Analysis of High-Dimensional Longitudinal Genomic Data for Monitoring Viral Infection

[1]Lawrence Carin, [2]Alfred Hero, III, [3]Joseph Lucas, [4]David Dunson, [1]Minhua Chen,
[3]Ricardo Heñao, [2]Arnau Tibau-Puig, [3]Aimee Zaas, [3]Christopher W. Woods, and [3]Geoffrey S. Ginsburg

[1]Electrical and Computer Engineering Department, Duke University
[2]Electrical Engineering and Computer Science Department, University of Michigan
[3]Institute for Genome Sciences and Policy & Department of Medicine, Duke University
[4]Department of Statistical Science, Duke University
POC: lcarin@duke.edu

I. Introduction

A. 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 missingness. 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].

In this paper we consider 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. For more details on the mRNA data, see [5].

B. 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], [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), 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.

II. Time-Dependent Factor Scores

A. Basic factor model

Let $\mathbf{X}_i \in \mathbb{R}^{P\times T}$ 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 jth 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

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

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

B. Shifted spline representation

Recall that individual i has data sampled at n_i time points; let $\boldsymbol{t}_i = (t_{i1}, t_{i2}, \dots, t_T)$ 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 $\boldsymbol{t}_i$.

Recall that $\mathbf{S}_{mi} \in \mathbb{R}^T$ 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 $\boldsymbol{t}_i$ ($\mathbf{S}_{mi}$ is a column vector, the transpose of $\mathbf{S}_{mi}^\top$ above). To model $\mathbf{S}_{mi}$, let $\boldsymbol{b}(t) \in \mathbb{R}^q$ 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 $\boldsymbol{b}(t-\tau) \in \mathbb{R}^q$ 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

$$\mathbf{S}_{mi} = \mathbf{B}(\boldsymbol{t}_i; \tau_{mi})\boldsymbol{w}_m + \boldsymbol{\epsilon}_{mi}, \quad (2)$$

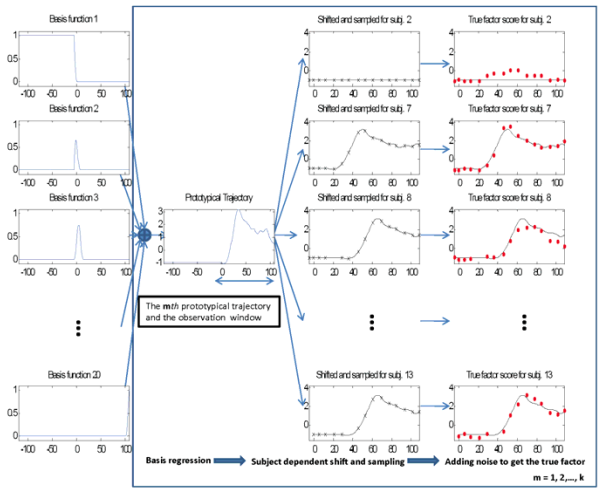

Fig. 1. Generative process for the factors $\mathbf{S}_{mi} = \mathbf{B}(\boldsymbol{t}_i; \tau_{mi})\boldsymbol{w}_m + \boldsymbol{\epsilon}_{mi}$. At left are shown the basis functions, corresponding to spline functions and a step function at earliest times (the latter represents the factor before the virus under study causes changes to the host). The basis functions are weighted by $\boldsymbol{w}_m$ and superposed, to constitute a continuous-time factor, termed here a "prototypical trajectory". For individual i, the trajectory is shifted by time τ_{mi}, and then sampled at the times defined by $\boldsymbol{t}_i$, manifesting the *discrete* samples in the second-to-last column. Finally, i.i.d. noise is added to each discrete observation, manifesting the final discrete individual-dependent factors for factor m (right-most column). The figures in the right two columns correspond to 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 basis functions (left column) are used for all factors $m \in \{1,\dots,k\}$, and separate weights $\boldsymbol{w}_m$ are used to yield the shifted factors within the box.

