Although the Department of Statistics at Berkeley decided they wanted to hire me in 1987, I didn’t take up my position there until 1989. I don’t have any recollection of meeting Terry when I interviewed, but, due in part to our shared Australian nationality, we became good friends shortly after I moved to Berkeley. Two years later, I jumped at the chance to move from my gloomy, north-facing office to one next to Terry’s. Its corner location with a view across the San Francisco Bay through the Golden Gate was merely an added inducement.

The resulting proximity led us to discuss scientific matters to a much greater extent. I was soon meeting with Terry and his students, serving on his students’ dissertation committees, and attending Terry’s weekly “statistics and biology” seminar. The thing that got me irredeemably hooked on the applications of statistics and probability to biology arose out of Terry’s work with his student Trang Nguyen on phylogenetics, the enterprise that seeks to reconstruct the evolutionary family tree of some collection of taxa (typically, species) using data such as DNA sequences. Phylogenetics was already a huge field in the early 1990s with a variety of statistical and non-statistical methods, and it has expanded greatly since then. Some idea of its scope may be gleaned from [42, 18, 19, 27].

Phylogenetic inference can be viewed as a standard statistical estimation problem [21]. One has a probability model for the observed DNA sequences that involves two kinds of parameters: those that define the mechanism by which DNA changes over time down a lineage and those that define the tree. The latter can be thought of as being further divided into a discrete parameter, the shape of the tree, and a set of numerical parameters, the lengths of the various branches (which represent either chronological time or evolutionary distance). In principle, the problem is therefore amenable to standard inferential techniques such as maximum likelihood or Bayesian methods. Unfortunately, likelihoods are somewhat expensive to compute for large numbers of taxa because they consist of large sums of products –
essentially, one has to sum over all the possibilities for the genetic types of the unobserved ancestors at each of the internal nodes of the tree. Even more forbidding is the fact that the number of possible trees for even a modest number taxa is enormous, so any approach that involves naively searching over tree space for the tree with maximal likelihood or summing and integrating over tree space to compute a posterior distribution will quickly become computationally infeasible, although there are widely used software packages that incorporate effective heuristics for maximizing the likelihood [20, 44, 43] and MCMC methods to computing posterior distributions [22, 23]. This computational difficulty is particularly galling because a significant proportion of the effort is expended to estimate the edge lengths of the tree and the parameters of the DNA substitution model, while the main object of interest is often just the shape of the tree.

Trang and Terry had come across an intriguing alternative approach to phylogenetic inference, the use of phylogenetic invariants, that had been proposed in [26, 11] and developed further in [9, 10]. The idea behind this approach is the following. Assume we have $N$ taxa. At any site there are $4^N$ possibilities for the nucleotides exhibited by the taxa. Each of these possibilities has an associated probability that is a function of the parameters in our model. It is usual to assume that these probabilities are the same for each site and that different sites behave independently. Suppose that for a given tree there is a collection of polynomial functions of these probabilities such that each function has the property it has value zero regardless of the values of the numerical parameters. Such functions are called phylogenetic invariants. Moreover, suppose that the values of these polynomials are typically non-zero when they are evaluated on the corresponding probabilities associated with other trees for generic values of the numerical parameters. The hope is that one can find enough invariants to distinguish between any pair of trees, estimate their values using the observed empirical frequencies of vectors of nucleotides across many sites, and then determine which tree appears to have the estimates of the values of “its” invariants close to zero and hence is a suitable estimate of the underlying phylogenetic tree.

In order to implement this strategy one needs ideally a procedure for finding all the invariants for a given tree. Because a linear combination of invariants is an invariant and the product of an invariant and an arbitrary polynomial is an invariant, the invariants form an ideal in the ring of polynomials, and so one actually wants to characterize an algebraically independent generating set. When Terry and I discussed this problem, we realized that the models of DNA substitution for which others had been successful in finding specific examples of invariants by ad hoc means were ones in which there is an underlying group structure. More specifically, if one identifies the nucleotides $\{A, G, C, T\}$ with the elements of the abelian group $\mathbb{Z}_2 \otimes \mathbb{Z}_2$ in an appropriate manner, then the substitution dynamics are just those of a continuous time random walk (that is, a processes with stationary independent increments) on this group. This suggested that we should attack the problem with Fourier theory for abelian groups – I should note that similar observations about the substitution models were made by others such as [48] around the same time. We found in our joint paper reproduced in this volume that the algebraic structure of
the likelihoods looks much simpler in “Fourier coordinates” and that one can determine a generating set for the family of invariants of a given tree using essentially linear algebra. We also proposed some conjectures on the number of “independent” invariants for various models that were verified subsequently in [16, 17].

