Taking variation of evolutionary rates between sites into account in inferring phylogenies

Joseph Felsenstein

Department of Genetics, University of Washington,

Box 357360, Seattle, Washington 98195-7360

Abstract. As methods of molecular phylogeny have become more explicit and more biologically realistic following the pioneering work of Thomas Jukes, they have had to relax their initial assumption that rates of evolution were equal at all sites. Distance matrix and likelihood methods of inferring phylogenies make this assumption; parsimony, when valid, is less limited by it. Nucleotide sequences, including RNA sequences, can show substantial rate variation; protein sequences show rates that vary much more widely. Assuming a prior distribution of rates such as a gamma distribution or lognormal distribution has deservedly been popular, but for likelihood methods it leads to computational difficulties. These can be resolved using Hidden Markov Model (HMM) methods which approximate the distribution by one with a modest number of discrete rates. Generalized Laguerre quadrature can be used to improve the selection of rates and their probabilities so as to more nearly approach the desired gamma distribution. A model based on population genetics is presented predicting how the rates of evolution might vary from locus to locus. Challenges for the future include allowing rates at a given site to vary along the tree, as in the "covarion" model, and allowing them to have correlations that reflect three-dimensional

structure, rather than position in the coding sequence. Markov Chain Monte Carlo likelihood methods may be the only practical way to carry out computations for these models.

Keywords: Phylogeny – evolutionary rate – molecular evolution – maximum likelihood – distance – parsimony --invariants

The development of phylogeny algorithms was given a great stimulus by the rise of molecular evolution. Thomas Jukes played a central role in the growth and self-definition of this field. His work with Charles Cantor, setting forth the Jukes-Cantor model of molecular evolution (Jukes and Cantor 1969) was a very small part of his wider studies on the genetic code and protein evolution, but it was an essential development, serving as a starting point for later studies. His influence on molecular evolution (a field he named) was exerted through his many papers, his editorial work on this Journal, and his support of molecular evolution studies in Berkeley. The University of California at Berkeley became the most influential center for the empirical study of molecular evolution with the work of Jukes and his friend Allan Wilson.

Theoretical work in Berkeley has been less widely noted. It started with the Jukes-Cantor distance and the stochastic model derived from it. Shortly afterward, Berkeley's famous statistician Jerzy Neyman became involved, owing to contact with Allan Wilson. He was the first to describe maximum likelihood phylogeny methods for molecular sequence data, a development for which I often mistakenly get the credit. Other Berkeley theoretical work included Sarich and Wilson's (1973) relative rate test, and Wilson's “winning sites test (Prager and Wilson 1987).

It is easy to forget how many barriers faced the development of probabilistic models in molecular evolution. Foremost among these was the skepticism of molecular biologists, who raised the argument from total realism. We can now see that it was essential to start with oversimplified models and gradually make them more realistic. This was less obvious to the molecular biologists, who were apt to demand of a model that it take all possible real phenomena into account at the start. Tom Jukes once told me that the reason the Jukes-Cantor model was buried in the midst of a large empirical paper was that this was the only way to get it published. He felt that if he had attempted to publish it on its own, it would have been rejected by editors as idle and oversimplified speculation.

One of the greatest oversimplifications of molecular evolution models has been the assumption that all sites change at the same expected rate. This paper will review this assumption and the various methods that have been put forward to relax it. Along the way, I will describe an improved method of calculating likelihoods for Yang's (1995) discrete gamma approximation, and a population-genetic model for variation of evolutionary rates among loci. At the end of the paper, the new era of Markov Chain Monte Carlo methods will make a brief appearance on the horizon.

The constant rate assumption

We start by examining which methods of inferring phylogenies assume constancy of evolutionary rate. We will see that distance and likelihood methods, in their simplest forms, do make this assumption, but that the issue is subtler with parsimony methods.

Distance methods

Jukes and Cantor's (1969) distance is based on the simplest possible stochastic model of molecular change. Changes occur at all sites at the same expected rate, independently. When each nucleotide changes it has an equal probability of ending up at each of the other three nucleotides. This assumption of equal rates at all sites was carried over into Kimura's 2-parameter model (Kimura 1980) and appeared in many later models as well. Relaxation of this assumption has come from two sources. Nei et. al. (1976) used a gamma prior distribution of rates across loci in computing a genetic distance from gene frequencies. Olsen (1987) used a lognormal prior on rates across sites for the Jukes-Cantor DNA distance. He was not able to provide a closed-form formula for his genetic distance -- a numerical integration was needed. Jin and Nei (1990) used the gamma prior distribution for the Jukes-Cantor distance, for which a closed-form formula can be derived. Waddell et. al. (1996) showed how to apply a class of prior distributions to the LogDet distance, using the inverse of the Laplace transform of the distribution of rates. This inverse is known for the gamma distribution but not for the lognormal. For the gamma prior, we can rather easily correct most other distances. With the exception of the very general distances such as the Generalized Time Reversible (Lanave et. al. 1984), LogDet and general 12-parameter distributions, all other widely-used distances have transition probabilities of the form

$$\text{Prob}(x \mid y, r, t) \quad = \quad a_{xy}^{(0)} + \sum_{i=1}^{3} a_{xy}^{(i)} e^{-\lambda_i rt} \tag{1}$$

A gamma prior with mean 1 and squared coefficient of variation C simply causes the $e^{-\lambda_i rt}$ to be replaced by

$$(1 + \lambda_i t / C)^{-C}.$$

The two quantities approach each other as *C* decreases.

Although the replacement may result in not being able to find a closed-form formula for the distance, this need not be a serious problem. For models such as the Hasegawa-Kishino-Yano (1985) model, we do not have a closed-form formula for the maximum likelihood estimate of the distance. This does not prevent its rapid estimation by numerical iteration.