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

C. 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 ($t = -5$ hours) to constitute a baseline signature, and another (distinct) blood sample is drawn just before inoculation (the latter occurring at what is defined as time $t = 0$ hours). The vector $\boldsymbol{t}_i$ 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 $\boldsymbol{t}_i$, and since the time of inoculation is by definition at $t = 0$, the first element in $\boldsymbol{t}_i$ 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 $\mathbf{B}(\boldsymbol{t}_i; \tau_{mi} = 0)\boldsymbol{w}_m$ has a constant form for $t \leq -5$

hours (with value of the constant inferred via the analysis), this representing the background/baseline (pre-inoculation) 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.

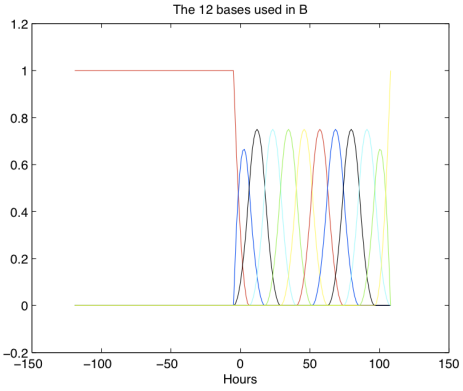

Fig. 2. Basis functions used for modeling the time dependence of the factor scores.

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 $\boldsymbol{\epsilon}_{mi}$). 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 $m \in \{1, \ldots, k\}$ implies that individual i is asymptomatic. Further, if a particular factor $m \in \{1, \ldots, k\}$ 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.

D. Sparse Factor Loadings

In many biological applications it is desirable to impose that the factor-loading matrix is sparse [9]. In the case of gene-expression data, the mth factor can be viewed as measuring overall expression of the mth pathway, with the non-zero elements in the mth 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

$$\mathbf{L} = \mathbf{V} \circ \mathbf{Z} \tag{3}$$

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 mth factor represents the mth dish. If gene g selects dish (factor) m, then $Z_{gm} = 1$, and otherwise $Z_{gm} = 0$.

E. Example results

Of the k factors, one of them manifested a time trajectory $\mathbf{B}(\boldsymbol{t}_i; \tau_{mi})\boldsymbol{w}_m$ that was closely aligned with the clinical scores, and 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.

III. Order preserving factor analysis

A. Dictionary learning and factor analysis

In addition to Bayesian factor analysis, we have also examined the data using non-Bayesian dictionary learning. The two

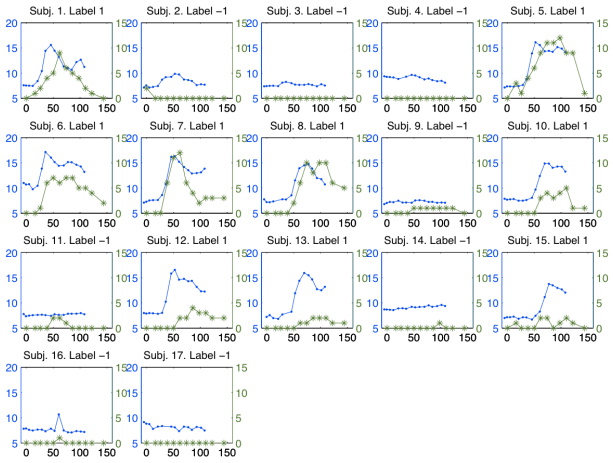

Fig. 3. Subject-dependent plots of average $x_g^{(m)}(t_{ij}) = Z_{gm}A_{gm}[\epsilon_{mi}(t_{ij}) + \sum_{l=1}^{q} w_{ml}B_l(t_{ij};\tau_{mi})]$, for gene RSAD2 from the factor linked to H3N2 (blue), as well as the clinically observed symptom score (green). We consider RSAD2 gene, for which $Z_{gm} = 1$. The horizontal axes correspond to time from inoculation, in hours, and the vertical axes correspond to factor (left) or clinical (right) 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 $t = 0$ hour 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 factor score (left) or the clinical score (right).

