Merging Multiple Data Sets to Study HCV

Jungmin Han[1], Rabab Ali[2], Grace Zhang[2], Elizabeth Townsend[2], Gabriella Quinn[2], Kareen Hill[2],
Ohad Etzion[2], Theo Heller[2], and Vipul Periwal[1]

Short Abstract — **Hepatitis C virus (HCV) is a pervasive public health problem that infects three to four million every year worldwide, many of whom will develop cirrhosis [1]. We aim to better understand the key attributes of HCV infection, which in turn may give insight into the disease prognosis. Here, we show how we applied similarity network fusion (SNF), a recently developed computational method that is particularly suited to obtaining a comprehensive integrated understanding of multiple types of measurements, to HCV clinical data collected from 29 patients.**

Keywords — **HCV infection, similarity network fusion, cirrhosis.**

I. Purpose

Approximately 170 million people suffer from chronic hepatitis C infection worldwide. There is no preventive vaccine for HCV. Although in some the infection has a good prognosis, a significant number of infected people develop more lethal health conditions such as cirrhosis or liver cancer [1]. To better understand the pathogenesis of HCV infection, we collected a range of clinical data from a group of patients with a spectrum of disease stages. We not only measured several different classes of biomarkers including cytokines, metabolites, and bacterial genomics, but also documented the patients' diet records. Given the diverse data types, it was not obvious how to collate embedded information and draw integrated conclusions. In order to overcome this challenge, we employed SNF, a method developed for coalescing multiple data sets emerged from a common source [2, 3, 4].

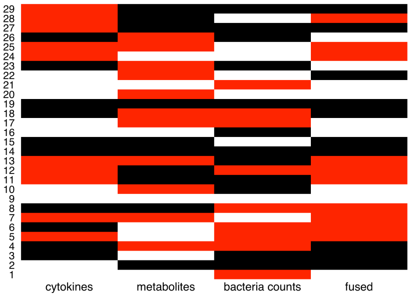
For the preliminary analysis, we focused our attention to the three biomarkers mentioned above and inspected for any separations within the group using spectral clustering. There was no common clustering pattern across the data sets. However, some of the clusters showed a positive association with the degree of liver damage, which emphasizes the necessity of an exhaustive analysis.

II. Data and Method

The data set consisted of traces of 64 different cytokines, the total of 1546 metabolites from 25 subgroups, and bacteria counts spanning 13 phyla. There were some missing entries, which were imputed using the R software package, *softImpute* [3, 5]. Since the metabolite data were partitioned into 25 subgroups, they required amalgamation into a single network through SNF, before it could be combined with other networks.

For each biomarker data set, the method computes an affinity matrix that quantifies the global similarities between the subjects and a kernel matrix representing local similarities between *k*-neighboring patients. As the final step, the SNF method iteratively updates each affinity matrix through a matrix multiplication of its kernel matrix with the other affinity matrices to obtain a unified network that conveys both shared and complementary information. All the sets had an equal weight towards the final network. We then applied the spectral clustering algorithm on the fused network to determine clusters among patients, and the same method was used on each data set for the comparison.

Figure 1. The clustering patterns within each biomarker data set, which are not consistent across the whole set. The clusters are represented in three colors. The last column shows the clusters identified from the converged network of all three data sets. The numbers on the left are patient numbers.



III. Conclusion

The information in the cytokine data set seems to be well carried through, compared to the other data sets. This suggests that the network structure underlying the cytokine data is dominant over the others.

Acknowledgements: This work was funded by Intramural Research Program of the NIH, NIDDK.

[1]Laboratory of Biological Modeling, National Institute of Diabetes and Digestive and Kidney Diseases, NIH. E-mail: jungmin.han@nih.gov, vipulp@niddk.nih.gov

[2]Liver Disease Branch, National Institute of Diabetes and Digestive and Kidney Diseases. E-mail: rabab.ali@nih.gov, grac3.zhang@gmail.com, ltowns22@gmail.com, gabriella.quinn@nih.gov, kareen.hill@nih.gov, ohadet34@yahoo.com, theoh@intra.niddk.nih.gov

References

[1] WHO (2017) *Global Hepatitis Report 2017*. Geneva: World Health Organization.

[2] Wang B, et al. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nature methods* **11**, 333-337.

[3] R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[4] Wang B, et al. (2017) SNFtool: Similarity Network Fusion. R package version 2.2.1. https://CRAN.R-project.org/package=SNFtool

[5] Hastie T and Mazumder R (2015) softImpute: Matrix Completion via Iterative Soft-Thresholded SVD. R package version 1.4. https://CRAN.R-project.org/package=softImpute