



Published in final edited form as:

IEEE Signal Process Mag. 2012 January 1; 29(1): 108–123. doi:10.1109/MSP.2011.943009.

High-Dimensional Longitudinal Genomic Data:

An analysis used for monitoring viral infections

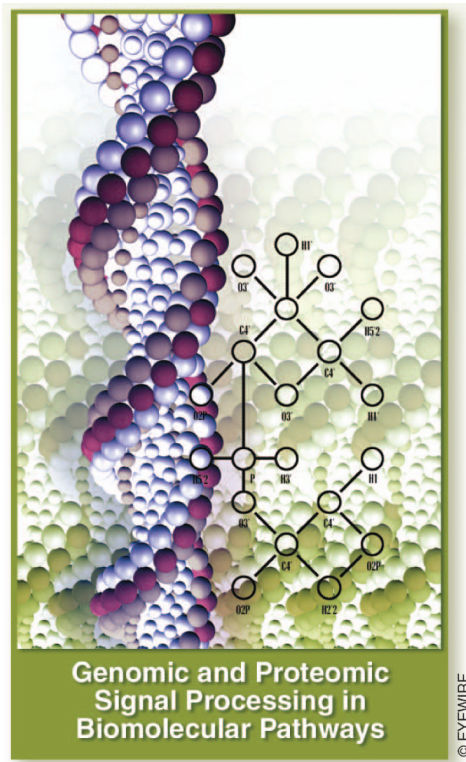
Lawrence Carin, Alfred Hero III, Joseph Lucas, David Dunson, Minhua Chen, Ricardo Heñao, Arnau Tibau-Puig, Aimee Zaas, Christopher W. Woods, and Geoffrey S. Ginsburg

In this article, we review recent progress in using high-dimensional longitudinal genomic data collected from virus challenge studies, performed with healthy human volunteers. The focus of the article is on statistical signal processing, but we also show how the results may be used to yield biological insights.

INTRODUCTION

We consider the analysis of time-evolving gene-expression and proteomic data. The proposed models explicitly address issues associated with inferring the time shift between biological times and “wall-clock” time, inferring the subject-dependent character of the former. We employ factor-analysis and related dictionary-learning-based approaches. The use of such methods obviates the need for explicit clustering [1] of genes.

The analysis techniques reviewed here are motivated by and demonstrated with a novel data set we have measured in recent challenge studies. Specifically, after receiving appropriate Institutional Review Board approval, we performed separate challenge studies in which human volunteers were inoculated with two strains of influenza (H3N2 and H1N1), human rhino virus (HRV), and respiratory syncytial virus (RSV). For each such challenge study, roughly 20 healthy individuals were inoculated with a particular influenza virus and blood samples were collected at regular time intervals until the individuals were discharged. These data provide a unique opportunity to examine the time-evolving host response to such viruses. The mRNA expression levels in blood were assayed with Affymetrix GeneChip Human Genome U133A 2.0 Arrays to constitute gene-expression values for 12,023 genes. See [5] for more details on the messenger ribonucleic acid (mRNA).



MOTIVATION

There is often interest in predicting an individual's latent health status based on high-dimensional genomic biomarkers that vary over time, for example, gene-expression and proteomic data. Motivated by novel longitudinal gene-expression and proteomic data we have collected in several viral challenge studies, performed with healthy human volunteers, we present signal processing methods for analysis of time-evolving genomic biomarkers. We consider this problem from multiple perspectives related to factor analysis and dictionary learning, and in each the high-dimensional data trajectories are related to a relatively low-dimensional vector of latent factors or dictionary elements. The multiple analyses are employed for cross validation, to assure that the inferred biological processes are meaningful and uncovered via distinct models. The models infer genes and proteins in the viral response pathway as well as variability among individuals in infection times. The inferred low-dimensional space in which the high-dimensional data resides is used to provide biological interpretation of the inferred viral response pathways.

There has been much recent interest in the analysis of dynamic biological processes, particularly with data from DNA gene-expression microarray chips [1], [2]. Appropriately analyzing the trajectories as multivariate functional data is challenging due to the massive dimensionality, few observations in time, low signal-to-noise ratio, and missing data. Ideally, methods would allow building a full joint model that allows each biomarker (e.g., gene or protein) to have its own trajectory, while accommodating dependence in these trajectories across biomarkers within shared pathways and variability across individuals. In such time-dependent modeling, one must often distinguish the observed ("wall clock") time at which a measurement was performed from the (latent) biological-clock time, and the difference between these two must be inferred (since the offset between the two is typically subject dependent) [3], [4].

EXISTING METHODS FOR TIME-COURSE ANALYSIS

There have been numerous previous studies on the analysis of time-course gene-expression data [1], [2] and almost all of these employ a clustering of the genes. To model the continuous time dependence of the gene expression, researchers have employed the Gaussian process [3] as well as spline basis functions [1]. Most of these methods employ mixed-effects models, where the fixed-effects component corresponds to clusters, with the genes clustered among one of C different classes or clusters. The random effect term typically has a continuous time dependence that is a function of the specific gene and subject (inferred, for example, using a spline expansion). One may also employ hierarchical clustering of the genes [6].

Additional examples of such a mixed-effect clustering model applied to time-course gene-expression data include [7] and [8]. While this approach has been applied successfully in many settings, it has limitations that restrict its utility. For example, we are typically interested in over 10,000 genes when performing microarray analysis, and therefore the number of spline-based expansions that must be fit is significant. Additionally, for the application of interest here, we have on the order of 20 different subjects, each manifesting a distinct time-course profile.

The proposed approaches avoid the need to explicitly perform clustering (it is done implicitly within the factor modeling and dictionary learning procedures), and the proposed models infer subject-dependent shifts in the latent biological turn-on time. Typically only a small fraction of the genes contribute to the biology under study, and in the context of the factor analysis these genes are inferred by imposing a sparseness constraint. One need only model the time dependence of the factor scores, rather than separately model the time evolution of each individual gene or protein.

TIME-DEPENDENT FACTOR SCORES

BASIC FACTOR MODEL

Let $\mathbf{X}_i \in \mathbb{R}^{p \times n_i}$ represent observed biomarkers (e.g., gene-expression data) for individual i , considering P markers, collected at T time points (the number of time points could also be subject-dependent); the j th column of \mathbf{X}_i corresponds to the P biomarkers measured at time t_{ij} , for $j \in \{1, \dots, n_i\}$. We assume a total of S individuals/subjects, constituting cumulative data $\{\mathbf{X}_i\}_{i=1, S}$. We consider a factor model with k factors

$$\mathbf{X}_i = \mathbf{L} \mathbf{S}_i + \mathbf{E}_i = \sum_{m=1}^k \mathbf{L}_m \mathbf{S}_{mi}^T + \mathbf{E}_i; \quad i=1, 2, \dots, S, \quad (1)$$

where $\mathbf{L} \in \mathbb{R}^{p \times k}$ is the factor loading matrix and \mathbf{L}_m is the m th column of \mathbf{L} ; the factor scores for individual i are $\mathbf{S}_i \in \mathbb{R}^{k \times n_i}$, and \mathbf{S}_{mi}^T is a row vector (m th row of \mathbf{S}_i) of time-varying scores for the i th individual and m th latent factor. The factor loadings are assumed fixed in time, while we allow the latent factors, \mathbf{S}_{mi}^T , to vary dynamically. The matrix $\mathbf{E}_i \in \mathbb{R}^{p \times n_i}$ is the additive noise or residual.

SHIFTED SPLINE REPRESENTATION

Recall that individual i has data sampled at n_i time points; let $\mathbf{t}_i = (t_{i1}, t_{i2}, \dots, t_{in_i})$ denote the time points at which data were collected for individual i (in units of minutes/hours, etc.), with respect to a time reference shared by all S individuals. Note that these are observed times, on a universal clock, to be distinguished from the latent biological clock of the system under investigation (in our specific example this corresponds to the host response to a virus),

which is generally individual dependent. The rows of $\mathbf{S}_i(t)$ are a continuous function of time, and the matrix \mathbf{S}_i represents each such row sampled at the T time points represented by t_i .

Recall that $\mathbf{S}_{mi} \in \mathbb{R}^{n_i}$ represents the factor score associated with factor $m \in \{1, \dots, k\}$ for subject $i \in \{1, \dots, I\}$, evaluated at the T discrete time points in t_i (\mathbf{S}_{mi} is a column vector, the transpose of \mathbf{S}_{mi}^T above). To model \mathbf{S}_{mi} , let $\mathbf{b}(t) \in \mathbb{R}^q$ represent a column vector, corresponding to evaluating each of q spline functions at any time t over the support of the splines [1], defined here by the time window over which data are collected. The number of splines q and their composition depend upon the specific application. The function $\mathbf{b}(t - \tau) \in \mathbb{R}^q$ corresponds to realigning the spline functions to have the time origin shifted forward by $\tau \in \mathbb{R}$. We allow a time shift τ_{mi} specific to latent factor m and individual i by characterizing the factor score trajectories as

$$\mathbf{S}_{mi} = \mathbf{B}(t_i; \tau_{mi}) \mathbf{w}_m + \varepsilon_{mi}, \quad (2)$$

where $[\mathbf{B}(t_i; \tau_{mi})]^T = [\mathbf{b}(t_{i1} - \tau_{mi}), \dots, \mathbf{b}(t_{in_i} - \tau_{mi})]$ $[\mathbf{B}(t_i; \tau_{mi})]^T$ is the transpose of $\mathbf{B}(t_i; \tau_{mi})$, and $\mathbf{w}_m \in \mathbb{R}^q$ corresponds to the spline coefficients for the m th latent factor. An illustration of this generative process is presented in Figure 1.

TEMPORAL SHIFT AND DISTINGUISHING HOST-RESPONSE FACTORS

In our motivating application, all individuals are inoculated with a virus at the same time. Blood is drawn from all subjects at a specified time prior to inoculation ($t = -5$ h) to constitute a baseline signature, and another (distinct) blood sample is drawn just before inoculation (the latter occurring at what is defined as time $t = 0$ h). The vector t_i is defined such that increasing element index corresponds to increasing time; this vector records the times at which blood samples were collected. Therefore, each individual shares the same first two time points in t_i , and since the time of inoculation is by definition at $t = 0$, the first element in t_i corresponds to negative time.

Our objective is to study the host (body) response to the virus, and therefore the spline-based construction for the time-dependent factors is constituted as in Figure 2. Note that the function $\mathbf{B}(t_i; \tau_{mi} = 0) \mathbf{w}_m$ has a constant form for $t = -5$ h (with value of the constant inferred via the analysis), this representing the background/baseline (preinoculation) factor score for a (presumably) healthy individual. Consequently, with application to our challenge studies, the shift τ_{mi} may be viewed as the delay between inoculation of subject i and the time at which factor m changes from its background (“normal”) value; i.e., this is the host response time for pathway m , which is expected to vary between subjects.