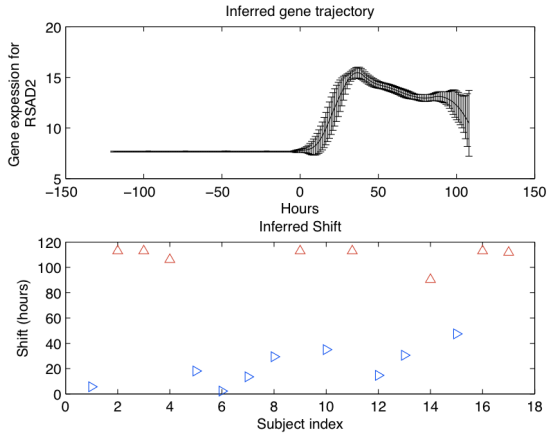

Fig. 4. Top: Inferred average trajectory for the (presumed) factor associated with the time-dependent host response to H3N2, $Z_{gm}A_{gm}\sum_{l=1}^{q} w_{ml}B_l(t_{ij};\tau_{mi} = 0)$ (with standard-deviation error bars), corresponding to the gene RSAD2 ($Z_{gm} = 1$). Bottom: Inferred shifts for all individuals. Note that the shifts cluster naturally into two groups (red: asymptomatic, blue:symptomatic), consistent with the clinical label information.

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 Section II, with $\mathbf{X}_i^\top$ from Section II 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

$$\mathbf{Y}_i = \mathbf{M}\mathbf{A}_i + \epsilon_i,\ i = 1, \ldots, S \tag{4}$$

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 $\mathbf{A}_i$ 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), that accounts for temporal misalignment, incorporates smoothness constraints on the factor loadings, and preserves their relative ordering over the subject population.

B. 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

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

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, $\mathbf{d}^i \in \left\{\mathbf{d} = [d_1, \ldots, d_f] : \cap_{j=2}^{f}\{d_j > d_{j-1}\}\right\}$ 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 non-convex optimization problem of the form

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

where minimization is performed over vectors of delays $\mathbf{d}^i$ that lie in the order-preserving set for all subjects and over non-negative matrices $\mathbf{F}$, $\mathbf{A}_i$. The non-negativity 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 sparse columns of $\mathbf{A}_i$. 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].

C. Example results

Our formulation (6) of order preserving factor analysis can be interpreted as an extension of Sparse Factor Analysis (SFA) [16], parallel matrix factorization (PARAFAC) [17], and non-negative 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 non-negative matrices, respectively. These matrix factorization methods are highly sensitive to misalignments of the factors. The order preserving restriction of OPFA overcomes this misalignment sensitivity as illustrated in Fig. 5 for a toy example.

Figure 6 shows the result of applying OPFA to real data, in particular the set of 9 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 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

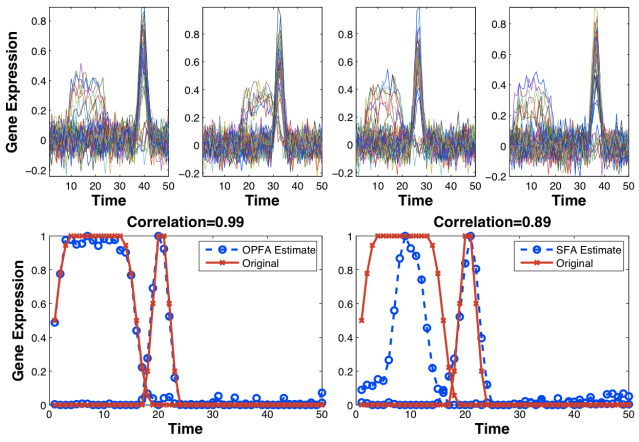

Fig. 5. Illustration of effect of temporal misalignment on order preserving factor analysis for a synthetic example with ten subjects (only four shown), 50 genes, 50 time points and $f = 2$ factors in a low-noise environment (SNR=10dB). The top plot shows the gene trajectories of four of the subjects, the misalignment of the signal features is evident. The left-bottom plot shows the OPFA estimated and original factors, after realignment to a common reference time-point. The right-bottom plot shows the same for Sparse Factor Analysis (SFA), a model that does not account for the order-preserving misalignments. The OPFA estimates correlate signficantly better with the original factors than the SFA ones.

