o: \sigma^2_{NR,\text{Compound}} = \sigma^2_{NR,\text{Control}}$ are tested using a Wald test. We found that treatment and control have the same variance irrespective of their responsive statuses ($p_1 = 0.521$ and $p_2 = 0.189$). Therefore, we refit the model while assuming different error variances for responsive and the non-responsive cell line groups. The null hypothesis $H_o: \sigma^2_{\epsilon NR} = \sigma^2_{\epsilon R}$ is tested against alternative hypothesis $H_a: \sigma^2_{\epsilon NR} \neq \sigma^2_{\epsilon R}$ using a Wald test. We found that responsive and non-responsive cell line groups have significantly different error variances ($p = 0.0001$).

Further, we tried to reduce the proposed model by reducing the cell line-specific as well as the replicate-specific random effects structures. We test the null hypotheses $H_o: \sigma^2_{b2} = 0$ and $H_o: \sigma^2_{c2} = 0$ separately by using approximate likelihood ratio tests. However, testing for the variance components has to be done with care because the parameter of interest is on the boundary of the parameter space (Pinheiro and Bates, 2000; Verbeke and Molenberghs, 2000). To account for the boundary problem under H_0, we can approximate the null distribution of the test statistic with a mixture of χ^2-distributions (Verbeke and Molenberghs, 2000). It is found that both variance components are significant (with P-values $P_1 = 0.0001$ and $P_2 = 0.0001$), implying the need for the random effects b_{2ig} and $c_{2j(ig)}$. Therefore, the proposed model (3) is not reduced. Supplementary Table S1 shows the restricted maximum likelihood estimates for the fixed effects and variance components for the model (3). Figure 2 shows the observed time series profiles and

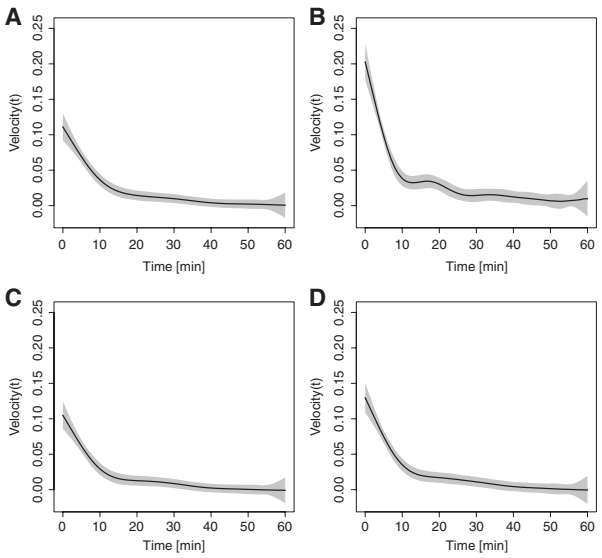

Fig. 3. Estimated velocity and the 95% pointwise confidence bands using the penalized thin plate regression splines for peptide 'TYRO3_679_691_Y686'. Note, that roughness penalty is set to $m=3$. The solid black curve indicates the estimated velocity and the gray color indicates the confidence bands. (**A**) NR–Control; (**B**) R–Control; (**C**) NR–Compound; (**D**) R–Compound.

the estimated smoothers for each group. The solid thick curve indicates the group-specific smoothing curve and gray lines indicate the observed time series profiles for all replicates. Note, that the replicate specific curves suggest a multimodal distribution of the data for the control, which can be modeled by the hierarchical structure of the random effects in Model (3). This is illustrated in Supplementary Figures S8 and S9. There is a huge variation among replicate specific profiles and among cell line-specific profiles. These variations are incorporated in our model by replicate-specific and cell line-specific random effects. Further, the variation between the different 96-well plates is captured with the plate-specific random effects. In Supplementary Figure S10, all fitted replicate specific profiles for all cell lines are displayed along with the data. It can be seen that the model captures the variability between the profiles adequately.

