Explorations in Very Early Prognosis of the Human Immune Response to Influenza

Mmanu Chaturvedi
Department of Computer Science
Colorado State University
Fort Collins, Colorado
mmanu.chaturvedi@gmail

Tomojit Ghosh
Department of Computer Science
Colorado State University
Fort Collins, Colorado
tomojit@rams.colostate.edu

Michael Kirby
Department of Mathematics
Colorado State University
Fort Collins, Colorado
kirby@math.colostate.edu

Xiaoyu Liu
Department of Mathematics
Colorado State University
Fort Collins, Colorado
arielliu@rams.colostate.edu

Xiaofeng Ma
Department of Mathematics
Colorado State University
Fort Collins, Colorado
ma@math.colostate.edu

Shannon Stiverson
Department of Mathematics
Colorado State University
Fort Collins, Colorado
stiverson@math.colostate.edu

ABSTRACT

We conduct machine learning experiments on time-dependent gene expression measurements associated with the immune response to influenza in humans. We employ three partitions of the two data sets focusing on H1N1 only, H3N2 only and H1N1 and H3N2 combined. From a total set of 1439 known biological pathways, we identify the most discriminatory, potentially capable of providing a very early prognosis of infection, focusing on the time period $t \leq 29$ hours post infection. We apply a suite of different machine learning algorithms to these partitions including linear, nonlinear, and sparse support vector machines. In addition, we use artificial neural networks (ANN), k-nearest neighbors and classification on Grassmann manifolds. The cAMP Signaling pathway and the genes PAPSS1 and PAPSS2 appeared to play central role in the very early prognosis problem.

Categories and Subject Descriptors

G.1.2 [**Numerical Analysis**]: Approximation—*Nonlinear approximation*; G.1.6 [**Numerical Analysis**]: Optimization—*Constrained optimization, Linear programming*; I.5.1 [**Pattern Recognition**]: Models—*Geometric, Neural nets*; J.3 [**Life and Medical Sciences**]: Medical Information Systems

Terms

Algorithms,Performance

Keywords

Complex data, biological pathway analysis, Grassmannian classification, sparse support vector machines.

BCB '16 October 02-05, 2016, Seattle, WA, USA

ACM ISBN 978-1-4503-4225-4/16/10.

DOI: http://dx.doi.org/10.1145/2975167.2985686

1. INTRODUCTION

Human influenza A viral infection, which includes the H1N1 and H3N2 strains, is one of the chief culprits of acute respiratory infections (ARIs) worldwide [1]. Viral ARIs are generally self-limited, and infected patients generally recover within one or two weeks without treatment. However, viral ARIs such as influenza can lead to disease exacerbation among individuals with prior pulmonary disease, and severe cases can lead to mortality in elderly and immunocompromised individuals [2]. Moreover, significant health care and social costs are associated with influenza epidemic due to the excessive hospitalizations and the need for production of a large amount of vaccines [3].

The evolution of the virus may escalate the global trend of clinical influenza infections and may even result in periodic epidemics [4]. Under these circumstances, early diagnosis of influenza A is essential to facilitating individual treatment decisions as well as aiding in forecast of an epidemic [5]. Traditionally, the test for human influenza A infection is based on pathogen detection [6], a method which has shown its limitations since the first outbreak of H1N1 in 2009.

Several previous studies have explored the potential for diagnosis of infection with H1N1 and H3N2 by extracting whole blood RNA samples to monitor changes in host gene expression in response to infection [4, 7, 8]. There is evidence that host gene signatures can provide a reliable method of pre-symptomatic detection for viral ARIs, including influenza A [9]. In two previous studies, healthy volunteers were inoculated with either H3N2 [7] or H1N1 [4] and monitored for development of illness. Blood RNA samples were taken for both groups prior to infection and again periodically after inoculation in order to monitor the temporal response of the host to each disease. This genetic data was analyzed for possible markers to differentiate infected individuals from controls [4, 7, 10].

The goal of this study is to explore the ability of different machine learning algorithms to discriminate control samples collected prior to inoculation, and samples associated with symptomatic subjects at the very earliest stages of infection. We focus on developing a pathway based approach that assembles collections of genes associated with a particular biological mechanism. Using the previously mentioned human

viral challenge data sets, we focus on distinguishing between infected and healthy individuals using data collected during the first 29 hours of both studies. This time-frame is well before the onset of the symptoms that generally occur at 40-60 hours.

2. THE INFLUENZA DATA SETS

Table 1: Distribution of subjects from the H1N1 and H3N2 data sets. For each subject, there are 2 pre-infection samples and 14 time dependent post-infection samples.

Data Set	Controls (C^-)	Symptomatic Infected (C^+)
H1N1	30 samples	9 samples
H3N2	29 samples	9 samples
H1N1 & H3N2	59 samples	18 samples

We conduct data learning experiments on two microarray data sets collected in association with disease challenges with human subjects, as summarized in Table 1. The first experiment consists of 24 human subjects inoculated with the H1N1 strain of influenza A (A/Brisbane/59/2007) [4]. The second challenge involved 17 human subjects who were inoculated with H3N2 (A/Wisconsin/67/2005) [7]. In both studies, it was deemed that certain data samples had irregularities, i.e. the clinical presentation was inconsistent with the results of viral diagnostic tests. Such data has been removed from consideration. Thus, as done in previous studies, for H1N1 a total of nine subjects were omitted, while for H3N2 two subjects samples were not included in the analysis [4]. Each data set thus consists of 9 symptomatic infected and 6 asymptomatic uninfected subjects. The samples from symptomatic uninfected and asymptomatic infected subjects were discarded. All subjects had peripheral blood samples taken prior to inoculation with virus ($t = -5$ and $t = 0$ hours) and at specific intervals following inoculation. For both H1N1 and H3N2, the actual time points of data sampling belong to the set $t = \{-5, 0, 5, 12, 21.5, 29, 36, 45.5, 53, 60, 69.5, 77, 84, 93.5, 101, 108\}$ hours.[1]

For the control group, we use pre-infection data consisting of samples taken at $t = -5$ and $t = 0$ for all subjects in both studies. The controls are referred to as the C^- class. The data used for the infected group, or C^+ class, consists only of subjects classified as both symptomatic and infected. Experiment one uses C^+ data taken from the H1N1 data set only, experiment two uses data from the H3N2 data set only, and experiment 3 uses C^+ data from both data sets. C^+ is analyzed for $t = 5$, 21.5, and 29 hours.

