

Research Statement

Deena Rae Schmidt
drs32@cornell.edu

My research area is applied probability. In particular, I'm interested in problems in probability theory that arise from genetics. The first project described below is work that was motivated by a Research Experience for Undergraduates (REU) led by my advisor, Rick Durrett, at Cornell University for which I was the teaching assistant. We examined selection pressures acting on zinc finger genes, an important gene family that has experienced many gene duplications. The second project investigates DNA regulatory sequence evolution in the context of how long it takes for these sequences to appear in the human population. The third project is a current collaboration between Jason Schweinsberg (UCSD), Rick Durrett, and me in which we extend the regulatory sequence project to handle other organisms with larger population sizes. Lastly, I note my involvement in various activities and discuss possible directions for future research.

Gene duplication

In the human genome, there are hundreds of zinc finger genes organized into more than a dozen different families. In Schmidt and Durrett (2004) we focused on the C_2H_2 type zinc finger gene, a 28 amino acid motif that contains DNA binding sites and is tandemly repeated at the end of zinc finger genes. The number of repeats per gene ranges from two up to three dozen or more. The total number of zinc finger genes appears to have increased through evolution. There are approximately 700 in humans compared with about 300 in *D. melanogaster* (fruit flies), 100 in *C. elegans* (worms), and 40 in *S. cerevisiae* (yeast). The average number of finger repeats per gene has also increased, numbering 8, 3.5, 2.5, and 1.5, respectively, in the four species just mentioned. In addition, some regions of the human genome contain many zinc finger genes with no homologs in rodents. These observations along with the clustered organization of these genes in humans suggest that gene duplication has played an important role in expanding this family of genes. Using maximum likelihood methods developed by Yang, we investigated four clusters of zinc finger genes on human chromosome 19 and found evidence that positive selection was involved in diversifying the family of zinc finger binding motifs.

DNA regulatory sequence evolution

At a genetic level humans and chimpanzees are closely related, with 98.7% of their DNA identical. It has long been speculated that many of the obvious differences between the two

species are due to changes in regulatory sequences that control how genes are expressed. A regulatory sequence is a short sequence of DNA (in vertebrates many are 6-9 nucleotides long) which is a binding site for transcription factors that promote or inhibit transcription of DNA to make proteins. Given what is known about transcription factor binding sites, this motivates the following probability question: Given a 1000 nucleotide region in our genome, how long does it take for a specified six to nine letter word to appear in that region in some individual?

Stone and Wray (2001) computed 5950 years as the answer for a given six letter word to appear in a 2000 nucleotide region. As MacArthur and Brookfield (2004) have already pointed out, there is a serious problem with this computation: they assumed that DNA sequences of different individuals are independent to get results for the population using simulations for a single individual. In Durrett and Schmidt (2007), we used Markov chains to analyze the waiting time problem, assuming non-independence of individuals. We showed that if we assume a mutation rate per nucleotide per generation of $\mu = 10^{-8}$, then for words of length 6, the average waiting time is 100,000 years. For words of length 8, the waiting time has mean 375,000 years if there is a 7 out of 8 letter match in the consensus sequence for the population, but if not the mean waiting time is 650 million years. Fortunately, in biological reality, the match to the target word does not have to be perfect for binding to occur. In other words, a transcription factor can still bind to a regulatory sequence even if one or more nucleotides do not agree with its target word. If we model this by saying that one mismatch is good enough, then the mean for words of length 8 becomes about 60,000 years. Thus, the moral of the story is that the inexactness of transcription factor binding is important to allow regulatory sequences to evolve at a reasonable rate in humans.

Extensions to *Drosophila*, yeast, and *E. coli*

In collaboration with Jason Schweinsberg, we are extending the previous project to handle populations of size $N = 10^6$ or larger such as *Drosophila*, yeast, and *E. coli*. The results for humans depend heavily on the fact the effective population size is $N = 10^4$. More precisely, one of our key approximations requires that $N^3\mu^2$ is small. This is false for the three organisms listed above. *Drosophila* have an effective population size of roughly $N = 10^6$ and a mutation rate $\mu = 10^{-8}$. The effective population size is about 6×10^7 for yeast (Fay and Benavides 2005) and 2×10^8 for *E. coli* (Hartl, Moriyama, and Sawyer 1994), while the mutation rates for these organisms per base pair per generation are 2.2×10^{-10} and 5.4×10^{-10} , respectively (Drake et al. 1998).