The phosphorylation rate, velocity, in model (3) is available over time and therefore the initial velocity can be extracted very easily by evaluating the first derivative of $\widehat{S(t)}$ at $t=0$. Further, standard errors for the estimated velocity are also available. We construct the confidence bands around the group-specific velocity profiles using the methodology discussed in Wood (2006) and Ruppert *et al.* (2003). The estimated velocity profiles and 95% pointwise confidence bands for each group are shown in Figure 3. These are the first-order derivatives of the fitted group-specific curves. The phosphorylation rate is higher at the beginning of the experiment and reaches a plateau after a few minutes. The estimated group-specific mean curve for the R–Control flattens slightly between time interval 5–25 min (Fig. 2). This corresponds to the shape on the corresponding first derivative in Figure 3. This might seem unusual for individual profiles if phosphorylation follows first-order kinetics. However, the relationship might be obscured, because we measure phosphorylation indirectly by using noisy intensities corresponding to fluorescence of labeled antibodies that bind to the phosphorylated peptides. Moreover, the average profile should not necessarily have the same functional relationship as the individual curves. Although the group specific mean curve flattens slightly between 5 and 25 min, this does not happen for the observed and fitted individual kinase activity profiles: they seem to reflect a realistic physiochemical behavior (see Fig. 2 and Supplementary Figure S10, respectively). However, the individual profiles reveal a huge variation among the different cell lines within this time interval. Therefore, the shape in the group-specific mean velocity profile might be attributed to the averaging process in a region where huge differences occur among the cell lines.

Supplementary Figure S11 illustrates velocity profiles for different groups and the differences in initial velocities between the non-responsive and responsive groups. Later we show that the difference for signature peptide `TYRO3_679_691_Y686` is statistically significant.

5.2 Selecting the signature peptides

This section describes the identification of signature peptides based on both initial velocity and the entire velocity profile using the proposed model. Similar to Section 4.3, our aim is to identify peptides with velocities that are different across the groups. The hypotheses in (4) are tested at the initial time point as well as for the entire velocity profile.

The rate of phosphorylation is expected to be higher at the start of the experiment than at later time points. Thus, the test is first performed using the initial velocities only. The hypothesis given in (4) can be used to test for time $t=0$ with contrast matrix $\boldsymbol{L}=[S_1'(0)\ -S_2'(0)\ S_3'(0)\ -S_4'(0)]$. Second, we update this hypothesis for drawing inferences based on entire velocity profiles (Section 4.3). The updated contrast matrix $\boldsymbol{L}$ has as many rows as the number of basis functions in the smoother described in Section 4.3. The proposed model is fitted for each peptide and the testing procedure is repeated. *P*-values are corrected for multiplicity using the false discovery rate (FDR) procedure of Benjamini and Hochberg described in Benjamini and Hochberg (1995).

Based on the tests using the entire velocity profile, 30 signature peptides are discovered and tabulated in Supplementary Table S2 with the FDR corrected *P*-values. Supplementary Table S3 summarizes the maximum likelihood estimates of initial velocities of significant peptides along with their standard errors for each group. *P*-values for all peptides are summarized in Supplementary Figure S12.

6 DISCUSSION

Protein kinases are key regulators of important cellular processes, including growth, stress response, differentiation and apoptosis. For drug development, signal transduction research and clinical research, in-depth insight in the effects of stimuli on multiple protein kinases is often required in order to fully understand the effects of kinase inhibitors. The novel PamChip® technology enables us to monitor kinase activity in a highly multiplex setting with wide range of peptides. We have proposed a semiparametric mixed model to analyze PamChip® array data. Our aim was to model the kinase activity profiles and estimate the rate of phosphorylation (velocity) over time. The proposed approach is capable of modeling these

activity profiles while incorporating the correlation structure of the data. Moreover, our approach does not require restrictive parametric model assumptions and thus allows greater flexibility by estimating phosphorylation profiles with thin plate regression splines. Hence, the risk of bias is greatly reduced as the estimates depend more on empirical data and less on *a priori* assumptions. Further, estimation and inference is performed in an unified way within the mixed model framework. This approach makes it easy to fit more complex models by including additional fixed effects, random effects or interactions. Moreover, various error structures may be formulated for the errors to account for spatial or longitudinal correlations (Pinheiro and Bates, 2000; Verbeke and Molenberghs, 2000; Wood, 2006).

The main interest of this study was to estimate the phosphorylation rate. Therefore, a smooth first derivative of the fitted phosphorylation profiles is desirable. We have investigated different sets of basis functions for model (3) and thin plate regression splines with a smoothness penalty on the third derivative had the best performance (Supplementary Section 9). Thin plate regression splines have the additional advantage that basis functions, number of knots and position of the knots naturally arise from optimizing the objective function given in (2). The proposed model belongs to the additive mixed model family and can be fitted using standard mixed model software, e.g. with the *gamm* procedure in the R package *mgcv* (Wood, 2006) or with procedure GLIMMIX in SAS.