In Figure 1, we visualize trajectories of the cAMP signaling pathway associated with asymptomatic (blue) and symptomatic (red) subjects. There is evidence that this pathway plays a role in the immune response; in this paper, we report that it serves as a signature for H1N1 infection using nonlinear SVM with polynomial kernel of degree 2 at $t = 5$ hours with 100% accuracy on the available data.

Biological Pathways

As originally proposed in [11], our data analysis is based on the exploration of the time evolution of biological pathways rather than single gene expression levels. A pathway consists of approximately 10 - 100 genes, encapsulated into a single high dimensional trajectory that evolves in time. The motivation for this stems from the fact that pathways, as multivariate biological units reflecting specific function, capture a stronger signal than single genes and, based on our experience, enable higher classification accuracy.

[1]These datasets may be downloaded at: http://people.ee.duke.edu/∼lcarin/reproduce.html.

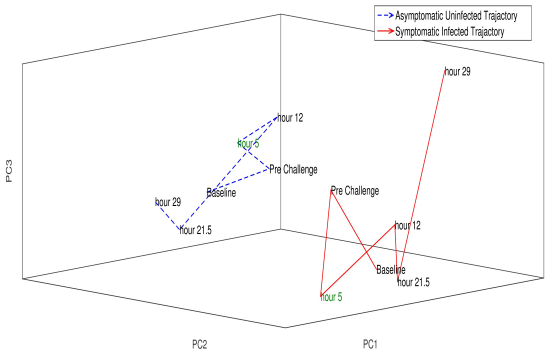

Figure 1: Trajectories of the cAMP pathway associated with asymptomatic (blue) and symptomatic (red) subjects. The coordinate system consists of the first three principal components of PCA.

A total of 1439 time-evolving pathways are assembled from the micro-array data sets. This is accomplished by mapping the genes from the available data matrix to the appropriate subsets of genes which comprise the biological pathways. These pathways capture the complex interactions between genes associated with biological processes, including metabolism and the immune response to infection. The pathways that form the backbone of our analysis are based on information from the Broad institute database [12, 13], comprising:

- 217 BioCarta pathways
- 295 KEGG pathways
- 10 Matrisome pathways
- 196 Pathway interaction database (PID) pathways
- 674 Reactome pathways
- 10 SigmaAldrich pathways
- 8 Signaling gateway pathways
- 28 Signal Transduction KE pathways
- 1 SuperArray pathway

The information required to assemble these pathways is available at http://software.broadinstitute.org.

3. METHODOLOGY

Here we briefly overview the methodology employed for classifying the data. It is interesting that no single method is superior in every experiment, underlining the need for a full toolbox of approaches to explore high-dimensional data sets.

One of the most powerful methods involves the mathematical framework of the Grassmann manifold and the computation of angles between subspaces. We use the known technique of support vector machine (SVM), both linear with C = 1 [14, 15] and nonlinear with a polynomial kernel of degree two [16]. In addition, we apply a feature selection version of SVM which we refer to as *sparse* support vector machines. This linear technique performs an *in situ* subset selection of the genes performing the classification.

3.1 Sparse Support Vectors Machines

An arbitrary norm separating hyperplane was proposed in [17]. We employed the ℓ_1-norm hyperplane to induce sparsity and *in situ* feature selection [18, 19, 11]. Here we are interested in solving the l_1 norm support vector machine optimization problem

$$\text{minimize} \, \|w\|_1 + Ce^T y \tag{1}$$

subject to the constraints $D(Zw - ge) + y \geq e$, $y \geq 0$, where $w \in \mathbb{R}^n, g \in \mathbb{R}$, and $e, y \in \mathbb{R}^m$. The class combined data matrix is $Z \in \mathbb{R}^{m \times n}$, where each row is an observation, and D is the matrix of binary class labels.

We use the substitutions $|w_i| = w_i^+ + w_i^-$ and $w_i = w_i^+ - w_i^-$ with the imposed constraints $w_i^+, w_i^- \geq 0$; see [19]. Since the scalar g is a free variable, we rewrite it $g = g^+ - g^-$ where $g^+, g^- \geq 0$. Our decision variable is then

$$x^T = [w^+ w^- g^+ g^- y] \in \mathbb{R}^{2n+2+m}$$

The associated cost vector is $c^T = [e_n \; e_n \; 0 \; 0 \; Ce_m]$ So we are solving the linear programming problem

$$\text{maximize} \, c^T x$$

subject to the side conditions $Ax \geq b, \; x \geq 0$ where

$$A = [DX \;\; -DX \;\; -De \; De \; I] \tag{2}$$

and $b = e_m$. See also [20]. The linear optimization problem which is solved by SSVM can be solved in polynomial time by interior point methods [21].

3.2 Classification on Grassmannian

The geometric framework of the Grassmann manifold has proven very effective for capturing the variability in large, complex data sets [22, 23, 24, 25, 26]. For example, one subspace angle is enough to separate the CMU-PIE data set as described in [27]. Facial recognition is possible at ultra low-resolution on the Grassmannian [28]. We now briefly describe the methodology, leaving the reader seeking more details to consult the references above.

