IMPaC-TB: Integrated analysis and dynamical systems modeling of experimental TB immunology data

Brooke Anderson PhD, Mike Lyons PhD, Amy Fox, Burton Karger

2019-12-12

Contents

This is a laboratory guide book for the CSU IMPAC-TB experiments.

Chapter 1

Introduction

The overall objective of the data analysis and mathematical modeling component for the CSU mouse immunology experimental studies is to develop an iterative framework to identify key biological components of immunity and to quantify their relationships to one another in both data-driven and mechanistic models for the purpose of evidence-based decision-making for tuberculosis (TB) vaccine development. The data sharing plan (DSP) will include the experimental data, the quantitative analysis framework, and an application programming interface (API) to the results. As detailed in our respective biosketches, we have established expertise in all critical areas of this data analysis project.

The host immune response to TB vaccination and infection is complex and involves interactions between large networks of molecular and cellular constituents that vary in time and location within the host. The experimental data will be generated from a wide range of measurements and across multiple scales; including measures of disease pathology, cellular and chemical measurements for cell type and cytokine concentrations, and intracellular measurements involving RNA expression and proteomics. The conceptual basis for our proposed data analysis and modeling framework is described in a summary of the recent National Institute of Allergy and Infectious Diseases (NIAID) workshop, 'Complex Systems Science, Modeling and Immunity' [1]. The major components of this framework are illustrated in Figure 1, where our approach will integrate experimental data with data-driven modeling to identify significant correlations and possible causal structures among the data elements, and with mechanistic modeling of cell-mediated immunity that translates biologically-based hypotheses into a dynamical system of time-dependent mathematical equations that can be used to simulate and test these hypotheses and to inform the design of subsequent experiments.

The proposed work plan begins with the collection and organization of quantitative and qualitative CSU generated experimental data that will then be used for data-driven and mechanistic modeling, with the analysis results and software

modeling tools being made available through a web-based API. The milestones of this project are: (1) establishing protocols and standardized documentation for data collection and pre-processing from each CSU experimental type, (2) construction of the relational database (RDB) for CSU-generated experimental data, (3) collection of qualitative data describing key immune features as input for mechanistic modeling, (4) development of single-type data-driven analysis tools for each separate experimental system, (5) development of integrated data-driven analysis tools for the combined experimental data, (6) development of a dynamical systems model of cell-mediated immunity based on qualitative analysis results, (7) development of parameter estimation and model calibration procedures for the dynamical systems model, (8) development of software tools that provide for a start-to-finish process framework, and (9) development of an API for public access to relevant data and results. For each milestone, the gates for Go/No Go decisions will be based on the positive reproducibility of the major results by each of the individual CSU investigators. This approach will ensure the integration and quality control of each component within the entire framework.

Immunology data is collected from each CSU mouse experiment as both quantitative measurements and qualitative data that includes hypotheses regarding key biological constituents in the context of TB vaccine development. An RDB will be developed to provide access and queries to all combined data sets. Data-driven modeling will proceed directly from the quantitative data while mechanistic modeling will begin with the qualitative data, with the two modeling approaches increasingly informing each other as analyses proceed. The data-driven integrated data analysis will include visualization and statistical analyses and will also inform parameter estimation for the dynamical systems model. Software tools will be developed for all quantitative data and results, including user testing. A tailored user interface will provide access to all data and analysis results.

Methods. A series of three dedicated network database and computational data analysis servers will be configured for both internal (CSU) and remote access to experimental data and to the generated software. The first in this series will be a low cost, minimal architecture, development server for use during the base period. This development server will be then be scaled for intermediate development to establish resource requirements for an expected maximum workload, followed by a final production server running the completed workflow applications as the project deliverable. During the base funding period, we will be collecting sample datasets from each of the mouse TB infection experiments to catalog the data and metadata that are being collected. In subsequent funding periods, we will standardize the (meta)data collected from these groups and devise a database strategy to store these research outputs. This will include appropriate scaling for computer hardware and data storage options. Once the (meta)data is harmonized, we will construct the relational database. We intend to leverage open source database software like QPortal [5] and openBIS [6] which are designed to house this type of multi-omics data, to streamline this process.

The final version of the web server will be housed centrally by Colorado State University's Academic Computing and Networking Services division and maintained by qualified IT staff. These systems will also provide a web interface and an API interface that will allow users to query the data generated by this project, satisfying the data sharing plan. As data are ready to

be published, they will be deposited to the appropriate NCBI Database (such as Bioproject, biosample, and SRA) to increase discoverability. One possible unusual expense would occur in the event that the data stored locally is lost. In this case, we would incur egress fees from the cloud storage provider of 1-3 cents per gigabyte of data, depending on the speed of retrieval. The data-driven modeling will proceed during the base funding period to develop protocols for computationally reproducible data collection, pre-processing, analysis, and data-driven modeling of all data to be collected under the grant. These protocols will be developed using the rmarkdown [7] framework, to combine code with documentation describing all processing, analysis, and data-driven modeling choices. These protocols will be made publicly available and, once established, will be used throughout the project to ensure all experimental data are consistently and reproducibly processed and analyzed. These protocols will incorporate existing open source software tools, including xcms [8], ramclustR [9], flowCore [10], and openCyto [11]. Further, we will develop our own software tools to complement these existing tools for the purposes of our research, including novel tools for visualizing and integrating different types of data (e.g., metabolomics and flow cytometry). The final result of this data-driven modeling component will be data type-specific (e.g., identification of key metabolites) and integrative across data types (e.g., quantification of notable associations between specific metabolites and cell populations).