Our semiparametric model overcomes several drawbacks of the methods (Versele *et al.*, 2009 and Hilhorst *et al.*, 2009). With our model, we found 30 significant peptides using tests based on the entire profiles, and 25 based on the initial velocities, 10 of which were also found with the Versele method. The differences between both methods can be explained as follows. Our single stage approach captures the complex dependence structure of PamChip® data and accounts for heteroscedasticity between responsive and non-responsive cell lines. Therefore, it correctly incorporates the uncertainty in the final statistical inference. The Versele and Hilhorst methods, however, are two-stage approaches that summarize information of the individual time profile into a single summary statistic (initial velocities) and use the initial velocities in the downstream analysis step without accounting for the uncertainty on these estimates. Hence, they ignore the within replicate variability and only capture the technical variation between the replicates. The outcome of the second stage analysis can be severely affected if the within variability of the replicates is large compared with the between variability or if the uncertainty on the initial velocity estimates differs among replicates, cell lines or treatments. In the Versele experiment, the uncertainty on the velocity estimates is bound to be different, because (i) some observations are missing and preprocessing steps are necessary to remove outliers, which results in a different number of observations for the different profiles and (ii) due to the inherent heteroscedasticity between the responsive and non-responsive cell lines. Hence, the model assumptions in the second stage analysis of Versele *et al.* (2009) were violated. Additionally, both Hilhorst and Versele approaches also ignore the correlation between treatments, cell lines and replicates that are assessed on the same 96-well plate, because each phosphorylation profile is preprocessed independently. The random plate effect was also not included in their second stage analysis. For our particular experimental design, the random plate effect does not affect inference on the contrasts between Compound and Control. However, we incorporate the random plate effect in our model for three reasons. First, to ensure that we capture the complex dependence structure of PamChip® data correctly: the data of different cell lines that are placed on the same plate are correlated. Second, to propose a more general model, which provides correct inference for other potential contrast of interest and for unbalanced designs. Third, to accommodate for experiments where the replicates of a specific cell line are placed on different plates. Based on all these arguments, the Hilhorst and the Versele methods are expected to suffer from a loss in power, efficiency and an incorrect control of the type-I error (Verbeke and Molenberghs, 2000).

The exponential model ($Y_{ij} = Y_o + Y_{\max}[1-\exp(-\eta t_j)]$) that is suggested by Hilhorst *et al.* (2009) can be integrated within a single stage non-linear mixed model by specifying random effects on the exponential model parameters:

$$\begin{aligned} Y_{ijlg} = & b_{0ig} + c_{0j(ig)} + a_l + Y_{og} + \\ & (b_{1ig} + c_{1j(ig)} + Y_{\max,g}) \times \\ & \left\{1 - \exp\left[-(\eta_g + b_{2ig} + c_{2j(ig)})t_j\right]\right\} \end{aligned} \tag{5}$$

where b_{0ig}, b_{1ig}, $c_{0j(ig)}$, $c_{1j(ig)}$ and a_l are cell line, replicate and plate specific random effects, respectively. We adopted the model in R and encountered issues related to numerical stability for some peptides. In model (5), we explicitly assume that the individual time profiles follow the Hilhorst model. However, this assumption does not necessarily hold for the average group-specific phosphorylation profiles. Usually there is no closed form solution for the average evolution of non-linear mixed effects models (Fitzmaurice *et al.* 2008; Molenberghs and Verbeke 2005). Hence, it might not be sufficient to draw inference on the parameters $Y_{\max,g}$ and η_g when assessing differences in the marginal group-specific velocity profiles. This was confirmed in our analysis; only one significant peptide could be detected with contrasts on model parameters $Y_{\max,g}$ and η_g (results not shown). Inference at specific time points is even more challenging. The average velocity can be estimated at a particular time point, however, standard errors and estimates of the correlation between the velocities are lacking. In principle, they can be approximated by the delta method or by using bootstrap techniques. Alternatively, one could also perform Monte Carlo Markov Chain (MCMC) simulations to sample from the posterior mean of the velocity profiles. But, both bootstrapping and MCMC will lead to a tremendous increase in numerical complexity. From the discussion above, it is clear that inference on the average group-specific velocity profiles of non-linear mixed models is not straightforward. Another argument against non-linear mixed model (5) is that we feel that the parametric assumptions of the exponential model should be relaxed. The model is motivated by the assumption that peptide phosphorylation follows first-order reaction kinetics. In practice, this might be too restrictive and competing kinetic models might be more appropriate. The reaction kinetics, however, will also be obscured as the phosphorylation is followed indirectly by measuring the fluorescence of labeled antibodies that bind to the phosphorylated peptides. Similar to conventional microarray data, a log transformation of the intensities is beneficial to stabilize the variance, which again will affect the functional form of the profiles that have to be modeled. Our semiparametric method estimates the phosphorylation profiles in a data-driven way by using thin plate regression splines without imposing an *a priori* functional relationship. The thin plate regression splines can be represented as a