We assume the data comes from one of two classes: asymptomatic control subjects C^- and symptomatic subjects C^+. We associate with each experiment a data matrix $X = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(p)}]$ with C^+ samples and a data matrix $Y = [\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(q)}]$ with C^- samples. Here $\mathbf{x}^{(i)}$, $\mathbf{y}^{(j)} \in \mathbb{R}^m$ and m is the ambient dimension of the data. In the context of pathway analysis, the dimension m is the number of genes in the pathway. For each matrix X (and Y), we apply singular value decomposition (SVD) on them respectively, where

$$X = U_X \Sigma_X V_X^T \quad \text{and} \quad Y = U_Y \Sigma_Y V_Y^T$$

Here the columns of U_X and U_Y are the left singular vectors of X and Y, which form a basis of the subspace spanned by columns of X and Y, respectively. We then pick d columns of U_X and U_Y corresponding to the d largest singular values to represent the subspace spanned by columns of X and Y. Note that for this experiment, we pick d to be 5 (H1N1),4 (H3N2) and 4 (H1N1 and H3N2 combined). These d left singular vectors constitute the d dimensional optimal basis of that class, $\mathcal{B}_X$ and $\mathcal{B}_Y$.

To classify a test point $\mathbf{t} \in \mathbb{R}^m$, we calculate its principal angle from the optimal basis of X and Y, $\mathcal{B}_X$ and $\mathcal{B}_Y$, respectively. The point is classified in the class with which its principal angle is smaller. The accuracy showed in this paper is obtained by using leave-one-out method for each experiment and then computing the balanced success rate. Leave-one-out cross validation technique has been used with repetition of 30 times to calculate the balanced success rate.

Table 2: Balanced accuracy for H1N1.

Method	t = 5 h	t = 21.5 h	t = 29 h
Sparse SVM	96.66	89.44	96.66
Linear SVM	91.11	83.89	92.78
Nonlinear SVM	100	85.56	92.78
Subspace	88.0	82.67	87.78
ANN	92.39	85.94	94.67
k-nearest neighbors	83.88	76.66	81.66

Table 3: Balanced accuracy for H3N2.

Method	t = 5 h	t = 21.5 h	t = 29 h
Sparse SVM	94.44	92.02	89.27
Linear SVM	87.16	92.03	91.0
Nonlinear SVM	94.23	90.30	91.0
Subspace	82.0	89.66	89.27
ANN	85.38	95.23	93.16
k-nearest neighbors	81.60	88.57	90.99

3.3 Artificial Neural Networks

Artificial Neural Networks is a state-of-the-art machine learning technique for classification and dimensionality reduction [29, 30, 31]. In this paper, we used a network architecture with two hidden layers each comprising of 10 hidden units with hyperbolic tangent as the non-linear transfer function. In training phase, Scaled Conjugate Gradient method [32] is used to minimize the cross-entropy error [33] of the network. In any conjugate direction search method there is an extra overhead of line search to figure out the step size in each iteration. But in scaled conjugate gradient method the line search is avoided by using the Levenberg-Marquardt approach to scale up step size. This makes this algorithm considerably faster than other second order methods like BFGS, CGL etc.

4. RESULTS

We conduct three experiments to test the accuracy of the pathways for early prognosis. The machine learning exploration focuses on three tasks, i.e., Experiment 1: H1N1 data only, Experiment 2: H3N2 only and Experiment 3: H1N1 and H3N2 combined. In all our experiments, we use samples from 15 subjects at time -5, 0 and t_i hours, where t_i can be 5, 21.5 or 29 hours. We split the data into two classes, *healthy* and *sick*. We use the cross-validation leave-one-out

Table 4: Balanced accuracy for the combined data sets H1N1 and H3N2.

Method	t = 5 h	t = 21.5 h	t = 29 h
Sparse SVM	88.27	83.99	83.56
Linear SVM	82.72	79.36	91.05
Nonlinear SVM	84.17	80.30	85.40
Subspace	76.55	81.80	80.65
ANN	81.62	79.87	80.55
k-nearest neighbors	67.60	70.08	79.33

technique to test our models.

4.1 Classification rates

We show the balanced success rates in Tables 2-4 using six methods including linear, nonlinear, and sparse SVM, in addition to k-nearest neighbors, classification on the Grassmannian, and artificial neural networks (ANN). It is noteworthy that sparse and nonlinear SVM take turns at having the best performance, with the exception of $t = 21.5$ hours and $t = 29$ hours in the H3N2 experiment where ANN performs best. Linear SVM and ANN are both uniformly accurate across all experiments, while the subspace method and k-nearest neighbors are consistently less accurate. Experiment 1 with H1N1 is more accurate than Experiment 2 with H3N2, and both are more accurate than the combined experiment.

4.2 Pathway Analysis

In this section, we summarize the top pathways for each experiment at each time point. The best pathways at each time point are examined for all experiments below.

In the first experiment on the H1N1 data only, the cAMP signaling pathway classifies with 100% accuracy at $t = 5$ hours with the nonlinear SVM model. It has been established that the cAMP signaling molecule is a secondary signaling molecule involved in immune system regulation. It acts on protein kinase A [34]. The pentose and glucuronate interconversions pathway performed best at $t = 21.5$ hours, with a BSR of 89.44% for the sparse SVM method. This is a metabolic pathway for carbohydrates. The NF-kappa B signaling pathway achieved a BSR of 96.66% for the sparse SVM method at $t = 29$ hours. It includes of set of transcription factors that regulate genes involved in immune response. Nuclear factor-kappa B has been found to play a primary physiological role in the immune system [35].

For the second experiment, using only the H3N2 data, the top pathway at $t = 5$ hours is the reactome developmental biology pathway. It is selected by sparse SVM with a BSR of 94.44%. This pathway is involved in cell differentiation and transcriptional regulation of development of blood cell components. At $t = 21.5$ hours, the purine metabolism pathway is selected by sparse SVM with a BSR of 92.02%. This is a metabolic pathway for the nucleic acid purine that has been shown to play a role in immune response [36]. The selenocompound metabolism pathway is a metabolic pathway for building certain amino acids and is selected by both ANN and SSVM as the top pathway at $t = 29$ hours. It has a BSR of 89.27% using sparse SVM and 93.16% using ANN. It is a small pathway containing only 17 genes that has very little overlap with other pathways. However, this pathway shares the genes PAPSS1 and PAPSS2 with the purine metabolism pathway. These two genes are associated with the protein 3'-phosphoadenosine 5'- phosphosulfate synthesis, which produces a sulfotranferase involved in viral entry into cells [37].