An aside about distributions

The reader will notice that we have repeatedly invoked the gamma and lognormal distributions. This is not because there is any reason to believe that the distribution of rates across sites actually follows either of these distributions. They are simply the two best-known distributions on the interval $[0,\infty)$ that have parameters for their mean and variance. There are other possibilities, such as the inverse Gaussian. However I suggest that it may be hard to accumulate evidence favoring one of these distributions over another. When the coefficient of variation (the ratio of standard deviation to mean) is small, they all approach narrow normal distributions around the mean rate. When the coefficient of variation is large, they still look somewhat similar. Figure 1 shows the densities of the gamma and lognormal distributions when the coefficient of variation is 2, which corresponds to the gamma distribution having the parameter $\alpha = 1/4$. The tail of the lognormal falls to zero at the extreme left, while the gamma's tail rises to infinity. There are some differences between the two distributions, but it may take a very large amount of data to effectively discriminate between them.

Correcting maximum likelihood methods

Maximum likelihood methods, as originally formulated for molecular data (Neyman 1971; Felsenstein 1981a) used models similar to distance matrix methods in that they assumed equal rates of evolution at all sites. If there is reason to assume that rates at different sites differ by known multipliers, this can be readily incorporated into likelihood methods (as it can also be incorporated into distance matrix methods). For example, if there is reason to think that first, second, and third codon positions in a protein-coding region have evolutionary rates that are in ratios of 0.8 : 1.0 : 2.7, we could use those factors in a maximum likelihood phylogeny program.

In cases where the rates of evolution are not known a priori, one could imagine inferring a separate rate for each site. This is done in Gary Olsen's program `dnarates`. It leads us to have a number of parameters that rises as we consider larger and larger numbers of sites. The amount of data per parameter does not increase as we add sites. Under situations like this, likelihood methods can fall into an “infinitely-many-parameters” trap. They can fail to have desirable statistical properties such as consistency.

Another approach that avoids this problem is to assume that the rates are drawn from particular distributions. The distributions have a few parameters, and the number of parameters then does not increase as the number of sites analyzed increases. This avoids the infinitely-many-parameters trap.

We have already seen the gamma and lognormal distributions used for this. Yang (1993, 1994, 1995) has introduced the use of the gamma distribution in correcting for rate variation among sites in maximum likelihood phylogenies. As long as one has only a few species, it is possible for each site to integrate the likelihoods over the distribution of rates. Thus if the likelihood of the site for tree *T* with rate *r* is *L(T, r),* and the density function of the distribution of rates is *f(r)*, the likelihood for the site will be

$$L = \int_0^\infty f(r)\, L(T,r)\, dr \tag{2}$$

For example for 4 species with an HKY model of DNA substitution and an unrooted tree with 5 branches, the expression for *L(T, r)* will be of the form

$$L(T,r) = \sum_{i=1}^{3^5} a_i \exp(-r b_i T_i) \tag{3}$$

where there will be $3^5 = 243$ terms, each with a different exponent $b_i T_i$ which is a linear combination of branch lengths from tree *T*. The integration in equation (2) can then be carried out termwise with no difficulty: it results in the *i*-th term simply having $e^{-r b_i T_i}$ replaced by $(1+b_i\ T_i\ C)^{1/C}$.

The likelihood for the tree is computed by taking the product over all sites of these integrals at each site (alternatively its logarithm is the sum of the logarithms of the integrals for each site).

For larger numbers of species this becomes impractical, the number of terms blowing up as 3^{2n-5}. Yang (1994, 1995) dealt with this case by approximating the gamma distribution by a discrete distribution with a fixed number of rates, using a Hidden Markov Model (HMM). An HMM was

also introduced for evolutionary rates by Felsenstein and Churchill (1996). This approximates the density *f(r)* by a histogram of rates, each with an associated probability of occurrence:

$$L \;=\; \int_0^\infty f(r)\; L(T,r)\; dr \;=\; \sum_{i=1}^{n} p_i\; L(T,r_i)\,. \tag{4}$$

Yang's (1994) method of choosing the rates r_i was to break the desired gamma distribution into n regions of equal area, and have the r_i be the medians (or else the means) of those regions. Thus if $n = 6$ under the median method, the r_i would be the 1/12, 3/12, 5/12, 7/12, 9/12, and 11/12 quantiles of the gamma distribution. When this quantile method is used, the p_i are simply taken as equal probabilities. We shall see below an improvement on this method.

Yang (1995) and Felsenstein and Churchill (1996) allowed for autocorrelation of rates in nearby sites; this can be done without great difficulty in the Hidden Markov Model framework, with little increase in computational effort.

Parsimony methods

It may seem a bit strange to be discussing how to correct parsimony methods for unequal rates at different sites, when they have no explicit model with well-defined rates. But unequal weighting of sites has been discussed since very early in the parsimony literature (Farris 1969). The reason for weighting some characters less has always been the suspicion that they might be unreliable, mostly owing to having a higher rate of evolution which then makes homoplasy less surprising in them. Farris proposed a successive weighting scheme which used a weighting function that depended on

the observed number of changes of state in the character. Thus if character *i* had n_i changes, their weight should each be $w(n_i)$. Farris suggested using these successively, by computing the weights based on the numbers of changes of the character on one tree, and then using those weights to search for the next tree. When the process converges, as it does quickly, one has found a set of weights and a tree that are consistent with each other.

A more general treatment of this sort of weighting is given by Goloboff (1997). He uses nonsuccessive weighting. For each tree, as it is evaluated, we calculate the n_i, and from those the $w(n_i)$; the parsimony score of that tree is then the sum over characters of the $n_i w(n_i)$. This eliminates the dependence of the score of a tree on which tree was looked at before it. A particular case of a nonsuccessive weighting method, threshold parsimony, had already been given by me (Felsenstein 1981b).