Considering Figure 2, note that large shifts τ_{mi} imply that individual i has a near constant host response for factor m , as function of time (“near,” but not exactly constant because of the addition of the ε_{mi}). This model is consistent with our influenza challenge study data, as approximately half of the individuals did not become symptomatic, and for these all of the associated factor scores manifested very weak temporal changes. Therefore, the presence of large τ_{mi} for all factors $m \in \{1, \dots, k\}$ implies that individual i is asymptomatic. Further, if a particular factor $m \in \{1, \dots, k\}$ is not related to the host response to the virus for individual i , the associated τ_{mi} will be large, implying that \mathbf{S}_{mi} is nearly time invariant.

Concerning the subject and factor-dependent shift τ_{mi} , we consider a finite set of discretized τ_{mi} , finely sampled in time, and place a Dirichlet prior on the probability that each of these shifts are selected. The model infers an approximate posterior density function on which time shift is most appropriate for each subject and factor.

SPARSE FACTOR LOADINGS

In many biological applications, it is desirable to impose that the factor matrix is sparse [9]. In the case of gene-expression data, the m th factor can be viewed as measuring overall expression of the m th pathway, with the nonzero elements in the m th column of the loadings matrix \mathbf{L}_m corresponding to the genes in that pathway. Biologically, we would expect a small minority of the genes to play a role in any one pathway, implying sparsity. Hence, we model the loading matrix as

$$\mathbf{L} = \mathbf{A} \circ \mathbf{Z}, \quad (3)$$

where \circ represents a pointwise (Hadamard) matrix product between $\mathbf{A} \in \mathbb{R}^{p \times k}$ and $\mathbf{Z} \in \{0, 1\}^{p \times k}$. Binary matrix \mathbf{Z} is designed to be sparse, and therefore the factor loadings, defined by the columns of $\mathbf{A} \circ \mathbf{Z}$, are also sparse.

The binary matrix \mathbf{Z} is constituted via a so-called Indian buffet process (IBP) [10], implemented in practice by setting k large, and allowing the model to infer the number of needed factors. In the context of the IBP, each of the P genes are “customers” in a buffet restaurant, and the m th factor represents the m th dish. If gene g selects dish (factor) m , then $Z_{gm} = 1$, and otherwise $Z_{gm} = 0$.

EXAMPLE RESULTS

Of the k factors, one of them manifested a time trajectory $\mathbf{B}(t_i; \tau_{mi})\mathbf{w}_m$ that was closely aligned with the clinical scores, it is this factor that is examined in further detail, as it is deemed to be associated with the (time-evolving) host response to the virus. Results are shown for the gene g corresponding to RSAD2, for the H3N2 virus. This gene had the strongest contribution to the loading of this factor (largest $Z_{gm}|A_{gm}|$). All computations with this method are performed using a Gibbs sampler.

We compare the individual- and time-dependent factor score of this factor with clinical symptom score provided by medical doctors. The clinical symptom score was recorded twice daily using standardized symptom scoring [11]. The modified Jackson score requires subjects to rank symptoms of upper respiratory infection (stuffy nose, scratchy throat, headache, cough, etc.) on a scale of 0–3 of “no symptoms,” “just noticeable,” “bothersome but can still do activities,” and “bothersome and cannot do daily activities.” For all cohorts, modified Jackson scores were tabulated to determine if subjects became symptomatic from the respiratory viral challenge. A modified Jackson score of ≥ 6 over the quarantine period was the primary indicator of successful viral infection [12] and subjects with such a score were denoted as “symptomatic” the latter individuals are represented with blue points in Figure 3.

In Figure 3, we plot the inferred time-dependent factor score for each of the subjects as well as the clinical symptom scores. Note that the clinical symptom score generally tracks the inferred factor score well for this time-evolving factor. Additionally, for the asymptomatic $x_g^{(m)}(t_{ij}) = A_{gm} [\epsilon_{mi}(t_{ij}) + \sum_{l=1}^q w_{ml} B_l(t_{ij}; \tau_{mi})]$ is almost a constant with time, but it is not zero.

We now examine the inferred mean trajectory $A_{gm} \sum_{l=1}^q w_{ml} B_l(t_{ij}; \tau_{mi})$ of the (typical) individuals who became symptomatic ($Z_{gm} = 1$). In Figure 4 we depict the inferred host response for this factor. Note that this trajectory has a constant value at early time; it is used as a prototype trajectory for both symptomatic and asymptomatic subjects, and the two are distinguished by the manner in which the trajectory evolves with time and the inferred temporal shifts.

ORDER PRESERVING FACTOR ANALYSIS

DICTIONARY LEARNING AND FACTOR ANALYSIS

In addition to Bayesian factor analysis, we have also examined the data using non-Bayesian dictionary learning. The two distinct classes of models have inferred very similar underlying biological processes. The use of independent analyses is deemed important for accurately uncovering new biology based on limited high-dimensional data.

Dictionary learning refers to a class of methods that seek to represent data by sparse combinations of an overcomplete basis set, called a dictionary [13]. Note that here we take a different approach from the factor modeling in the section “Time-Dependent Factor Scores,” with \mathbf{X}_i^T from this section corresponding to \mathbf{Y}_i . Dictionary learning is also called sparse coding and is widely used in neuroscience, speech, audio, and image processing. On the surface, dictionary learning resembles factor analysis in that they both seek a factored representation for the data matrix. For example, when \mathbf{Y}_i is the matrix of subject i with rows corresponding to T time points and columns corresponding to P gene indices dictionary learning seeks a factorization of the form

$$\mathbf{Y}_i = \mathbf{M}\mathbf{A}_i + \epsilon_i, \quad i=1, \dots, S, \quad (4)$$

where S is the number of subjects, the f columns of matrix \mathbf{M} form the universal dictionary of basis elements, and \mathbf{A}_i is a sparse coefficient vector (the code) associated with subject i 's particular linear combination of dictionary elements composing \mathbf{Y}_i . In dictionary learning, as in factor analysis, both the dictionary and the coefficients are learned from the data $\{\mathbf{Y}_i\}$. However, while in standard factor analysis the objective is to find a low rank \mathbf{A}_i , in dictionary learning the objective is to find a sparse matrix \mathbf{A}_i .

For consistency, we will use the standard factor analysis terminology for the dictionary learning model (5): columns of \mathbf{M} and rows of \mathbf{A}_i will be called factor loadings and factor scores, respectively. While this model can be used for a wide range of applications it is not applicable when there are unknown delays among factor loadings shared by different subjects. We describe a variant of the dictionary learning model, called order preserving factor analysis (OPFA), which accounts for temporal misalignment, incorporates smoothness constraints on the factor loadings, and preserves their relative ordering over the subject population.

ORDER PRESERVING FACTOR ANALYSIS

The principle of evolutionary conservation suggests that major gene regulation mechanisms, such as cell growth and death, operate similarly over the human population. According to this principle, all healthy individuals share the same basic mechanisms of immune response. Systems biology models formalize this principle by modeling gene regulation according to a causal cascade of modules or pathways. Under such a model, signals associated with viral sensing and antigen presentation precede signals for the inflammation response to the virus. The order in which these signals occur is important to effective immune response while the precise timing of these signals may be less important. The order may in some cases be known, or hypothesized based on known biology, or it may be unspecified and learned from data [14].

The systems biology viewpoint motivates an order preserving modification of the dictionary learning model (5) that restricts immune-related signaling to occur in a (unknown) temporal precedence order. The modified model accounts for temporal misalignments between signals up to these order restrictions

$$\mathbf{Y}_i = \mathbf{M}(\mathbf{F}, \mathbf{d}^i) \mathbf{A}_i + \epsilon_i, \quad (5)$$

where \mathbf{d}^i is a vector of unknown delays that specifies the delay of each factor, a column of \mathbf{F} . When the factors satisfy a precedence order constraint the vector of delays will lie in a cone shaped region for all subjects i , for example,

$\mathbf{d}^i \in \left\{ \mathbf{d} = [d_1, \dots, d_f] : \bigcap_{j=2}^f \{d_j \geq d_{j-1}\} \right\}$ is the region where each factor precedes the next in the natural index order of the factors. The factors, delays and coefficients are estimated from the data by solving a nonconvex optimization problem of the form

$$\min_{\mathbf{F}, \mathbf{d}^i, \mathbf{A}_i} \sum_{i=1}^S \|\mathbf{Y}_i - \mathbf{M}(\mathbf{F}, \mathbf{d}^i) \mathbf{A}_i\|_F^2 + \lambda P_1(\mathbf{A}_1, \dots, \mathbf{A}_S) + \beta P_2(\mathbf{F}), \quad (6)$$

where minimization is performed over vectors of delays \mathbf{d}^i that lie in the order-preserving set for all subjects and over nonnegative matrices \mathbf{F} , \mathbf{A}_i . The nonnegativity constraint on the factors is natural since gene expression is measured in units of abundance of mRNA. The functions P_1 and P_2 are penalties that induce temporally smooth columns of \mathbf{F} and shared sparsity across rows of \mathbf{A}_i , for each i . For more details on the implementation of the optimization algorithm for solving the order preserving dictionary learning problem (6) the reader is referred to [15].

EXAMPLE RESULTS

Our formulation (6) of OPFA can be interpreted as an extension of sparse factor analysis (SFA) [16], parallel matrix factorization (PARAFAC) [17], and nonnegative matrix factorization (NMF) [18] that accommodates factor misalignment and unknown factor ordering common to all measurements. PARAFAC and NMF generalize PCA to higher dimensions (tensors) and to nonnegative matrices, respectively. These matrix factorization methods are highly sensitive to misalignments of the factors. The order preserving restriction of OPFA over-comes this misalignment sensitivity as illustrated in Figure 5 for a toy example.

Figure 6 shows the result of applying OPFA to real data, in particular the set of nine clinically sick subjects in the H3N2 challenge study described above. OPFA factors were discovered that correspond to three characteristic temporal profiles: suppressed response (factor 1 in red), suppressed response followed by recovery (factor 2 in blue), and enhanced response (factor 3 in green). The genes in these three groups are associated with the Jun N-terminal kinases (JNK) pathway (factor 1), ribosomal protein (RP) expression (factor 2), and interferon inducible (IFN) genes (factor 3). Furthermore, the factor order reconstructed by OPFA indicates that the gene onset times in factor 2 occur earlier than those in factor 3, a finding that is consistent with previous studies of temporal immune host response [19]. Remarkably, OPFA discovered this ordering de novo despite subject misalignments in the gene expression trajectories. Furthermore, even though clinically determined symptom onset times reported in [20] were not used by OPFA, the subject-dependent delays of OPFA factor 3 track these clinical symptom onset times. This provides independent confirmation of the power of the OPFA method.