For the combined H1N1 and H3N2 experiment, the PID CD8 TCR pathway performed best at $t = 5$ hours with a BSR of 88.27%; this pathway was selected by the sparse SVM method. It is related to T cells, which are involved in the suppression of excessive immune response [38]. The PID PDGFRB pathway was selected as the best pathway at $t = 21.5$ hours by sparse SVM, with a BSR of 83.99%. This protein is platelet-derived growth factor receptor beta, and the pathway is part of a system of negative immune system regulators that prevent the immune response from cascading out of control [39]. The acute myeloid leukemia pathway is the top pathway selected by linear SVM for the combined data experiment, with a BSR of 91.05%. It is associated with leukemia and contains proteins that up-regulate cell survival genes. The PID LIS1 pathway yields a BSR of 95.23% at $t = 29$ hours for ANN on the H3N2 data set. This pathway is related to motor proteins and neuronal migration and affects cranial development [40].

We note that each classification method addresses the geometry of the data differently and as a consequence may reveal distinct optimal pathways.

4.3 Sensitivity and Specificity

Sensitivity and specificity are basic descriptive statistics that measure the quality of a decision function, or model, for a two-class classification problem. In this problem, we refer to symptomatic infected samples as the positive class and the controls as the negative class. Sensitivity is a measure of the fraction of symptomatic infected samples that are correctly classified, while specificity is the fraction of control samples correctly classified. The average of the sensitivity and specificity is referred to as the balanced success rate (BSR) and is a measure of classification accuracy that is especially effective when the number of elements of one class differs substantially from that of the other.

In Figure 2, we see the distribution of sensitivities and specificities for the H1N1 pathway classifiers based on the sparse support vector machine model. Every point corresponds to a pathway in the sensitivity-specificity plane. The 9 levels of quantization along the y-axis are due to the fact that we have 9 symptomatic subjects while the 30 levels of quantization on the x-axis come from the 30 samples from all subjects at $t = -5, 0$. We highlight the top sensitive (red) and specific (blue) pathways. We see a general trend in this figure, as well as the other sensitivity-specificity plots, where the points are shifted to the right. This is because pathways that contain no information about the immune response can't be used discriminate sick and healthy subjects; these pathways simply indicate the subject is in the same class as the controls. This makes sense from the point of view that many pathways concern basic biological functions unrelated to the host immune response. In Figure 3, we see a similar pattern for H1N1 (above) and H3N2 (below) at time $t = 29$ hours.

4.4 Pathway variation

In Figures 4-6 we show the variation of the given pathway accuracies over the the life of the H1N1 experiment. In Figure 4, we select five pathways which are determined to be optimal using sparse SVM for $t = 5$ hours. We see that the

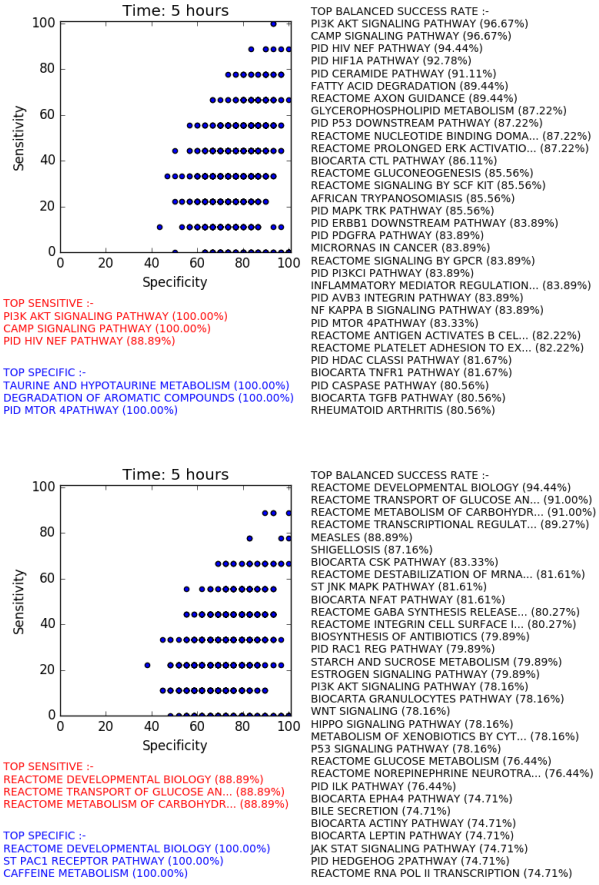

Figure 2: The sensitivity and specificity of biological pathways using a sparse support vector machine classifier to discriminate between symptomatic and control subjects at 5 hours post infection. Top: H1N1, Below: H3N2.

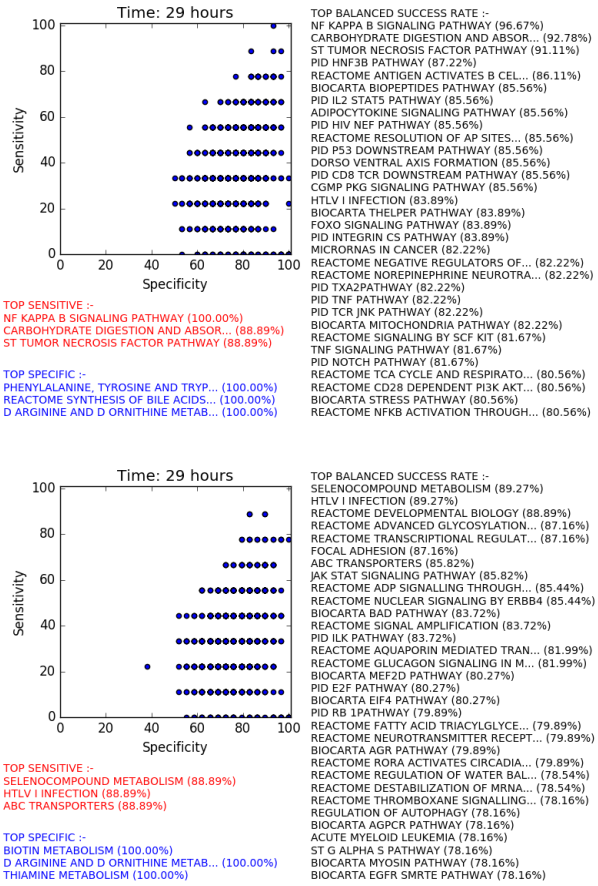

Figure 3: The sensitivity and specificity of biological pathways using a sparse support vector machine classifier to discriminate between H1N1 symptomatic and control subjects at 29 hours post infection. Top: H1N1, Below: H3N2.

accuracy of these pathways drops to 50%-60% by $t = 21.5$ hours. The situation is similar for top pathways at $t = 21.5$ hours, as seen in Figure 5, with the top five pathways again selected by sparse SVM classification. In Figure 6, the accuracy of the top pathways at $t = 29$ hours is lower at approximately 85%. This suggests that the question of prognosis is heavily dependent on what stage the subject is at, i.e. the best pathways depend on the time elapsed since exposure.