The mechanistic model component will be developed as a hierarchy of dynamical systems models of cell mediated immunity to TB infection with the basic approach described in previous models [2,3]. During the base funding period, qualitative data will be used to identify biologically important factors for the host immunity in each of the animal models under investigation. Common factors will be represented as graphs, and quantitative experimental data gaps needed to establish parameter estimates will be identified. Once appropriate experimental data is available, parameter estimation will be performed using Bayesian hierarchical modeling, and model simulations will be conducted using Monte Carlo methods to account for model uncertainty and population variability. Anticipated difficulties include additional gaps in experimental data that limit model identifiability, and hypotheses regarding the model elements that lead to model predictions that substantially disagree with corresponding experimental results. These issues will require additional iterations of model development, including possible new and targeted experimental studies. Schedule. The schedule for completion and delivery of items specified in the statement of work is shown in the table below, where arrows denote the duration and expected completion date.

References [1] Vodovotz Y, Xia A, Read EL, Bassaganya-Riera J, Hafler DA, Sontag E, Wang J, Tsang JS, Day JD, Kleinstein SH, Butte AJ, Altman MC, Hammond R, Sealfon SC. (2017) Solving Immunology? Trends Immunol. 38(2):116-127. [2] Friedman A, Turner J, Szomolay B. (2008) A model on the influence of age on immunity to infection with Mycobacterium tuberculosis. Exp Gerontol. 43(4):275-85. [3] Wigginton JE, Kirschner D. (2001) A model to predict cell-mediated immune regulatory mechanisms during human infection with Mycobacterium tuberculosis. J Immunol. 2001 Feb 1;166(3):1951-67. [4] Gideon HP, Skinner JA, Baldwin N, Flynn JL, Lin PL. (2016) Early Whole Blood Transcriptional Signatures Are Associated with Severity of Lung Inflammation in Cynomolgus Macaques with Mycobacterium tuberculosis Infection. [5] Mohr C, Friedrich A, Wojnar D, Kenar E, Polatkan AC, et al. (2018) qPortal: A platform for data-driven biomedical research. PLOS ONE 13(1): e0191603.https://doi.org/10.1371/journal.pone.0191603 [6] Barillari C, Ottoz DSM, Fuentes-Serna JM, Ramakrishnan C, Rinn B, Rudolf F. (2016) openBIS ELN- LIMS: an open-source database for academic laboratories. Bioinformatics, 32(4): 638–640, https://doi.org/10.1093/bioinformatics/btv606 [7] JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng and Winston Chang (2018). rmarkdown: Dynamic Documents for R. R package version 1.9. htt ps://CRAN.Rproject.org/package=rmarkdown [8] Smith, C.A., Want, E.J., O'Maille, G., Abagyan,R., Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. Analytical Chemistry, 78, 779–787. [9] Broeckling, C.D., Afsar, F.A., Neumann, S., Ben-Hur, A., Prenni, J.E. (2014). RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data. Analytical Chemistry, 86(14), 6812-6817. [10] Ellis B, Haaland P, Hahne F, Le Meur N, Gopalakrishnan N, Spidlen J, Jiang M (2018). flowCore: flowCore: Basic structures for flow cytometry data. R package version 1.46.1. [11] Finak, Greg, Frelinger, Jacob, Jiang, Wenxin, Newell, Evan W., Ramey, John, Davis, Mark M., Kalams, Spyros A., De Rosa, Stephen C., Gottardo, Raphael (2014). "OpenCyto: An Open Source Infrastructure for Scalable, Robust, Reproducible, and Automated, End-to-End Flow Cytometry Data Analysis." PLoS Computational Biology, 10(8), e1003806.

Chapter 2

Day -247

The purpose of ths timepoint is to identify the cage numbers with the treatments and create a short cage id.

Note: in future experiments, we will tag the mice at this timepoint.

The data template files can be found here

Chapter 3

Day -242

The purpose of this timepoint is to shave the mice before vaccination and to find the iniital total weight of the mice in the cage before vaccination.

The data template files can be found here

Chapter 4

Day -240

The purpose of this timepoint is to perform the first round of vaccinations.

The data template files can be found here

Chapter 5

Day -233

The purpose of this timepoint is to get the toal cage weight post-vaccination.
The data template files can be found here