# Research Statement

CANER KAZANCI

Extensive work has been done in mathematical modeling of biological systems, focusing on small subsystems that can be described using only a few differential equations. With recent advances in experimental techniques in biology, there is a need to extend mathematical modeling to include larger systems. One of the major challenges is understanding how cells and other large biological systems work. It is clear that only a few differential equations are not sufficient to capture the complexity of such systems. New mathematical and computational tools need to be developed.

My research focuses on developing new techniques for modeling and simulation of problems arising in systems biology. My current work focuses on large biochemical reaction systems: understanding their behavior, analyzing their statistical properties, reverse engineering, and generating faster numerical methods for their simulation. The biological motivation is to develop mathematical and computational methods that can cope with the complexity of representing living cells made up of tens of thousands of molecules, with hundreds of thousands of interactions.

## Statistical Properties of Large Biochemical Networks

Recent advances in data acquisition techniques, such as gene microarrays, result in vast amounts of quantitative information about biological systems to be analyzed and interpreted. Cells, the building blocks of living organisms, can be viewed as large, complex reaction systems with embedded subsystems such as gene regulatory networks, enzymatic pathways and protein-protein interactions. These complex cellular systems share many similar features in different organisms, and some aspects of their behavior can be studied in terms of abstract large networks that exhibit proper statistical properties. Examples of statistical properties that show universal behavior across organisms and tissues include the well known Zipf's law for chemical abundance [1] and the scale-free network structure [4].

A cell, which may be viewed as a large network of reactions, can be associated with a graph, representing interactions. Using reaction kinetics, this may be modeled as a large system of ordinary differential equations. This assumes that the abundance of molecules is sufficiently large, allowing for a continuous description; otherwise stochastic modeling is required. My work thus far is focused on modeling using ODE's.

$$\text{IVP}\left\{\begin{array}{l}\dot{x}=f(x)\\ x(0)=x_0\end{array}\right. \quad \text{where} \quad \begin{array}{l}x=(x_1,x_2,\cdots,x_n)\\ f(x)=(f_1(x),f_2(x),\cdots,f_n(x))\end{array} \tag{1}$$

where $x_i(t)$ represents the abundance or concentration of molecule $i$ at time $t$.

Recalling that we consider thousands of equations, a classical approach using phase diagrams may be insufficient in studying these systems. Due to the large number of elements, statistical interpretation, or even representation is necessary. Inspired by classical works like Boltzmann's treatment of gas behavior, I consider the statistical description of these systems in terms of the abundance distribution, defined as:

$$d(x,t)=\frac{1}{N}\sum_{i=1}^{N}\delta(\xi-x_i(t)) \tag{2}$$

The study of this distribution function raises some interesting questions. Unlike classical problems in chemistry or physics where interacting particles are identical, in biology there are many different types of interacting elements;

up to forty thousand proteins are expressed in complex organisms. The distribution function above ignores the details regarding different proteins, but still provides interesting information.

I have studied large random reaction networks consisting of more than fifty thousand reactions and ten thousand molecules with various statistical properties including network connectivity, initial conditions and reaction constants. I have developed a computational tool in C++ to automatically generate these systems, convert them to ODE's, simulate them and interpret them statistically. The simulation techniques involve $4^{th}$ order Runge-Kutta, adaptive Runge-Kutta-Fehlberg, and a stochastic method using the chemical Langevin equation [3]. Within this computational environment I have studied the abundance distribution function defined in (2). One major conclusion of this work is that in all cases, the abundance distribution function quickly reaches self similarity and then approaches its stationary limit. Surprisingly, this limiting abundance distribution is insensitive to the initial conditions, reaction constants and network connectivity. However, it does depend on the average number of reactions per molecule.

To explain the universality in the behavior of these systems, I have considered an evolution equation (PDE) for this distribution function, an approach that was partially successful. I computed transition probabilities between neighboring populations, those with abundance $\xi$ and abundance $\xi \pm \Delta\xi$. Through simulations I have observed that these transition probabilities vary slowly over time resulting in a PDE with time-dependent coefficients. Thus no Fokker-Planck type equation with time-independent coefficients exists to explain this global behavior.

Another observation is that in all cases, the tail of the abundance distribution function obeys a power law. Interestingly, experimental data regarding the abundance distribution of RNA in various organisms, from E. coli to H. sapiens, exhibit a power-law distribution. In other words, the probability that a gene has an expression level $k$, decays as a power law, $P(k) \propto k^{-r}$which coincides with my observations.