Figure 7 illustrates the utility of OPFA for improving cluster separation performance for unsupervised clustering. The two scatter plots show the coefficients of each gene over the first two OPFA factors and the two first principal components of the covariance matrix of the misaligned data, $\sum_{i=1}^9 \mathbf{X}_i \mathbf{X}_i^T$. Each gene is color coded according to the clusters found by performing hierarchical clustering on the OPFA-aligned and on the original misaligned data,

respectively. The scatter plots show how the OPFA coefficients show better separation between the up-regulation genes (reds) from the down-regulation genes (dark blues). In the OPFA scatter plot, the red and dark blue groups are separated by genes with weak temporal response (light blue). In contrast, the PCA-based representation of the raw misaligned data does not separate these groups well, reflecting the higher variance in the red and dark blue genes groups due to misalignment. This tightening of the clusters is further illustrated by comparing the far-left time profiles (cluster means after OPFA alignment) to the far-right time profiles (cluster means without first applying OPFA alignment).

METAPROTEIN EXPRESSION MODELING

MASS SPECTROMETRY PROTEOMICS

Unbiased mass spectrometry (MS)-based proteomics has made tremendous progress since initial studies using matrix-assisted laser desorption/ionization-time of flight (MALDI-ToF) machines in the late 1980s. Current machines are now capable of splitting samples according to a number of different features such as pK, hydrophobicity, and ion mobility, and the resolution of measurements of mass-to-charge ratios are now high enough to detect the difference between two polypeptides that are identical except for the inclusion of a single extra neutron. In this section, we present a different extension of the factor model used in [21] that is specific for the analysis of unbiased, label-free MS proteomics data. We incorporate multiple sources of information about correlation in the hierarchical structure of the model, and this leads to significant improvement in posterior estimation.

MS data may be summarized at a number of different levels, and the analysis of that data may be tailored to any of these summarizations. The smallest unit that is measured by liquid chromatography-tandem MS (LC-MS-MS) is a single peak, which is termed a feature, and there are typically on the order of 10^5 such features. This is a single peak in the two-dimensional surface over a plane defined by the retention time (amount of time a polypeptide takes to pass through the liquid chromatography column) and mass-to-charge ratio. The intensity of this feature is defined to be the volume under this peak. Because a certain percentage of carbon in nature has an extra neutron, each polypeptide leads to multiple features. The collection of features from a single polypeptide that differ in mass-to-charge ratio only by an integer number of neutrons is called an isotope group, and the intensity of the isotope group is the sum of the intensities of its associated features—this is the level at which we summarize our data in this article. In addition to differences in mass, polypeptides may accept a variable, integer number of protons during electrospray ionization. Thus there may be multiple isotope groups per peptide. Finally, for a collection of isotope groups that are known to originate from the same protein one might summarize the data at the protein level. We note that, in contrast to gene expression microarray data in which each spot on the array is fully characterized, the chemical species that make up a MS peaks are often unknown.

EXISTING METHODS FOR ANALYSIS

There are a number of different regression models designed for summarization of proteomics data at the protein level. The simplest such procedures involve direct summarization of all features/isotope groups/peptides that are identified for each protein. This may involve averaging or robust summarization based on quantiles [22]. In addition to these algorithms, there are a number of different analysis of variance (ANOVA) approaches that include fixed effects for protein, peptide, and experimental group [23]; include an additional random effect for cases in which subjects are measured in replicate [24]; or add additional interaction effects between treatment and feature [25]. These may assume constant or varying noise levels across isotope groups and have been shown to exhibit better

performance than naive summarization approaches that do not adjust for confounding factors [25]. While all generally acknowledge the existence of incorrect identifications, none of these approaches directly address this problem. In addition, other than our previous work which examines an earlier factor model in greater detail in a different biological context [26], we are unaware of any such techniques that utilize correlation between features/isotope groups/peptides in any way, nor do any of them utilize unidentified features in protein level quantitation. We review in this article a statistical model first described in [27] that allows the direct modeling of correlation structure and its deconvolution into separate protein and pathway effects. We do not examine any improvements that might be made by the inclusion of fixed and random effects associated with treatment group or replicate measurements of sample. However, the model we describe is a regression model, and would, therefore, be amenable to the inclusion of such effects.

FACTOR MODEL AND HIERARCHICAL STRUCTURE

There are two key sources of information we might use to collect isotope groups into coherent subsets—identifications and coexpression. The identifications tell us which isotope groups originate from the same protein, and if we assume that proteomics is actually measuring differential expression of proteins, then all of the isotope groups from the same protein should coexpress. However, identifications are incomplete, and while those that are obtained are reasonably high accuracy, there are still some mistakes. Additionally, there are a number of biological processes that add chemical modifications to specific regions of proteins. These modifications change the relative abundance of the unmodified regions of the proteins, which exerts strong effects on the resulting measurements (thus proteomics is actually measuring something more subtle than just differential expression of proteins).

Aside from biological processes that may lead to differential expression of individual proteins, there is technical variation that will lead to differential measurements of expression across large portions of the data set. We will utilize a factor model to represent the correlation structure present in the data, but we break that model into two parts, each of which will have its own hierarchical structure. We suppose that $\mathbf{X} \in \mathbb{R}^{p \times N}$ is a matrix of intensities, with columns X_i , where P is the number of measured isotope groups and N is the number of samples. We assume

$$\mathbf{X} = \mathbf{MA} + \mathbf{LS} + \mathbf{E}.$$

Note that the basic form of this model is related to the factor model in the section “Basic Factor Model,” but now \mathbf{X} corresponds to proteomic data. Both \mathbf{MA} and \mathbf{LS} describe latent factors, but we have split them because they describe different types of correlation with different hierarchical structures.

We represent technical noise with the \mathbf{MA} factor structure where $\mathbf{M} \in \mathbb{R}^{p \times d}$ is a factor loadings matrix and $\mathbf{A} \in \mathbb{R}^{d \times N}$ is a factor scores matrix. We assume that this noise is ubiquitous throughout all isotope groups and therefore do not impose sparsity, $\mathbf{M}_{i,j} \sim N(0, \tau_0)$. We may optionally include design variables in \mathbf{A} if we want to control for specific known features of data. This may include either known batch effects (although we find that these are captured well by latent factors) or known phenotypes of the samples. Otherwise, we assume latent factors such that $\mathbf{A}_{k,j} \sim N(0, 1)$.

The correlation structure that is present in the data because there are multiple isotope groups are derived from the same protein is modeled by a separate factor structure \mathbf{LS} . As before $\mathbf{L} \in \mathbb{R}^{p \times k}$ is a factor loadings matrix and $\mathbf{S} \in \mathbb{R}^{k \times N}$ is a matrix of factor scores. This structure of \mathbf{L} is expected to be sparse—correlation described by this structure should be

largely restricted to sets of isotope groups from the same proteins. We assume that every isotope group originates from a single protein, and therefore that every row of the loadings matrix \mathbf{L} contains only one nonzero element. Thus we introduce a latent variable z_i which identifies, for isotope group i , the metaprotein factor to which it belongs (i.e., which element of the i th row of \mathbf{L} is nonzero). Thus element $\mathbf{L}_{i,j} = 0$ when $j \neq z_i$ and $\mathbf{L}_{i,z_i} \sim \mathcal{N}(0, \tau_0)$. Our hierarchical prior for z_i is

$$z_i \sim \text{Multinomial}(1, q_i), q_i \sim \text{Dir}(a_i). \quad (7)$$

We utilize an informative choice of Dirichlet distribution parameters a_i in cases where we have prior information telling us in which protein isotope group i originated. Specifically, if isotope group i is from protein k then we assume that $a_i = (a_0, \dots, a_0, a_k, a_0, \dots, a_0)'$ where $a_k \gg a_0$.

In addition to correlation between isotope groups due to originating from the same protein, expression of the proteins themselves is also correlated due to that expression being regulated within the same biological pathways. Because we have information about the relationships between some isotope groups and proteins, we are able to deconvolute these two sources of structure in the data. To capture this “pathway level” correlation between proteins, we impose a binary tree model on the metaproteins. We suppose that each row, \mathbf{S}_k of \mathbf{S} (each metaprotein) identifies an expression pattern that is associated with a leaf in a binary tree. We define $t_{a \rightarrow b}$ to be the “distance” between node a and its child node, b , and \mathbf{w}_a to be an N -dimensional vector describing an expression pattern associated with node (or leaf) a . Then, assuming b is a child of a , we assume

$$\mathbf{w}_b \sim N(\mathbf{w}_a, t_{a \rightarrow b} \mathbf{I}_N).$$

Given any pair of leaves, we a priori assume that the distance between one of those leaves and the first node which is an ancestor of both is exponential with rate parameter 1. This is the Kingman’s coalescent [28]. It describes a uniform distribution on the space of binary trees and provides a proper prior distribution on child-to-parent distances, t .

We introduce a factor for each protein that has more than one identified isotope group in the data set. This model is conjugate, and we utilize Gibbs sampling in a Markov chain Monte Carlo (MCMC) algorithm to obtain posterior distributions for all model parameters. Sampling of coalescence times for the tree model is accomplished via belief propagation [29], [27]. Trees are constructed through the coalescence procedure described in [28] and in [27].

PLASMA PROTEOMICS DURING VIRAL INFECTION

Working with the same set of subjects discussed in the previous sections (inoculated with an H3N2 strain of viral influenza), we obtained LC-MS/MS proteomics data from blood plasma at baseline (time = 0), at the time of maximum symptoms (time = 1) and at times .2 and .8. The resulting MS traces were then used to estimate the concentrations of approximately 40,000 isotope groups in each sample (with $\approx 5\%$ overall missing data). Approximately 10% of these were identified—the amino acid sequence of the peptide was characterized and that sequence was mapped to a known protein.

In our viral infection data, this leads to a very sparse model with 109 latent factors, each nominally representing the expression of a particular protein. Due to uncertainties in identifications as well as biological perturbations of particular sections of the proteins, there are many peptides (approximately half) that do not follow a pattern of expression across the

samples that is consistent with the majority of peptides from that protein (see, for example, Figure 8).

We were able to identify a number of metaproteins that are associated with the disease state, the strongest of which is shown in Figure 9. We note that the majority of isotope groups in this factor are identified as belonging to the protein A2GL and that it is grouped together by the tree model with c-reactive protein and lipopolysaccharide binding protein (Figure 10), both of which are known to react to the presence of infection. It is informative to examine the full collection of isotope groups from A2GL (Figure 8). We note that, while around two thirds of the isotope groups show clear visible coexpression, the remainder show patterns that are not highly correlated. An understanding of this structure in the data provides a number of benefits. First, in cases where it is the bulk of the protein that shows differential expression that correlates with biological phenotypes, as is the case with A2GL and the “symptomatic versus asymptomatic” phenotype, the aggregate expression from the metaprotein model will provide a much stronger predictor than a summarization based on isotope group identifications. Second, if our goal is the development of biosignatures then we must be careful about which peptides, not just which proteins, we will use for that biosignature. Finally, in cases where we are looking for association between protein expression and phenotype data, we will be able to perform many fewer hypothesis tests, and have commensurate higher power, if we can perform those tests on just metaproteins rather than on peptides. This is also true of protein summarization approaches based on identifications as well, however, we find that there are approximately half as many metaproteins versus proteins, that the metaproteins are typically less correlated with each other than are proteins, and that we can include potentially informative but unidentified isotope groups in targeted studies when we select them based on the metaprotein model.