It is not immediately obvious what weighting function to choose. I have given (Felsenstein 1981b) a likelihood argument leading to the choice of a weighting function. According to that argument, changes ought to have weights that depended not only on the rate of evolution of the character. but also on the length of the branch in which they happen. The latter is not easily accommodated in a parsimony method, and this leads us onward toward likelihood methods. What is immediately clear, however, is that the common practice of weighting changes inversely proportional to the rate of change of that character is not correct. The correct weights depend on the distribution of rates in a complicated way. However the same likelihood-based weighting method does make it clear that in the limit in which all characters change at low rates (though different low rates) unweighted

parsimony is justified. If the rates differ but are all low, such that for any two characters *i* and *j* and any two branches *k* and *l*,

$$(r_i t_k)^2 < r_j t_l ,$$

then trees that require reconstructions of changes in which two changes occur in one character will always be less well supported than trees in which one change occurs in another character. The unweighted parsimony method will choose the same tree as maximum likelihood in this limiting case in which the rates of evolution, although different among sites, are all small. In this limiting case parsimony is expected to be robust against different evolutionary rates at different sites.

Invariants methods

Another family of phylogeny methods is invariants (or evolutionary parsimony), introduced by James Cavender (Cavender and Felsenstein 1987) and by James Lake (1987). Lake's invariants were intended to work even when evolutionary rates differ from site to site. For DNA under a simple model of base change, Lake showed that for a four-species tree of topology ((A,B),(C,D)) the linear combination of expected pattern frequencies (frequencies of outcomes at individual sites)

$$P_{xyxy} + P_{xyyx} + P_{xyzw} - P_{xyzz} - P_{xxzw} \qquad (5)$$

is zero, no matter what the branch lengths on the tree. In this computation these patterns include all those in which *x* and *y* are any two bases that are either both purines or both pyrimidines, and *z* and *w* are the other two bases. This relationship holds for the expected frequencies of the patterns. It is true for each site separately, even if the sites have different rates of change (and hence in effect different branch lengths). Being a linear combination, it then also holds for all sites combined, so that we can use the observed pattern frequencies such as n_{xyxy} summed over all sites.

The insensitivity of this relationship to variation of evolutionary rates from site to site is remarkable, but one pays a price for it. It uses only a few degrees of freedom out of many potentially available. Consequently in computer simulations (e.g. Huelsenbeck and Hillis 1993) Lake's invariants have proven to have very low power; although they work, they may require large amounts of sequence data to successfully infer the tree topology. Although they were developed for four-species cases. they could be extended to deal with larger trees. There has as yet been no attempt to develop nonlinear invariants to cope with rate variation. Being nonlinear, they do not cope with it automatically.

Laguerre Quadrature

The use of a gamma distribution of rates among sites in likelihood analysis of a tree with more than a few species leaves us with the issue of how many discrete rates to use to approximate the distribution. The more rates are used the better the approximation, but the slower the computation, as the computation time is proportional to the number of rates.

Fortunately the problem is simply one of numerical integration, and falls into the well-known class of numerical quadrature methods. Taking a numerical quadrature approach to the integration in equation (2) amounts to choosing rates r_i and probabilities p_i so as to most accurately approximate the integral. There are different numerical quadrature methods, depending on which weighting function *f(r)* is used. The gamma density corresponds precisely to a known numerical quadrature method, Generalized Laguerre Quadrature. This has a parameter α which corresponds to the shape

parameter of the gamma distribution. Having chosen a value of α and a number of points, *n*, the rates r_i are the roots of the Generalized Laguerre polynomial of degree *n*, and the weights p_i are can be computed as well (cf. Abramowitz and Stegun 1965, chap. 22).

Numerical quadrature methods are often quite accurate with a relatively modest number of points. Table 1 shows likelihoods achieved for a simulated data set under both methods. The data set had 10 species with 200 sites, simulated by a Jukes-Cantor (1969) model of base change. The first 100 sites had rate of change 0.2 per unit time, the second 100 had 0.4 per unit time. The data were analyzed using DNAML version 3.6 with gamma rate variation, coefficient of variation of rates 1.0 (corresponding to $\alpha = 1$), and a Jukes-Cantor model. The effect of using different numbers of rates is shown. The calculation with the quantile method was done with a modified version of DNAML written by Lindsey Dubb. Both analyses evaluate the same user-defined tree (it happens to be the ML tree for 9 rates for DNAML).

The results differ in log-likelihood. It is not clear whether this is a reason to prefer one or the other analysis, as they differ owing to the different scheme for choosing rates and probabilities. However it is evident that the log-likelihoods vary much less with different numbers of rates if the quadrature method is used than if the quantile method is used. The log-likelihood difference between 3 rates and 9 rates is 0.297 for quadrature, but 6.77 with the quantile method. The greatest difference between any two numbers of rates is 0.945 with the quadrature method, and 6.77 with the quantile method. Although the matter needs more careful examination, this example seems consistent with the expectation that the quadrature method will achieve a better approximation to the gamma distribution for any given number of rates.

The advantage of the quadrature method is that it approximates the gamma distribution by choosing both rates and probabilities. It is known that Generalized Laguerre Quadrature with *n* rates will be an exact method of integration with gamma distributed rates if the function *L(T, r)* (the likelihood for the site as a function of rate) is a polynomial of degree *2n-1* or less. Thus with 9 rates, it will give exact integration of a polynomial of degree 17 or less. It is able to do this because it is adjusting 9 rates and 8 probabilities (the ninth probability is determined, as they have to sum to 1). It is thus able to use a discrete distribution of rates that has its first 17 moments exactly match the first 17 moments of the gamma distribution. By contrast, the quantile method has only 9 parameters to adjust.

The two distributions of rates are quite different. Table 2 shows an example of rates and probabilities for the quantile method; Table 3 shows them for the quadrature method. It is quite noticeable that the quadrature method assigns a rather different distribution of rates. Some appear to be too high and too rare to be of use; nevertheless they seem to help the method make a better approximation of the gamma distribution. of these rates.