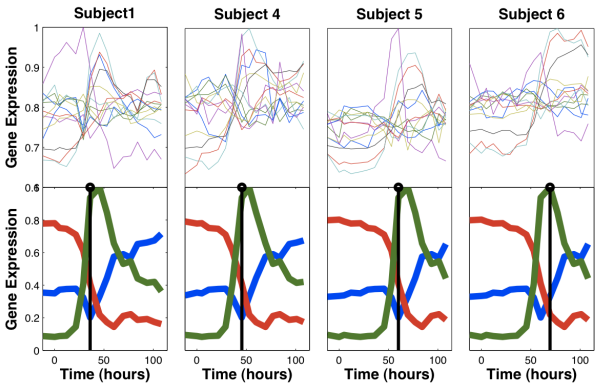

Fig. 6. Upper panel: gene expression trajectories of 13 highly-variant genes for 4 of the 9 symptomatic subjects in the H3N2 challenge study. Lower panel: OPFA discovers 3 factors (red blue and green), and their corresponding alignment parameters, that explain these gene expressions trajectories by solving 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.

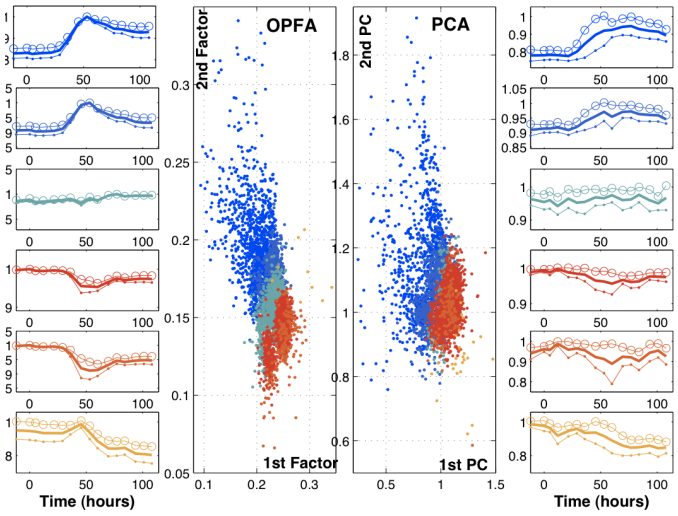

Fig. 7. Scatter plots: Representation of each gene trajectory on the first two OPFA coordinates (given by the first two columns of $\mathbf{A}_i$) and the coordinates obtained through a PCA analysis of the misaligned joint covariance $\sum_{i=1}^{9} \mathbf{X}_i \mathbf{X}_i^T$. The left-most and right-most figures show the correspondence between the color coding and the clusters obtained by doing hierarchical clustering on the OPFA re-aligned data (left) and the raw misaligned data (right). 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 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 obtained from OPFA re-aligned data (far left) are more concentrated than those obtained from raw misaligned data (far right), as can be seen from the tighter confidence envelopes on the OPFA cluster means.

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).

IV. Metaprotein expression modeling

A. Mass spectrometry proteomics

Unbiased mass spectrometry based proteomics has made tremendous progress since initial studies using MALDI-ToF (matrix-assisted laser desorption/ionization - time of flight) machines in the late 1980's. 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 mass spectrometry 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.

Mass spectrometry 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 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 2-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 w e summarize our data in this paper. 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 mass spectrometry peaks are often unknown.

B. 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 ANOVA approaches which 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 i ncorrect identifications, none of these approaches directly address this problem. In addition, other than our previous work which examines an earlier factor model in greater detal 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 paper 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.

