C:/ITOOLS/WMS/CUP-NEW/12557476/WORKINGFOLDER/WINTHER/9781107111721C30.3D 323 [323–331] 24.2.2018 11:13AM

Anthony Edwards, Luca Cavalli-Sforza, and phylogenies

JOSEPH FELSENSTEIN

PROOF

Encountering Trees and Encountering Anthony

I want to concentrate here on Anthony Edwards's work with Luca Cavalli-Sforza on inferring phylogenies. I became aware of Anthony's work on this topic in 1966, when I first met him. I cannot remember seeing the 1963 abstract ([25]) or the 1964 paper ([27]) before then.

I was a graduate student in Richard Lewontin's laboratory at the University of Chicago, studying theoretical population genetics. I had become distracted from that work by being asked to help two faculty members, Jack Hubby and Lynn Throckmorton, analyze data on protein banding patterns in *Drosophila* species. They showed me Robert Sokal and Peter Sneath's important 1963 book Principles of Numerical Taxonomy; based on it, I had written a clustering program. I became intrigued by reconstructing trees from data. In late 1965, Robert Sokal, with Joseph Camin, published a paper in *Evolution* (Camin and Sokal 1965) on the reconstruction of phylogenies by parsimony methods from characters with discrete states. Shortly after that, Sokal visited our university, partly to see his former thesis advisor, the noted termite systematist Alfred E. Emerson. After hearing his seminar, I wrote a computer program to carry out a version of Camin and Sokal's method. About then, Harold Voris, a student of Throckmorton's, introduced me to the folks at the Division of Reptiles and Amphibians at the Field Museum of Natural History, especially Robert Inger and Hymen Marx, and I attempted an analysis of some of their data.

I was working on phylogenies, but not yet on their statistical analysis. In June 1966, I attended the Third International Congress of Human Genetics. This involved no travel or housing expenses, since it was held at the University of Chicago. At the meeting I met Anthony Edwards. He took me across the Midway to meet Luca Cavalli-Sforza, who had a room at the Center for Continuing Education. Luca wanted to know where Lewontin was. My answer bordered on the absurd: I had to tell him that Lewontin was then in Italy, on his way to a meeting. In fact, he

C:/ITOOLS/WMS/CUP-NEW/12557476/WORKINGFOLDER/WINTHER/9781107111721C30.3D 324 [323–331] 24.2.2018 11:13AM

visited Pavia, where he tried to look up Luca, only to be told that Luca was in Chicago.

From Edwards and Cavalli-Sforza's talk at the Congress, and from talking to Anthony, I became aware of their work, and of their 1964 paper ([27]).

Rereading that paper now, I am struck by how hard it works to justify interest in the phylogeny. They understood their audience, who were predominantly systematists. Overwhelmingly, systematists in the 1960s were focused on finding the correct classification. The dominant school of systematics was the 'evolutionary systematics' school led by Ernst Mayr and George Gaylord Simpson. They wanted to define groups that had some consistency with the phylogeny. They allowed monophyletic and paraphyletic groups, but not polyphyletic. The classification system of amniotes then included separate classes for birds, for mammals, and for reptiles. The former two are monophyletic groups, the latter was long known to be a paraphyletic group. The ancestor of Class Reptilia was a member of that group, but was also known to be an ancestor of birds and of mammals.

Today, the phylogeny needs no justification. There is no longer a single Class Reptilia, as paraphyletic groups are not allowed by the current consensus on classification. Currently, the inference of phylogenies is central to systematics and evolutionary biology. Although present-day systematists are reluctant to admit it, the classification system above the species level is fading in importance. But, back then, the phylogeny was, at most, an aid to classification. Edwards and Cavalli-Sforza spend much of the 1964 paper arguing the relevance of the phylogeny, knowing that the audience for the volume consists of systematists.