One limitation of the quadrature method is that it is not easy to see how to apply it to other distributions, particularly the lognormal. The quantile method can be adapted to the lognormal, but until we know what are the orthogonal polynomials that correspond to a lognormal weighting function, we will not have a quadrature method for lognormal distributions of rates. This is as far as we can go with the quadrature approach for now -- *Laguerre est fini*.

Rates varying across loci

A puzzling issue that arises when multiple loci are analyzed together is how to allow for the variation of rates of evolution from locus to locus. Yang (1996) suggests estimating a separate rate for each locus. One potential problem with this approach is that as the number of loci rises, the number of parameters being estimated rises proportionately. We are then at risk of falling into the infinitely-many-parameters trap, as we were when different rates were estimated for each site.

One could also assume that locus-wide rates were drawn from a distribution, as we do with sites. If both loci and sites within loci are vary in their rates of evolution, the issue arises as to how to combine these rates. A natural assumption is the multiply the two rates, so that the locus rate is in effect a multiplier affecting all sites in the locus. If both the sites and the loci have rates drawn from a lognormal distribution, it is a convenient property of this family of distributions that their product will also be lognormally distributed. There is no counterpart to this property for gamma distributions.

However, using the locus rates as a multiplier is an arbitrary procedure. It is illuminating to consider a model of rate variation and see what rules emerge from that.

A population-genetic model of rate variation

Imagine a locus with new mutants constantly arising, with the fitness of those new mutants in heterozygote drawn from a distribution. To be realistic, the distribution should have most mutants

with fitness at or below 1, so that there is only a small chance of an advantageous mutant. Of course, if there are *n* sites in the gene, there are only *3n* single-step base change mutations possible, which means that a continuous distribution of fitness may not be appropriate. However we could imagine that changes in the environment mean that different occurrences of the same base change might have different fitnesses. One must also be careful in translating this distribution into distributions of fitnesses of mutants at a single site, and working out the resulting rate of substitution at the site.

Suppose that the fitness of mutants comes from a normal distribution with mean μ and variance σ^2. We cannot simply translate this distribution into a distribution of rates of evolution, since natural selection intervenes and decides whether the mutant will or will not substitute. Kimura (1957, 1962) gave formulas for the fixation probabilities of mutants whose initial gene frequency is *p* and whose selection coefficient is *s* in a finite diploid population of population size *N*:

$$U(s, p) = \frac{1-\exp(-4Nsp)}{1-\exp(-4Ns)} \quad (6)$$

We are interested in the case where there is only one initial mutant copy, so that *p = 1/(2N)*, in which case the formula becomes

$$U(s) = \frac{1-\exp(-2s)}{1-\exp(-4Ns)} \quad (7)$$

To have some idea of how evolutionary rates will differ across sites, we should have some idea how fitness of various new mutants will differ. Suppose that we (arbitarily) decide that the fitness of new mutants at a locus varies according to a lognormal distribution with mean less than 1 and a small enough variance that only a tiny fraction of the distribution has fitness exceeding 1. We use a lognormal distribution rather than a normal distribution because the latter would allow some

mutants to have negative fitness. Figure 2 shows a lognormal distribution with mean 0.99997 and standard deviation 0.00001. Only a small fraction of new mutants have fitness greater than 1, and thus positive selection coefficients.

Figure 3 shows the fixation probabilities for mutants of different selection coefficients, calculated from Kimura's formula for a population with $N=10^6$. Note that although mutants with negative selection coefficients are overwhelmingly eliminated, a small fraction of ones that are slightly deleterious will be able to fix (for a review of this see Ohta 1996).

If we apply this fixation probability to the fitness distribution, we get as the distribution of selection coefficients of those mutants that succeed in substituting the distribution in Figure 4. Note the greatly different vertical scale. The original distribution was scaled to have area 1. The survivors have not been rescaled so that they have area 1 -- they are a small corner of the original distribution. The elimination of most of the mutants by natural selection is dramatic.

This still does not get us the distribution of evolutionary rates across sites. The distribution across mutations can be obtained by transforming the distribution of selection coefficients into the distribution of evolutionary rates. If $\phi(s)$ is the density function of selection coefficients, (easily obtained from the density function of fitnesses of the mutants) the density function of evolutionary rate will be

$$f(r) = \phi(s)\left(\frac{dU}{ds}\right) \tag{8}$$

where the derivative is easily obtained from Kimura's formula. More precisely, the evolutionary rates are the product of the mutation rate and *U(s)*.

An important concern about this calculation is that it gives us only the distribution of rates among mutants. There are three possible mutants at each site. To the extent that these may have different selection coefficients, we would need to modify this calculation to average rates over the possible mutants at the site.

How will the rates vary across loci? We may feel that some loci will be more carefully scrutinized by natural selection than others. A naive model of "rounds" of selection, with fitness interpreted as viability, will lead us to expect that if at one locus each mutant is scrutinized four times as severely as at another locus, the fitness distribution at the first locus will the distribution of w^4, the probability of the mutant surviving all four rounds of selection. It is here that assuming a lognormal distribution helps. If *w* is lognormally distributed, w^4 will also be lognormally distributed. In fact, since $\ln(w^4) = 4\ln(w)$, the distribution will simply be shifted on the logarithmic scale.

Figure 5 shows two rate distributions computed from equation (8). One (the dashed line) has *w* raised to the fourth power compared to the other. The results are reminiscent of gamma distributions with large coefficients of variation (small values of the α shape parameter).

This model of rate variation is admittedly crude, and the choices of a lognormal distribution of fitnesses of mutants is arbitrary. The matter needs more careful development before it can be said to provide a justification for the use of gamma distributions of rates. To make a full Hidden Markov

Model of variation of evolutionary rates among loci, we would also need to add a distribution of amounts of evolutionary surveillance (the power 4 in our example) among loci.

