Integrating Residual Networks and Symbolic Regression for Biological ODE System Modeling

Abstract

This research presents a novel approach to biological Ordinary Differential Equation (ODE) system modeling, combining Residual Networks (ResNet) with symbolic regression techniques. Our methodology is designed to simulate biological systems more realistically than traditional numerical fitting methods. By utilizing logarithmic and exponential functions within the ResNet, our model effectively mirrors manual equation formulation while maintaining compatibility with neural network principles. The model's performance was evaluated using data from two distinct studies, demonstrating its ability to accurately capture both linear dynamics and complex, periodic patterns in gene regulatory networks. This innovative approach offers a promising path for exploring systems biology, highlighting the potential of merging diverse methodologies to advance our understanding of complex biological systems.

Introduction

The endeavor to decipher the intricacies of complex biological systems remains a pivotal challenge in the realm of systems biology. This quest has historically necessitated the formulation of ordinary differential equations (ODEs) to simulate these systems. Such an approach, however, often encounters obstacles due to the intrinsic complexity of biological processes and the presence of unascertained variables. [1] Complications arising from a partial comprehension of biological networks have led to the creation of models that might not accurately reflect the actual dynamics of these systems. [2]

The field of systems biology has undergone significant evolution, integrating a range of innovative methodologies to enhance modeling methods for biosystems. A key advancement in this realm is the adoption of symbolic regression. This technique plays a crucial role in discerning complex patterns and relationships within intricate datasets. Symbolic regression has shown effectiveness in uncovering meaningful correlations in biological data, thereby enriching our understanding of biological processes. Its application extends to equation-free modeling, a method enabling the study of multiscale phenomena without the reliance on detailed models. [3-4]

Recent advancements in systems biology have been characterized by a blend of data-driven inverse Ordinary Differential Equation (ODE) methodologies and the integration of artificial neural networks (ANNs), including deep neural networks (DNNs). [5-8] These innovative approaches, diverging from traditional modeling techniques, use observational data to deduce and replicate ODEs governing biological phenomena, offering new insights in biological modeling. Simultaneously, the rise of ANNs, particularly DNNs, has revolutionized data analysis in multiple disciplines by adeptly identifying complex patterns in

extensive datasets, thus significantly enhancing our understanding of biological systems.

In this study, a novel method is presented that integrates physical principles with neural network and symbolic regression techniques to address traditional challenges in Ordinary Differential Equation (ODE) modeling in biology. The approach utilizes a log and exp function pair in a residual network (ResNet), reflecting the intrinsic properties of symbolic regression and physical insights. Despite its relatively straightforward architecture and training process, this model effectively models ODE equations, offering valuable insights and predictions for biological systems.

Methodology

In developing a neural network model for biological Ordinary Differential Equation (ODE) systems, the primary challenge lies not in performance but in ensuring compatibility across diverse fields such as biology, mathematics, and computer science. Biological systems are often modeled using ODEs, with their equations, formats, and symbols providing a robust representation of these systems. Our model framework is designed to closely resemble traditional human-written ODE functions, focusing more on realistic modeling than on mere numerical fitting. Methods like Long Short-Term Memory (LSTM) or attention-based mechanisms, which alter their behavior based on previous data, are not suitable for our purposes, as the underlying mechanisms in biological systems do not change due to past states. By employing symbolic regression, our model yields outputs that are not only understandable but also physically valid.

