A Bayesian hierarchical model for identifying epitopes in peptide microarray data

SERENA ARIMA*
Dipartimento di metodi e modelli per l’economia, il territorio e la finanza, Sapienza Università di Roma, via del Castro Laurenziano 9, Rome 00161, Italy
serena.arima@uniroma1.it

JING LIN
Division of Pediatric Allergy & Immunology and Jaffe Institute for Food Allergy, The Mount Sinai School of Medicine, New York, NY 10029-6574, USA

VALENTINA PECORA
Dipartimento di Allergologia, Università Cattolica Del Sacro Cuore, Policlinico A. Gemelli, Rome 00168, Italy

LUCA TARDELLA
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome 00185, Italy

SUMMARY
Peptide Microarray Immunoassay (PMI for brevity) is a novel technology that enables researchers to map a large number of proteomic measurements at a peptide level, providing information regarding the relationship between antibody response and clinical sensitivity. PMI studies aim at recognizing antigen-specific antibodies from serum samples and at detecting epitope regions of the protein antigen. PMI data present new challenges for statistical analysis mainly due to the structural dependence among peptides. A PMI is made of a complete library of consecutive peptides. They are synthesized by systematically shifting a window of a fixed number of amino acids through the finite sequence of amino acids of the antigen protein as ordered in the primary structure of the protein. This implies that consecutive peptides have a certain number of amino acids in common and hence are structurally dependent. We propose a new flexible Bayesian hierarchical model framework, which allows one to detect recognized peptides and bound epitope regions in a single framework, taking into account the structural dependence between peptides through a suitable latent Markov structure. The proposed model is illustrated using PMI data from a recent study about egg allergy. A simulation study shows that the proposed model is more powerful and robust in terms of epitope detection than simpler models overlooking some of the dependence structure.

Keywords: AR model; Bayesian inference; Epitope detection; Hidden Markov process; Peptide microarray.

*To whom correspondence should be addressed.
1. Introduction

The advent of microarray technology has enabled biomedical researchers to monitor changes in the expression levels of thousands of genes, and a rich statistical literature is now available (see among the others Parmigiani and others, 2003; Do and others, 2006) and is still under development (Xiao and others, 2009; Lo and Gottardo, 2007). Nowadays, the availability of complete genome sequences of several living organisms with advanced and empirically validated statistical tools has encouraged researchers to shift the attention toward the identification and characterization of all gene products, represented by proteins. Researchers are interested in understanding the biological functions of proteins and characterizing interactions among them. Peptide Microarray Immunoassay (PMI for brevity) is a rapidly emerging tool that provides both largescale and high-throughput capabilities for protein detection and activity studies (Cretich and Chiari, 2009). Once the protein of interest has been identified and partitioned in sequence of amino acids, PMI allows investigators to gain insight into the recognition of specific amino acids subsequences of interest called peptides and to understand the immune system functioning in detail, looking specifically at those epitopes which are recognized in the presence of a pathological status. Epitopes are the part of a protein antigen, that is, recognized by the immune system and are formed by localized subchains of consecutive peptides. A PMI is built using a library of peptides synthesized from the primary sequence of the protein. Each peptide is made of a sequence of amino acids: Consecutive peptides in the primary structure of the protein consist of a sequence of partially overlapping amino acids. Hence, these peptides, sharing a common subsequence of amino acids, are somehow structurally dependent on each other (details in Section 1 of the supplementary material available at Biostatistics online. In order to avoid confounding the dependence along the primary structure of the protein with possible artifacts due to the spatial arrangement of peptides on the array surface, peptides are randomly spotted around the microarray surface and incubated with serum samples. The fluorescence coming from the binding of the serum antibody with the protein is quantified by a scanner, and it is interpreted as a measure of the binding activity determined by the antibody reaction towards the protein. The main aim of the analysis is to identify which epitope regions show evidence of significant binding or differential binding under alternative experimental conditions. These issues are similar to those widely discussed in genomic literature for detecting differentially expressed genes (see for instance Lo and Gottardo, 2007). However, these methods originated within the quantitative genomic literature cannot properly account for the structural dependence among peptides and, as they stand, are not completely suitable for analyzing peptide microarray data. We remark that the dependence we are concerned about is not to be confused with the spatial arrangement of peptides on the array surface. Indeed, there is some recent literature addressing a different issue of spatial dependence among genes on the chromosome or among probes on the array (Xiao and others, 2009). We will see that for our data, dependence due to the spatial arrangement is a minor concern since no significant spatial patterns are detected. On the other hand, the overlapping design of peptide synthesis represents a relevant structural information which can explain some dependence in the binding signals corresponding to peptides which are consecutive in the protein primary structure. Indeed a limited number of approaches are currently available for PMI data (Flinterman and others, 2008; Cerecedo and others, 2008). Although taking into account some peculiarity of PMI data, such as the dependence among peptides, both methods can be considered only a first oversimplified attempt to exploit all the quantitative information of peptide microarray experiments. We propose a flexible Bayesian hierarchical model which allows one to detect recognized peptides and bound epitope regions in a single framework, taking into account the dependence between peptides through a continuous latent autoregressive structure combined with a discrete switching hidden Markov component. The paper is organized as follows: Details on the motivating study and dataset are given in Section 2. Our proposed model for PMI data analysis and decision rules for epitope regions detection are presented in Section 3. Implementation of our model on our data is described in Section 4. Section 5 highlights the advantages of the proposed model with respect to simpler models using both real and simulated data. We conclude with a brief discussion in Section 6.
2. Data description