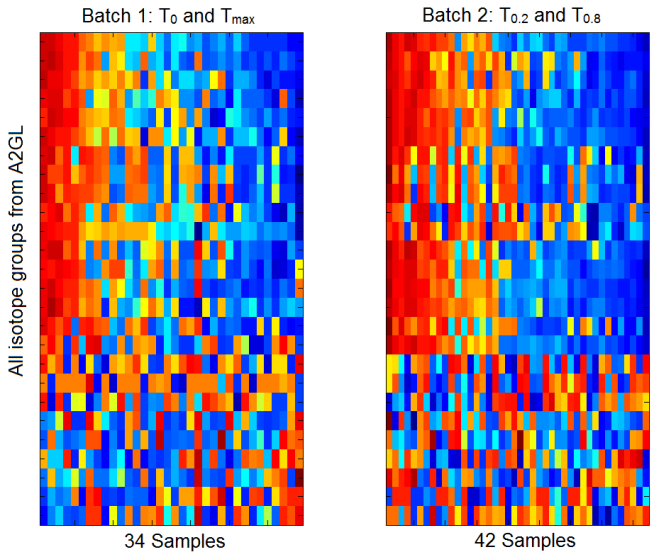

Fig. 8. All isotope groups originating from the protein A2GL. The columns have been ordered to make the first principal component monotone (independently in each heatmap) and the rows have been ordered from top to bottom in order of decreasing correlation with the first principal component in the left heatmap, with that ordering preserved in the right heatmap. 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.

C. Factor model and hierarchical structure

There are two key sources of information we might use in order 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

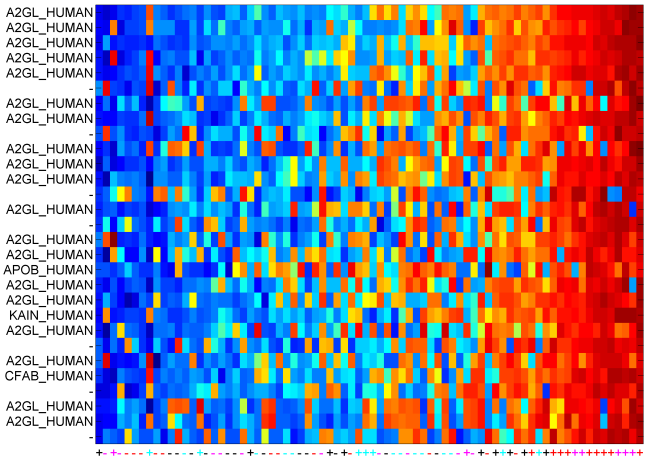

Fig. 9. A heatmap of the metaprotein showing the strongest association with disease. Each row is an isotope group and each column is a sample. Note that the majority of the peptides are from the protein A2GL (Leucine-rich alpha-2-glycoprotein), but that there are peptides that were identified as belonging to other proteins such as Apolipoprotein B-100 (APO B), Complement factor B (CFAB), Kallistatin (KAIN) and other unidentified isotope groups. The red color represents a relatively high concentration of the isotope group in the sample while blue represents low (each row has been standardized to have mean zero and variance 1). Samples from subjects who became symptomatic are labeled with + and those who remained assymptomatic are labeled with -. The label colors, black, blue, pink and red represent times 0, .2, .8 and 1 respectively. The samples are ordered so that the associated factor is increasing, and because almost all samples from symptomatic individuals at times .8 and 1 (red and pink +'s) are at the far right we can see that this factor clearly distinguishes sick from healthy individuals.

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 Section II-A, 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 $\mathbf{M} \in \mathbf{R}^{P\times d}$ is a factor loadings matrix and $\mathbf{A} \in \mathbf{R}^{d\times N}$ 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 the 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 non-zero element. Thus we introduce a latent variable $\boldsymbol{z}_i$ which identifies, for isotope group i, the metaprotein factor to which it belongs ($i.e.$, which element of the ith row of $\mathbf{L}$ is non-zero). Thus element $\mathbf{L}_{i,j} = 0$ when $j \neq z_i$ and $\mathbf{L}_{i,\boldsymbol{z}_i} \sim N(0, \tau_0)$. Our hierarchical prior for $\boldsymbol{z}_i$ is

$$\boldsymbol{z}_i \sim \text{Multinomial}(1, \boldsymbol{q}_i) \quad , \quad \boldsymbol{q}_i \sim \text{Dir}(\boldsymbol{a}_i). \tag{7}$$

We utilize an informative choice of Dirichlet distribution parameters $\boldsymbol{a}_i$ 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 $\boldsymbol{a}_i = (a_0, \cdots, a_0, a_k, a_0, \cdots, a_0)'$ where $a_k >> a_0$.

