Florida International University

FIU Digital Commons

7-11-2016

Predicting symptom severity and contagiousness of respiratory viral infections

Medhini Narasimhan
National Institute of Technology, India, mnarasim@fiu.edu

Giuseppe Vietri
Bioinformatics Research Group, Florida International University, gvietri@fiu.edu

Vanessa Aguiar-Pulido
Bioinformatics Research Group, Florida International University

Arpit Mehta
Bioinformatics Research Group, Florida International University

Farid Rajabli
Bioinformatics Research Group, Florida International University

See next page for additional authors

Follow this and additional works at: https://digitalcommons.fiu.edu/com_facpub

Part of the Medicine and Health Sciences Commons

Recommended Citation

Authors

Medhini Narasimhan, Giuseppe Vietri, Vanessa Aguiar-Pulido, Arpit Mehta, Farid Rajabli, Kalai Mathee, and Giri Narasimhan

Predicting Symptom Severity and Contagiousness of Respiratory Viral Infections

Medhini Narasimhan[1,2], Giuseppe Vietri[2], Arpit Mehta[2], Farid Rajabli[2], Vanessa Aguiar- Pulido[2], Kalai Mathee[3], Giri Narasimhan[2]

[1]National Institute of Technonology Karnataka, India
[2]Bioinformatics Research Group (BioRG), School of Computing and Information Sciences, Florida International University
[3]Herbert Wertheim College of Medicine, Florida International University

INTRODUCTION

Infections due to respiratory viruses affect millions of people all over the world and have a huge economic impact. While the proc ess of immune clearance allows most people to combat these infections, for many others viral exposure causes a variety of symptoms including runny nose, stuffed nose, cough, sore throat, headache, fever, myalgia and general malaise. These symptoms can vary in severity and have different onset and recovery times. To make matters worse, the viruses reproduce and "shedding" ensues, whereby the viral progeny are expelled making the host contagious. **The goal of this work is to build predictive models for both severity of symptoms and contagiousness, given gene expression time series data recorded over a multi-day period starting prior to exposure, and measured at different intervals following exposure.**
Our predictive models resulting from data from prior to exposure performed nearly as well as reported models with data from 29 hours post infection. Performance rose to 100% when using later time points. We have identified several biomarkers, which emerged as being significant from models for multiple time points.
The potential biomarkers obtained with the proposed approach need to be investigated further as vaccine and therapy targets.

DATA

- Data represents **4 different respiratory viruses**, including Respiratpry Syncytial Virus (**RSV**), H3N2, H1N1 and Rhinovirus.
- Healthy volunteers were followed for seven to nine days following controlled nasal exposure to one respiratory virus. Subjects enrolled into these viral challenge experiments had to meet several inclusion and exclusion criteria.
- Nasal lavage samples and symptom data and were collected from each patient on a repeated basis over the course of 7-9 days. Viral infection was quantified by measuring release of viral particles from nasal passages ("viral shedding") as assessed from nasal lavage samples via qualitative viral culture and/or quantitative influenza RT-PCR.
- Symptomatic data was collected through self-report on a repeated basis. Symptoms were assessed via modified Jackson score which assessed the severity of 8 upper respiratory symptoms (runny nose, cough, headache, malaise, myalgia, sneeze, sore throat and stuffy nose) and integrates daily scores over 5-day windows.
- Blood was collected and gene expression of peripheral whole blood was performed 1 day (24 to 30 hours) prior to exposure, immediately prior to exposure, and at regular intervals following exposure. All patients challenged with influenza (H1N1 or H3N2) received oseltamivir 5 days post-exposure. Rhinovirus additionally includes 7 volunteers who were exposed to sham rather than active virus. Below is a summary of the number of subjects for each catgory along with the number who showed viral shedding and were symptomatic or asymptomatic.
- The above data was provided by the DREAM Respiratory Virus Challenge [http://dreamchallenge.org]

Virus	Subjects	Symptomatic	Viral Shedding
RSV	20	10	13
H1N1	36	19	21
H3N2	30	16	18
Rhinovirus	38	25	22