INFERRED BIOLOGY

As summarized in the sections “Time-Dependent Factor Scores” and “Order Preserving Factor Analysis,” two very distinct techniques were employed to analyze the time-course gene-expression data (in the section “Time-Dependent Factor Scores,” a fully Bayesian approach was employed, while in the section “Order Preserving Factor Analysis,” a non-Bayesian optimization approach was employed). It is encouraging that these approaches agreed on the following 50 genes as being important to the host response to the virus (these contribute significantly to the factor linked to the host response to the virus): RSAD2, OAS1, IFI44L, RTP4, IFIT3, IFITM1, IFI44, PLSCR1, LY6E, ISG15, P2RX5, IFI27, GBP1, KIAA0125, APOBEC3A, EPB41L3, IFIT1, XAF1, PSMB9, TRIM22, SERPING1, HERC5, OASL, SCO2, IFI6, DDX60, BLK, MS4A4A, TNFRSF9, BLVRA, LOC26010, MX1, C1QA, OAS3, IRF7, VAMP5, IFIT5, SMPDL3A, FER1L3, UBE2L6, SIGLEC1, C13orf18, PSME2, IFI35, C1QB, BST2, OAS2, PNOC, RRAS, and SRBD1. These same genes were found to be important to all viruses we have studied (H3N2, H1N1, HRV, and RSV), and therefore we refer to these as constituting a “pan-viral” factor. Further, we emphasize that we only list 50 genes for brevity, but hundreds of other genes are also inferred to play a role in the host response. In the context of the factor analysis in the section “Time-Dependent Factor Scores,” for example, these genes are those that contribute appreciable amplitude to the factor loading highlighted in Figures 3 and 4. In a third distinct analysis (not covered in this article), based upon elasticnet and Bayesian elastic-net analyses [30], these genes were again found to play principal roles in the host response; we therefore emphasize that these genes have been analyzed and reanalyzed from multiple statistical perspectives, and their robustness suggests biological importance.

These findings are also supported by our proteomics data. Three of the proteins (CRP, A2GL, and CO9), and 35 of the 50 top genes, have in their promoter regions binding sites

for interferon regulatory factor 1. This suggests that activation of this interferon pathway is critical for an active response to infection, although this has yet to be tested thoroughly.

In Figure 11, we relate the aforementioned genes to an inferred pathway. This pathway is deemed to be of high accuracy as the strength of association of the multitask gene list with this pathway is quite robust (z-score [an indication of how many genes in the gene list are represented in a particular network] 76.83). The top represented pathway, the ISG15 pathway in Figure 11, is highly involved in viral immunity as it is activated by initial viral sensing and subsequent interferon production. ISG15 is known to target the influenza A protein NS1 and result in limitation of viral replication [31].

SUMMARY

An underlying theme of the statistical analysis is constituted by use of factor analysis to yield a small number of factors responsible for the high-dimensional data. This framework significantly aids analysis, as we typically have far fewer samples than genes and proteins, and therefore dimensionality reduction is essential. The factor analysis has been implemented from various perspectives. Specifically, in one analysis of the gene-expression data, and when analyzing the proteomic data, the factor loadings were related to the genes/proteins, and the factor loadings were assumed sparse, as to infer the low-dimensional set of genes/proteins responsible for biological pathways. In a distinct factor analysis, related to dictionary learning, the factor loadings were employed to model the time dependence of the gene expressions, and in this case the loadings are not sparse, but smooth. In addition to these different usages of the underlying model, we also employed Bayesian and non-Bayesian inference methods. It is highly encouraging that these very different methods yielded very similar biological interpretations, concerning the genes that play a pivotal role in the host response to virus.

We have focused here on the H3N2 influenza virus, to simplify the discussion. However, we have performed related analyses on all virus investigated in our challenge studies, and we found consistent host responses and underlying genes/proteins across all of them. This has led us to constitute what we term a “pan-viral” factor, with an associated pathway we have briefly discussed.

Acknowledgments

The research reported here was supported by the Defense Advanced Research Projects Agency, under the Predicting Health and Disease program. The findings of this article are those of the authors only.

Biography

Lawrence Carin (lcarin@duke.edu) earned the B.S., M.S., and Ph.D. degrees in electrical engineering at the University of Maryland, College Park, in 1985, 1986, and 1989, respectively. In 1989, he joined the electrical engineering department at Polytechnic University (Brooklyn) as an assistant professor and became an associate professor there in 1994. In September 1995, he joined the electrical engineering department at Duke University, where he is now the William H. Younger Professor of Engineering. He is currently the Chair of the Electrical and Computer Engineering Department at Duke. He is a cofounder of Signal Innovations Group, Inc. (SIG), a small business, where he serves as the director of technology. His current research interests include signal processing, sensing, and machine learning. He has published over 200 peer-reviewed papers, and he is a member of the Tau Beta Pi and Eta Kappa Nu honor societies.

Alfred Hero, III (hero@eecs.umich.edu) received the B.S. (summa cum laude) degree from Boston University (1980) and the Ph.D. degree from Princeton University (1984), both in electrical engineering. Since 1984, he has been with the University of Michigan, Ann Arbor, where he is the R. Jamison and Betty Williams Professor of Engineering. His primary appointment is in the Department of Electrical Engineering and Computer Science and he also has appointments, by courtesy, in the Department of Biomedical Engineering and the Department of Statistics. In 2008, he was awarded the Digiteo Chaire d'Excellence, sponsored by Digiteo Research Park in Paris, located at the Ecole Supérieure d'Electricité, Gif-sur-Yvette, France. Several of his research articles have received best paper awards from signal processing and medical journals. He received the IEEE Signal Processing Society Meritorious Service Award (1998) and the IEEE Third Millennium Medal (2000). He was president of the IEEE Signal Processing Society (2006–2008) and sits on the Board of Directors of the IEEE (2009–2011), where he is the Division IX director (Signals and Applications).

Joseph Lucas (joe@stat.duke.edu) is an assistant research professor at the Duke Institute for Genome Sciences and Policy. He received master's degrees in mathematics and computer science from the University of Pennsylvania and completed his Ph.D. degree in statistics at Duke University in December 2006. After completing his studies, he worked as a postdoctoral fellow in the Duke Department of Statistical Science studying the statistical issues related to high-dimensional, low sample size data sets. Such data sets are generated in abundance in the proteomics, microarray, metabolomics, and sequencing cores. Since starting as an assistant professor, he has worked to build a group that is specifically capable of operating in highly collaborative research environments. His group is focused on generating novel statistical analyses of new biological data sets and facilitating the use of those analyses in a wide variety of collaborations.

David Dunson (dunson@stat.duke.edu) is a professor of statistical science at Duke University. His research focuses on Bayesian statistical theory and methods motivated by high-dimensional and complex data sets. A particular emphasis is on nonparametric probability models and adaptive low-dimensional factorizations of high-dimensional and structured data and functions. He has made broad contributions to the literature on nonparametric Bayesian methods for multivariate, longitudinal, and functional data motivated by applications in epidemiology, genetics, and signal processing. He is a fellow of the American Statistical Association and of the Institute of Mathematical Statistics. He received the 2007 Mortimer Spiegelman Award for the top public health statistician, the 2010 Myrto Lefkopoulou Distinguished Lectureship at Harvard University, and the 2010 COPSS Presidents' Award for the top statistician under 41.

Minhua Chen (minhua.chen@duke.edu) received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 2004 and 2006, respectively. In 2009, he received another M.S. degree from Duke University, Durham, North Carolina. He is currently working toward the Ph.D. degree at Duke University. His research interest is in machine learning and signal processing.

Ricardo Heñao (rhenao@binf.ku.dk) received the bachelor's degree in electronic engineering and master's degree in industrial automation from the Universidad Nacional de Colombia, Manizales, in 2002 and 2004, respectively. He completed his Ph.D. at DTU Informatics, Technical University of Denmark in 2011. He worked as a postdoctoral researcher at the Bioinformatics Centre, University of Copenhagen during the first half of 2011, and he is currently a postdoctoral associate at the Duke Institute for Genome Sciences and Policy. His research interests include graphical model learning and supervised modeling with applications in bioinformatics.

Arnau Tibau-Puig (arnau.tibau@gmail.com) received the electrical engineering and Diplôme d'Ingénieur degrees from the Technical University of Catalonia (UPC) and the École Supérieure d'Électricité (Supélec) in 2007. He is currently pursuing the Ph.D. degree at the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, in partnership with the Signal Processing and Electronic Systems Department, Supélec, Gif-sur-Yvette. In 2007, he performed research in blind estimation for passive radar with the SONDRALaboratory, National University of Singapore. His current research interests are in efficient methods for high-dimensional time-series analysis.

Aimee Zaas (aimee.zaas@duke.edu) is an associate professor of medicine in the Division of Infectious Diseases and International Health, Department of Medicine at Duke University Medical Center. She received her M.D. degree from Northwestern Feinberg School of Medicine, completed internal medicine training at The Johns Hopkins Hospital, and a fellowship in infectious diseases at Duke University. Her research interests include utilization of gene expression as a diagnostic tool in infectious diseases.

Christopher W. Woods (woods004@mc.duke.edu) received the M.D. degree and postdoctoral training in medicine and medical microbiology at Duke University and an M.P.H. degree from The University of North Carolina at Chapel Hill. He is the codirector of the Hubert-Yeargan Center for Global Health, which is a member of both the Institute for Genome Science and Policy and the Duke Global Health Institute, where he is the director of graduate studies for the M.Sc. in global health and lead on the institute's signature research initiative on emerging infections. He is an associate professor with both the Departments of Medicine and Pathology, Duke University, Durham, North Carolina; an adjunct professor in epidemiology with the University of North Carolina at Chapel Hill School of Public Health; and chief of infectious diseases for the Durham Veterans Affairs Medical Center. He has a particular interest in the development of medical microbiology capacity in the developing world and the epidemiology of emerging and re-emerging infectious diseases.