Variation between branches

The simplifying assumption that rates should be the same in all branches of the tree also needs to be relaxed. Sanderson (1997) did so using a rate-smoothing method which constrained rates in neighboring branches of the tree to be similar. Thorne et. al. (1998) have used Markov Chain Monte Carlo (MCMC) methods to integrate likelihoods over a phylogeny on which branch lengths can vary among branches, but in the context of an approximate clock. Bickel and West (1998) have used a fractal Poisson process, Cutler (2000) a general stationary process, and Huelsenbeck et. al. (2000) have a compound process approximation. All of these allow variation of evolutionary rates from branch to branch. Huelsenbeck et. al. used MCMC methods to infer the tree. We can also allow the rates to vary among sites and among branches of the tree, with which sites have the highest rates of change varying from one part of the tree to another. This is the covarion model of Fitch and Markowitz (1970). While they adduced evidence for it, use in practice has been greatly impeded by the difficulty of computing likelihoods under it. The rise of Markov Chain Monte Carlo methods, which would randomize over assignments of rates of rates to sites and to branches of the phylogeny, may have finally created the conditions for the covarion model to be of practical use.

The future

With Hidden Markov Model and Markov Chain Monte Carlo methods becoming widely used, the incorporation of rate variation into inference of phylogenies has become common. MCMC methods will perhaps allow rates to be correlated, not only along the nucleotide or amino acid sequence, but also in three-dimensional space in a molecular structure.

It seems evident that the dominant framework in which these developments will occur is maximum likelihood inference (and/or the related Bayesian approach). Currently there is much interest in rapid distance matrix methods for inferring phylogenies with large numbers of species. However distance methods have a severe limitation: they cannot cope efficiently with rate variation among sites. As we have seen, rate variation can be taken into account in distance methods, but this happens separately for each pair of species. Distance methods cannot carry over from one pair of species to another the information as to which parts of the molecular have the highest rate of evolution. We are thus unable to accumulate, from one part of the tree to another, an assessment of which parts of the molecule have high rates of change.

Likelihood methods can do this, so that they can deal efficiently with the accumulation and disposition of this evidence. There is thus a tension between rapid computation and efficient inference. If the ability of computers to carry out computations rises faster than an appropriate power (approximately, the cube) of the number of species in a typical data set, we can expect the competition to favor the more efficient methods.

Acknowledgments

I am indebted to Monty Slatkin for knowing the pun involving Laguerre, to Lindsey Dubb for use of his program which computes likelihoods under the quantile method, and to Leona Wilson for insights into the history of Allan Wilson's work. Research for this paper was funded by NIH grants GM-51929 and HG-01989, and NSF grant DEB-9815650.

References

Abramowitz, M and Stegun, IA (1965) *Handbook of Mathematical Functions*. Dover Publications, New York.

Bickel, DR and West, BJ (1998) Molecular evolution modeled as a fractal Poisson process in agreement with mammalian sequence comparisons. Mol Biol Evol 15:967-977

Cavender, JA and Felsenstein, J (1987) Invariants of phylogenies in a simple case with discrete states. J Classif 4:57-71.

Cutler, DJ (2000) Estimating divergence times in the presence of an overdispersed molecular clock. Mol Biol Evol 17:1647-1660

Farris, JS (1969) A successive approximations approach to character weighting. Syst Zool 18:374-385

Felsenstein, J (1981a) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-376

Felsenstein, J (1981b) A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol J Linnean Soc 16:183-196

Felsenstein, J and Churchill, GA (1996) A Hidden Markov Model approach to variation of evolutionary rates among sites. Mol Biol Evol

Goloboff, PA (1997) Self-weighted optimization: Tree searches and character state reconstructions under implied transformation costs. Cladistics 13:225-245

Hasegawa, M, Kishino, H, and Yano, T (1985) Dating of the human-ape splitting by a molecular clock of mitchondrial DNA. J Mol Evol 22:160-174

Huelsenbeck, JP and Hillis, DM (1993) Success of phylogenetic methods in the 4-taxon case. Syst Biol 42:247-264

Huelsenbeck, JP, Larget, B, and Swofford, D (2000) A compound poisson process for relaxing the molecular clock. Genetics 154:1879-1892

Jin, L and Nei, M (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol 7:82-102

Jukes, TH and Cantor, C (1969) Evolution of protein molecules. pp. 21-132 *in* M. N. Munro, ed. *Mammalian Protein Metabolism*. Academic Press, New York.

Kimura, M (1957) Some problems of stochastic processes in genetics. Ann Math Stat 28:882-901

Kimura, M (1962) On the probability of fixation of mutant genes in a population. Genetics 47:713-719

Kimura, M (1980) A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111-120

Lake, JA (1987) A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol 42:167-191

Lanave, C, Preparata, G, Saccone, C and Serio, G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86-93

Nei, M, Chakraborty, R, and Fuerst, PA (1976) Infinite allele model with varying mutation rate. Proc Natl Acad Sci USA 73:4164-4168

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

Ohta, T (1996) The current significance and standing of neutral and nearly neutral theories. Bioessays 18:673-677 (discussion, p. 683)

Olsen, GJ (1987) Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp Quant Biol 52:825-837

Prager, EM, and Wilson, AC (1988) Ancient origin of lactalbumin from lysozyme: Analysis of DNA and amino acid sequences. J Mol Evol 27:326-335

Sanderson, MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 14:1218-1231

Sarich, VM, and Wilson, AC (1973) Generation time and genomic evolution in primates. Science 179:1144-1147

Thorne, JL, Kishino, H, and Painter, IS (1998) Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15:1647-1657

Waddell, P. J. and M. A. Steel. 1997. General time-reversible distances with unequal rates across sites: mixing gamma and inverse Gaussian distributions with invariant sites. *Mol Phylog Evol* 8:398-414