Table 1. Clinical data summary

APPROACH

PLS-DA

Random Forest Classifier

| Time Series Gene Expression from Subjects | Feature Selection | Training the Classifier | Prediction | Annotation Enrichment Analysis |

Feature Selection

- Partial Least Squares Discriminant Analysis (PLS-DA) is an extension of the multiple linear regression models and is a supervised method to filter relevant genes as biomarkers for estimating the symptom score and viral shedding.
- PLS-DA sharpens the separation between groups of labeled observations by rotating the frame of reference to a direction that maximizes the separation between the groups. It also provides information on the class separating variables.
- In this work, PLS-DA was used to score the genes that contribute to the best separation between groups of subjects, in our case symptomatic vs asymptomatic and shedding vs non-shedding. A threshold was used to filter relevant genes to be used by the classifier.

PLS-DA feature selection for H1N1

Classification

- In this work, a random forest classifier was used whose input was the gene list obtained after the filtering in feature selection. It is an ensemble method that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and to control over-fitting.
- 10-Fold Cross validation was used.

RESULTS

- Table on the left shows genes that act as biomarkers at different time points during infection by the RSV virus, as identified by PLS-DA.
- Rows correspond to important genes and columns correspond to time points.
- Each entry in the table is the rank of that gene in the list of importance scores. Brighter colors indicate higher ranks. "NA"s indicate that the gene did not receive a sufficiently high importance score.
- Some genes act as biomarkers across the entire time range of study. Others are either early or late biomarkers.
- The first column gives the name of the gene in question.
- The last column represents the functional annotations of the genes.
- Related annotations are given the same color.

- Table on the right shows accuracies of the random forest classifiers constructed for different time points.
- Accuracy and the number of genes used to build the model are shown for three sample time points and for the time point at which the most accurate model was constructed.

		0 Hours		10 Hours		36 Hours		Best Time		
	Virus	Accuracy	# Genes	Accuracy	# Genes	Accuracy	# Genes	Time Point	Accuracy	# Genes
Viral Shedding	Rhinovirus	85.2	138	90.2	9	78.3	397	84	94.7	1000
	H1N1	81.5	1000	78.7	1000	78.9	1000	94	99.9	834
	H3N2	89.0	216	77.8	1000	95.0	98	34	100.0	3
	DEE1 RSV	86.8	592	87.8	58	90.3	1000	53	95.3	149
Symptom	Rhinovirus	74.3	875	84.1	26	78.8	4	108	95	912
	H1N1	86.3	166	80.4	1000	82.2	82	132	100	10
	H3N2	87.9	393	88.9	146	95.3	18	69.5	100	70
	DEE1 RSV	87.0	79	94.5	1000	95.0	234	69.5	99	149

Enrichment Analysis for GO:0006952~defense response

	RSV Day 3 PI	H1N1 Day 4 PI	H3N2 Day 2-3 PI	Rhinovirus Day 4 PI
Symptoms	1.3E-05	3.7E-4	9.2E-7	1.5E-7
Shedding	2.0E-3	3.7E-4	5.0E-4	1.4E-10

The table on the left shows Enrichment Analysis for a functional term of interest for all four viruses. The scores represent the Benjamini Hochberg Corrected P-values.

CONCLUSIONS

- Predictive models for viral shedding and symptoms were constructed.
- While predictions are near perfect at later time points, they are reasonably high even at much earlier time points.
- Significant genes were detected as early as 5 and 10 hours post infection (PI), as compared to prior work that did so at 29 hours PI.
- Biomarkers were identiied for all viruses, both unique and shared.
- Genes for defense and immune response were differentially expressed in all four viruses.
- The functional annotation term "cell division" was significant for H3N2. Also, translational genes were differentially expressed.
- Genes annotated with proteolysis were differentially expressed in subjects with and without shedding.
- For Rhinovirus, innate immune response genes are activated early.

REFERENCES

1. Barker, Matthew, and William Rayens. "Partial least squares for discrimination." Journal of chemometrics 17.3 (2003): 166-173.
2. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
3. DAVID TOOLS: https://david.ncifcrf.gov/
4. G-Profiler: http://biit.cs.ut.ee/gprofiler/