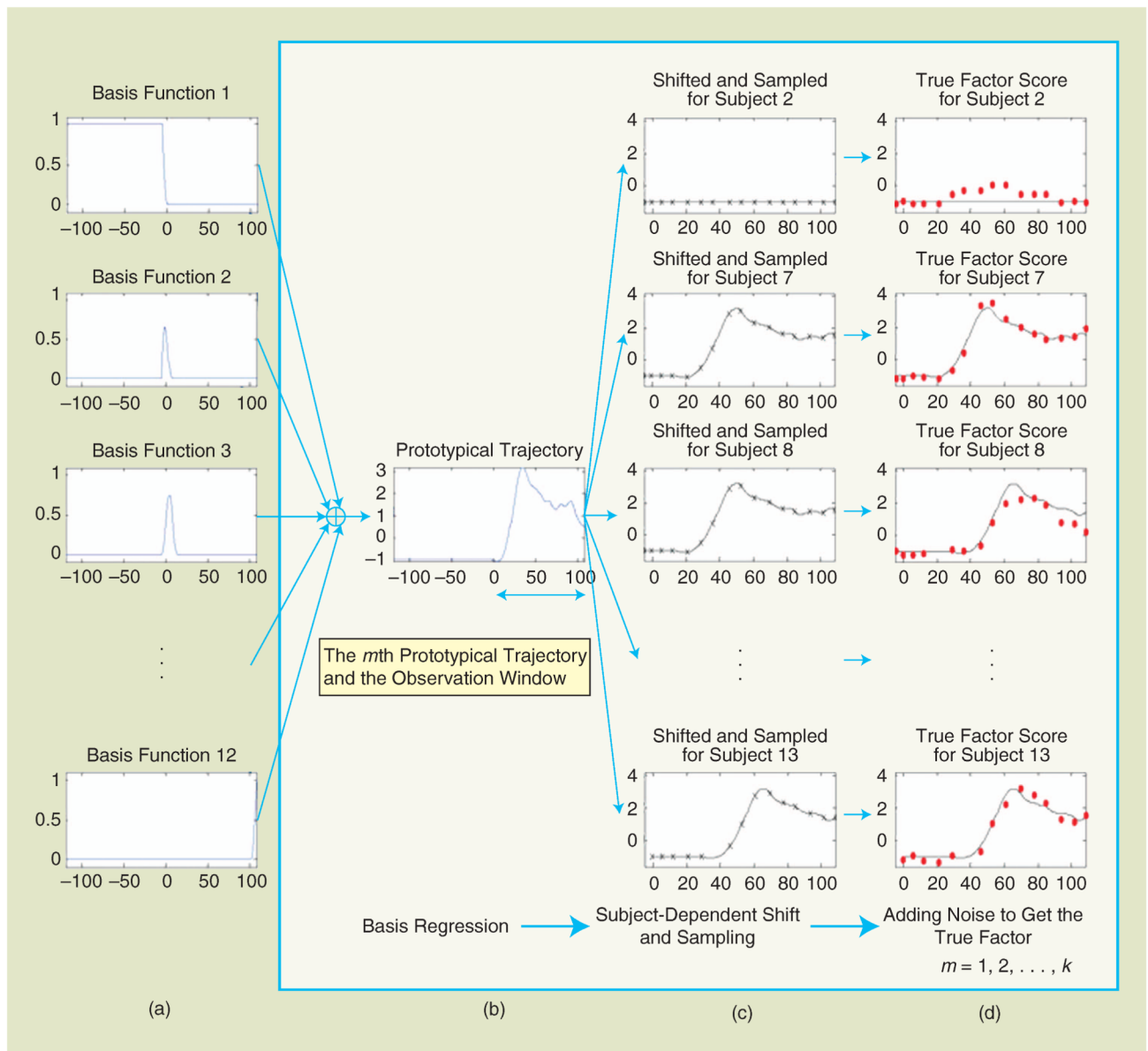
Geoffrey S. Ginsburg (ginsb005@win.duke.edu) is the director of the Center for Genomic Medicine, Duke Institute for Genome Sciences and Policy as well as professor of medicine and pathology at Duke University Medical Center. His research focuses on the development of genomics-based predictive and diagnostic biomarkers for health and disease. He was previously on the faculty of Harvard Medical School and was vice president of molecular and personalized medicine at Millennium Pharmaceuticals, Inc. He is a founding member of the Personalized Medicine Coalition; editorial advisory board member for science translational medicine; senior consulting editor for *The Journal of the American College of Cardiology*; editor-in-chief of *The HUGO Journal*; and coeditor of *Genomic and Personalized Medicine*. He is a member of the Board of External Experts for the National Heart, Lung, and Blood Institute; the Institute of Medicine's Roundtable on Genome-Based Research to Human Health; the National Advisory Council for Human Genome Research; and the External Scientific Panel for the Pharmacogenomics Research Network.

REFERENCES

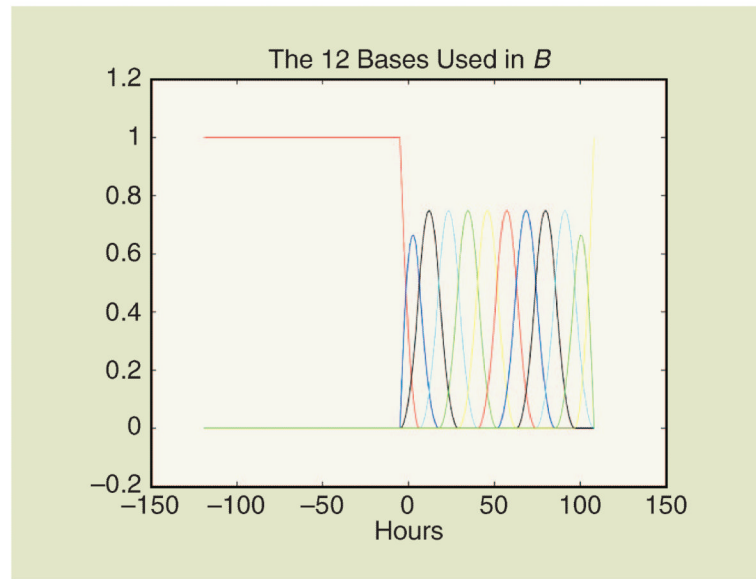
- [1]. Bar-Joseph Z, Gerber G, Gifford D, Jaakkola T, Simon I. Continuous representations of time-series gene expression data. *J. Comput. Biol.* 2003; 10(3-4):341–356. [PubMed: 12935332]
- [2]. Bar-Joseph Z. Analyzing time series gene expression data. *Bioinformatics.* 2004; 20(16):2493–2503. [PubMed: 15130923]
- [3]. Liu Q, Lin K, Andersen B, Smyth P, Ihler A. Estimating replicate time shifts using Gaussian process regression. *Bioinformatics.* 2010; 26(6):770–776. [PubMed: 20147305]

4. James G, Hastie T. Functional linear discriminant analysis for irregularly sampled curves. *J. Royal Stat. Soc. Ser. B.* 2001; 63(3):533–550.
5. Huang Y, Zaas A, Rao A, Dobigeon N, Woolf P, Veldman T, Øien N, McClain M, Varkey J, Nicholson B, Carin L, Kingsmore S, Woods C, Ginsburg G, Hero AIII. Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. *PLoS Genet.* 2011; 7(8):e1002234. [PubMed: 21901105]
6. 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. *J. Amer. Stat. Assoc.* 2006; 101(473):18–29.
7. Scharl T, Grun B, Leisch F. Mixtures of regression models for time-course gene expression data: Evaluation of initialization and random effects. *Bioinformatics.* 2010; 26(3):370–377. [PubMed: 20040587]
8. Wang L, Chen G, Li H. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics.* 2007; 23(12):1486–1494. [PubMed: 17463025]
9. Carvalho C, Chang J, Lucas J, Nevins J, Wang Q, West M. High-dimensional sparse factor modelling: Applications in gene expression genomics. *J. Amer. Stat. Assoc.* 2008; 103(484):1438–1456. [PubMed: 21218139]
10. Griffiths, T.; Ghahramani, Z. Infinite latent feature models and the indian buffet process. In: Weiss, Y.; Scholkopf, B.; Platt, J., editors. *Advances in Neural Information Processing Systems*. MIT Press; Cambridge, MA: 2005. p. 475–482.
11. Brieland J, Essig D, Jackson C, Frank D, Loebeberg D, Menzel F, Arnold B, DiDomenico B, Hare R. Comparison of pathogenesis and host immune responses to *Candida glabrata* and *Candida albicans* in systemically infected immunocompetent mice. *Infect. Immun.* 2001; 69(8):5046–5055. [PubMed: 11447185]
12. Turner R. Ineffectiveness of intranasal zinc gluconate for prevention of experimental rhinovirus colds. *Clin. Infect. Dis.* 2001; 33(11):1865–1870. [PubMed: 11692298]
13. Lewicki M, Sejnowski T. Learning overcomplete representations. *Neural Comput.* 2000; 12(2): 337–365. [PubMed: 10636946]
14. Lee M, Kim Y. Signaling pathways downstream of pattern recognition receptors and their cross talk. *Biochemistry.* 2007; 76(1):447–480.
15. Tibau-Puig, A.; Wiesel, A.; Zaas, A.; Ginsburg, GS.; Fleury, G.; Hero, AO. Order preserving factor analysis; *Proc. IEEE Workshop Sensor, Array and Multichannel Signal Processing (SAM)*; Jerusalem, Israel. Sept. 2010; p. 209–212.
16. Jenatton R, Obozinski G, Bach F. Structured sparse principal component analysis. *Proc. AISTATS.* 2010; 9:366–373.
17. Kolda T, Bader B. Tensor decompositions and applications. *SIAM Rev.* 2009; 51(3):455–500.
18. Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization. *Nature.* 1999; 401(6755):788–791. [PubMed: 10548103]
19. Taylor M, Tsukahara T, McClintick J, Edenberg H, Kwo P. Cyclic changes in gene expression induced by Peg-interferon alfa-2b plus ribavirin in peripheral blood monocytes(PBMC) of hepatitis C patients during the first 10 weeks of treatment. *J. Transl. Med.* 2008; 6(1):1–15. [PubMed: 18171482]
20. Zaas A, Chen M, Lucas J, Veldman T, Hero A, Varkey J, Turner R, Oien C, Kingsmore S, Carin L, Woods C, Ginsburg G. Peripheral blood gene expression signatures characterize symptomatic respiratory viral infection. *Cell Host Microbe.* 2009; 6(3):207–217. [PubMed: 19664979]
21. Chen M, Zaas A, Woods C, Ginsburg G, Lucas J, Dunson D, Carin L. Predicting viral infection from high-dimensional biomarker trajectories. *J. Amer. Stat. Assoc.* to be published.
22. Polpitiya A, Qian W-J, Jaitly N, Petyuk V, Adkins J, Camp D, Gordon A, Smith R. Dante: A statistical tool for quantitative analysis of proteomics data. *Bioinformatics.* 2008; 24(13):1556–1558. [PubMed: 18453552]
23. Karpievitch Y, Stanley J, Taverner T, Huang J, Adkins JN, Ansong C, Heffron F, Metz TO, Qian W-J, Yoon H, Smith RD, Dabney AR. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics.* 2009; 25(16):2028–2034. [PubMed: 19535538]

