Genetic Algorithms for model refinement and rule discovery in a high-dimensional agent-based model of inflammation

Chase Cockrell and Gary An, Department of Surgery, University of Vermont

The University of Vermont
LARNER COLLEGE OF MEDICINE

Abstract

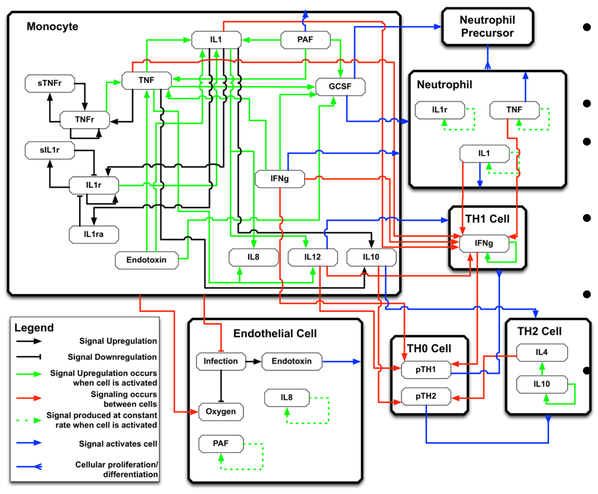
Introduction: Agent-based modeling frequently used modeling method for multi-scale mechanistic modeling. However, the same properties that make agent-based models (ABMs) well suited to representing biological systems also present significant challenges with respect to their construction and calibration, particularly with respect to the large number of free parameters often present in these models. The challenge of dealing with parameters is further exacerbated due to the fact that a great deal of phenotypic and clinical heterogeneity can be attributed to intrinsic genetic/epigenetic variation manifesting as functional parameter variation. We propose that various machine learning (ML) and evolutionary computing approaches (such as genetic algorithms (GAs)) can be used to more effectively and efficiently deal with parameter space characterization.

Methods: This project uses a GA to fit a previously validated ABM of acute systemic inflammation, the Innate Immune Response ABM (IIRABM) to clinical time series data of systemic cytokine levels. The genome for the GA is a vector generated from the IIRABM's Model Rule Matrix (MRM), which is a matrix representation of not only the constants/parameters associated with the IIRABM's cytokine interaction rules, but also the existence of rules themselves. Capturing heterogeneity is accomplished by a fitness function that incorporates the sample value range ("error bars") of the clinical data.

Results: The GA-enabled parameter space exploration resulted in a set of putative MRM parameterizations which closely (though not perfectly) match the cytokine time course data used to design the fitness function.

Conclusion: We present an HPC-enabled evolutionary computing approach that utilizes a GA to calibrate a complex ABM to clinical data while preserving biological heterogeneity. The integration of machine learning/evolutionary computing, HPC and multi-scale mechanistic modeling provides a pathway forward to more effectively represent the heterogeneity of clinical populations and their data.

Innate Immune Response Agent-Based Model (IIRABM)

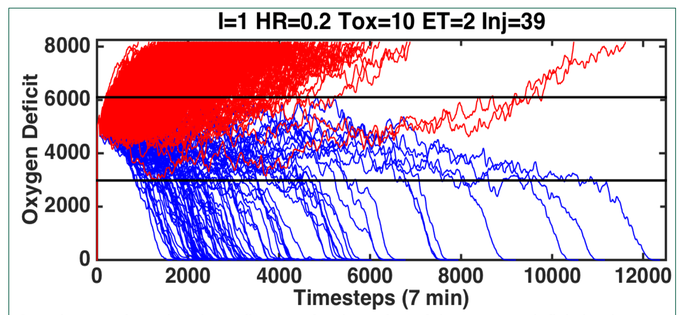


- Agent-Based Model of the human innate immune response to insult/infection/injury
- Spatially Explicit, Stochastic, Nonlinear
- Calibrated to represent clinical trajectory of patient with penetrative, infectious trauma
- Includes endothelial cells, macrophages, neutrophils, Th0, Th1, and Th2 cells, as well as their associated precursors
- Cytokine signaling network used to determine cellular interactions and behaviors

Used to investigate sepsis, a dysregulation of inflammatory signaling network dynamics, affecting ~1 million people/year with a mortality rate of approximately 40%

Background – Modeling Philosophy

- There are two principal factors which lead to heterogeneity in biological data: genetic variability between individuals and stochastic responses to identical stimuli.
- In order for a computational model to provide maximum utility, it must be able to recreate the same range of variability *in silico* that is seen *in vivo*.
- To capture this heterogeneity, the dependence of model output on both the model content (parameterization of internal model rules) and model context (description of the environment in which a biomedical simulation operates) must be quantified.