The paper describes the reconstruction of a tree of characters undergoing a Brownian-motion process from data on the tips of the tree. The Brownian-motion process is intended to approximate changes of p gene frequencies by random genetic drift. Edwards and Cavalli-Sforza give an expression for the likelihood of the change along one branch, where the elapsed time (or branch length) is t and the distance that the Brownian-motion process moves is d, the probability density being calculated from a p-variate normal distribution with independent variates that have an equal rate of change by Brownian motion. The resulting density depends only on the elapsed time, and on the distance d between the gene-frequency coordinates at the start and finish of this change. They also discuss the calculation of a distance from gene-frequency data, using a square-root transformation to make the Brownian-motion process more accurately approximate the genetic drift of gene frequencies.

These would seem to be straightforward, but there is a hidden problem. We do not actually observe the gene-frequency coordinates produced by the Brownian-motion process at the end of each branch. Edwards and Cavalli-Sforza proposed solving this problem by estimating these gene-frequency

C:/ITOOLS/WMS/CUP-NEW/12557476/WORKINGFOLDER/WINTHER/9781107111721C30.3D 325 [323–331] 24.2.2018 11:13AM

branch coordinates at the ends of branches, considering them as parameters of the problem. In fact, they were not able to do this. This is not discussed in [27]. The tree that they present for human blood groups is one derived from a different method, a "minimum-evolution" approach – in effect by gene-frequency parsimony.

The paper and the project underlying it were pioneering. For the first time, the inference of the tree was presented as a statistical problem. Maximum likelihood was proposed as one solution, and progress made on the likelihood function. Parsimony was proposed as another solution. I am not aware of any earlier attempt to use either method for the inference of phylogenies.

A similar development is given by Cavalli-Sforza, Barrai, and Edwards in the Cold Harbor Spring Symposium paper ([28]). There (p. 10), it is acknowledged that "Estimation by maximum likelihood is still causing some difficulties, because of the very high number of parameters to be estimated and some difficulties of the system." As in As in the Edwards and Cavalli-Sforza 1964 paper ([27]), gene-frequency parsimony is used for the computations instead. Less time is spent there on transformation of gene frequencies to make them be more accurately approximated by Brownian motion. There is an extensive discussion of the role of migration in villages of the Upper Parma Valley where the authors have collected blood-group data. But no method is given for combining migration with trees in a single analysis. That difficult problem is only now being tackled.

The Difficulty with Their Likelihood

Edwards and Cavalli-Sforza were finding their likelihood method ill-behaved, and they discussed this in several places (their 1966 and 1967 papers, [35] and 37]). It is now easy to see the source of the problem. They computed the likelihood as a product of terms, one for each branch in the tree, assuming that they were able to infer the gene-frequency coordinates at the ends of the branches. Taking these gene-frequency coordinates at the start and end of the branch, they could compute a distance d between those points. If the length of the branch in time was t, the term in the likelihood expression for that branch would simply be the density of the normal distribution

$$\frac{1}{\sqrt{2\pi t}} \exp(-d^2/(2t))$$

so that the corresponding term in the log-likelihood would be

PROOF

C:/ITOOLS/WMS/CUP-NEW/12557476/WORKINGFOLDER/WINTHER/9781107111721C30.3D 326 [323–331] 24.2.2018 11:13AM

$$-\frac{1}{2}\ln(2\pi t) - d^2/(2t)$$

What Edwards and Cavalli-Sforza found was that, if they made t approach zero, and the inferred coordinates at the end of the branch approached those at the start of the branch, these terms blew up in their faces. Likelihoods went off to infinity. The above expression shows why. If we took the estimates of the coordinates at the two ends of the branch to be the same, then $d = 0$.

Then, as $t \to 0$, the logarithm approaches $-\infty$ and so the whole expression rises to infinity. At first this seems strange. But recall that the term calculates, not a probability, but a probability density. Speaking informally, the probability of the change along that branch is the normal density with mean zero and variance t, multiplied by p infinitesimal quantities dx, where p is the number of dimensions in which the Brownian motion occurs. Once t reaches zero, the probability of no change becomes 1, and is not multiplied by infinitesimals. (Of course, I realize that all this is not strictly speaking mathematically kosher, but abi gezint[1]).