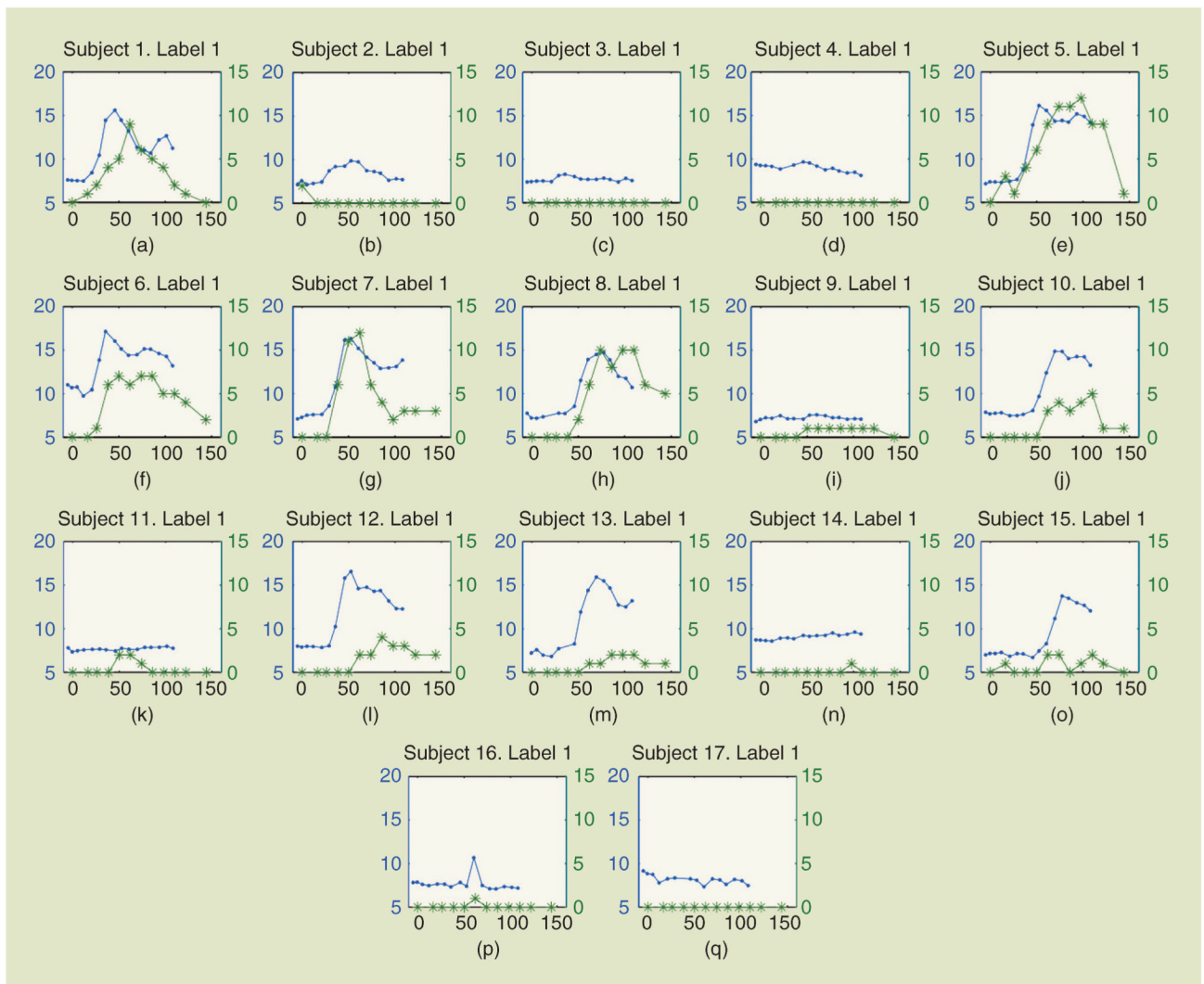
24. Daly DS, Anderson KK, Panisko EA, Purvine SO, Fang R, Monroe ME, Baker SE. Mixed-effects statistical model for comparative LC-MS proteomics studies. *J. Proteome Res.* 2008; 7(3):1209–1217. [PubMed: 18251496]
25. Clough T, Key M, Ott I, Ragg S, Schadow G, Vitek O. Protein quantitation in label-free LC-MS experiments. *J. Proteome Res.* 2009; 8(11):5275–5284. [PubMed: 19891509]
26. Lucas J, Thompson J, Dubois L, McCarthy J, Patel K, Tillman H, Thompson A, McHutchison J, Moseley M. Metaprotein expression modeling for label-free quantitative proteomics. *J. Proteome Res.* 2011 RECOMB Satellite Conf. on Computational Proteomics. 2011 [Online]. Available: <https://www.murdock-study.com/about-the-study/horizon-1/liver-disease-hepatitis-c/Lucas%20et%20al%20chemPaper.pdf>.
27. Henao R, Thompson JW, Moseley M, Ginsburg G, Carin L, Lucas J. Latent protein trees. Working Paper. 2011 [Online]. Available: http://people.genome.duke.edu/_jel2/workingPapers/latentProteinTrees.pdf.
28. Kingman JFC. The coalescent. *Stochast. Processes Applicat.* [Online]. 1982; 13(3):235–248. Available: <http://www.sciencedirect.com/science/article/B6V1B-45FT7RY-V/2/9aa9d856ebcd608a17099d47bb63319b>.
29. Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann; San Francisco, CA: 1988.
30. Chen M, Carlson D, Zaas A, Woods C, Ginsburg G, Hero A, Lucas J, Carin L. Detection of viruses via statistical gene expression analysis. *IEEE Trans. Biomed. Eng.* 2011; 58(3):468–479. [PubMed: 20643599]
31. Zhao C, Sun G, Li S, Lang M-F, Yang S, Li W, Shi Y. MicroRNA let-7b regulates neural stem cell proliferation and differentiation by targeting nuclear receptor TLX signaling. *Proc. Nat. Acad. Sci.* 2010; 107(5):1876–1881. [PubMed: 20133835]
32. Versteeg G, Hale B, van Boheemen S, Wolff T, Lenschow D, Garca-Sastre A. Species-specific antagonism of host ISGylation by 11 the influenza B virus NS1 protein. *J. Virol.* 2010; 84(10): 5423–5430. [PubMed: 20219937]

**[FIG1].**

Generative process for the factors $S_{mi} = B(t_i; \tau_{mi}) w_m + \varepsilon_{mi}$. Part (a) shows the basis functions, corresponding to spline functions and a step function at earliest times (the latter represents the factor before the virus under study causes changes to the host). The basis functions are weighted by w_m and superposed, to constitute a continuous-time factor, termed here a (b) “prototypical trajectory.” For individual i , the trajectory is shifted by time τ_{mi} , and then sampled at the times defined by t_i , manifesting the discrete samples in (c). Finally, i.i.d. noise is added to each discrete observation, manifesting the (d) final discrete individual-dependent factors for factor m . Part (d) corresponds to factor scores for actual samples from the H3N2 challenge study (microarray data), with the “prototypical trajectory” representing the inferred typical host response, apart from the individual-dependent shift τ_{mi} . The (a) basis functions are used for all factors $m \in \{1, \dots, k\}$, and separate w_m and shift time τ_{mi} are used to yield the shifted factors within the box.

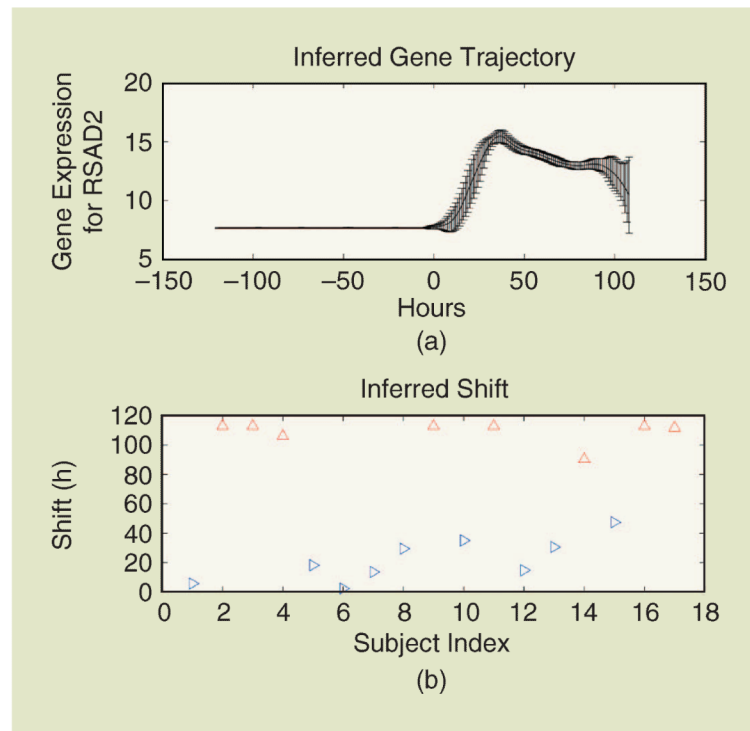


[FIG2].
Basis functions used for modeling the time dependence of the factor scores.

**[FIG3].**

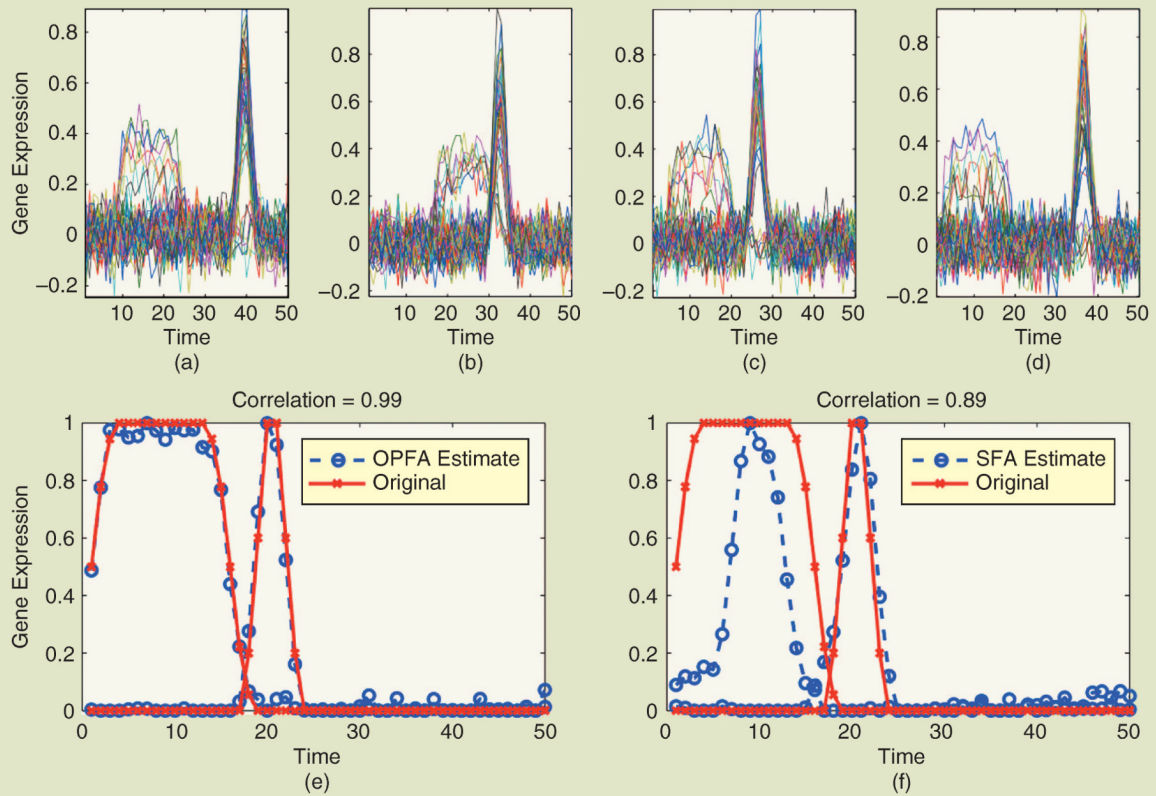
Parts (a)–(q) show subject-dependent plots of average