Yang, Z (1993) Maximum-likelihood-estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396-1401

Yang, Z (1994) Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites - approximate methods. J Mol Evol 39:306-314

Yang Z (1995) A space-time process model for the evolution of DNA sequences. Genetics 139:993-1005

Yang Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42:587-596

Table 1. Likelihoods achieved for different numbers of rates in a simulated 10-species data set with 200 sites and a true coefficient of variation of 0.333 among sites. Likelihoods for a Jukes-Cantor model of base change and coefficient of variation of rates among sites 1 ($\alpha = 1$).

number of rates	ln L	
	quadrature method	quantile method
3	-1909.49687	-1934.41557
4	-1908.84881	-1936.97424
5	-1909.08548	-1938.49597
6	-1909.40984	-1939.50557
7	-1909.62518	-1940.22490
8	-1909.73959	-1940.76365
9	-1909.79412	-1941.18232

Table 2. The rates and probabilities chosen by the quantile method for 6 rates and coefficient of variation of rates among sites 1 ($\alpha = 1$).

	probability	rate
1	1/6	0.092
2	1/6	0.305
3	1/6	0.571
4	1/6	0.928
5	1/6	1.469
6	1/6	2.634

Table 3. The rates and probabilities chosen by the quadrature method for 6 rates and coefficient of variation of rates among sites 1 ($\alpha = 1$).

	probability	rate
1	0.278	0.264
2	0.494	0.898
3	0.203	1.938
4	0.025	3.459
5	0.00076	5.617
6	0.000003	8.823

Figure Captions

Fig. 1. The gamma and lognormal densities with a coefficient of variation of 2 and expectation 1.

Fig. 2. A lognormal distribution of fitnesses of new mutants, with only a small fraction of advantageous mutants. The horizontal scale is not fitness *w* but selection coefficient *s*, where $w = 1+s$.

Fig. 3. Fixation probabilities of single mutants with different selection coefficients, computed from Kimura's formula in equation (7).

Fig. 4. Distribution of selection coefficients of mutants that succeed in substituting, when the original distribution of mutants is as in Figure 2.

Fig. 5. Distributions of rates of substitution among mutants for two distributions of fitnesses of new mutants, one with four times as much selection on new mutants as in the other.