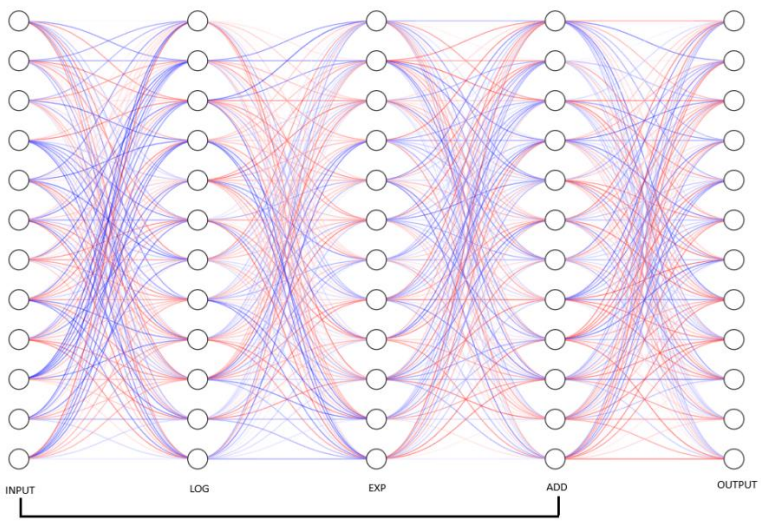

Figure 1: the model demo. The log-exp pair layers within our model can mimic various terms in ODE functions simply by adjusting the weights of each layer, thereby altering the interactions. This approach also facilitates relatively straightforward training. The size of the model is determined by the dimensions of the log-exp pair layers. A width of 1 implies a layer size equivalent to the input size; as the input size increases, the complexity of the ODE operations also grows. A depth of 1 indicates a single pair of log-exp layers.

From the perspective of computer science and neural networks, an ODE system modifies its derivatives based on the current state and incorporates these derivatives into the subsequent state. This characteristic is well-aligned with the principles of a residual neural

network. To address the challenges of emulating manual-style equation formulation and to ensure compatibility with neural networks, we utilize a combination of logarithmic and exponential (log-exp) functions. The outputs generated by these functions are then added to the inputs to form the next state of the ODE function, serving as the subsequent input for the model.

When there is no operation in log-exp pair:
If w = 1,

$$, \exp(w \log(node)) = node$$

When there is a need of do complicated multiplication or division:
If w1 = 1, w2 = 2, w3 = 3,

$$\exp(w_1 \log(node_1) + w_2 \log(node_2) - w_3 \log(node_3)) = \frac{[node_1] * [node_2]^2}{[node_3]^3}$$

When there is a need of trigonometric or log or exp:

$$sine\ (x) = Taylor\ expansion\ of\ sine\ (x)$$

While many symbolic operations cannot be perfectly represented or approximated by log-exp pairs, this configuration suffices for most common interactions in biological systems. Additionally, many ODE systems can be approximated by dynamically adjusting the weights and biases in the log-exp pair layers, allowing both symbolic representation and physical meaning to be captured through a flexible adjustment of matrix weights.

Data preparation involves introducing a certain number of random changes to initial points – for example, altering [1,2,3] to [1,2,5], and then increasing the number of elements changed progressively. As the number of element changes increases, the count of initial states naturally decreases, reducing workload and simulating the effect of introducing various agents like drugs or RNAi into an environment.

Training is based on a small batch size (typically 2), inputting one of the initial states and requiring the model to predict subsequent data points, before switching to another batch/pair of data. During evaluation, the batch size is adjusted to encompass the full data time interval (e.g., 500 data points). Apart from the initial state, the model generates outputs based on its own predictions, ensuring that test data is never used during training.

Results

In our study, we utilized data from the research article "The Rate of Immune Escape Vanishes When Multiple Immune Responses Control an HIV Infection." This provided us with a robust framework of 19 equations for our analysis. For training, we followed the previously outlined methodology, resulting in a dataset of 894 samples. These samples were subjected to a 1X1 model configuration across five epochs at a learning rate of 1e-5. Consistent with our training approach, the model was tasked with predicting the next point in a sequence, whereas the testing phase involved a continuous evaluation over 500 time points.