In biological systems, responses to perturbations are often very important. Examples include the effects of drugs and responses to a chemical insult or a pathogen. This motivates my study of perturbations in these large networks, including changes in initial conditions, reaction constants and molecule deactivation. The main result is that a perturbation in a single molecule affects the system as a whole. Moreover, the resulting changes in the abundance of different molecules are proportional to their steady-state abundance. This mathematical result agrees with experimental results on organisms ranging from E. coli to humans [5].

One major property of biochemical reaction systems is that their connectivity distribution function obeys a power law [5]. I investigated the relationship between the connectivity distribution function and the system response to perturbations. I computed the eigenvalues of the linearized system at steady state with special emphasis on the values that are close to zero. In a random reaction network, where the connectivity is a normal distribution function, only a few eigenvalues were close to zero. However, in reaction networks with a connectivity distribution that obeys a power law, a significant fraction of the eigenvalues are close to zero. This implies that in systems obeying a power law, response to perturbations is much more global. This theoretical result has implications for some important problems in biomedicine. For example, many drugs affect an organism on a global level, and therefore drug design targeting specific molecules may be too simplistic, requiring a system level approach instead.

## Reverse Engineering of Reaction Systems

One of the challenges spanning all branches of biology is inferring biological knowledge from experimental results. New high throughput technologies which produce large amounts of data make this a much more difficult task. I am interested in understanding the extent to which this process of biological discovery can be automated or become computer assisted. In this context, biological knowledge is understood in terms of interactions, which we model as differential equations. Mathematically, this is an inverse problem in which we try to identify all interactions as well as their rates from experimental observations. To gain some insight into this difficult problem I have developed a network reconstruction algorithm, whose input is the abundance of molecules in the system at different time points.

Networks were created randomly and simulated to generate artificial experimental data. I begin with the steady state and consider the response to various perturbations and their time evolution in the reconstruction process.

The mathematical foundation for the reverse engineering algorithm that I have developed is as follows. Let $x^* = (x_1^*, x_2^*, \cdots, x_n^*)$ be the steady state solution for (1). A change in the abundance of molecule $i$ in the system will result in a response function $y(t)$. Changing one molecule at a time, we have, $y(0) = c\,e_j$, and

$$\dot{y} = (x^* + y)' = f(x^* + y) \approx \underbrace{f(x^*)}_{=0} + \underbrace{\nabla_x f(x^*)}_{=A}\, y = Ay\,.$$

An approximate solution to this new IVP is

$$y(t) \quad \approx \quad c\,e_i + c\,t\,A\,e_i\,,$$

with the $i^{\text{th}}$ column of the matrix $A$ approximately equal to

$$\frac{y(t) - ce_i}{ct}\,.$$

Note that it is possible to reconstruct the whole reaction system using this matrix.

Using this idea, I have experimented with different systems of up to one thousand molecules. Depending on the complexity of the system, the algorithm produces between 60% and 98% of the original reaction equations.

## Future Work

This work has several natural extensions. The most obvious one is introducing different classes of molecules into the network, representing genes, RNA and proteins. There are interesting challenges in this extension since genes are either on or off (0 or 1), so I need to simultaneously combine discrete and continuous variables. Such models will be more realistic and will enable the study of new questions, such as the limitations in reverse engineering when only RNA is measured (gene microarray technology [2]). The previously studied questions related to abundance distribution, system response to perturbations, and reverse engineering are also still relevant. Overall, the computational tools that result from this work will provide a broader understanding of system level behavior at cellular and sub-cellular levels.

# References

[1] C. Furusawa. Zipf's law in gene expression. *Phys. Rev. Lett.*, 90:088102, 2003.

[2] T. S. Gardner, D. Bernardo, D. Lorenz, and J. J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301:102, 2003.

[3] D. T. Gillespie. The chemical langevin equation. *J. Chem. Phys.*, 113:297, 2000.

[4] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651, 2000.

[5] H. R. Ueda, S. Hayashi, S. Matsuyama, T. Yomo, S. Hashimoto, S. A. Kay, J. B. Hogenesch, and M. Iino. Universality and flexibility in gene expression from bacteria to human. *PNAS*, 101:3765, 2004.