5. PRIOR WORK

We now briefly review the statistical and machine learning literature related to experiments that have been conducted on the H1N1 and H3N2 data sets considered here. We note that some of these experiments include additional data not considered in the current investigation. The focus of this paper exclusively concerns influenza. The experiments can be divided into those that ignore temporal evolution and those that account for the time-dependent nature of the data.

5.1 Time-Independent Studies

Zaas et al. [7] used the H3N2 dataset in development of

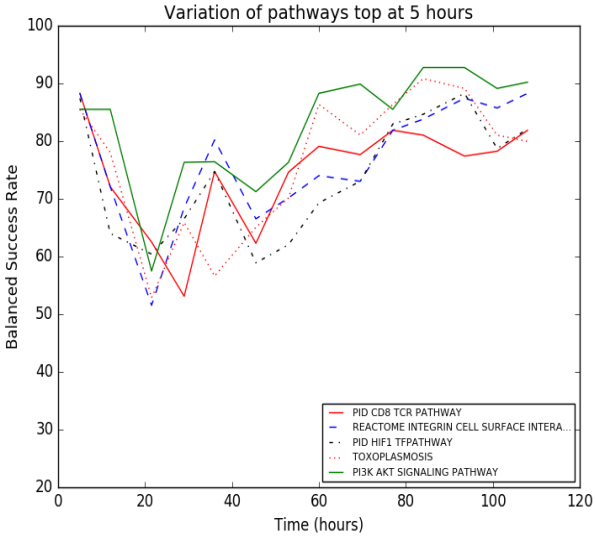

Figure 4: The balanced success rate as a function of time post infection for the pathways that are optimal at $t = 5$ hours post infection by H1N1. Pathways selected using sparse SVM.

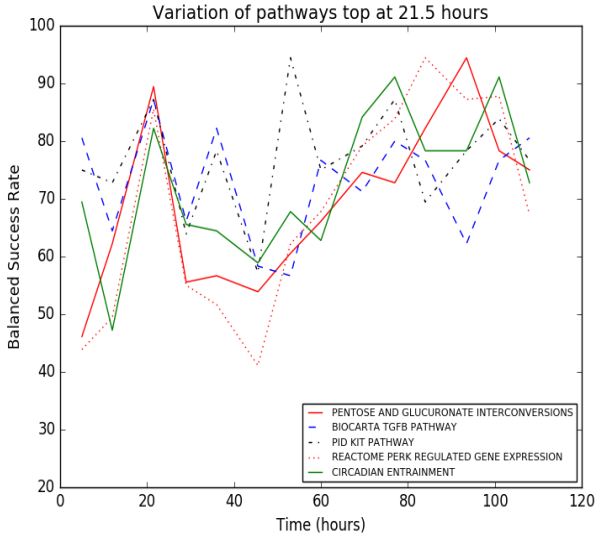

Figure 5: The balanced success rate as a function of time post infection for the pathways that are optimal at $t = 21.5$ hours post-infection by H1N1. Pathways selected by sparse SVM.

a gene expression signature for identifying HRV, RSV, and influenza. All data sets were combined and analyzed as a whole using leave-one-out cross validation, then further analyzed by training on one data set and testing on the other two. When validated on the pediatric influenza data set from [41], the gene signature was able to accurately classify all subjects. The ability of the gene expression signature to differentiate between influenza and bacterial infection was also tested, with a reported accuracy of 80% [7].

A later study of the data sets from Zaas et al. [7] utilized a biomarker discovery method and SVM models to classify subjects as infected or uninfected using 10-fold-cross-validation [42, 43]. Estimated predicted accuracy of this method was stated to be 0.94 AUC. Analysis of all data samples produced a 12 gene signature with a predicted accuracy of 0.99 AUC [42].

Davenport et al. [10] performed PCA on data from their own experiments involving H3N2 vaccination and infection and developed a six gene signal for classification and tested it on the H3N2 dataset from [8], using data ranging from 48 to 69 hours for infected individuals. This signal was able to correctly classify all controls and 89% of the infected individuals [10].

In 2013, Zaas et al. [44] used both the H1N1 and the H3N2 data set to develop and test an RT-PCR assay classifier for influenza. The classifier was first trained and validated separately on each data set using leave-one-out cross-validation. This resulted in 0% error for H3N2 and 13% error for H1N1. The classifier was then trained using one data set and validated on the second data set. Training on H3N2 and testing on the full H1N1 set yielded a 17% classification error, and testing on the H1N1 data set with five ambiguous individuals omitted yielded 6.7% error. Training on the full H1N1 data set and testing on H3N2 resulted in a 0% error, whereas training on the H1N1 data set with the same five individuals removed resulted in 13% error. Finally, the two data sets were combined and randomly partitioned 100 times, with 50% used for training and the rest used for testing. The average AUC was 0.975 for all runs [44]. The classifier was then tested on a set of 102 individuals, including 35 controls, 28 with viral infections, and 39 with bacterial infections, to determine its ability to correctly identify respiratory viral infections. Of those with viral infections, 25 had H1N1 and the remaining 3 were infected with rhinovirus. 89% of the subjects with viral infections were identified correctly, and 10.3% of the bacterial infections and 3% of the controls were incorrectly labeled as viral [44].