When the length of the branch, in time, reaches zero, the likelihood for the whole tree, which was a density function, is now a density function in p fewer dimensions, thus multiplied by p fewer infinitesimal quantities. So it is thereby infinitely higher. In effect, what we have allowed the problem to do is to merge two points that it is estimating. Since a zero branch length means that the coordinates at the two ends of that branch must be the same, the estimation is estimating one fewer internal point on the tree. In effect, the estimation problem is allowed to make its task simpler, and is infinitely strongly attracted to doing so.

A Solution

In 1967, when I was writing my PhD thesis (Felsenstein 1968), I was inspired by my contact with Edwards and Cavalli-Sforza to take a crack at the problem. Instead of trying to estimate the gene-frequency coordinates at all interior nodes of the tree, I tried to write down the density function for the tip coordinates only. This leaves out estimating the coordinates at the interior nodes; in effect, integrating over all possible interior node coordinates.

If we have a Brownian-motion process, the change in each branch is drawn from a normal distribution. The changes in each branch are independent, and the gene-frequency coordinates at a tip are then the sums of changes in the branches leading up to that tip. The set of the coordinates at the tips are drawn from a multivariate

[1] In my father's first language, Galitzian Yiddish, this means "as long as you're healthy," and is to be said with a resigned and accepting shrug.

C:/ITOOLS/WMS/CUP-NEW/12557476/WORKINGFOLDER/WINTHER/9781107111721C30.3D 327 [323–331] 24.2.2018 11:13AM

normal distribution, whose means are equal, and whose covariances can be easily worked out from the tree topology and branch lengths. The likelihood function can be written straightforwardly as the density function of a multivariate normal distribution, which involves evaluating a matrix inverse.

There also turns out to be a simple linear transformation, which can be read off the tree by a recursive algorithm, which diagonalizes the covariance matrix. This makes the computation of the likelihood linear in the number of species on the tree and linear in the number of dimensions. All of this eliminates the infinite likelihoods and makes the inference relatively well-behaved. I later published this work (Felsenstein 1973).

Distances Too

In 1967, three papers appeared that proposed distance matrix methods for inferring phylogenies. Most widely noticed was the paper by Fitch and Margoliash (1967), which appeared in *Science* at the beginning of the year. Later in the year, two papers appeared – or perhaps it's three. The ambiguity is because Cavalli-Sforza and Edwards published the same paper ([37]) twice that year. It was submitted to *Evolution*, and appeared in their October issue. But it was also delivered at a symposium of the American Society of Human Genetics, which required publication in *American Journal of Human Genetics*. The required mutual permissions were arranged, and it appeared in their May issue, so that the same paper appeared in two journals in the same year. Finally, in the November issue of *Evolution*, a paper by Sandra Horne (1967) appeared. All three of these papers proposed choosing the tree whose predicted distances best fit a table of pairwise distances between species. Fitch and Margoliash's paper proposed a weighted least-squares fit, and the other two papers proposed unweighted least-squares methods.

At the time of the publication of [27] and [28], Cavalli-Sforza and Edwards already had their least-squares method, but they did not mention it in that paper. It was Fitch and Margoliash's paper that caught the attention of molecular evolutionists. Distance matrix methods became an important tool, particularly in the form of the neighbor–joining method of Saitou and Nei (1987), which was faster than least-squares methods, and which found trees that were close approximations to those produced by least squares.

The Number of Trees

In [37], there is also a combinatorical calculation of the number of different rooted bifurcating trees for n labelled tips. This turns out to be the product of odd integers from 1 to $2n - 3$:

PROOF

C:/ITOOLS/WMS/CUP-NEW/12557476/WORKINGFOLDER/WINTHER/9781107111721C30.3D 328 [323–331] 24.2.2018 11:13AM

$$1 \times 3 \times 5 \times 7 \times \ldots \times (2n-3) = \frac{(2n-3)!}{(n-2)!2^{n-2}}$$

This is a particularly straightforward expression. The quantity was earlier derived by Ernst Schröder (1870) using generating function methods, without an explicit expression for the result. The Cavalli-Sforza and Edwards result is derived much more simply. They go on to give the number of unrooted bifurcating trees and the number of tree topologies. That latter quantity had also been derived by Wedderburn (1922). Cavalli-Sforza and Edwards's 1967 paper ([37]) is the start of the modern literature on enumeration of trees with labeled tips and enumeration of tree topologies.