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. In order 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 $t_{a \to b}$ to be the "distance" between node a and its child node, b, and $\boldsymbol{w}_a$ 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

$$\boldsymbol{w}_b \quad \sim \quad N(\boldsymbol{w}_a, t_{a \to b} I_N)$$

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 algorithm to obtain posterior distributions for all model parameters. Sampling of coalescense times for the tree model is accomplished via belief propogation [29], [27]. Trees are constructed through the coalescence procedure described in [28] and in [27].

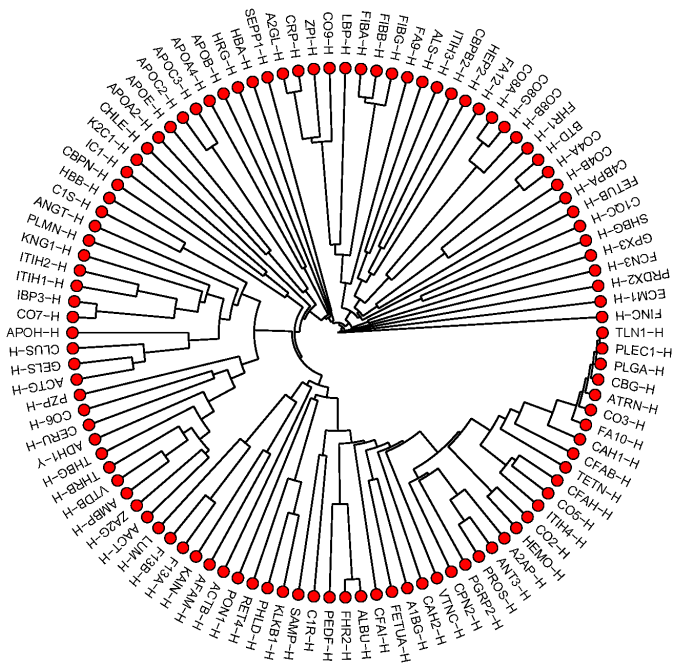

Fig. 10. The tree with the highest posterior likelihood from among those that were visited during the MCMC chain. A2GL is shown at the top in a sub-tree with c-reactive protein and lipopolysaccharide binding protein, both of which are known to react to the presence of infection.

D. Plasma proteomics during viral infection

Working with the same set of subjects discussed in the previous sections (innoculated 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 missingness). 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 meta-proteins 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 assymptomatic" 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.

V. Inferred Biology

As summarized in Sections II and III, two very distinct techniques were employed to analyze the time-course gene-expression data (in Section II a fully Bayesian approach was employed, while in Section III a non-Bayesian optimization approach was employed). It is encouraging that these approaches agreed on the 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 Section II, 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 paper), based upon elastic-net 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 re-analyzed from multiple statistical perspectives, and their robustness suggests biological importance.

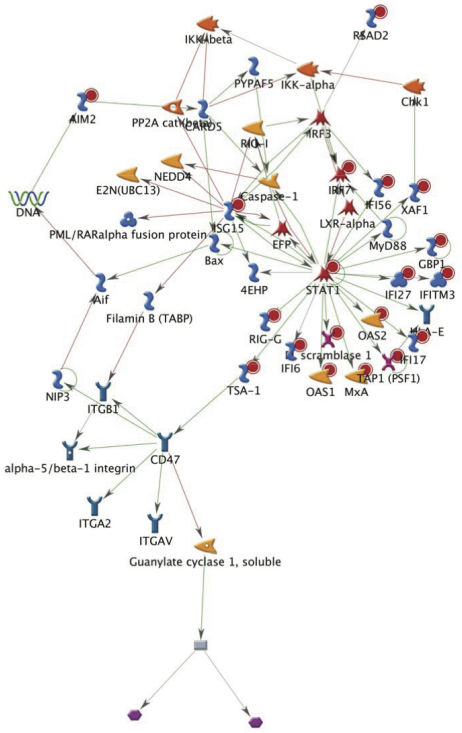

Fig. 11. Genes identified from a key anti-viral immune pathway. Pathway analysis (www.genego.com) illustrates the ISG15 pathway, with over-representation of genes identified by the multi-task elastic net. ISG15 is a ubiquitin-like modifier that is induced by interferon to restrict viral replication [32]. Downstream elements of ISG15 activation include activation of STAT-1.