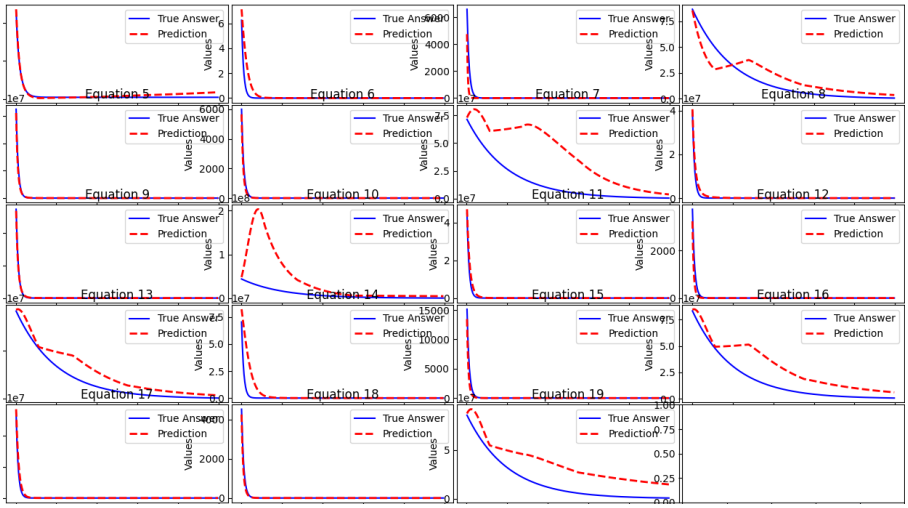

Figure 2: HIV model tested on one of the test sets.

Upon testing multiple datasets, we observed largely consistent results with minor variations. This consistency could be attributed to the inherent dynamics of the ODE system under study, which does not exhibit drastic changes. Most of the equations, with the exception of equations 11 and 14 that showed slightly poorer fits, were moderately to well fitted by the model.

Further, we extended our analysis to data from the study "Using single-index ODEs to study dynamic gene regulatory network." This data set comprised 7 equations, and we generated 247 training samples as per our established methodology. The training was executed on a 2X1 model (featuring 14 nodes wide with a pair of log and exp layers) over five epochs with a learning rate of 1e-4. As with our prior approach, the model predicted the next data point during training, but was tested against a sequence of 500 time points for continuity. This test served not only to assess accuracy but also to demonstrate the model's capability in utilizing Taylor expansion or polynomial approximation for addressing trigonometric challenges.

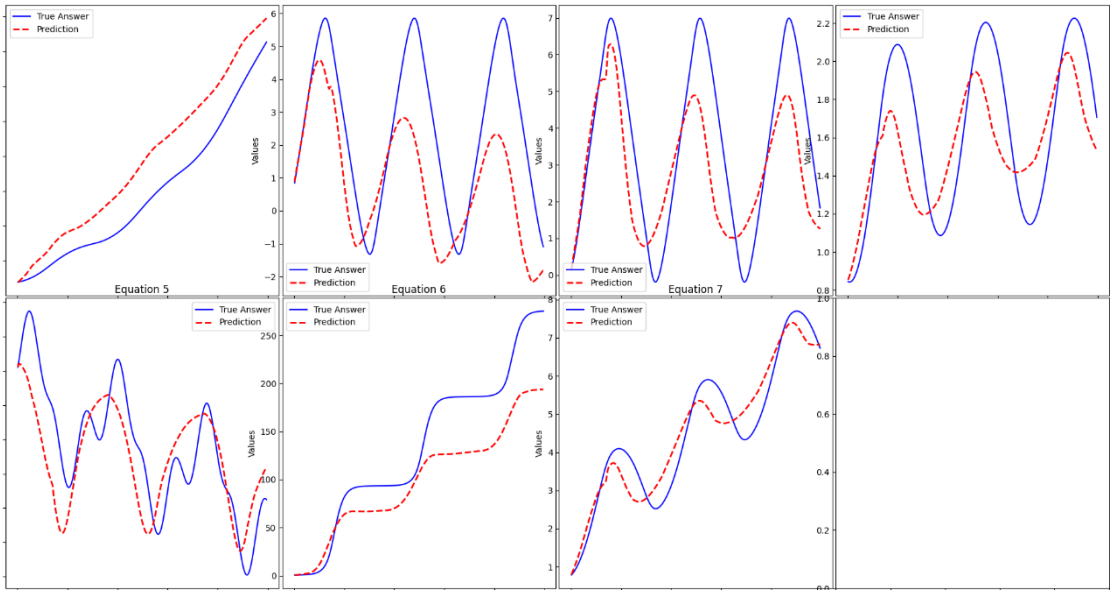

Figure 3: GRN model tested on one of the test sets.