A More General Formulation

Anthony's exploration of the inference of phylogenies culminates in his 1970 paper ([46]) in Series B of the *Journal of the Royal Statistical Society*. Considering the Brownian-motion process, he backs up one step further, and considers the tree as arising from a pure-birth process (a Yule process) with a parameter λ for the rate of branching. This gives a prior distribution on the topologies of the tree and on its branch length. That, taken together with the likelihood functions for the data, should permit finding the posterior distribution of trees. The existence of λ makes the inference of the tree in effect an empirical Bayesian inference, since it uses a prior that has a parameter.

The result is a general expression for the empirical Bayesian inference of the tree. It abandons any attempt to estimate the coordinates at the interior nodes of the tree, instead integrating over the values at those nodes. There, he makes use of my result that the joint density function of the coordinates at the tips is multivariate normal, with a covariance matrix derived from the tree.

In the end, he retreats from this empirical Bayesian treatment, using an approximate simulation method to estimate the branching rate λ. In those years, we had no way to do integrals needed in a Bayesian treatment. Markov Chain Monte Carlo methods were years in the future. The density function of the branching times for the tree, conditional on the number of tips, would be obtained by Elizabeth Thompson in her PhD work under Anthony's supervision, and published in her Smith's Prize monograph in 1975.

A Remarkable Achievement

Taken together, the papers on phylogenies by Anthony Edwards and Luca Cavalli-Sforza, published over a period of only seven years, introduce, for phylogenies,

PROOF

C:/ITOOLS/WMS/CUP-NEW/12557476/WORKINGFOLDER/WINTHER/9781107111721C30.3D 329 [323–331] 24.2.2018 11:13AM

parsimony methods, distance matrix methods, likelihood methods, and Bayesian inference. That all the major methods now in use for inferring phylogenies should spring from one collaboration is remarkable. Of course, some are easier to invent than others – parsimony apparently also occurred to Joseph Camin and Robert Sokal at nearly the same time; the development of distance methods was undoubtedly influenced by the spread of numerical clustering algorithms in the late 1950s and early 1960s. Thus, it is not surprising that Walter Fitch and Sandra Horne also published least-squares methods, and before Cavalli-Sforza and Edwards did, though Cavalli-Sforza and Edwards had the idea earlier.

In the chapter on history in my 2004 book, *Inferring Phylogenies*, I tried tell this remarkable story. What is most remarkable about it is that it is so little known. Why? Why did it take until 2015, half a century later, for Anthony to be elected to membership in the Royal Society?

I can think of a number of reasons:

1. Anthony and Luca each moved the main focus of their work elsewhere. Luca has of course become the leader of the effort to use molecular data to illuminate human prehistory. Anthony was inspired by the problem of inferring phylogenies to consider statistical inference more generally, and he became a leading advocate of likelihood methods.
2. The data that Edwards and Cavalli-Sforza used were gene frequencies derived from blood-group loci. Starting in 1966, gene frequencies from electrophoretic variation became available, though without any way of ensuring that alleles in different species were analogous. In the 1970s, the focus shifted strongly toward variation in DNA sequences, starting with restriction-fragment and restriction-site variation. The need for Brownian-motion models largely disappeared.
3. During the 1970s work on evolution of morphological characters became focused on non-statistical parsimony approaches. All of the methods used in that field became attributed to the German entomologist Willi Hennig. Hennig's methods were not numerical or statistical, so in those controversies, Brownian-motion models were not considered.

In the long run, the centrality of their work has been acknowledged:

1. In my own papers on inferring phylogenies using the Brownian-motion model (Felsenstein 1973, 1981), and in review articles that I wrote (Felsenstein 1982, 1983), I discussed Edwards and Cavalli-Sforza's work as well as subsequent work on statistical inference with molecular sequences (Neyman 1971, Kashyap and Subas 1974; Felsenstein 1981). Interest in gene-frequency trees waned; it was the analysis of molecular-sequence data that now interested most readers.