There is a growing body of experimental evidence that in *Drosophila* and yeast, significant changes in gene regulation can occur in a short amount of time. Ludwig et al. (2005) noticed that the genes regulating the even-skipped stripe 2 enhancer in *Drosophila* are different in closely related species. Prudhomme et al. (2006) observed that regulatory changes have caused two independent gains and five losses of wing pigmentation spots in the *Drosophila melanogaster* species group. Tsong et al. (2006) have shown that the gene regulatory circuitry that governs mating type in yeast species *Saccharomyces cerevisiae* and *Candida albicans* are

different despite the fact that the overlying phenotypes remain similar.

To understand the mechanisms at work in the evolution of gene regulation for organisms with large population sizes, we consider the mathematical problem: How long does it take in a population of N diploid individuals until some individual has accumulated k specified mutations? In this general setting, we do a similar analysis of the waiting time problem for these organisms. Let τ_k denote the waiting time. The number of copies of a mutant in the population is $o(N)$, but over time it behaves like a critical branching process which has a total progeny distribution with $P(X > m) \sim Cm^{-1/2}$. Thus, if mutations are nonoverlapping in time, the total number of mutant births in the first M mutant families will be $O(M^2)$ and we will see the first double mutant when $M^2\mu = O(1)$, i.e., $M = \mu^{-1/2}$. In the case of *Drosophila*, $\mu = 10^{-8}$ and $N = 10^6$ which gives $M = 10^4$. Mutations occur at rate $2N\mu = .02$ per generation, so 10^4 mutations yield a waiting time (τ_2) of about 500,000 generations or 50,000 years. This is consistent with the experimental observation that in *Drosophila*, we frequently see regulatory sequence changes that turn off one binding site and turn on another.

Repeating the calculation above, we see that $M^2\mu$ double mutants will have of order $(M^2\mu)^2$ total offspring, so we can expect triple mutants to occur when $M^4\mu^3 = O(1)$ or $M = \mu^{-3/4}$. As long as $N \gg \mu^{-3/4}$, then an individual with three mutations will occur before the first fixation. For *Drosophila*, $N = \mu^{-3/4}$ so fixation of one mutation followed by two new mutations, or an individual gaining three mutations before fixation are both possible scenarios, so the waiting time (τ_3) in this case has a complicated limiting distribution. We formulate a mathematical limit theorem to properly delineate the transitions between regimes and then generalize it to handle yeast and *E. coli*.

Future Research

In my research thus far, I have worked mainly with discrete models. To learn more about Brownian motion and stochastic calculus, I have started working on two new research projects dealing with evolutionary models of gene expression and a genetic phenomenon known as Hill-Robertson interference.

Models for the evolution of gene expression

Several recent studies have examined differences in gene expression patterns in closely related species and have reported conflicting results. Khaitovich et al. (2004) measured gene expression levels in samples from six humans, three chimpanzees, one orangutan and one rhesus macaque. Observing that expression differences have accumulated linearly with time, they suggest that the majority of gene expression changes follow a neutral model of evolution. However, this study relied on microarray data that used exclusively human-based gene probes. Gilad et al. (2006b) argued that in such studies, sequence mismatches caused by species divergence can affect hybridization intensity and bias results. To correct this problem, they used multi-species microarrays and did not find a linear trend of divergence with

time for primate gene expression, suggesting widespread stabilizing selection. This is consistent with results of Rifkin et al. (2003, 2005) who investigated expression patterns in several closely related *Drosophila* species. Finding a dominant signature for stabilizing selection, they report that gene expression does not evolve according to strictly neutral models.