The “Egg Project” is a recent study on egg allergy coming from a collaboration between the Allergy Department, Catholic University (Rome, Italy), and the Pediatric Allergy Clinics, Mount Sinai School of Medicine (New York, USA). Egg allergy is mediated by IgE and IgG4 antibodies directed toward specific epitopes of some characteristic egg proteins. IgE binding to one or more localized sequential epitopes produces allergic reactions (Patriarca and others, 2009). The Egg Project was aimed at identifying ovomucoid- and ovalbumin-specific IgE and IgG4 epitopes recognized by patients with persistent allergy and studying the immunological changes that occurred after a new oral desensitization treatment. The study enrolled 20 patients, 5 negative nonallergic controls and 15 allergic patients. The allergic patients were exposed to a desensitization treatment, and 11 of them successfully reintroduced the assumption of eggs into their regular diet. In this paper, we limit ourselves to analyze ovalbumin IgE and IgG4 data corresponding to 11 allergic patients who successfully reintroduced the assumption of eggs in their regular diet. PMI that has been used for ovalbumin antigen in this study has been manufactured with a library of 125 peptides. The primary structure of the protein naturally induces a sequence of consecutive peptides formed, in our case, by 15 amino acids overlapping by 12 (Figure 1 in the supplementary material available at Biostatistics online). Each peptide has been spotted 6 times, and triplets are randomly arranged on the array surface. For each array, the fluorescence signals have then been collected in a $125 \times 6$ data matrix whose rows are ordered according to the primary structure. In the following, we refer to consecutive peptides as peptides which are neighboring in the primary structure. A PMI experiment has been performed with serum from each patient before and after the desensitization treatment. Following Nahtman and others (2007), we account for the background noise using the ratio of foreground intensity and local background intensity. The adjusted signal will be hereafter named Signal-To-Noise ratio (SNR) for brevity. Nahtman and others (2007) have shown that such SNR is a reliable normalized measure of peptide binding. Other standard data quality diagnostics for our data are reported in Section 2 of the supplementary material available at Biostatistics online. We have basically verified that within-array variability is reasonably moderate, and there is no significant evidence of spatial patterns over the chip surface. On the other hand, the dependence among consecutive peptides is more than evident. In the first row of Figure 1, one can immediately grasp that the first lags of autocorrelation function of the series of the average peptide SNRs for an allergic patient (right panel) are significantly different from zero with a different pattern from that observed in the negative controls (left panel). A similar pattern is present in all allergic patients as opposed to all negative controls as shown in the second row of Figure 1. In Figure 2, we show SNRs (averaged over the replicates) of the 11 allergic patients together with those relative to 5 negative controls used just as benchmark. Only SNRs of the 11 allergic patients will be used as input of our statistical model.

The main goals of the study are identifying in allergic patients which are the recognized peptides and the corresponding epitope regions and comparing epitope regions under different experimental conditions, namely before and after the treatment. Peptides belonging to epitopes are characterized by local intensification (peaks) of their SNRs: The more the peptide is recognized by the immune response, that is, IgE antibodies, the greater is the binding and the corresponding SNR peak.