$x_g^{(m)}(t_{ij}) = Z_{gm} A_{gm} [\epsilon_{mi}(t_{ij}) + \sum_{l=1}^q w_{ml} B_l(t_{ij}; \tau_{mi})]$, for gene RSAD2 from the factor linked to H3N2 (blue) as well as the clinically observed symptom score (green). We consider RSAD2 gene, for which $Z_{gm} = 1$. The horizontal axes correspond to time from inoculation, in hours, and the vertical axes correspond to (left) factor or (right) clinical score. The subjects with a +1 label (top of each subfigure) corresponds to individuals who became symptomatic, and those with -1 labels were asymptomatic. Time $t = 0$ h corresponds to when the virus inoculation occurred. To reduce clutter in the figures, the axes are not labeled; the horizontal axes correspond to time in hours, and the vertical axes represent the (left) factor score or the (right) clinical score.



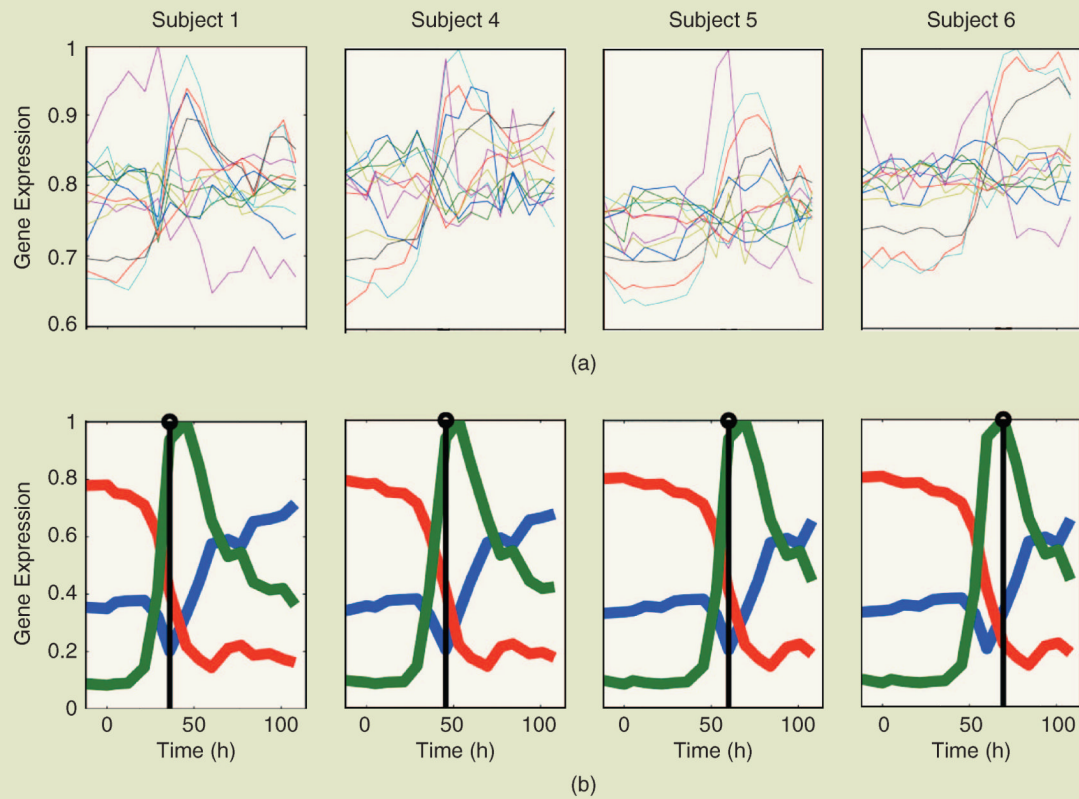
[FIG4].

(a) Inferred average trajectory for the (presumed) factor associated with the time-dependent host response to H3N2, $Z_{gm}A_{gm}\sum_{l=1}^q w_{ml}B_l(t_{ij};\tau_{ml}=0)$ (with standard-deviation error bars), corresponding to the gene RSAD2 ($Z_{gm} = 1$). (b) Inferred shifts for all individuals. Note that the shifts cluster naturally into two groups (red: asymptomatic, blue: symptomatic), consistent with the clinical label information.



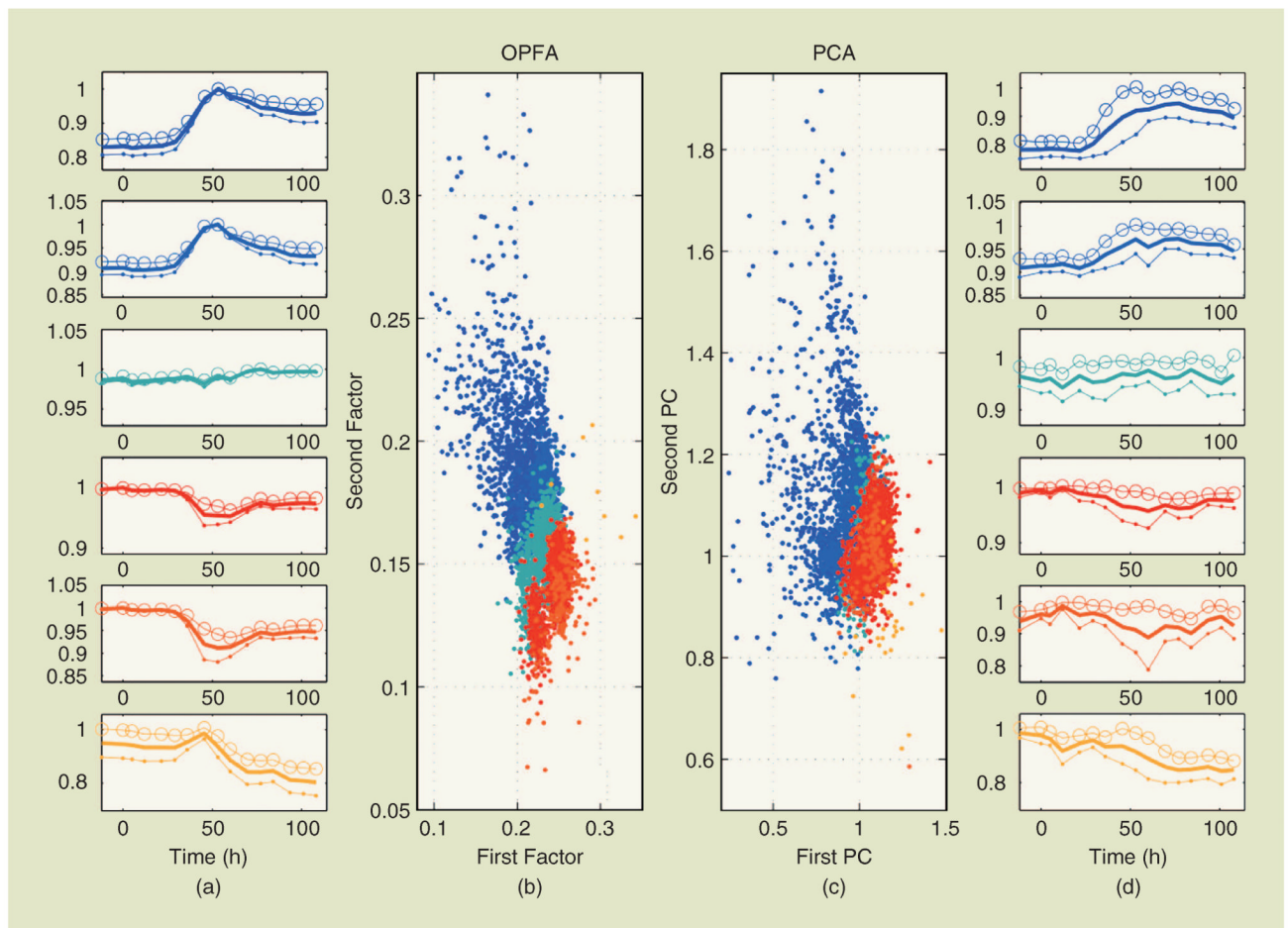
[FIG5].

Illustration of the effect of temporal misalignment on OPFA for a synthetic example with ten subjects (only four shown), 50 genes, 50 time points and $f = 2$ factors in a low-noise environment ($\text{SNR} = 10\text{dB}$). Parts (a)–(d) show the gene trajectories of four of the subjects; the misalignment of the signal features is evident. Part (e) shows the OPFA estimated and original factors, after realignment to a common reference time-point. Part (f) shows the same for SFA, a model that does not account for the order-preserving misalignments. The OPFA estimates correlate significantly better with the original factors than the SFA ones.

