Research Statement
Deena Schmidt

My research is driven by a desire to understand the roles of stochasticity, structure, and evolution in shaping the dynamics of biological systems. I develop and analyze mathematical models, combining methods from probability, dynamical systems, and random graph theory, to shed light on biological issues while generating new mathematical questions. Stochasticity is fundamental to gene expression and ion channel gating; it also plays an under-appreciated role in physiological regulation at larger scales, such as the regulation of sleep and wake behavior. This behavior is governed by networks of neurons in the brain with underlying structures that likely influence dynamics, suggesting a prominent role for stochastic processes evolving on networks. To what extent are the dynamics that we observe shaped by the stochastic nature of biochemical reactions, by physiological structure such as neuronal networks, and by the constraints of evolution? A particular challenge of this question is the need to bridge multiple spatial and temporal scales: from microscopic to macroscopic, from millisecond to millions of years. Below I describe my work using methods such as stochastic processes on graphs and stochastic models in genetics to understand interlocking large and small facets of biological regulation.

Large facets: Processes on networks

Random graphs with deterministic or stochastic processes on the nodes is becoming a very active area of research in neuroscience, cell signaling (e.g., gene regulatory networks), and other areas of computational biology. Most studies rely on numerical explorations or mean field approximations. The projects described below pair biologically-motivated numerical studies with mathematical analysis and a framework to extend the mathematical theory of processes on random graphs.

Quantitative measure of edge importance

A major source of noise in neural systems is random gating of ion channels, and such systems are typically represented as a Markov process on a graph. Recently, Schmandt and Galán (PRL, 2012) introduced the stochastic shielding approximation, a new approach to efficient, accurate simulation of Markov chain models. They show that neglecting a carefully chosen subset of noise sources within the Markov process leads to significantly faster simulations with practically no observable error.

Inspired by their work and in collaboration with Peter Thomas (CWRU) [1], I conducted a thorough mathematical analysis of this stochastic shielding heuristic, both in toy models, in the Hodgkin-Huxley ion channel model, and in a broad class of random graph models. This method is based on replacing a high-dimensional stochastic process defined on a graph with a lower-dimensional process on the same graph, rather than replacing a complex network with a simpler one. We show that this form of model reduction can be represented as a mapping from the original process to an approximate process defined on a significantly smaller sample space. Our analysis results in a new, quantitative measure of the importance of individual edges (reactions) within a Markov process on a graph. Our measure not only confirms the optimality of the stochastic shielding approximation, but also sheds new light on to the contributions of different ion channel transitions to the variability of neural systems.

References refer to publications in my CV.
Extending “edge importance” measures

We obtained mathematically rigorous results in [1] by exploiting the stationarity of the stochastic ion channel process, corresponding to a common experimental condition known as “voltage clamp” of an electrically active nerve cell. A major challenge that I plan to address is to extend our analysis to the nonstationary case, corresponding in experimental terms to “current clamp”. Under this condition, the cell voltage is allowed to vary over a significant range as the cell “spikes” or produces electrical discharges. The transition rates of the stochastic process are voltage dependent, and hence, the ion channel process becomes strongly nonstationary. Nevertheless, Schmandt and Galán showed empirically that the same stochastic shielding heuristic works equally well under stationary conditions (for which we have rigorous results [1]) and nonstationary conditions. In order to address the latter case, we need to develop new mathematical ideas.

Network structure and dynamics of neural systems

A fundamental question regarding neural systems and other applications involving networks is the extent to which the network architecture may contribute to the dynamics occurring on the network. Recent studies have found interesting structure in data arising from biological systems, such as data with power law distributions. Several questions naturally arise: If trying to understand data produced by a process on a network, should one study the process or the network? How can one understand the contributions of the network topology and of the process itself in shaping the data?

In collaboration with Janet Best (OSU) and Mark Blumberg (Univ. of Iowa) [5], I explored the relative contributions to network dynamics made by the graph structure and by the nature of the stochastic process occurring on that graph. Primarily using degree distribution as a marker of graph structure, we considered randomly generated graphs on \( N \) nodes with varied degree distributions; we further varied the clustering occurring in the graph by considering some small-world networks. We considered two different stochastic processes on the graphs: (i) a percolation type process of activity and (ii) a neural spiking process. Our results suggest that memoryless processes such as percolation-like processes and spiking of integrate-and-fire neurons may reflect the degree distribution of the graph. We find that processes with memory, such as those in which nodes change their firing rate in response to inputs, can robustly produce heavy-tailed distributions of activity, including power laws. These results are a first step toward understanding the influence of neural network structure on the dynamics of sleep-wake regulation in the brain.