The test results predominantly showcased the periodic nature of the data. Both in terms of direction and dynamics, the fits appeared to be precise, indicating the model's effectiveness in capturing the essential features of the datasets.

Conclusion

In this research, we have developed a novel approach using Residual Networks (ResNet) for the symbolic regression of biological ODE systems, effectively capturing their dynamics. This methodology, harmonizing principles from biology, mathematics, and computer science, exemplifies a versatile and multifaceted utility. Unlike common symbolic fitting methods, our model adheres closely to the framework of biological ODEs, allowing it to capture more than just numerical data. This alignment with the Bio ODE framework ensures that our model not only predicts but also reflects the intrinsic nature of the biological systems it represents.

Further validation of our model's capabilities was demonstrated through its application to gene regulatory network data, which involved complex trigonometric functions. By employing Taylor expansion and polynomial approximation techniques, our model adeptly handled these complexities, as evidenced by its accurate depiction of periodic data features. This shows that our approach is not only accurate in representing biological data but also adaptable in handling mathematical challenges.

The successful implementation of this model marks a significant advancement in the field of systems biology. By combining the physical realism of traditional methods with the flexibility and scalability of modern neural networks, our approach opens up new avenues for understanding and predicting the behavior of complex biological systems. This has far-reaching implications, potentially impacting various areas of scientific research and therapeutic development.

As a result of our study, the model we have developed is expected to stabilize over time, becoming a valuable tool for comparison and analysis in specific fields of study. This methodological innovation signifies a promising trajectory for future exploration in systems biology, highlighting the potential of integrating diverse methodological concepts to deepen our understanding of complex biological systems.

Reference

1. Kohl, P., Crampin, E. J., Quinn, T. A., & Noble, D. (2010). Systems biology: an approach. Clinical Pharmacology & Therapeutics, 88(1), 25-33. DOI: 10.1038/clpt.2010.92
2. Zupanic, A., Bernstein, H. C., & Heiland, I. (2020). Systems biology: current status and challenges. Cell & Molecular Life Sciences, 77(3), 379-380. DOI: 10.1007/s00018-019-03410-z
3. Džeroski, S., & Todorovski, L. (2008). Equation discovery for systems biology: finding the structure and dynamics of biological networks from time course data. Current Opinion in Biotechnology, 19(4), 360–368. https://doi.org/10.1016/j.copbio.2008.06.008
4. Guimerà, R., Reichardt, I., Aguilar-Mogas, A., Massucci, F. A., Miranda, M., Pallarès, J., et al. (2020). A Bayesian machine scientist to aid in the solution of challenging scientific problems. Science Advances, 6(5), eaav6971. https://doi.org/10.1126/sciadv.aav6971

5. Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2020). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045
6. Hasenauer, J., Waldherr, S., Doszczak, M., Radde, N., Scheurich, P., & Allgöwer, F. (2011). Identification of models of heterogeneous cell populations from population snapshot data. BMC Bioinformatics, 12(1), 125. https://doi.org/10.1186/1471-2105-12-125
7. Liepe, J., Kirk, P., Filippi, S., Toni, T., Barnes, C. P., & Stumpf, M. P. (2014). A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. Nature Protocols, 9(2), 439-456. https://doi.org/10.1038/nprot.2014.025
8. Zhan, C., Yeung, L. F., Wang, X., & Chung, F. L. (2011). Parameter estimation in systems biology models using spline approximation. BMC Systems Biology, 5(1), 14. https://doi.org/10.1186/1752-0509-5-14