ENHANCED SYMBOLIC REGRESSION TO INFER BIOCHEMICAL NETWORK MODELS

Nicole Beauregard, University of Connecticut, USA
nicole.beauregard@uconn.edu
Mark Marten, University of Maryland Baltimore County, USA
Steven Harris, University of Manitoba, Canada
Ranjan Srivastava, University of Connecticut, USA

Key words: symbolic regression, genetic programming, biological networks, systems biology

Biological systems can be represented as complex networks illustrating the relationships and connections among biochemical species. Complex networks can uncover vital information regarding specific pathways or network bottlenecks, helping to reveal novel discoveries relevant to a variety of applications. Biological networks, however, are often highly interconnected and non-linear in nature making development of a comprehensive model challenging. Large amounts of data can be acquired to elucidate specific pathways, but deducing the entire network topology requires more rigorous computational techniques. There are *in silico* techniques, including evolutionary algorithms, to predict network topologies using information from experimental data. Biological networks can be decomposed into a system of differential equations under mass action kinetics assumptions describing the rate of change of the various biochemical species in the network. Symbolic regression can be used to generate a system of equations from acquired data.

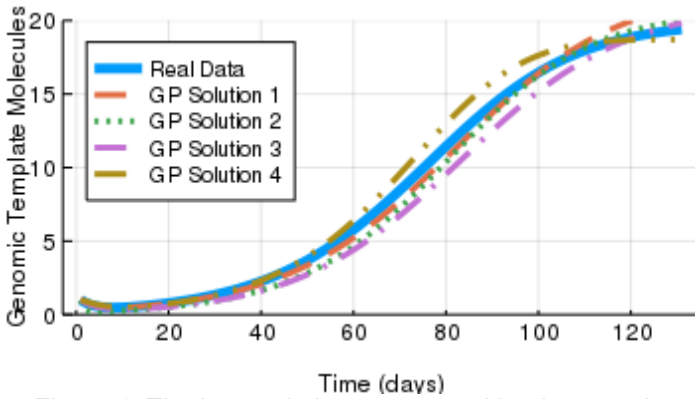

Figure 1: The best solutions generated by the genetic program for viral DNA replication dynamics show good agreement with the data generated by the original model.

In this work, we employed *genetic programming*, a stochastic optimization method, to generate an ensemble of symbolically regressed equations describing intracellular viral kinetics. Due to the highly inter-connected and nonlinear nature of these systems, it can often be computationally infeasible to find a solution for an unconstrained system. To address these hurdles, the complexity of the differential equations and the search space for the kinetic parameters were constrained. First, we assumed that the kinetic equations regressed could only be zero, first, or second order. Higher order equations are rarely observed in nature and therefore excluded to narrow the pool of potential reaction combinations. To limit the search space for rate constants, data acquired from the system were used to make an assisted stochastic guess. From the data, in addition to the actual value for each species at each time point, it was possible to determine an approximate derivative value at each time point. It was then possible to backcalculate a range of potential rate constants and randomly choose one from that range. An initial guess selected from the ranges generated were used to estimate the parameters using simulated annealing for the entire differential equation system.

Instead of only considering the best solution from each simulation, information about the system from an ensemble of models was gathered. Once the genetic program completed, stability analysis was employed to extract only stable and practical solutions from the set of best models. As shown in Figure 1, results of the best models generated via genetic programming bounded the results of the original model. By evaluating the ensemble of equations, it was possible to look for terms that appeared in the majority of these equations. The more frequently terms appeared, the more confident we were that the relationship was part of the real network. Our simulations were run for 100 generations and were completed on a standard desktop computer in less than a day using simulated annealing for parameter estimation. Other parameter estimation algorithms tested took significantly longer, including the Nelder-Mead algorithm which took approximately seven times as long on the same computer. In the future, we plan to extend this algorithm to more complicated systems including multi-omic networks where extensive data can be collected to discover or corroborate different multi-omic mechanisms.