Linear mixed models have been used to quantify the contributions of individuals, populations, and species to the observed differences, and to identify trends in interspecies comparisons (see Hsieh et al. 2003). However, to test hypotheses we need a model of gene expression changes. Khaitovich, Pääbo, and Weiss (2005) model the changes in the logarithm of gene expression levels as random walk along lineages. This is a generalization of the stepwise mutation model (SMM) for which they use an extreme value distribution in addition to the normal as the distribution of individual changes. The problem with the SMM, which came up when it was used to model DNA repeat sequences (see Kruglyak et al. 1998), is that a random walk does not have a stationary distribution so the distribution becomes more and more spread out as time evolves. This cannot happen with gene expression levels. There are limits to the possible changes in gene expression so ultimately the SMM breaks down. Most genes are part of large networks so expression changes in one gene can have a cascade of effects. A challenge in modeling the evolution of gene expression levels is dealing with boundary conditions which exist at both ends of the spectrum (Gilad et al. 2006a). Expression levels cannot go below zero, and energetic costs of transcription and physical limitations of transcriptional machinery put an upper limit on expression levels. In order to study this rigorously, we need to develop neural models that incorporate boundary effects of gene expression levels.

Gu (2004) introduced a model in which changes in log expression levels along a lineage are a Brownian motion. This model also lacks a stationary distribution, but has the advantage that it can handle stochastic changes in expression levels due to environmental fluctuations. Environmental factors have a considerable influence on gene expression so a realistic model should take this into account. It is an interesting mathematical problem to compute the predictions of Gu's model for a sample of size n and to see what happens when an Ornstein-Uhlenbeck process, which does have a stationary distribution, is used instead of Brownian motion. The goal is to refine these models and advance toward a realistic evolutionary model of gene expression changes.

Hill-Robertson Interference

Hill-Robertson interference refers to the fact that two advantageous mutations can interfere with each other if the second one enters the population before the first one reaches fixation. To explain this, suppose that a beneficial allele A is present at frequency x in a population when one copy of a second advantageous allele B is introduced. Assume that A has relative fitness $1 + s_1$ compared to its alternative allele a (which has relative fitness 1), and that B has relative fitness $1 + s_2$ compared to its alternative allele b . We consider two cases: upon entering the population, B is linked to the unfavorable allele a or to the favorable allele A . When $s_1 = s_2$ and the genotype frequencies of Ab , aB and ab are x_1 , x_2 and $1 - x_1 - x_2$,

we have been able to prove a new result for the first case. The probability of fixation of aB starting with initial frequencies x_1 and x_2 is

$$h(x_1, x_2) = \frac{1 - e^{-\gamma(x_1+x_2)}}{1 - e^{-\gamma}} \cdot \frac{x_2}{x_1 + x_2}$$

where the first factor is the probability that ab is lost from the population, and competition of the other two alleles is a fair game. At this point, we have not had much success in studying the other case in which the second mutation produces the genotype AB . The Hill-Robertson interference problem has been extensively studied in the genetics literature (see eg McVean and Charlesworth 2000 and Kim 2004). However, most of these studies are based on simulations or propose approximate solutions. Looking at this analytically will lead to novel problems concerning multidimensional diffusion processes.

Noteworthy Activities

A broader impact of my research is my involvement with graduate, undergraduate, and high school students. I participated in the 2004 IMA Math Modeling in Industry graduate student summer workshop, collaborating with six other grad students and a mentor from 3M on a project entitled “Data to knowledge in pharmaceutical research”. The gene duplication project mentioned above was motivated by an REU led by my advisor at Cornell for which I was the TA. Extending my involvement with undergraduate research, I have been invited by the University of Akron to direct a project for their mathematics REU this summer. The topic I’ve chosen looks at basic models involved in various problems in population genetics. Back in April 2005, I was the invited guest speaker for the “Women in Mathematics Program” held at the University of Akron. As a graduate of Akron and an active participant in this program, I was thrilled to go back and talk with the female math community that initially sparked my interest in pursuing graduate school. This program aims to make careers in math and science attractive and accessible to female students by providing them with a strong network of support. I enjoyed bringing my interdisciplinary talk, “Mathematical Approaches to Problems in Genetics,” to a school where there isn’t much collaboration between math and science. I plan to present my REU experience to the Women in Math Program next fall.

I enjoy teaching, and I have had the opportunity to teach both undergraduates and high school students. At Cornell, I have taught two courses: first semester calculus for non math majors and finite mathematics for the life sciences. At Ithaca High School, I designed and taught a course for their Senior Seminar focusing on probability theory with applications ranging from gambling to genetics. Geared toward advanced students, planning this course involved writing my own lectures, creating interesting and challenging problems to work on during class, and helping the students organize original research projects. I have also volunteered with the national program Expanding Your Horizons which encourages middle school aged girls to pursue their interests in math and science. The event is an opportunity for girls and their parents to spend a day attending three hands-on workshops relating to

fields in math and science. My favorite workshop was called "Math - It's Contagious!" which involved modeling the spread of infectious diseases.