Tzu-Yu Liu et al. [45] designed a reference-aided classification algorithm based on the viral challenge study model by learning sparse linear score functions in a multi-block multiclass SVM and found a smaller test panel without sacrifice of classification accuracy. Each subject in the dataset was designated as a symptomatic subject (Sx) or an asymptomatic subject (Asx) and as an infected subject (Inf) or uninfected subject (UnInf) on the basis of symptom scores and the viral shedding measurements. Then, each of them was labeled as one of the five stages: i) baseline before inoculation, ii) Asx and UnInf, iii) Sx and pre-acute, iv) Sx and acute and v) Sx and post-acute. The reference-aided predictor achieved an average (cross-validated) state prediction accuracy improvement of: 14% for RSV, 13% for H3N2, 9% for HRV, and 6% for H1N1 with the highest prediction accuracy of: 36.5% for RSV, 61.4% for H3N2, 51.7% for HRV, and 52% for H1N1. Additionally, this gain in accuracy was achieved with a smaller panel of genes: 60% fewer for RSV, 39% fewer for H3N2, 20% fewer for HRV, and 31% fewer for H1N1 [45].

5.2 Time-Dependent Studies

The studies above are primarily focused on examining the data sets as a whole. The experiments described in this

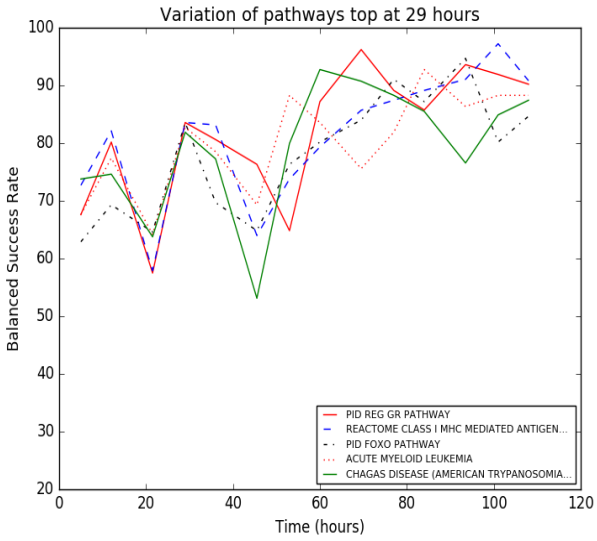

Figure 6: The balanced success rate as a function of time post-infection for the pathways that are optimal at $t = 29$ hours post infection by H1N1. Pathways selected by sparse SVM.

section examine the role of the time dependence of the gene expression data.

Woods et al. [46] identified gene signatures for both H1N1 and H3N2. Without accounting for time, applying the H3N2 factor to the H1N1 data set resulted in correct identification of 93% of infected subjects, and applying the H1N1 factor to the H3N2 data correctly identified 100% of the infected subjects. The H3N2 factor has a sensitivity of 89% near 53 hours and increases to 100% by 69 hours. H1N1 factor achieves 89% sensitivity by around 60 hours. The influenza factor was validated on patients hospitalized with H1N1-like symptoms as well as healthy individuals for the control group, and was able to correctly identify 92% of the infected subjects and 93% of the controls.

Rose et al. [47] developed a viral gene expression factor based on the data from [7] using data from 3 to 4 days after infection. Temporal changes in the viral expression factor were then compared with changes of a platelet gene expression signature in both H1N1 and H3N2 and examined for correlations [47].

Huang et al. [8] use the H3N2 data set to analyze temporal changes in gene expression during infection using self-organizing map clustering to identify significant genes.

A recent publication by McClain et al. [9] uses the H3N2 data set to identify a gene signature that is significant as early as 24 hours after infection. Expression of this gene signature was monitored in two groups of patients that received either standard treatment for influenza or early treatment. Expression of this gene signature peaks around 50 hours for both groups, but declines more rapidly for the early treatment group [9]. This suggests that pre-symptomatic detection of influenza using host genetic markers is possible and that early treatment affects disease progression.

Linel et al. [48] constructed ODE models for dynamic response genes (DRGs) and observed clear differences in the number of significant DRGs between the symptomatic and asymptomatic subjects (in symptomatic case, the number of DRGs is significantly larger). DRG signatures for symptomatic subjects with influenza infection were identified using the ODE model. The false discovery rate is controlled at 0.05.

In more recent work, we have demonstrated that anomaly detection algorithms provide a promising technique of analysis to reveal signatures of the immune response to respiratory viruses [49]. That work is distinguished from this current study in that no class labels were used from the symptomatic patients to build the models. In contrast, here we use labels from both classes in our supervised data learning algorithms.

6. CONCLUSIONS

We conducted three experiments for evaluating the potential of early prognosis of infection using a pathway based analysis and supervised learning. We establish that classification with high accuracy is possible at the earliest stages of infection including as soon as 5 hours after exposure to the pathogen. The prognosis pathway found at $t = 5$ is the cAMP signaling pathway known to be involved with immune system regulation. Surprisingly, we observe that classification accuracy actually goes down over the first 24 hours, in particular with with nonlinear and sparse SVM. This may be due to complex interactions among multiple biological pathways. Further analysis is required to better understand this downward evolution of classification accuracy.

The top two pathways for detecting H3N2 at $t = 21.5$ hours and $t = 29$ hours, purine metabolism and selenocompound metabolism, respectively, have the genes 3'-phosphoadenosine 5'-phosphosulfate synthesis 1 and 2 (abbreviated as PAPSS1, PAPSS2) in common. These genes have been found to be involved with viral entry into cells as described above

The results here provide evidence consistent with the idea that the immune response to infection is really a cascade of biological defense mechanisms. We found the top pathways were very much time dependent. The pathways used for prognosis at $t = 5$ hours will be different than the pathways used for prognosis at $t = 21.5$ hours or $t = 29$ hours. More work needs to be done to establish the pattern of the evolution of pathway activity and to determine signatures characteristic of successful defense against the invading pathogen.