A limited number of approaches are available for analyzing PMI data. Flinterman and others (2008) have proposed a descriptive approach that aims at identifying which peptides belong to an epitope. In order to take into account the overlapping nature of the data, they propose to normalize the SNR by considering a trimmed mean of the previous 2 and subsequent 2 peptides. A peptide is classified as belonging to an epitope if its normalized SNR exceeds a fixed cutoff value. This cutoff value is computed as the mean SNR of control spots containing only printing buffer. The analysis in Cerecedo and others (2008) aims at comparing, through a formal statistical test, the SNRs of 2 different groups of patients, reactive and tolerant: a region is defined as an epitope if it is statistically associated with reactive groups and recognized by at least 75% of reactive patients. Statistical significance is assessed by computing an F-test between
Fig. 1. Autocorrelation functions (ACFs) of the series of consecutive peptide average SNRs for negative controls and allergic patients. The first row shows the ACFs of a single randomly selected allergic patient (right panel) and for a single negative control (left panel). In the second row, boxplots of the first 5 autocorrelations of all negative controls (left panel) and all allergic patients are shown. These graphs clearly suggest to avoid the assumption of independence of SNRs over consecutive peptides, possibly due to the overlapping of the sequences.

Fig. 2. IgE SNRs of 125 ovalbumin peptides of 16 patients averaged over 6 technical replicates. Different lines label the 5 negative controls and the 11 allergic patients who successfully reintroduced the egg in their diet (successes). Notice that consecutive peptides are ordered according to the primary structure (e.g. OVA-01 and OVA-02 are consecutive in the primary structure and share 12 amino acids).
the 2 groups. Lin and others (2009) proposed a standardized Z-score computed with log₂ transformation of the normalized data. We believe that although taking into account the peculiarity of the data, these methods can be considered only a first simplified attempt to exploit all the quantitative information of peptide microarray experiments and can be improved on with more appropriate inferential tools derived from a comprehensive statistical model.

Since Bayesian hierarchical models have become increasingly popular in the analysis of gene expression data (Newton and others, 2004; Lewin and others, 2006) and also in the analysis of more complex ChIP-chip data (Gottardo and others, 2008), we believe that they can make the best of available prior information while borrowing strength from the data when estimating the quantities of interest.

3. BAYESIAN HIERARCHICAL MODEL

A model-based analysis of peptide microarray data is to be preferred in order to account for several sources of variability, such as technical replicates, patient-to-patient variability, as well as for the peculiar structural dependence among consecutive peptides. We now introduce the specific features of the proposed Bayesian hierarchical model here presented for simplicity for one experimental condition. Let \( y_{prc} \) be SNR of peptide (spot) \( p \) (\( p = 1, \ldots, P \)), replicate \( r \) (\( r = 1, \ldots, R \)), and patient \( c \) (\( c = 1, \ldots, C \)). We stress once again that \( y_{prc} \) and \( y_{p+1rc} \) are signals corresponding to 2 peptides that are consecutive along the primary protein sequence and share 12 amino acids. The following 3 equations define the first level of hierarchy of the proposed model:

\[
\begin{align*}
    y_{prc} & \sim N(\mu_p + \beta_{pc}, \sigma^2_{pc}), \\
    \mu_p & = a_0 + \sum_{k=1}^{K} a_k (\mu_{p-k} - a_0) + \gamma_p \delta_p + \varepsilon, \\
    \beta_{pc} & = a_0c + a_1c \mu_p + a_2c \mu^2_p,
\end{align*}
\]

where \( \varepsilon \sim N(0, \tau^2) \);

with the constraint \( \sum_{c=1}^{C} \beta_{pc} = 0 \) for identifiability reasons. Equation (3.1) states that for a single patient \( c \), SNR expectation is modeled as a sum of 2 components: \( \mu_p \) and \( \beta_{pc} \). The first component \( \mu_p \) defines the overall peptide mean effect, while the second component \( \beta_{pc} \) can account for patient-to-patient variability.

The overall mean effect \( \mu_p \) is in turn modeled in (3.2) as a sum of 2 components: a latent \( K \)th order autoregressive Gaussian process and a switching factor \( \delta_p \) multiplied by a positive peak size \( \gamma_p \). The autoregressive component accounts for the interior spot dependence mainly due to the overlapping nature of the signal. The inclusion of this autoregressive component is strongly suggested by the autocorrelations of available data displayed in Figure 1. For our Egg Project data, after comparing different autoregressive orders, \( K = 4 \) has been chosen. Indeed, it sounds also \textit{a priori} a reasonable choice in light of the overlapping pattern. In fact, each peptide is made of 15 amino acids, 12 of which overlapping its neighbors so that 4 is actually the minimum lag which makes consecutive peptides nonoverlapping. We assume that \( \delta_p \) is binary, \( \delta_p \in \{0, 1\} \), and