As an interdisciplinary graduate student, I currently weave between the departments of Mathematics, Genetics and Development, Molecular and Cellular Biology, and Biological Statistics and Computational Biology at Cornell. I participate in weekly meetings with Chip Aquadro's lab (my minor advisor in genetics), engage in the population genetics journal club, and attend departmental seminars. I enjoy collaborating with both mathematicians and biologists. My goal is to expand my knowledge of math and bioscience, enhance my ability to conduct successful research within the two fields, integrate research and education, and ultimately attain a tenure-track position.

References

- Drake, J.W., Charlesworth, B., Charlesworth, D., and Crow, J.F. (1998) Rates of spontaneous mutation. *Genetics*. 148: 1667–1686.
- Durrett, R. and Schmidt, D. (2007) Waiting for regulatory sequences to appear. *Ann. Appl. Prob.* 17: 1–32.
- Durrett, R., Schmidt, D., and Schweinsberg, J. A waiting time problem arising from the study of multi-stage carcinogenesis. (In preparation)
- Fay, J.C., and Benavides, J.A. (2005) Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genetics*. 1: e5.
- Gilad, Y., Oshlack, A., and Rifkin, S.A. (2006) Natural selection on gene expression. *Trends in Genetics*. 22: 456–461.
- Gilad, Y., Oshlack, A., Smyth, G.K., Speed, T.P., and White, K.P. (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature*. 440: 242–245.
- Gu, X. (2004) Statistical framework for phylogenetic analysis of gene family expression patterns. *Genetics*. 167: 531–542.
- Hartl, D.L., Moriyama, E.N., and Sawyer, S.A. (1994) Selection intensity for codon bias. *Genetics*. 138: 227–234.
- Hsieh, W.P., Chu, T.M., Wolfinger, R.D., and Gibson, G. (2003) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics*. 165: 747–757.
- Khaitovich, P., et al. (2004) A neutral model of transcriptome evolution. *PLoS Biology*. 2: 682–689.
- Khaitovich, P., Pääbo, S., and Weiss, G. (2005) Toward a neutral evolutionary model of gene expression. *Genetics*. 170: 929–939.

- Kim, Y. (2004) Effect of strong directional selection on weakly selected mutations and linked sites: implication for synonymous codon usage. *Mol. Biol. Evol.* 21: 286–294.
- Kruglyak, S., Durrett, R.T., Schug, M.D., and Aquadro, C.F. (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *PNAS* 95: 10774–10778.
- Ludwig, M.Z., Palsson, A., Alekseeva, E., Bergman, C.E., Nathan, J., and Kreitman, M. (2005) Functional evolution of a *cis*-regulatory module. *PLoS Biology*. 3: 588–598.
- MacArthur, S., and Brookfield, J.F. (2004) Expected rates and modes of evolution of enhancer sequences. *Mol. Biol. Evol.* 21: 1064–1073.
- McVean, G., and Charlesworth, B. (2000) The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variability. *Genetics*. 155: 929–944.
- Prudhomme, B., et al (2006) Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. *Nature*. 440: 1050–1053.
- Rifkin, S.A., Kim, J., and White, K.P. (2003) Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nature Genetics*. 33: 138–144.
- Rifkin, S.A., Houle, D., Kim, J., and White, K.P. (2005) A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature*. 438: 220–223.
- Stone, J.R., and Wray, G.A. (2001) Rapid evolution of *cis*-regulatory sequences via local point mutations. *Mol. Biol. Evol.* 18: 1764–1770.
- Schmidt, D. and Durrett, R. (2004) Adaptive evolution drives the diversification of zinc-finger binding domains. *Mol. Biol. Evol.* 21: 2326–2339.
- Tsong, A.E., Tuch, B.B., Li, H., and Johnson, A.D. (2006) Evolution of alternative transcriptional circuits with identical logic. *Nature*. 443, 415–420.