7. ACKNOWLEDGMENTS

This paper is based on research partially supported by the National Science Foundation under Grants No. DMS-1513633, and DMS-1322508. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] Nelson Lee, Chun Kwok Wong, Paul KS Chan, Martin CW Chan, Rity YK Wong, Samantha WM Lun, Karry LK Ngai, Grace CY Lui, Bonnie CK Wong, Sharon KW Lee, et al. Cytokine response patterns in severe pandemic 2009 h1n1 and seasonal influenza among hospitalized adults. *PLoS One*, 6(10):e26050, 2011.

[2] Noelle-Angelique M Molinari, Ismael R Ortega-Sanchez, Mark L Messonnier, William W Thompson, Pascale M Wortley, Eric Weintraub, and Carolyn B Bridges. The annual impact of seasonal influenza in the us: measuring disease burden and costs. *Vaccine*, 25(27):5086–5096, 2007.

[3] Sebastian L Johnston. Natural and experimental rhinovirus infections of the lower respiratory tract. *Am J Respir Crit Care Med*, 152:546–S52, 1995.

[4] Christopher W Woods, Micah T McClain, Minhua Chen, Aimee K Zaas, Bradly P Nicholson, Jay Varkey, Timothy Veldman, Stephen F Kingsmore, Yongsheng Huang, Robert Lambkin-Williams, et al. A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza h1n1 or h3n2. *PloS one*, 8(1):e52198, 2013.

[5] Matthew J Memoli, David M Morens, and Jeffery K Taubenberger. Pandemic and seasonal influenza: therapeutic challenges. *Drug discovery today*, 13(13):590–595, 2008.

[6] Alfredo Chiarini, Angelo Palmeri, Teresa Amato, Rita Immordino, Salvatore Distefano, and Anna Giammanco. Detection of bacterial and yeast species with the bactec 9120 automated system with routine use of aerobic, anaerobic, and fungal media. *Journal of clinical microbiology*, 46(12):4029–4033, 2008.

[7] Aimee K Zaas, Minhua Chen, Jay Varkey, Timothy Veldman, Alfred O Hero, Joseph Lucas, Yongsheng Huang, Ronald Turner, Anthony Gilbert, Robert Lambkin-Williams, et al. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell host & microbe*, 6(3):207–217, 2009.

[8] Yongsheng Huang, Aimee K Zaas, Arvind Rao, Nicolas Dobigeon, Peter J Woolf, Timothy Veldman, N Christine Øien, Micah T McClain, Jay B Varkey, Bradley Nicholson, et al. Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. *PLoS Genet*, 7(8):e1002234, 2011.

[9] Micah T McClain, Bradly P Nicholson, Lawrence P Park, Tzu-Yu Liu, Alfred O Hero, Ephraim L Tsalik, Aimee K Zaas, Timothy Veldman, Lori L Hudson, Robert Lambkin-Williams, et al. A genomic signature of influenza infection shows potential for presymptomatic detection, guiding early therapy, and monitoring clinical responses. In *Open forum infectious diseases*, volume 3, page ofw007. Oxford University Press, 2016.

[10] Emma E Davenport, Richard D Antrobus, Patrick J Lillie, Sarah Gilbert, and Julian C Knight. Transcriptomic profiling facilitates classification of response to influenza challenge. *Journal of Molecular Medicine*, 93(1):105–114, 2015.

[11] Stephen O'Hara, Kun Wang, Richard A Slayden, Alan R Schenkel, Greg Huber, Corey S O'Hern, Mark D Shattuck, and Michael Kirby. Iterative feature removal yields highly discriminative pathways. *BMC genomics*, 14(1):832, 2013.

[12] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[13] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, et al. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267–273, 2003.

[14] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[15] Terrence S Furey, Nello Cristianini, Nigel Duffy, David W Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[16] Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support vector machines and kernels for computational biology. *PLoS Comput Biol*, 4(10):e1000173, 2008.

[17] OL Mangasarian. Arbitrary-norm separating plane. *Operations Research Letters*, 24(1):15–23, 1999.

[18] Kun Wang, Vineet Bhandari, Sofya Chepustanova, Greg Huber, O Stephen, SO Corey, Mark D Shattuck, Michael Kirby, et al. Which biomarkers reveal neonatal sepsis? *PloS one*, 8(12):e82700, 2013.

[19] Sofya Chepushtanova, Christopher Gittins, and Michael Kirby. Band selection in hyperspectral imagery using sparse support vector machines. In Miguel Velez-Reyes and Fred A. Kruse, editors, *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XX*, volume 9088 of *Proc. of SPIE*.

[20] Michael C Ferris, Olvi L Mangasarian, and Stephen J Wright. *Linear programming with MATLAB*, volume 7. SIAM, 2007.

[21] Arkadi S Nemirovski and Michael J Todd. Interior-point methods for optimization. *Acta Numerica*, 17:191–234, 2008.

[22] Bruce Draper, Michael Kirby, Justin Marks, Tim Marrinan, and Chris Peterson. A flag representation for finite collections of subspaces of mixed dimensions. *Linear Algebra and its Applications*, 451:15–32, 2014.

[23] Tim Marrinan, Bruce Draper, J Ross Beveridge, Michael Kirby, and Chris Peterson. Finding the subspace mean or median to fit your need. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1082–1089. IEEE, 2014.

[24] Tim Marrinan, J Ross Beveridge, Bruce Draper, Michael Kirby, and Chris Peterson. Flag manifolds for the characterization of geometric structure in large data sets. In *Numerical Mathematics and Advanced Applications-ENUMATH 2013*, pages 457–465. Springer, 2015.

[25] Kun Wang, Josh Thompson, Chris Peterson, and Michael Kirby. Identity maps and their extensions on

parameter spaces: Applications to anomaly detection in video. In *Science and Information Conference (SAI), 2015*, pages 345–351. IEEE, 2015.

[26] Sofya Chepushtanova, Michael Kirby, Chris Peterson, and Lori Ziegelmeier. Persistent homology on grassmann manifolds for analysis of hyperspectral movies. In *International Workshop on Computational Topology in Image Context*, pages 228–239. Springer, 2016.