\[
    \mathbb{P}(\delta_p = j|\delta_{p-1} = i) = p_{ij},
\]

with \( j \in \{0, 1\} \), that is, \( \delta_p \) is a two-state homogeneous Markov process, known in econometric literature as switching factor (McCulloch and Tsay, 1990). The latent binary switch \( \delta_p \) is interpreted as a “peak indicator”: when \( \delta_p = 1 \), a peptide \( p \) is classified as belonging to an epitope region. The parameter \( \gamma_p \) specifies the “peak size”: the larger \( \gamma_p \), the higher the peak corresponding to peptide \( p \). Notice that we assume that the process is Markovian in order to account for the dependence between peptides belonging
to the same epitope region. In fact, since epitope regions consist of neighboring peptides, the probability that a peptide belongs to an epitope region is influenced by the probability of its neighbors.

For the peak size $γ_p$, a 2 component mixture of exponential distributions is assumed, that is, $γ_p \sim w\text{Exp}(λ_1) + (1-w)\text{Exp}(λ_2)$. This choice has been motivated by the heterogeneous nature of the peak size, and pilot runs of the model implementation have confirmed that it allows one to take into account more flexibly for the asymmetry in the peak size distribution. In turn, $λ_1$ and $λ_2$ are assumed to come from 2 flat lognormal distributions. The weight of the mixture component $w$ is modeled through a uniform prior. Notice that with the mixture exponential peak size, we can capture the typical pattern of binding variability as a monotone increasing function of the average binding, as also typically experienced in gene expression microarray data.

The main parameters of interest are $μ_p$, $δ_p$, and $γ_p$. Somehow they summarize the information coming from the main innovative aspects of our model: The latent autoregressive average peptide binding $μ_p$ accounts for the structural dependence, the Markovian structure of the latent binary switch $δ_p$ identifies epitope regions, and $γ_p$ scales for the different degrees of binding. We will see in Section 5 that both Markovian and autoregressive dependence are useful and relevant in order to suitably analyze our data.

The second component $β_{pc}$ in (3.3) is the “patient/array effect”: It can account for variability among patients (arrays) within the same experimental condition. Notice in fact that in our data, we have a single array for each patient. Hence, the patient effect and the array effect cannot be separated out. However, when more than one array is replicated for each patient, a simple extension can be easily accommodated in (3.3) so that the 2 components can be distinguished. We model $β_{pc}$ as a function of the overall mean effect as suggested in Lewin and others (2006). We use a second-order polynomial function of the overall mean effect. The random intercept $a_{0c}$ defines the constant patient/array-specific effect, while $a_{1c}$ and $a_{2c}$ characterize the patient/array effect as nonlinear function of the overall mean $μ_p$. The constraint $Σ_{c=1}^C β_{pc} = 0$ ensures the identifiability of the model parameters. More complex functions can be indeed considered. We prefer a polynomial function to keep it simple, while the order of the polynomial has been empirically validated (see Section 5.1). Equations (3.1), (3.2), and (3.3) define the first level of the hierarchical model. At the second level, information is shared between peptides to stabilize the variances. The variances $σ^2_{pc}$ are modeled as exchangeable within each patient. They are assumed to come from a common distribution, chosen here to be lognormal: $σ^2_{pc} \sim \text{logNorm}(ν_c, η^2_c)$. For the hyperparameters, we borrow strength across patients considering $ν_c \sim N(b_1, b_2)$ and $η^2_c \sim IG(b_3, b_4)$, where $b_1, b_2, b_3$, and $b_4$ are common to all patients.