Network structure and dynamics of sleep-wake regulation

Recent studies of sleep-wake dynamics have shown that, in several mammalian species, periods of wakefulness (or wake bouts) follow a power law distribution while sleep bouts follow an exponential distribution. Previous work by Badal Joshi and Janet Best considered a model of three interacting neuronal populations (or nodes in a 3-node graph) that mimics the dynamics of sleep-wake cycling in developing rats: it produces short, exponentially distributed wake and sleep episodes with power law distributed wake bouts.

The spiking model we considered in [5] can be viewed as expanding the wake-active population node from the model by Joshi and Best into a network of individual spiking neurons where the activity of each node is described by a stochastic differential equation. Currently, we are extending our model to link network structure and stochastic dynamics to neural activity patterns involved in sleep-wake regulation [2]. Focusing on the study of two interacting networks, such as the dynamics between sleep-active and wake-active neural networks, we have a model that robustly produces
wake bout durations distributed as a power law followed by an exponential tail for the case where both networks have degree distributions with low mean degree. The size of the power law region and the variability of the power law exponent across different realizations of the model depends on the network structure. The distribution of wake bout durations also depends on the age of the mammal so a related question is: what developmental changes are influencing the distribution of wake bouts? To that end, we are working with Mark Blumberg to design experiments that will allow us to test model predictions and further develop the model.

**Theoretical foundations in network dynamics**

In collaboration with Boris Pittel (OSU) and Janet Best, I am working to extend the mathematical theory of processes on random graphs. This work will generalize earlier results concerning bootstrap percolation, a type of deterministic diffusion of activity, on random graphs. In that case, the basic problem is to identify, for a given random graph, probabilities $p_-$ and $p_+$ such that, if each node is initially active with probability $p < p_-$ (respectively, $p > p_+$), the probability that all vertices are eventually active is very close to 0 (respectively, 1). Motivated by neuronal activity in sleep-wake regulation, we ask related questions such as: for a given family of random graphs and assuming that $p > p_+$, what is the distribution of times to fixation in the active state? What can we say if the percolation process is random instead of deterministic? What if active nodes can become inactive with some positive probability, so that the process is no longer monotonic? Such questions arise naturally in the study of neuronal networks, but they are beyond the reach of current mathematical theory. This presents an exciting opportunity to extend the mathematical underpinnings of stochastic processes on random graphs so that this growing area of research can move beyond numerical explorations.

**Small facets: Stochastic models in genetics**

From cancer to sleep, these projects explore how genes can regulate macroscopic behavior and how the DNA itself changes. A central difficulty in such problems lies in capturing the essential components of the process in a mathematical model amenable to analysis, computation, and prediction.

**Stochastic gene regulation**

Gene regulatory networks dynamically orchestrate the level of expression for each gene in the genome by controlling how vigorously that gene will be expressed. Gene expression is inherently stochastic due to the small number of regulatory molecules typically present within a cell. Genetic feedback loops in cells break detailed balance and involve bimolecular reactions and, hence, exact solutions revealing the nature of the stochastic fluctuations in these loops are lacking. In collaboration with Ramon Grima (Univ. of Edinburgh) and Timothy Newman (Univ. of Dundee) [3], I analyzed a stochastic model of a gene regulatory feedback loop using the chemical master equation (Markov process theory) for theoretical analysis. In this model, a gene produces protein which then binds to the promoter of the same gene and regulates its expression. The protein degrades in its free and bound forms. This network breaks detailed balance and involves a single bimolecular reaction step. We provide an exact solution of the steady-state master equation for arbitrary values of the parameters, and present simplified solutions for a number of special cases. We verify the full parametric dependence of the analytical non-equilibrium steady-state probability distribution
by direct numerical solution of the master equations. Our results emphasize the importance of stochastic effects in modeling gene regulation.

**Waiting times for de novo generation of gene regulatory sequences**
A regulatory sequence is a short sequence of DNA (in vertebrates many are 6-10 nucleotides long) located near a gene where regulatory proteins known as transcription factors bind preferentially, controlling the expression of that gene. Mutations in regulatory sequences can cause changes in gene expression which in turn can lead to phenotypic evolution. Hence, one possible explanation for the substantial differences between humans and chimpanzees is that there have been extensive changes in regulatory sequences, rather than in genes themselves. This raises quantitative questions about the rate of regulatory sequence evolution in humans and whether it is sufficiently rapid to contribute to the differences between humans and chimps. Using this biological framework, we considered the following probability problem: Given a gene regulatory region of length $L$ in the human genome, how long does it take for a word of length $w$ to appear (by random mutation) in that region in some individual?