One thousand stochastic replicates of trajectories of the oxygen deficit (an inverse measure of human health) are displayed for a single model parameterization and identical injury. Trajectories diverge over time due to stochastic variabilities in model responsiveness.

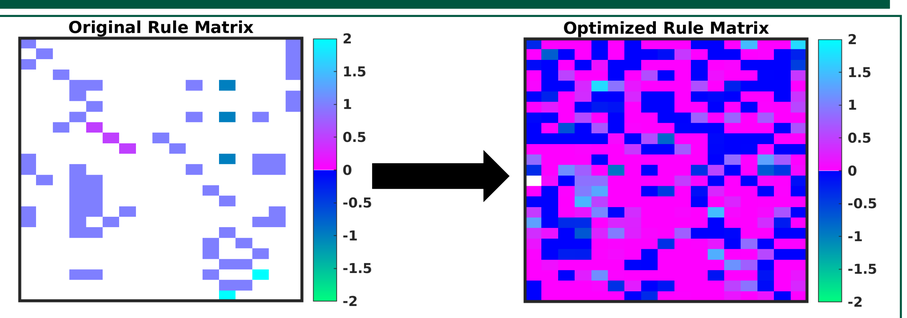
Genetic Algorithms on Model Rule Matrix

Model rules are represented in the model rule matrix (MRM): rules are rows, the magnitude of contributions from relevant entities (i.e., a cytokine concentration at a given location) are represented in the columns

- Hypothetical rule set: $\left.\begin{matrix} IL10_{t+1} = IL10_t + TNF_t \\ TNF_{t+1} = -IL10_t + IFNg_t \end{matrix}\right\} \Rightarrow \begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$
- Decompose matrix into vector to apply genetic algorithm
- Set fitness function to maximize data coverage: $F = \sum_{i,t} |\max(c_{i,t}^{e}) - \max(c_{i,t}^{m})| + k|R_e - R_m|$
- Employ nonviability filter: if a given MRM parameterization leads to death prior to the first experimental time point for all replicates, that parameterization is discarded prior to breeding
- Employ elitism: on each generation, discard the least fit 10% of breeders, replace with most fit 10%
- Use standard continuous GA crossover operator: $C_{1,i} = \beta P_{1,i} + (1-\beta)P_{2,i}$, $C_{2,i} = \beta P_{2,i} + (1-\beta)P_{1,i}$