At the upper level of the model, we need to specify prior distributions for all the remaining parameters. Our choices are generally intended to be noninformative. In particular, transition probability $p_{ij}$ is assumed to come from a Dirichlet distribution $p_{ij} \sim \text{Dirichlet}(1, 1)$. With respect to the autoregressive component, in the Bayesian framework, the estimation of the autoregressive parameters is not limited to the stationarity region: The debate about the prior distributions of the autoregressive parameters for non-stationary process is still open and a universally preferred solution has not yet agreed upon. However, as stated in Congdon (2001), if a stationarity constraint is not imposed, then a flat prior may be chosen for the $a_1, \ldots, a_K (k = 1, \ldots, 4)$ but also proper priors provide a relatively vague alternative. A common choice is $a_k \sim N(0, 1)$: Eliciting priors with larger variances would neglect the typical pattern of autoregressive coefficients on the endogenous variable, with values exceeding one being uncommon, except in short-term explosive series. For the Egg Project data, it turns out that there is some evidence of stationarity since the posterior distribution of the autoregressive parameters is well concentrated within the stationarity region (see Section 4 of the supplementary material available at Biostatistics online. Flat normal prior has been considered for $a_0, a_1, a_2$, and $b_1$. Flat lognormal prior distributions have been used for $b_2, b_3, b_4$, and $τ^2$. One can easily understand that the proposed model framework is amenable to further generalizations. For our data, we have slightly extended the model formulation to take into account the presence of 2 different experimental conditions (before and after the desensitization treatment). In that case, one can add an extra
index \( s = 1, 2 \) labeling the experimental condition and considering the main parameters of interest distinct for each condition \((\mu_{ps}, \delta_{ps}, \gamma_{ps})\). Other interesting extensions of the model framework illustrated in this section will be sketched in the final remarks.

3.1 Parameter estimation and implementation

We have considered a fully Bayesian approach for estimating model parameters. For actual implementation, we use WinBUGS software (Spiegelhater and others, 1999) which performs Monte Carlo Markov chain simulations approximately from the posterior distribution. We allow 10,000 iterations for the sampler to converge with burn in of 5000. Posterior samples are thinned by 10 for estimating the posterior distribution of quantities of interest. For the full model processing, 10,000 iterations took 3 h on a dual processor 2.4 GHz machine running version 1.4 of WinBUGS under Windows. Simulation studies show that while the computational time scales linearly with the number of peptides, it scales quadratically with the number of patients involved. When the number of peptides does not exceed 300, and the patients recruited do not exceed 50, WinBUGS implementation can be an effective tool. However, when more patients are involved, it is certainly necessary to convert it in some more efficient language. WinBUGS code and additional computational details regarding processing time, prior distributions, constraints formulation, and convergence diagnostics are available in Sections 3 and 4 of the supplementary material available at Biostatistics online.

3.2 Identification of epitope regions

Once simulations from the joint posterior distribution of all parameters have been obtained, we are ready to pursue our inferential goals. We exploit the joint posterior distribution of the parameters of interest to answer biological questions with formal inferential tools based on decision-theoretic criteria. In particular, in order to select which peptides belong to an epitope region, we focus on the binary switch \( \delta_p \) and use its posterior expectation \( \hat{\delta}_p = P(\delta_p = 1|\text{data}) \) estimated by the simulated chain as \( \hat{\delta}_p \equiv \frac{1}{T} \sum_{t=1}^{T} I(\delta_p^{(t)}) = 1 \) defining the following decision rule and the corresponding epitope region: We say that a peptide \( p \) belongs to an epitope region if \( \hat{\delta}_p > \delta_{\text{cut}} \) so that the collection of all epitope regions can be denoted as follows: \( S_{\delta_{\text{cut}}} = \{: p: \hat{\delta}_p > \delta_{\text{cut}} \} \). To calibrate an appropriate cutoff value \( \delta_{\text{cut}} \), we account for the presence of multiple decisions concerning all epitope regions and estimate the false discovery rate (FDR) as in Newton and others (2004), by

\[
\text{FDR}(\delta_{\text{cut}}) = \frac{1}{|S_{\delta_{\text{cut}}}^\delta|} \sum_{p \in S_{\delta_{\text{cut}}}^\delta} (1 - \hat{\delta}_p),
\]  

(3.4)

where \(|S_{\delta_{\text{cut}}}^\delta|\) denotes the cardinality of the epitope region.

Indeed for the Egg Project, data allergists are particularly interested in the differential IgE binding pattern corresponding to 2 different experimental conditions namely, before \((s = 1)\) and after \((s = 2)\) the desensitization treatment. In that case, we have considered a slightly more general model than that detailed in Section 3 with an additional index \( s \) labeling the experimental condition. To investigate the differential behavior, one needs to restrict the attention to those peptides which belong to an epitope region in either one of the 2 conditions and look at the following posterior probability

\[
\hat{\mu}_p = P(\mu_{p1} > \mu_{p2}|\{\hat{\delta}_p > p_{\text{cut}}\} \cup \{\hat{\delta}_p > p_{\text{cut}}\}, \text{data}),
\]  