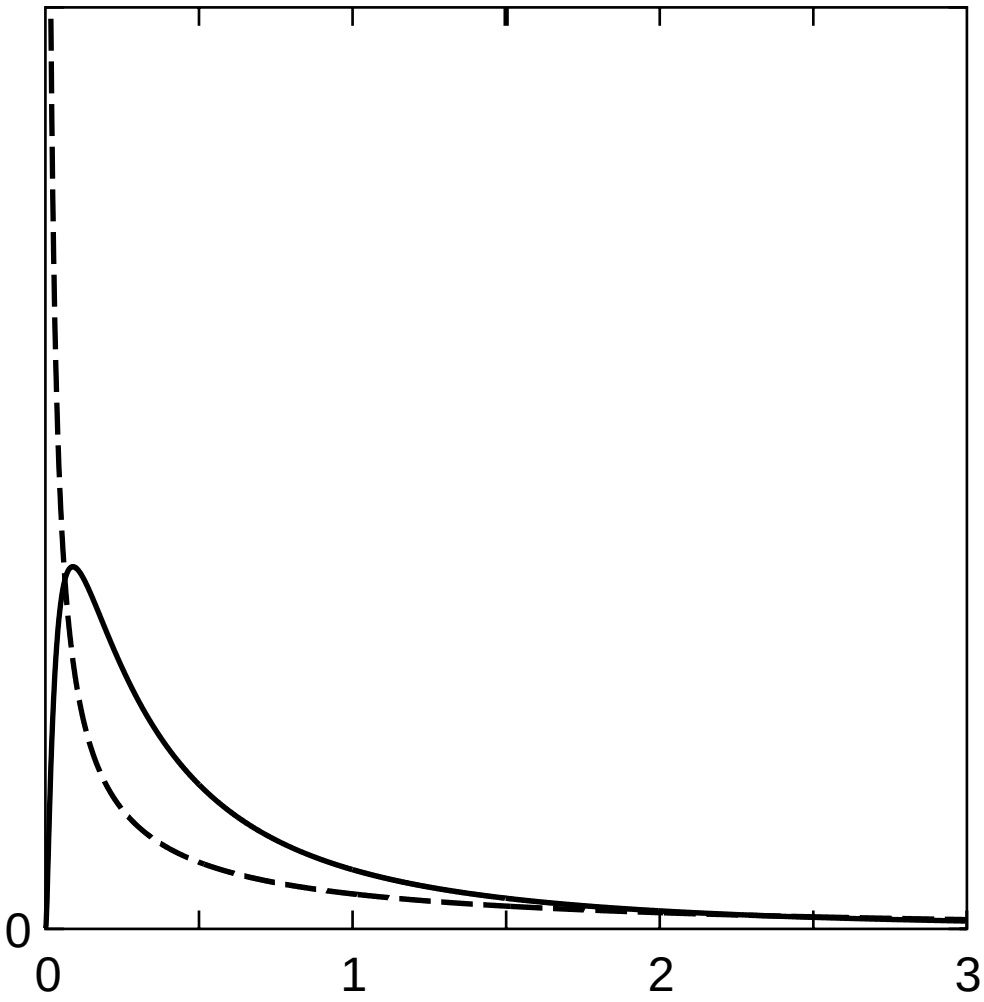

Figure 1

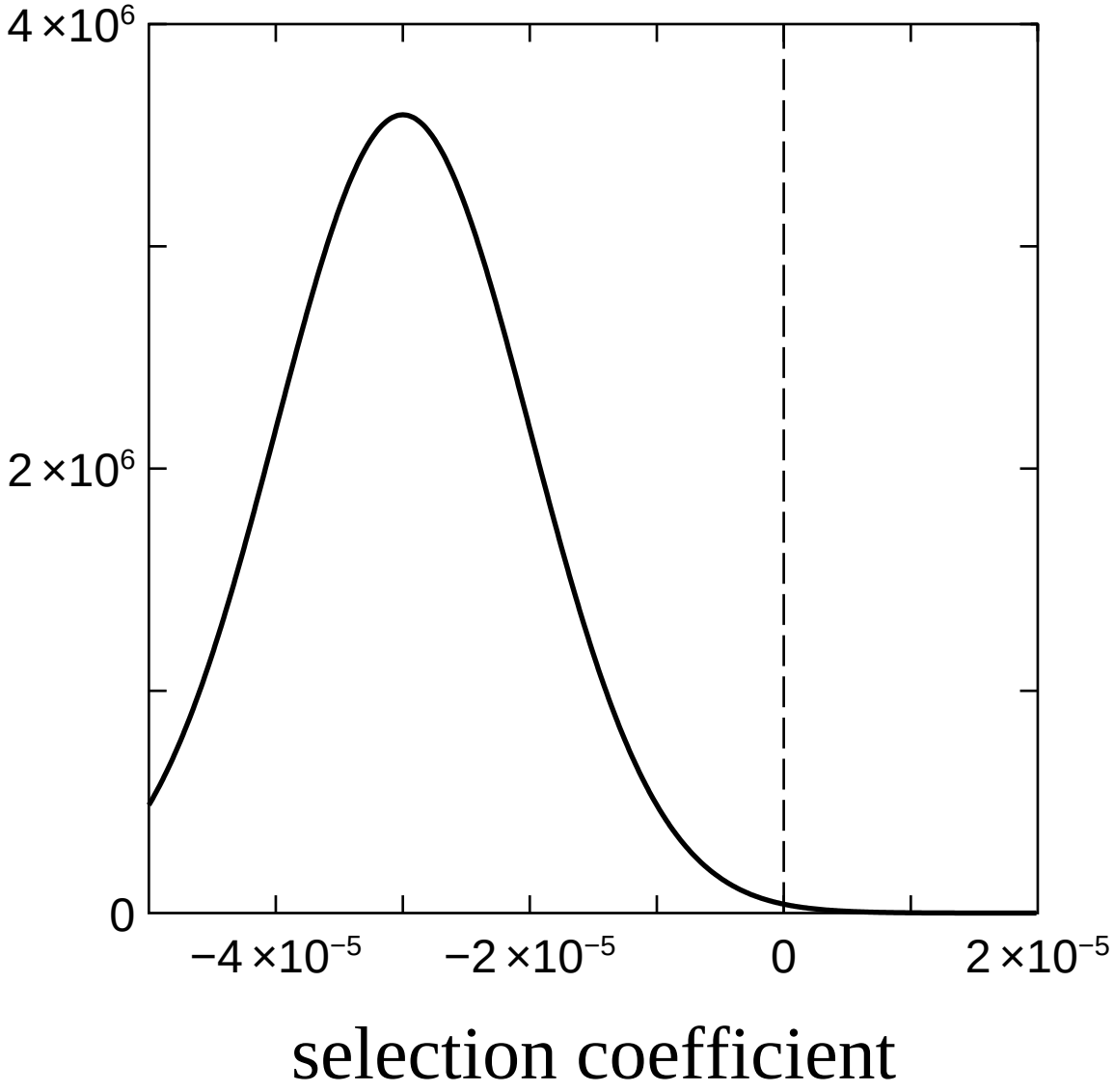

Figure 2

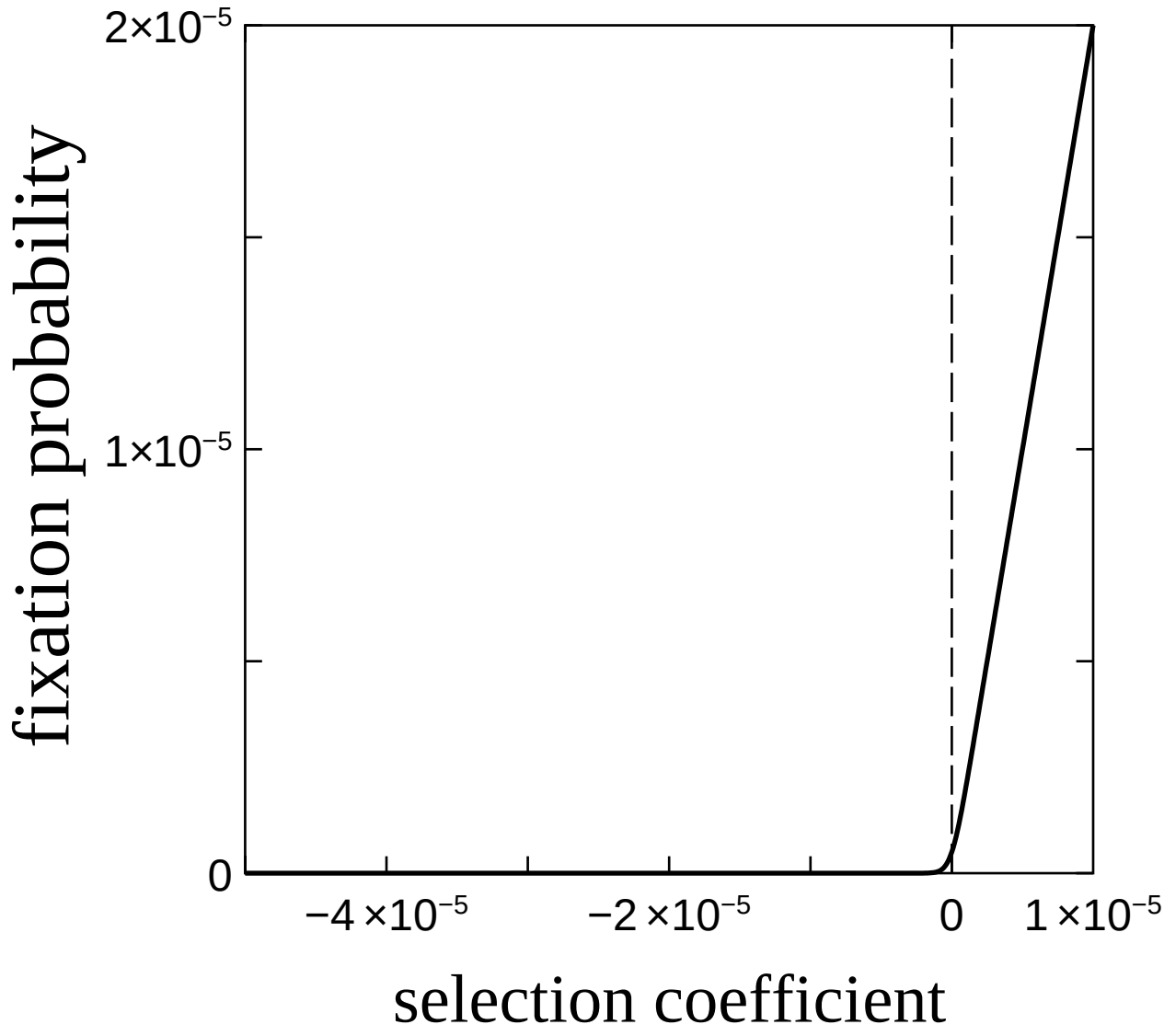
2×10⁻⁵
1×10⁻⁵
0
fixation probability
−4 ×10⁻⁵
−2 ×10⁻⁵
0
1 ×10⁻⁵
selection coefficient