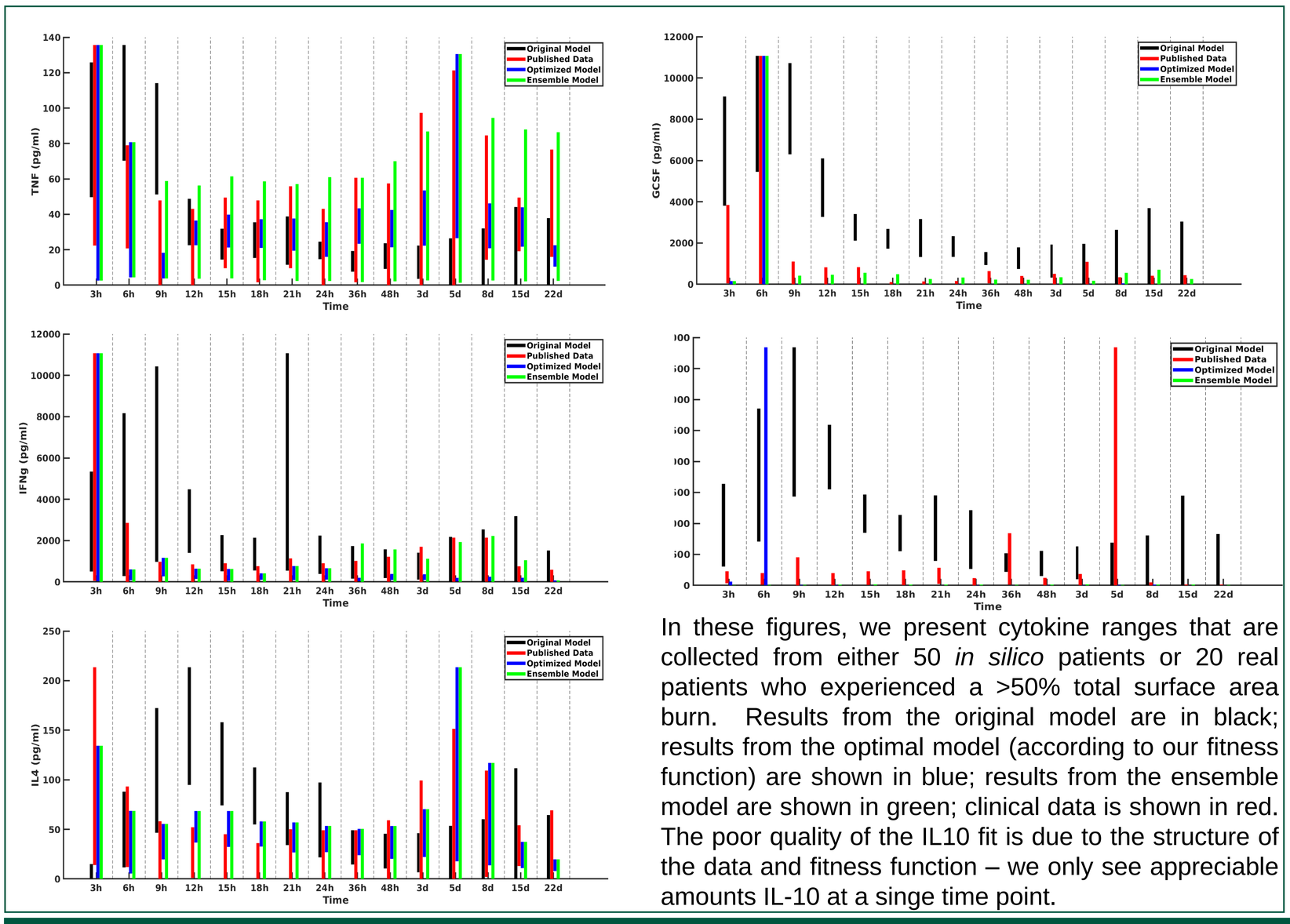
Ensemble Construction

- A single MRM parameterization is analogous to a single genome –thus, all variability comes from inherent stochasticity
- In order to construct a more biologically realistic cohort, we store any MRM parameterization discovered during the GA optimization which lies within the boundaries of clinical data

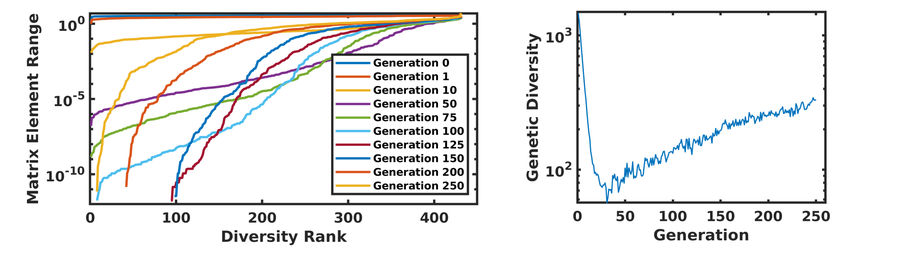


Above, we compare the original rule matrix to the optimized rule matrix. The optimized matrix has a much more connected structure, and is a dense matrix, as opposed to the sparse original rule matrix. There are not any matrix elements with a value of 0 in the optimized matrix, though there are many elements with comparatively small values. This structure is similar to what is seen in experimental bioinformatic studies; all of the cytokines in this network appear to be connected to each other, at least to a small degree, while a smaller number of strong connections (which could also be considered correlations) provide the majority of the influence on the system dynamics.

Results



In these figures, we present cytokine ranges that are collected from either 50 *in silico* patients or 20 real patients who experienced a >50% total surface area burn. Results from the original model are in black; results from the optimal model (according to our fitness function) are shown in blue; results from the ensemble model are shown in green; clinical data is shown in red. The poor quality of the IL10 fit is due to the structure of the data and fitness function – we only see appreciable amounts IL-10 at a singe time point.



On the left, the ranges for each individual element in the MRM are displayed at various generations during the GA. On the right, total diversity, expressed as the sum of ranges of matrix elements, is displayed as a function of the generation. We use a mutation rate that increases with time in order to combat premature convergence and to increase the volume of space explored. An increasing mutation rate combined with elitism and the ensemble criterion ensures we keep the best solutions while maximizing the volume of space explored.

Conclusions

Machine-learned models are constrained by the data upon which they are trained. In order to develop clinically useful and generalizable ML models, near-comprehensive data is required; however, it is not tractable to acquire this data experimentally. In order to capture, *in silico*, the heterogeneity that is seen in actual biomedical systems, machine learning techniques (in this case, GA), can be used in conjunction with high-performance computing and agent-based modeling to optimize a fitness function predicated upon capturing heterogeneity for an agent-based model. Future work will focus on refining the construction of appropriate fitness functions such that viable parameterizations are not excluded.

Acknowledgments

This work was supported by MSM U01 Grant Number: **EB025825-01**. Additionally, this research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.