(3.5)

where \( \mu_{ps} \) and \( \hat{\delta}_p \) are, respectively, the overall mean expression of peptide \( p \) and the estimated probability of its peak indicator in the experimental condition \( s \) \((s = 1, 2)\). Moreover, a peptide \( p \) is considered as differentially bound if \( p_{\text{cut}} > \hat{\mu}_p \). Cutoff values \( \mu_{\text{cut}} \), similarly to \( \delta_{\text{cut}} \), are calibrated controlling the FDR as in (3.4).
4. Empirical findings

Here, we illustrate the usefulness of the inferential output of the proposed model applied to the Egg Project data described in Section 2 and discuss the main biological findings resulting from the defined decision rules. To select an appropriate cutoff value \( p_{\text{cut}} \), we estimate the FDR as in (3.4). An FDR of 10% corresponds to a probability cutoff of 0.82. The number of peptides obtained using this cutoff is 20 and 5 for IgE bindings, respectively, before and after the desensitization treatment, while for IgG4, 15 and 24 peptides have been recognized, respectively, before and after the desensitization treatment. From a biological standpoint, the allergists had a chance to analyze the affinity of antigen-specific IgE and the competition between IgE and IgG4 for the egg protein (ovalbumin) in egg allergic patients during the desensitization treatment (see Figure 3). For IgE, the successful treatment yields a remarkable reduction in the number of recognized peptides from the previous 20 to 5 which are located in 2 of the epitope regions already recognized. This reduction is expected as a consequence of the reduced reaction of the immune system, and allergists and biologists are particularly interested in localizing, characterizing, and

Fig. 3. Epitopes 59–65: IgE and IgG4 model-based recognized peptides before and after the desensitization treatment and peptide differential binding. In the upper portion of the diagram, downward triangles highlight IgE peptides which are found to be recognized before and after the treatment. The treatment yields a remarkable reduction of the number of recognized peptides from 6 to 2. In the lower band of the diagram, upward triangles highlight IgG4 peptides which are recognized before and after the treatment. The treatment yields an increase in the number of peptides recognized from 3 to 5. In the middle, posterior distributions of peptide differential binding before versus after the treatment are displayed for IgE (boxplots over the null line) and IgG4 (boxplots under the null line). The boxplots highlight a consistent decrease of the average IgE binding after the treatment. An opposite trend is detected for IgG4 for which the treatment increases the average differential binding so that boxplots lie under the null line.
understanding the differential epitopes. For those peptides, a reduced average fluorescence signal is also statistically validated by the model. On the other hand, for IgG4, successful treatment resulted in an increase of the number of identified peptides from 15 before the treatment to 24 after the treatment and in a significant increase average fluorescence signal. The different trends of the specific IgE and IgG4 binding patterns observed in the 2 groups confirmed some of the biological hypotheses of the medical doctors involved in the project and gave insights on the current understanding. Figure 3 shows one of the epitope regions identified by applying the model separately to IgE and IgG4 before \((s = 1)\) and after \((s = 2)\) the treatment for the 11 allergic patients.

5. MODEL DISCUSSION AND DIAGNOSTICS

We now go back to the proposed model to highlight and evaluate empirically the relevance of the main innovative aspects introduced in our model formulation: the latent hidden Markov process \(\delta_p\) for identifying peptides belonging to epitope regions and the latent autoregressive component \(\mu_p\). Indeed, there are some connection between our innovative features and recent literature on similar contexts. The first aspect is not completely new in proteomics literature. A similar idea has been already proposed by Gottardo and others (2008) for modeling IP enrichment effect in ChIP-chip data: The authors proposed a Bayesian hierarchical model in order to account for the probe dependence due to the proximity of the probes on the genome. However in that case, the latent autoregressive component, accounting for bivariate spatial dependence, was regulating the mixing probability of a 2 component peak size. Differently from Gottardo and others (2008), our autoregressive component is present in the whole underlying expected SNRs. The introduction of latent autoregressive component for modeling the relationships among consecutive peptides is a novel aspect which, to our knowledge, has not been used so far in proteomics literature. As we will show, it plays a crucial inferential role in detecting epitope regions in our peptide microarray data. The introduction of these innovative aspects in the model is evaluated considering alternative models and model comparison tools. The impact of these innovative features is also discussed from a predictive point of view with our real data (Section 5.1) as well as with simulated data (Section 5.2).

