MODELS AND THEIR EVOLUTION

JEFFREY L. THORNE

Jeff Thorne is a professor in the Genetics and Statistics Departments of North Carolina State University. He was born in 1963 and spent most of his childhood in Wisconsin. His undergraduate degrees were in Molecular Biology and in Mathematics (University of Wisconsin, Madison). In 1991, he received a Ph.D. in Genetics from the University of Washington. His research concentrates on the development of statistical techniques for studying DNA sequence evolution. – Address: 1507 Partners II, Bioinformatics Research Center, North Carolina State University, Box 7566, Raleigh, NC 27695-7566, USA.
E-mail: thorne@statgen.ncsu.edu

Long ago, my postdoctoral mentor told me that the one thing that statisticians and artists have in common is that neither should fall in love with their models. This advice has long stuck with me and I wondered who originated it. A perfunctory Google search did not answer the question, but it did support a suspicion that I have long had regarding the field of economics …

"Economists, like artists, tend to fall in love with their models – with decidedly less enjoyment, I imagine." (p. 437)[1].

Unfortunately, I am prone to sharing this weakness with economists. In my case, I rely upon probabilistic models of DNA sequence evolution as a central tool of my research. Throughout my four months at the Wissenschaftskolleg, I constantly tried to remind my-

[1] Leamer, E. E. 1993. "Factor-Supply Differences as a Source of Comparative Advantage." *American Economic Review* 83, 2: 436–439.

self that the statistical descriptions of molecular evolution, although useful, are crude descriptions of reality. During the course of these self-reminders, I noticed that others – even my three-year-old daughter Evelyn – also tended to put too much faith in unjustified models. In Evelyn's case, the models were rather parsimonious and she came to the widely held conclusion that truth can be created via repeatedly stating a belief. She explained again and again that Hohenzollerndamm was her favorite street because it was noisy, Warmbrunner Straße was my favorite street because it was quiet, and Im Dol was her mother's favorite street because it had a funny name. Alternative explanations are possible. For example, Hohenzollerndamm also happens to host a bakery where Evelyn and I stopped each day on our way home from her preschool.

The Wissenschaftskolleg Fellows relied upon more elaborate models than did Evelyn, and the very best part of my experience in Berlin was the opportunity to be exposed to this group of excellent scholars and their intellectual frameworks. Long ago when I made the decision to pursue an academic career, I had the naive idea that cross-disciplinary interactions would be part of the daily routine for university professors. For me, this was a major attraction of academia. In reality, I found that university life affords plenty of contact with those who have a similar research focus, but that disciplinary boundaries are not easily transcended. I am accustomed to interactions with biologists and statisticians and found the Wissenschaftskolleg experience to be a powerful complement to my usual environment.

The Wissenschaftskolleg Fellows were not simply a group of highly accomplished scholars with diverse expertise. This was a group of impressive individuals who were keen to learn about disciplines in which they were not formally trained. Too often, academics are narrow-minded and chauvinistic regarding their chosen field. The selection process employed by the Institute is a mystery to me, but it succeeded wonderfully. The other key factor in my Wissenschaftskolleg enjoyment was the personalized and intellectual environment of the Institute. The Wissenschaftskolleg staff deserves tremendous credit for fostering this ambiance. I had never experienced such a positive intellectual climate in my career and I am very grateful that I had the chance to be afforded this luxury.

I also very much enjoyed the vibrant collection of biologists that the Wissenschaftskolleg put together. I particularly appreciated Arne Mooers and Wayne Maddison. Prior to my arrival in Germany, I already knew of their work in phylogenetics and the respect it had earned. Soon after my arrival in Germany, I realized that these were true Renaissance men. Collectively, these two possess a knack for gourmet cooking, artistic talent (including the

best pumpkin-carving technique to which I have been a witness), impressive people skills, an understanding of international policy, and a worrisome affection for milk chocolate.

One of the best features of the Wissenschaftskolleg is the Tuesday colloquium. I was intrigued that almost every talk by a social scientist began with an admission that each researcher brings his or her personal biases to the topic being studied. The social scientist would then explain that they thought they could bring some insight to the topic despite their inherent biases. Among biologists and others who specialize in the natural and physical sciences, I believe the point about how research can be influenced by the biases of the researchers would be generally accepted, but this issue of researcher bias seems to get much more emphasis in the social sciences. I wonder whether the difference in emphasis stems from differences in the nature of the topics being studied or from the differences in the cultures of those who study the topics. Does the extreme awareness of how personal biases can affect conclusions make researchers hesitant to propose detailed models?

Without doubt, aversion to models can be a good thing. Models are inevitably oversimplifications of reality and are therefore almost guaranteed to be technically incorrect. Failure to recognize the limitations of model-based approaches can lead to serious mistakes.

This is particularly true regarding my own research field of evolutionary genetics. To enable inferences about evolutionary process and history from DNA sequence data, I construct probabilistic models of how sequences change over time. These models are inevitably flawed, and hence the aforementioned warning about how one should not fall in love with models applies.

However, one hopes that the model one adopts features the most important elements of the evolutionary process. An advantage of explicit probabilistic models is that assumptions can be statistically assessed. Assumptions that are particularly flawed can be replaced by better ones. The study of evolution with DNA is only a few decades old, but the evolutionary models being explored today are inarguably more realistic than those used in the past. On the other hand, there is no denying that the evolutionary models being explored today remain overly simplistic.

During my Wissenschaftskolleg tenure, my emphasis was on developing probabilistic models of DNA sequence change that improve treatment of the relationship between genotype and phenotype. In the jargon of biology, DNA represents "genotype" because it is the genetic material transmitted from parent to offspring, whereas characteristics that describe an organism's appearance or what it does are known as "phenotype". A principal aim of the field of genetics is to elucidate the connection between genotype and phenotype.

In general, genotype-phenotype connections are not well understood. At the same time, these connections are rapidly becoming better understood and can now be exploited to better characterize evolution.

For the three-dimensional structure of proteins and for a few other aspects of phenotype, the genotype-phenotype relationship is determined well enough to permit automated predictions of phenotype from genotype (i. e., DNA sequence). Although these prediction systems tend to be far from perfect, they have some merit. If a DNA sequence is predicted to encode a deleterious phenotype, then we can predict that natural selection will reduce the probability that the DNA sequence is found in a genome.

Conventionally, evolutionary biologists rely upon models of DNA sequence change that ignore the impact of phenotype on genotype. Prior to my arrival in Berlin, my collaborators and I had already been working with statistical procedures for making inferences when aspects of the phenotype influenced evolution of the genotype. Although we could estimate values for parameters in the mode, we wanted to be able to have more biologically meaningful interpretations of the parameters. This was the main goal that I set for my research while I was in Germany.

I was partially successful. We can now interpret values of model parameters in terms of how they affect the relative fitness of sequences. With this population genetic interpretation, we can estimate how a change in phenotype affects the rate of sequence evolution and we can estimate how a change in a DNA sequence affects the relative number of progeny that an organism is expected to have. A limitation of our population genetic interpretation is that it only applies to certain situations, such as a low mutation rate. One of the goals of our current work is to relax these assumptions.

As I write this, it is July and most of the Fellows whom I was lucky enough to meet are finishing their stay in Berlin. My stay was unfortunately much shorter and I have been practicing my usual routine since my return from Berlin at the end of January. My Wissenschaftskolleg experience seems now to be a very special time that occurred very long ago, but which I will always treasure.