linear combination of a set of non-linear basis functions. Therefore, our model is very flexible while remaining linear in the model parameters. This approach is very attractive: it relaxes the model assumptions and simplifies statistical inference as it can build upon the standard theory of linear mixed models.

7 CONCLUSION

We have applied penalized thin plate regression splines within a linear mixed model framework to PamChip® microarray data. We discovered 30 signature peptides that are significantly phosphorylated by the novel multitargeted kinase inhibitor. From the 30 peptides, 10 are also identified by the Versele *et al.* (2009). We demonstrate how one can obtain the group-specific velocities and provide inference on velocity contrasts at individual time points or based on the entire profile. Our flexible modeling approach can easily be adapted to PamChip® array data with various experimental designs.

ACKNOWLEDGEMENT

We would like to thank the three anonymous referees for very useful comments and suggestions in an earlier version of this article.

Funding: This research was supported by IAP research network grantnr. P6/03 of the Belgian government (Belgian Science Policy) and by SymBioSys, the Katholieke Universiteit Leuven, Center of Excellence on Computational Systems Biology, (EF/05/007).

Conflict of Interest: none declared.

REFERENCES

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

Bowman,A.W. and Azzalini,A. (1997) *Applied Smoothing Techniques for Data Analysis*. Oxford Science Publications, New York.

Durbán,M. *et al.* (2004) Simple fitting of subject-specific curves for longitudinal data. *Stat. Med.*, **24**, 1153–1167.

Fitzmaurice,G. *et al.* (eds) (2008) *Longitudinal Data Analysis.* Chapman & Hall/CRC, New York.

Hilhorst,R. *et al.* (2009) Peptide microarrays for detailed, high-throughput substrate identification, kinetic characterization, and inhibition studies on protein kinase A. *Anal. Biochem.*, **387**, 150–161.

Kondapalli,L. *et al.* (2005) The promise of molecular targeted therapies: Protein kinase inhibitors in the treatment of cutaneous malignancies *J. Am. Acad. Dermatol.*, **53**, 291–302.

Maringwa,J. *et al.* (2008) Analysis of cross-over designs with serial correlation within periods using semi-parametric mixed models. *Stat. Med.*, **27**, 6009–6033.

Nagaoka,T. *et al.* (2005) Use of a three-dimensional microarray system for detection of levofloxacin resistance and the mec A gene in Staphylococcus aureus. *J. Clin. Microbiol.*, **43**, 5187–5194.

Molenberghs,G. and Verbeke,G. (2005) Models for discrete longitudinal data. *Springer Series in Statistics*. Springer, New York.

Perera,T. *et al.* (2006) JNJ-26483327 is a novel multi-targeted tyrosine kinase inhibitor with cellular activity against EGFR, Her2, Src and VEGFR3. *EJC*, **4**, 178.

Pinheiro,J.C. and Bates,D.M. (eds) (2000) *Mixed-Effects Models in S and S-PLUS.* Springer Series in Statistics and Computing. Springer, New-York.

Ramsay,J.O. and Silverman,B.W. (eds) (2005) *Functional Data Analysis*, 2nd edn. New York, Springer.

Ruppert,D. *et al.* editors (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Verbeke,G. and Molenberghs,G. (2000) *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer, New York.

Versele,M. *et al.* (2009) Response prediction to a multitargeted kinase inhibitor in cancer cell lines and xenograft tumors using high-content tyrosine peptide arrays with a kinetic readout. *Mol. Cancer Therap.*, **8**, 1846–1855.

Wood,S.N. (2006) *Generalized Additive Models: An Introduction with R* . Chapman & Hall/CRC, New York.