It turned out that Terry and I had been like Molière’s Monsieur Jourdain in Le Bourgeois Gentilhomme who “for more than forty years” had been “speaking prose without knowing it.” The simple structure we observed after the passage to Fourier coordinates is an instance of a toric ideal, and we had unwittingly reproduced some of the elementary theory related to such objects. This connection was made in [45] and it led to a large amount of work using tools from commutative algebra to construct and analyze phylogenetic invariants in a number of different settings [1, 2, 7, 14, 6, 4, 5, 3, 8, 13]. Even tools from the representation theory of non-abelian groups have turned out to be useful in this context [46, 47]. Moreover, the investigation of phylogenetic invariants led in part to an appreciation of the extent to which many statistical models could be profitably studied from the perspective of commutative algebra and algebraic geometry, and this point of view is the basis of the field of algebraic statistics [36, 37, 40, 15].

An extremely important observation in phylogenetics is that evolution occurs at the level of genes and that different genes can have evolutionary family trees that disagree with the associated species tree. For example, genes can be duplicated and the duplicate can mutate to take on a new function, sometimes resulting in the loss of another gene that originally performed that function. Also, the lineages of orthologous genes (that is, genes descended from a common ancestral gene) in two taxa will split some time before the corresponding split in the species tree, and if this difference is sufficiently great the shape of the tree for a given gene will differ from that of the species tree. This means that in constructing a species tree one needs to resolve the incompatibilities observed between the trees constructed for various genes. On the other hand, if one has an accepted species tree, then it is desirable to reconcile a discordant gene tree with the species tree by describing how the above phenomena might have conspired to produce the differences between the two. This general problem is discussed in [39, 38, 35, 31].

The papers by Bourgon et al. and Wilkinson et al. reproduced in this volume carry out the task of clarifying the connection between a gene tree and a species tree in two important instances, the evolution of the serine repeat antigen in various Plasmodium species (including P. falciparum, the parasite responsible for the most acute form of malaria in humans) and the evolution of relaxin-like peptides across species ranging from humans to the zebra fish and the African clawed frog.

There has been considerable fascinating theoretical research on the problem of constructing species trees from gene trees, some of it showing quite paradoxical behavior; for example, the most likely gene tree can differ from the species tree and inferring a species tree by concatenating the sequences of several genes and treating the result as one gene can lead to an incorrect species tree with high probability [41, 12, 30, 25, 33, 32]. Some recent approaches to constructing well-behaved estimates of species trees using data from several genes are [29, 28, 24, 34].
The last of Terry’s work on molecular evolution reproduced here is his paper with Sidow and Nguyen on estimating invariable codons using capture-recapture methods. Invariable codons are those that are conserved across different species because of structural or functional constraints. In essence, they are codons that are prevented from changing because any change would have fatal biochemical consequences. It is not possible to observe which codons are invariable by simply looking at sequence data because some codons might be conserved by chance across all species even though there is no biochemical reason preventing a change, and so the invariable codons form some unknown fraction of the conserved ones. This paper is another example of Terry at his best: it provides answers of genuine scientific importance using simple, sensible statistical ideas that are normally not associated with the analysis of molecular data and which he probably learned from his extensive teaching and consulting experience.

Working with Terry has been one of the high points of my career at Berkeley. He has affected deeply the areas in which I have worked and my general attitude to research. Perhaps more importantly, by my good fortune of being his neighbor for around twenty years I have had an unrivaled opportunity to witness the humanity, dedication and commitment that he always shows to his students and collaborators. I may not have always lived up to the wonderful example he continues to set, but that does not make me any the less grateful for it.

References