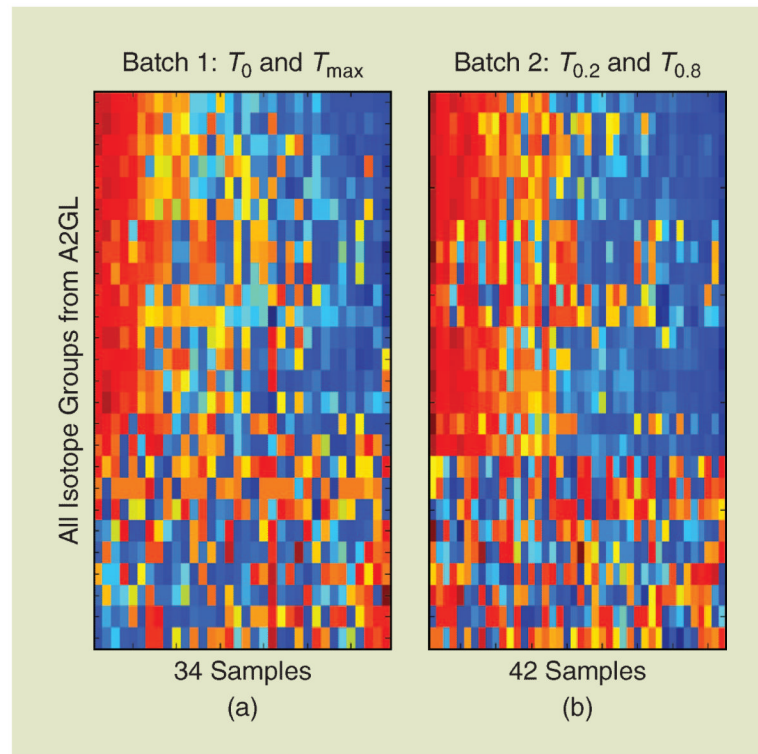


[FIG6].

(a) Gene expression trajectories of 13 highly variant genes for four of the nine symptomatic subjects in the H3N2 challenge study. (b) OPFA discovers three factors (red, blue, and green), and their corresponding alignment parameters, that explain these gene expression trajectories by solving the optimization problem in (6). The clinical onset times, determined by physicians, are shown in black. It is clear that the peak of the green factor predicts the onset time and that precedence order among the three factors is consistent across subjects.

**[FIG7].**

Scatter plots: Representation of each gene trajectory on the first two OPFA coordinates (given by the first two columns of A_i) and the coordinates obtained through a PCA analysis of the misaligned joint covariance $\sum_{i=1}^9 X_i X_i^T$. Parts (a) and (d) show the correspondence between the color coding and the clusters obtained by doing hierarchical clustering on the (b) OPFA realigned data and the (c) raw misaligned data. The down-regulated genes (reds) are clearly clustered away from the up-regulated genes (dark blues) in the OPFA representation, and both groups are separated by the genes that show little or no variation (light blue). The PCA analysis (c) suffers from misalignments and the first two principal components fail to separate the up-regulated genes (reds) from the down-regulated genes (dark blues). Furthermore, the temporal clusters in (a), obtained from OPFA realigned data, are more concentrated than those in (d), obtained from raw misaligned data, as can be seen from the tighter confidence envelopes on the OPFA cluster means.

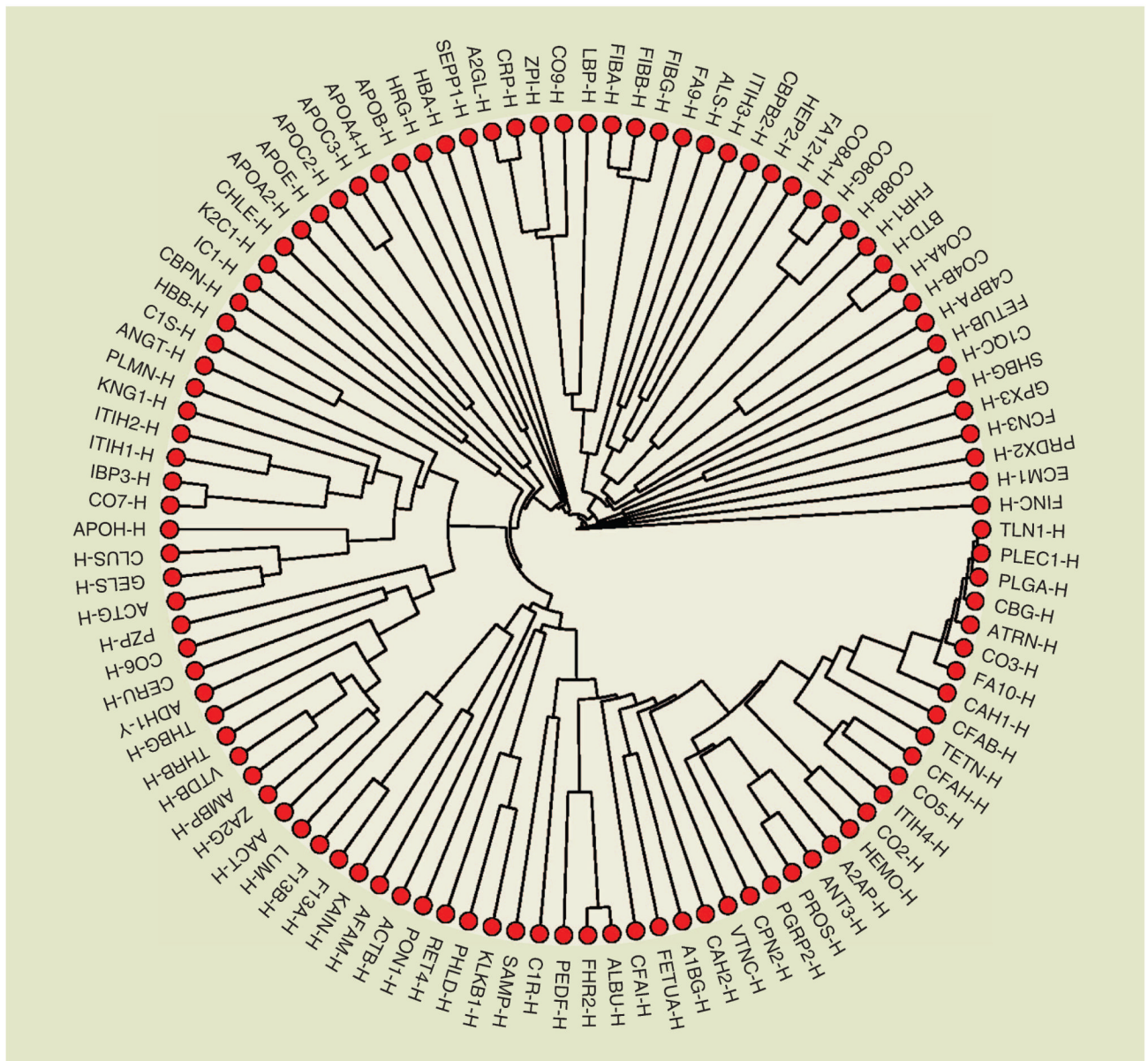


[FIG8].

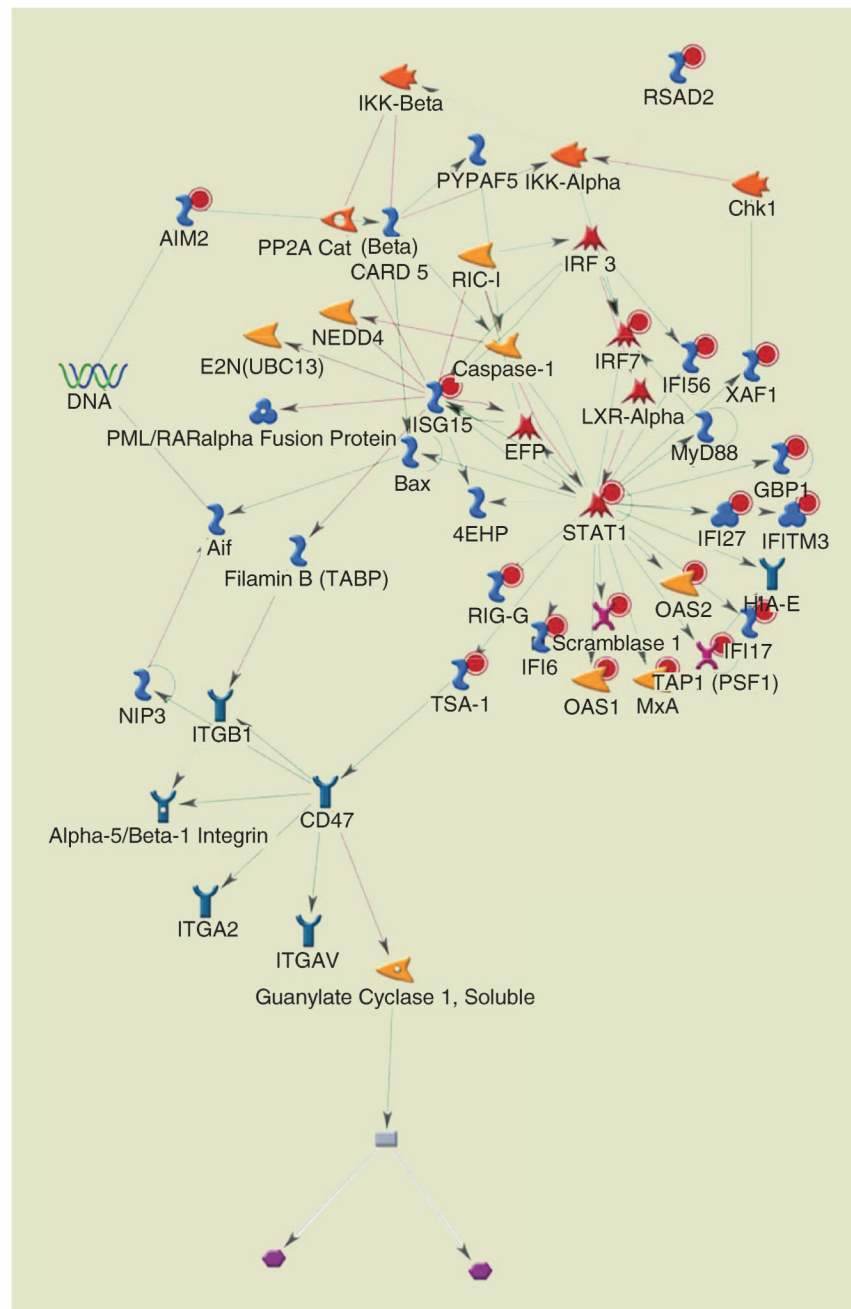
All isotope groups originating from the protein A2GL. The columns have been ordered to make the first principal component monotone (independently in each heatmap) and the rows have been ordered from top to bottom in order of decreasing correlation with the first principal component in (a), with that ordering preserved in (b). We have broken the figure by batch to demonstrate that the correlation structure is preserved even when the experiment is repeated on different samples months apart.



IEEE Signal Process Mag. Author manuscript; available in PMC 2014 March 25.



[FIG10]. The tree with the highest posterior likelihood from among those that were visited during the MCMC chain. A2GL is shown at the top in a subtree with c-reactive protein and lipopolysaccharide binding protein, both of which are known to react to the presence of infection.



[FIG11].

Genes identified from a key antiviral immune pathway. Pathway analysis (www.genego.com) illustrates the ISG15 pathway, with over-representation of genes identified by the multitask elastic net. ISG15 is a ubiquitin-like modifier that is induced by interferon to restrict viral replication [32]. Downstream elements of ISG15 activation include activation of STAT-1.