[27] J.R. Beveridge, Bruce Draper, Jen-Mei Chang, Michael Kirby, Holger Kley, and Chris Peterson. Principal angles separate subject illumination spaces in ydb and cmu-pie. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29, 2008.

[28] Jen-Mei Chang, Michael Kirby, Holger Kley, Chris Peterson, Bruce Draper, and J Ross Beveridge. Recognition of digital images of the human face at ultra low resolution via illumination spaces. In *Computer Vision–ACCV 2007*, pages 733–743. Springer, 2007.

[29] David E Rummelhart, James L McClelland, PDP Research Group, et al. Parallel distributed processing. explorations in the microstructure of cognition. volume 1: Foundations, 1986.

[30] James L McClelland, David E Rumelhart, PDP Research Group, et al. *Parallel distributed processing*, volume 2. MIT press Cambridge, MA, 1987.

[31] Michael Kirby. *Geometric data analysis: an empirical approach to dimensionality reduction and the study of patterns*. John Wiley & Sons, Inc., 2000.

[32] Martin Fodslette Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, 6(4):525–533, 1993.

[33] Douglas M Kline and Victor L Berardi. Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Computing & Applications*, 14(4):310–318, 2005.

[34] Sarah E Fiedler, Amelia R Kerns, Catherine Tsang, Vivian Tsang, Dennis Bourdette, and Sonemany Salinthone. Dimethyl fumarate activates the prostaglandin ep2 receptor and stimulates camp signaling in human peripheral blood mononuclear cells. *Biochemical and biophysical research communications*, 475(1):19–24, 2016.

[35] MS Hayden, AP West, and S Ghosh. Nf-κb and the immune response. *Oncogene*, 25(51):6758–6780, 2006.

[36] HA Simmonds, GS Panayi, and V Corrigall. A role for purine metabolism in the immune response: Adenosine-deaminase activity and deoxyadenosine catabolism. *The Lancet*, 311(8055):60–63, 1978.

[37] Eli Chapman, Michael D Best, Sarah R Hanson, and Chi-Huey Wong. Sulfotransferases: structure, mechanism, biological activity, inhibition, and synthetic utility. *Angewandte Chemie International Edition*, 43(27):3526–3548, 2004.

[38] Shimon Sakaguchi, Tomoyuki Yamaguchi, Takashi Nomura, and Masahiro Ono. Regulatory t cells and immune tolerance. *Cell*, 133(5):775–787, 2008.

[39] Jingjing Tang, Koichi Kozaki, Andrew G Farr, Paul J Martin, Per Lindahl, Christer Betsholtz, and Elaine W Raines. The absence of platelet-derived growth factor-b in circulating cells promotes immune and inflammatory responses in atherosclerosis-prone apoe-/- mice. *The American journal of pathology*, 167(3):901–912, 2005.

[40] Ji Yeoun Lee, Ae-Kyung Park, Eun-Sun Lee, Woong-Yang Park, Sung-Hye Park, Jung Won Choi, Ji Hoon Phi, Kyu-Chang Wang, and Seung-Ki Kim. mirna expression analysis in cortical dysplasia: regulation of mtor and lis1 pathway. *Epilepsy research*, 108(3):433–441, 2014.

[41] Octavio Ramilo, Windy Allman, Wendy Chung, Asuncion Mejias, Monica Ardura, Casey Glaser, Knut M Wittkowski, Bernard Piqueras, Jacques Banchereau, A Karolina Palucka, et al. Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood*, 109(5):2066–2077, 2007.

[42] Alexander Statnikov, Lauren McVoy, Nikita Lytkin, and Constantin F Aliferis. Improving development of the molecular signature for diagnosis of acute respiratory viral infections. *Cell host & microbe*, 7(2):100, 2010.

[43] Alexander Statnikov, Nikita I Lytkin, Lauren McVoy, Jörn-Hendrik Weitkamp, and Constantin F Aliferis. Using gene expression profiles from peripheral blood to identify asymptomatic responses to acute respiratory viral infections. *BMC research notes*, 3(1):264, 2010.

[44] Aimee K Zaas, Thomas Burke, Minhua Chen, Micah McClain, Bradly Nicholson, Timothy Veldman, Ephraim L Tsalik, Vance Fowler, Emanuel P Rivers, Ronny Otero, et al. A host-based RT-PCR gene expression signature to identify acute respiratory viral infection. *Science translational medicine*, 5(203):203ra126–203ra126, 2013.

[45] Tzu-Yu Liu, Thomas Burke, Lawrence P Park, Christopher W Woods, Aimee K Zaas, Geoffrey S Ginsburg, and Alfred O Hero. An individualized predictor of health and disease using paired reference and target samples. *BMC bioinformatics*, 17(1):1, 2016.

[46] Christopher W Woods, Micah T McClain, Minhua Chen, Aimee K Zaas, Bradly P Nicholson, Jay Varkey, Timothy Veldman, Stephen F Kingsmore, Yongsheng Huang, Robert Lambkin-Williams, et al. A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1 or H3N2. *PloS one*, 8(1):e52198, 2013.

[47] Jason J Rose, Deepak Voora, Derek D Cyr, Joseph E Lucas, Aimee K Zaas, Christopher W Woods, L Kristin Newby, William E Kraus, and Geoffrey S Ginsburg. Gene expression profiles link respiratory viral infection, platelet response to aspirin, and acute myocardial infarction. *PloS one*, 10(7):e0132259, 2015.

[48] Patrice Linel, Shuang Wu, Nan Deng, and Hulin Wu. Dynamic transcriptional signatures and network responses for clinical symptoms in influenza-infected human subjects using systems biology approaches. *Journal of pharmacokinetics and pharmacodynamics*, 41(5):509–521, 2014.

[49] Kun Wang, Stanley Langevin, Corey S. O'Hern, Mark D. Shattuck, Serenity Ogle, Adriana Forero, Juliet Morrison, Richard Slayden, Michael G. Katze,

and Michael Kirby. Anomaly detection in host signaling pathways for the early prognosis of acute infection. *PLoS ONE*, 11(8):1–26, 08 2016.