PROOF

C:/ITOOLS/WMS/CUP-NEW/12557476/WORKINGFOLDER/WINTHER/9781107111721C30.3D 330 [323–331] 24.2.2018 11:13AM

2. In the decade of the 2000s, single nucleotide polymorphism (SNP) sites started to be used intensively for inference of trees of descent of human populations. Likelihood inference for SNP sites is difficult, though progress has been made (RoyChoudhury *et al*. 2008, Bryant *et al*. 2012). As a check on the result, RoyChoudhury *et al*. used a Brownian-motion process as a quick approximation to the full likelihood analysis. Pickrell and Pritchard (2012) also turned to Brownian-motion processes to approximate gene-frequency changes in SNPs in their more widely noticed work using trees with added loops to model admixtures between human populations.
3. Phylogenies became more important as time passed. Until the late 1970s, systematists usually assumed that they would be inferring the classification, but that they would not have good enough information to infer the phylogeny. As molecular evolutionists inferred ever more phylogenies, this information began to be used routinely. By the 1990s, even genomicists were forced to admit that they needed phylogenies and coalescents to compare multiple genomes and investigate comparative genomics. We now realize that phylogenies are not simply an odd side-issue of little interest, but the central structure that is needed to study evolution above the species level.

Gradually, the work on inferring phylogenies using Brownian-motion models is becoming known, and with it, the work of Edwards and Cavalli-Sforza in the 1960s is coming to be appreciated. This is as it should be – it is the major starting point for all work on statistical inference of phylogenies, and that problem in turn is central to evolutionary biology.

References

Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution* 29, 1917–1932.

Camin, J. H. and R. R. Sokal. 1965. A method for deducing branching sequences in phylogeny. *Evolution* 19, 311–326.

Felsenstein, J. 1968. *Statistical inference and the estimation of phylogenies*. PhD Thesis, Department of Zoology, University of Chicago.

Felsenstein, J. 1973. Maximum likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics* 25, 471–492.

Felsenstein, J. 1981a. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17, 368–376.

Felsenstein, J. 1982. Numerical methods for inferring evolutionary trees. *Quarterly Review of Biology* 57, 379–404.

Felsenstein, J. 1983. Statistical inference of phylogenies. *Journal of the Royal Statistical Society. Series A (General)* 146, 246–272.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

C:/ITOOLS/WMS/CUP-NEW/12557476/WORKINGFOLDER/WINTHER/9781107111721C30.3D 331 [323–331] 24.2.2018 11:13AM

Fitch, W. M. and E. Margoliash. 1967. Construction of phylogenetic trees. *Science* 155, 279–284.

Horne, S. L. 1967. Comparisons of primate catalase tryptic peptides and implications for the study of molecular evolution. *Evolution* 21, 771–786.

Kashyap, R. L. and S. Subas. 1974. Statistical estimation of parameters in a phylogenetic tree using a dynamic model of the substitutional process. *Journal of Theoretical Biology* 47, 75–101.

Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems. In *Statistical Decision Theory and Related Topics*, eds. S. S. Gupta and J. Yackel, New York: Academic Press, pp. 1–27.

Pickrell, J. K. and J. K. Pritchard. 2012. Inference of population splits and mixtures from genome-wide gene frequency data. *PLoS Genetics* 8, e1002967.

RoyChoudhury, A., J. Felsenstein, and E. A. Thompson. 2008. A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics* 180, 1095–1105.

Saitou, N. and M. Nei. 1987. The neighbor–joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4, 406–425.

Schröder, E. 1870. Vier combinatorishe Probleme. *Zeitschrift für Mathematik und Physik* 15, 361–376.

Sokal, R. R. and P. H. A. Sneath. 1963. *Principles of Numerical Taxonomy*. San Francisco, CA: W. H. Freeman.

Thompson, E. A., 1975. *Human Evolutionary Trees*. Cambridge: Cambridge University Press.

Wedderburn, J. H. M. 1922. The functional equation $g(x^2) = 2ax + [g(x)]^2$. Annals of Mathematics 24, 121–140.

PROOF