5.1 Model diagnostics

In this section, we focus on discussing some model features with respect to our real data such as the overall peptide mean \(\mu_p\) and the array effect \(\beta_{pc}\). With respect to the overall peptide mean, we compare the proposed model, labeled \(M^*\) with 2 alternative models \(M_1\) and \(M_2\) which differ from the proposed model in the following aspects:

- \(M_1\): \(\mu_p \sim N(v_0 + \delta_p \gamma_p, \tau)\) with no latent autoregressive component, while \(\delta_p\) is a Markov process as in the proposed model;
- \(M_2\): \(\mu_p \sim N(v_0 + \delta_p \gamma_p, \tau)\) with no latent autoregressive component and \(\delta_p \sim \text{Bernoulli}(\theta)\).

In other words, \(M_1\) differs from our \(M^*\) in the mean modeling where the autoregressive parameters are set equal to zero, while the binary switch is modeled as in \(M^*\). On the other hand, \(M_2\) differs from the proposed model for both aspects: The \(\mu_p\)'s are modeled as independent Gaussian variables and \(\delta_p\)'s are modeled as independent Bernoulli variates.

The Bayesian setting allows us to discuss various aspects of the model from a predictive point of view in the same spirit of Meng (1994). Replicated data \(y_{REP}\) are generated from the posterior predictive distribution and compared with the observed data in terms of some interesting summary of the data. More specifically, in order to understand to what extent the proposed models can properly reproduce and account for the observed dependence structure, we consider as summary statistics the autocorrelation

\[
\hat{\rho}_{kc} = \frac{\sum_{p=1}^{P-k} (\tilde{y}_{p+kc} - \bar{y}_c)(\tilde{y}_{pc} - \bar{y}_c)}{\sum_{p=1}^{P} (\tilde{y}_{pc} - \bar{y}_c)^2},
\]
Table 1. Overall fit indexes for models $M^*$, $M_1$, and $M_2$. $\Delta$AIC and $\Delta$BIC are the differences of AIC and BIC of models $M_1$ and $M_2$ with respect to the proposed model. The larger the differences, the more $M^*$ is favored.

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
<th>$\Delta$AIC</th>
<th>$\Delta$BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M^*$</td>
<td>-5265.59</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>$M_1$</td>
<td>-5038.77</td>
<td>159.69</td>
<td>148.66</td>
</tr>
<tr>
<td>$M_2$</td>
<td>-5008.05</td>
<td>202.72</td>
<td>185.75</td>
</tr>
</tbody>
</table>

AIC, Akaike information criterion; BIC, Bayesian information criterion; DIC, deviance information criteria.

where $\bar{y}_{pc}^{\text{REP}} = \frac{1}{R} \sum_{r=1}^{R} y_{pc}^{\text{REP}}$ and $\bar{y}_c^{\text{REP}} = \frac{1}{P} \sum_{p=1}^{P} \bar{y}_{pc}^{\text{REP}}$ are computed for each replicated vector $y_{pc}^{\text{REP}}$ simulated from the predictive posterior distribution.

For each patient $c$, the observed autocorrelations $\hat{\rho}_{ck}$ and the distribution of the predicted autocorrelation $\hat{\rho}_{kc}^{\text{REP}}$ have been compared. Figure 12 in supplementary material (available at Biostatistics online) compares the posterior distributions of the predicted autocorrelations (at the first 2 lags) for the 3 models for an allergic patient. For the proposed model, they are roughly peaked around their observed counterpart. This means that the model is well catching the correlation structure of the data induced by the primary structure. On the other hand, the other 2 models $M_1$ and $M_2$ tend to systematically underestimate the observed autocorrelation.

In order to assess the overall fit of the proposed model, we consider 3 widely used model comparison indexes: the Akaike index, the Bayesian information criteria, and the deviance information criteria (Spiegelhater and others, 2002). Table 1 summarizes the differences in terms of overall fit indexes for the proposed model $M^*$, the model $M_1$, and the simplest model $M_2$: Although the proposed model involves a larger number of parameters, all information criteria agree on the fact that our $M^*$ fits the data better than the other 2 models.

The array effect $\beta_{pc}$, accounting for the patient-to-patient variability, is defined as a second-order polynomial function of the overall mean effect. This choice is justified using several diagnostic tools. We have compared the posterior distribution of the array effect when it is specified as different functions of the mean effect. For our data, the quadratic effect results more appropriate than the linear and the cubic effect. A detailed discussion about the array effect choice can be found in Section 5 of the supplementary material available at Biostatistics online.