In collaboration with Rick Durrett (Duke Univ.) [9,10], I analyzed this problem using Markov chain models and a variety of probabilistic techniques, including Chen-Stein calculations, Aldous’ Poisson clumping heuristic, and mixing times for Markov chains. Our results yield innovative and novel approximations of the waiting time distribution. We also applied our theoretical results to estimate average waiting times for the generation of words of length 6-10 within a 1000 nucleotide regulatory region in humans. The waiting time follows a mixed exponential distribution that depends on the presence of an “almost match” in the population. Such flexibility in transcription factor binding is important to allow regulatory sequences to evolve at a reasonable rate in humans.

**Waiting times for regulatory sequence turnover**
In [8], we considered an improved mathematical model for regulatory sequence turnover within a species, given the results of our previous work. We estimated this rate of evolution and compared it to known divergence times between species for both humans and Drosophila (fruit flies). In particular, we examined the waiting time distribution for a pair of mutations: the first mutation (A) inactivates an existing transcription factor binding site and the second (B) creates a new active binding site within the regulatory region. This model is a general two-step process of binding site turnover where the mutations can occur in either order in an individual, but we focus on the sequence A then B. Using a population genetic model (i.e., continuous-time Moran model), we studied the time evolution of this mutation process under the effects of selection and random genetic drift in the human and Drosophila populations. We then quantified the mean waiting time for a type B mutation to arise in some individual, and also for the B mutation to become fixed in the population, under various fitness conditions for the A and B type mutation.

The results from our previous work assumed that the human ‘effective population size’ (i.e., number of breeding individuals in an idealized constant size population, a basic parameter in population genetic models) is 10,000 which is relatively small compared to that of organisms such as Drosophila and yeast whose effective population sizes are in the millions. We applied our results to find that a few million years is sufficient for the rate of binding site turnover in Drosophila, consistent with recent experimental observations, but in humans, this type of change would take more than 100 million years. Our analysis also exposed some flaws in the literature concerning mathematical limits to Darwinian evolution [7].
**Waiting times for multi-stage carcinogenesis**

Work in collaboration with Jason Schweinsberg (UCSD) and Rick Durrett generalized the waiting time results from the two-step process above, yielding new insights into the study of colon cancers and other diseases which are the result of a sequence of several mutations occurring in a collection of cells. In [6], we considered the mathematical problem: Given a population of $N$ individual cells, how long does it take until some individual has accumulated $m$ specified mutations? We formulated a general $m$-step model that applies both to cancer development and to gene regulatory sequence evolution (for the case $m = 2$). We assumed that the population evolves according to the Moran model and that individuals may acquire mutations during their lifetimes due to radiation or other environmental factors. We labeled each individual with a type $1 \leq j \leq m$ to denote the number of mutations acquired by that individual and assumed that individuals of type $j - 1$ mutate to type $j$ at rate $u_j$, given that initially all individuals are type 0. Let $X_j(t)$ be the number of type $j$ individuals at time $t$. For each positive integer $m$, let $\tau_m = \inf\{t : X_m(t) > 0\}$ be the first time at which there is an individual of type $m$ in the population. Clearly, $\tau_1$ is exponentially distributed with rate $Nu_1$. We computed the asymptotic distribution of $\tau_m$ for $m \geq 2$ as $N \to \infty$ under various conditions for how the population size $N$ scales with the mutation rates $u_j$.

**Genetic regulation of sleep**

Models of sleep-wake cycling typically invoke two distinct processes: a circadian rhythm arising from a clock-like transcription and translation feedback loop plus a poorly-understood homeostatic process reflecting an inclination to sleep following prolonged wakefulness. These processes are often regarded as independently tracking distinct aspects of sleep need, despite accounts of their nonlinear interactions. Moreover, biologists have recently discovered that several circadian clock genes are significantly involved in shaping the homeostatic properties of sleep. For instance, the clock gene transcription factor NPAS2 affects the rate at which mice accumulate the homeostatic need for sleep. This project uses stochastic process and differential equation models to understand how these clock genes can simultaneously participate in multiple, linked sleep processes and to uncover the relations between genetic differences and altered sleep dynamics.