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 multi-task 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].

VI. Summary

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

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. In addition to these different usages of the underlying model, we also employed Bayesian and non-Bayesian i nference 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.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Z. Bar-Joseph, G. Gerber, D. Gifford, T. Jaakkola, and I. Simon, "Continuous representations of time-series gene expression data," *Journal of Computational Biology*, vol. 10, pp. 3–4, 2003.

[2] Z. Bar-Joseph, "Analyzing time series gene expression data," *Bioinformatics*, vol. 20, pp. 2493–2503, 2004.

[3] Q. Liu, K. Lin, B. Andersen, P. Smyth, and A. Ihler, "Estimating replicate time shifts using gaussian process regression," *Bioinformatices*, vol. 26, pp. 770–776, 2010.

[4] G. James and T. Hastie, "Functional linear discriminant analysis for irregularly sampled curves," *Journal of the Royal Statistical Society Series B*, vol. 63, pp. 533–550, 2001.

[5] Y. H. *et al.*, "Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection," *PLoS Genetics*, p. 7(8): e1002234. doi:10.1371/journal.pgen.1002234, 2011.

[6] N. Heard, C. Holmes, and D. Stephens, "A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of bayesian hierarchical clustering of curves," *Journal of the American Statistical Association*, vol. 101, pp. 18–29, 2006.

[7] T. Scharl, B. Grun, and F. Leisch, "Mixtures of regression models for time-course gene expression data: evaluation of initialization and random effects," *Bioinformatics*, vol. 26, pp. 370–377, 2010.

[8] L. Wang, G. Chen, and H. Li, "Group SCAD regression analysis for microarray time course gene expression data," *Bioinformatics*, vol. 23, pp. 1486–1494, 2007.

[9] C. Carvalho, J. Chang, J. Lucas, J. Nevins, Q. Wang, and M. West, "High-dimensional sparse factor modelling: Applications in gene expression genomics," *Journal of the American Statistical Association*, vol. 103, pp. 1438–1456, 2008.

[10] T. Griffiths and Z. Ghahramani, "Infinite latent feature models and the indian buffet process," in *Advances in Neural Information Processing Systems*, 2005, pp. 475–482.

[11] J. Brieland, D. Essig, C. Jackson, D. Frank, D. Loebenberg, F. Menzel, B. Arnold, B. DiDomenico, and R. Hare, "Comparison of pathogenesis and host immune responses to Candida glabrata and Candida albicans in systemically infected immunocompetent mice," *Infect. Immun.*, vol. 69, pp. 5046–5055, 2001.

[12] R. Turner, "Ineffectiveness of intranasal zinc gluconate for prevention of experimental rhinovirus colds," *Clin. Infect. Dis.*, vol. 33, pp. 1865–1870, 2001.

[13] M. Lewicki and T. Sejnowski, "Learning overcomplete representations," *Neural computation*, vol. 12, no. 2, pp. 337–365, 2000.

[14] M. Lee and Y. Kim, "Signaling pathways downstream of pattern-recognition receptors and their cross talk," *Biochemistry*, vol. 76, no. 1, p. 447, 2007.

[15] A. Tibau-Puig, A. Wiesel, A. Zaas, G. S. Ginsburg, G. Fleury, and A. O. Hero, "Order preserving factor analysis," in *IEEE Workshop on Sensor, Array and Multichannel Signal Processing (SAM)*, Jerusalem, Israel, Sept 2010.

[16] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *Proc. AISTATS*, 2009.