5.2 Comparative performance with simulated data

We use a small simulation plan aimed at understanding which qualitative and quantitative features of the real data can be lost in the posterior analysis, once either one of the components is removed from the model. One could have hoped that only one autoregressive component would suffice to grasp the main dependence pattern. We simulated 100 data sets from the proposed model $M^*$ with 125 peptides and 6 replicates similar to our data. The proposed model and the alternative models $M_1$ and $M_2$ have been compared in terms of FDR, false nondiscovery rate, and percentage of true epitopes discovered. Simulation results are summarized graphically as in Lo and Gottardo (2007) (see Figures 15 and 16 in the supplementary material available at Biostatistics online). Detailed results about the simulation study can be found in Section 6 of the supplementary material available at Biostatistics online. This small simulation plan was useful to qualify the impact on the final inference of each autoregressive component. The binary Markov switch in model $M_1$ seems to be essential to capture some of the dependence in such a way that it correctly singles out part of the actual epitope regions. However, it usually enlarges the identified regions with some extra irrelevant peptides hence loosing control of the FDR (see Figure 14 in the supplementary material
available at *Biostatistics* online). The final addition of the autoregressive component in $M^*$ helps to control the extension of the epitope regions and captures some more regions of consecutive peptides which are not identified by $M_1$. Overall, $M^*$ proved itself robust and effective with respect to the other 2 models.

6. **DISCUSSION**

This paper is one of the first attempts to provide a coherent statistical framework for the analysis of peptide immunoassay data. To our knowledge, the spot dependence structure in peptide microarray experiments has not been formally addressed previously in the literature. PMI experiments allow to investigate the immune system behavior at a fine molecular scale looking at those peptides which are involved in the anticorpal response to a protein of interest in a pathological condition. The main aim of peptide microarray analysis is to identify which chains of consecutive peptides, called epitope regions, show evidence of significant binding or differential binding under alternative experimental conditions. Peptide microarray data call for suitable statistical methodology properly accounting for the peculiar structural dependence of SNR. Our Bayesian hierarchical approach accounts for the peculiar features of PMI experiments in a comprehensive flexible statistical framework which allows one to detect recognized peptides and bound epitope regions. The dependence between peptides is modeled through an autoregressive latent continuous structure combined with a discrete switching hidden Markov component. The autoregressive component models the overlapping nature of the signal and accounts for the structural dependence of neighboring peptides. On the other hand, the switch component models the possible addition of a binding peak which is present only if the peptide belongs to an epitope region. A hidden binary Markov process regulates the switch effect accounting for the fact that neighboring peptides are more likely to belong to the same epitope region. Both aspects—singly and jointly considered—resulted to play a key role in terms of overall model fit and peak detection (Section 5). With the new proposed model, we have analyzed PMI data coming from the Egg Project, a study aimed to evaluate the modifications of the immune responses against ovalbumin in patients affected by IgE- and IgG4-mediated allergy and undergone a desensitization treatment. Using decision criteria based on the posterior distribution of the parameter of interest, we have detected relevant peptides belonging to epitope regions and differential binding regions under the 2 experimental conditions. Our model has been successfully validated from a predictive point of view and compared with simpler models. Moreover, a simulation study showed that our model is more robust in terms of peak detection than other simpler competing models which overlook, completely or in part, the modeled dependence structure. Some more extensions can be embedded in the proposed model framework. In particular, when for the same patient multiple arrays are available, the model can be easily adjusted so that the array effect and the patient effect can be estimated separately. Spatial artifacts can be also embedded in the array effect component using coordinates of peptides on the array as additional covariates in the functional relation. Another possible extension which can be of interest in order to investigate possible interactions of the 2 antibodies and to better account for their competitive behavior is the simultaneous bivariate modeling of IgE and IgG4.

**SUPPLEMENTARY MATERIAL**

Supplementary material is available at http://www.biostatistics.oxfordjournals.org.

**ACKNOWLEDGMENTS**

The authors thank the associated editor and 2 anonymous referees for helpful comments which contributed to improve the article. The authors thank Prof. Giampiero Patriarca and Prof. Hugh A. Sampson, who supervised the Egg Project overall, for kindly providing the microarray data and Alessio Farcomeni and Antonella Cuteri for helpful discussions. **Conflict of Interest:** None declared.
References


[Received August 8, 2010; revised June 30, 2011; accepted for publication July 10, 2011]