[17] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[18] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[19] M. Taylor, T. Tsukahara, J. McClintick, H. Edenberg, and P. Kwo, "Cyclic changes in gene expression induced by Peg-interferon alfa-2 b plus ribavirin in peripheral blood monocytes(PBMC) of hepatitis C patients during the first 10 weeks of treatment," *Journal of Translational Medicine*, vol. 6, no. 1, p. 66, 2008.

[20] A. Zaas, M. Chen, J. Lucas, T. Veldman, A. Hero, J. Varkey, R. Turner, C. Oien, S. Kingsmore, L. Carin, C. Woods, and G. Ginsburg, "Peripheral blood gene expression signatures characterize symptomatic respiratory viral infection," *Cell Host & Microbe*, vol. 6, pp. 207–217, 2009.

[21] M. Chen, A. Zaas, C. Woods, G. Ginsburg, J. Lucas, D. Dunson, and L. Carin, "Predicting viral infection from high-dimensional biomarker trajectories," *submitted to J. Am. Statistical Association*, 2010.

[22] A. Polpitiya, W.-J. Qian, N. Jaitly, V. Petyuk, J. Adkins, D. Camp, A. Gordon, and R. Smith, "Dante: a statistical tool for quantitative analysis of -omics data," *Bioinformatics*, vol. 24, no. 13, pp. 1556–1558, 2008.

[23] Y. Karpievitch, J. Stanley, T. Taverner, J. Huang, J. N. Adkins, C. Ansong, F. Heffron, T. O. Metz, W.-J. Qian, H. Yoon, R. D. Smith, and A. R. Dabney, "A statistical framework for protein quantitation in bottom-up ms-based proteomics," *Bioinformatics*, vol. 25, no. 16, pp. 2028–2034, 2009.

[24] D. S. Daly, K. K. Anderson, E. A. Panisko, S. O. Purvine, R. Fang, M. E. Monroe, and S. E. Baker, "Mixed-effects statistical model for comparative lc-ms proteomics studies," *Journal of Proteome Research*, vol. 7, pp. 1209–1217, 2008.

[25] T. Clough, M. Key, I. Ott, S. Ragg, G. Schadow, and O. Vitek, "Protein quantitation in label-free lc-ms experiments," *Journal of Proteome Research*, vol. 8, pp. 5275–5284, 2009.

[26] J. Lucas, J. Thompson, L. Dubois, J. McCarthy, K. Patel, H. Tillman, A. Thompson, J. McHutchison, and M. Moseley, "Metaprotein expression modeling for label-free quantitative proteomics," *working paper*.

[27] R. Henao, J. W. Thompson, M. Moseley, G. Ginsburg, L. Carin, and J. Lucas, "Latent protein trees," *Working Paper*, 2011. [Online]. Available: http://people.genome.duke.edu/~jel2/workingPapers/latentProteinTrees.pdf

[28] J. F. C. Kingman, "The coalescent," *Stochastic Processes and their Applications*, vol. 13, no. 3, pp. 235 – 248, 1982. [Online]. Available: http://www.sciencedirect.com/science/article/B6V1B-45FT7RY-V/2/9aa9d856ebcd608a17099d47bb63319b

[29] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[30] M. Chen, D. Carlson, A. Zaas, C. Woods, G. Ginsburg, A. Hero, J. Lucas, and L. Carin., "Detection of viruses via statistical gene expression analysis," *Biomedical Engineering, IEEE Transactions on*, vol. 58, no. 3, pp. 468 –479, 2011.

[31] C. Zhao, G. Sun, S. Li, M.-F. Lang, S. Yang, W. Li, , and Y. Shi, "MicroRNA let-7b regulates neural stem cell proliferation and differentiation by targeting nuclear receptor TLX signaling," *Proc. Nat. Ac. Sciences*, vol. 107, pp. 1876–1881, 2010.

[32] G. Versteeg, B. Hale, S. van Boheemen, T. Wolff, D. Lenschow, and A. Garca-Sastre, "Species-specific antagonism of host ISGylation by

the influenza B virus NS1 protein," *J. Virology*, vol. 10, pp. 5423–5430, 2010.