Figure 3

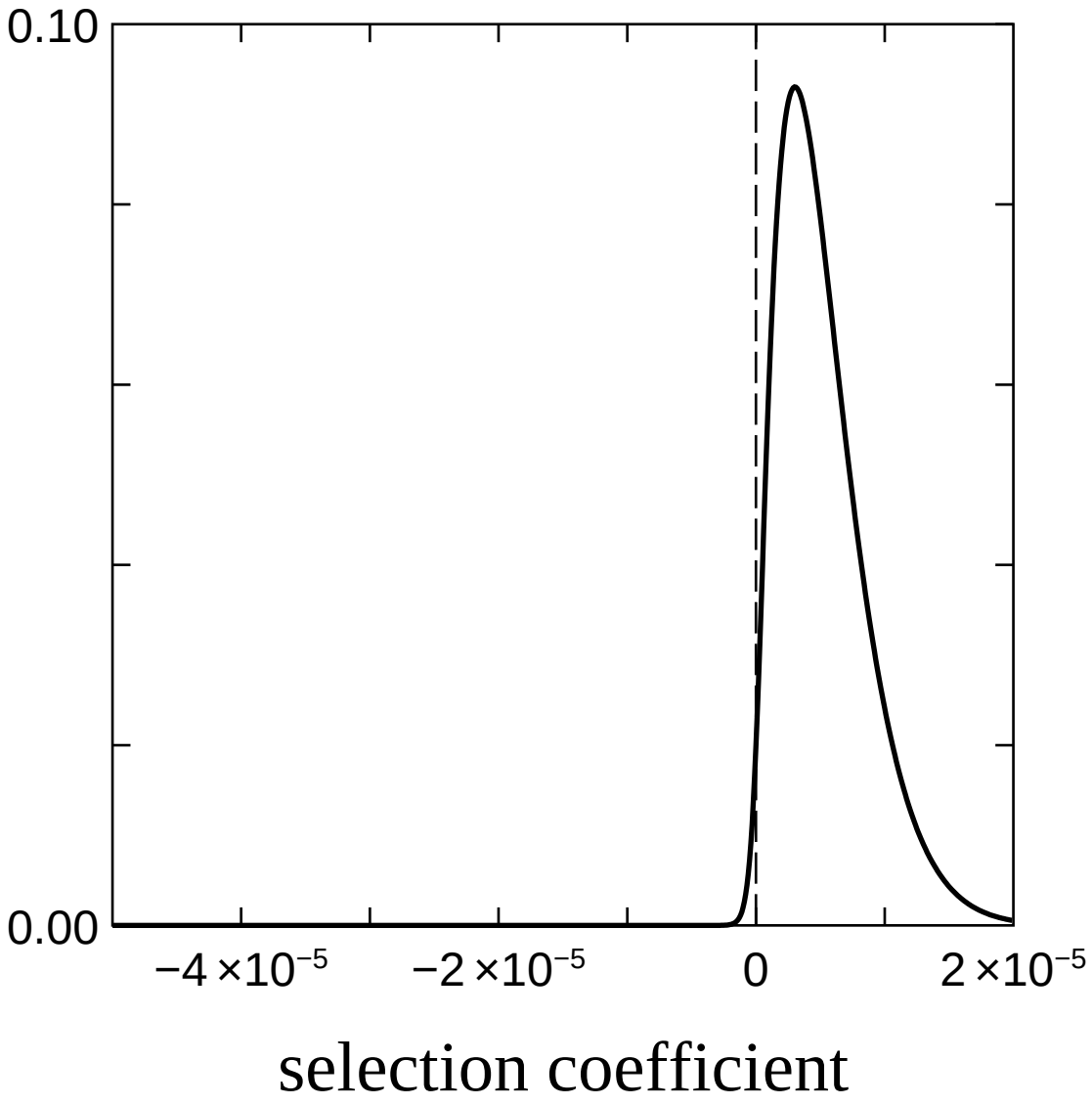

Figure 4

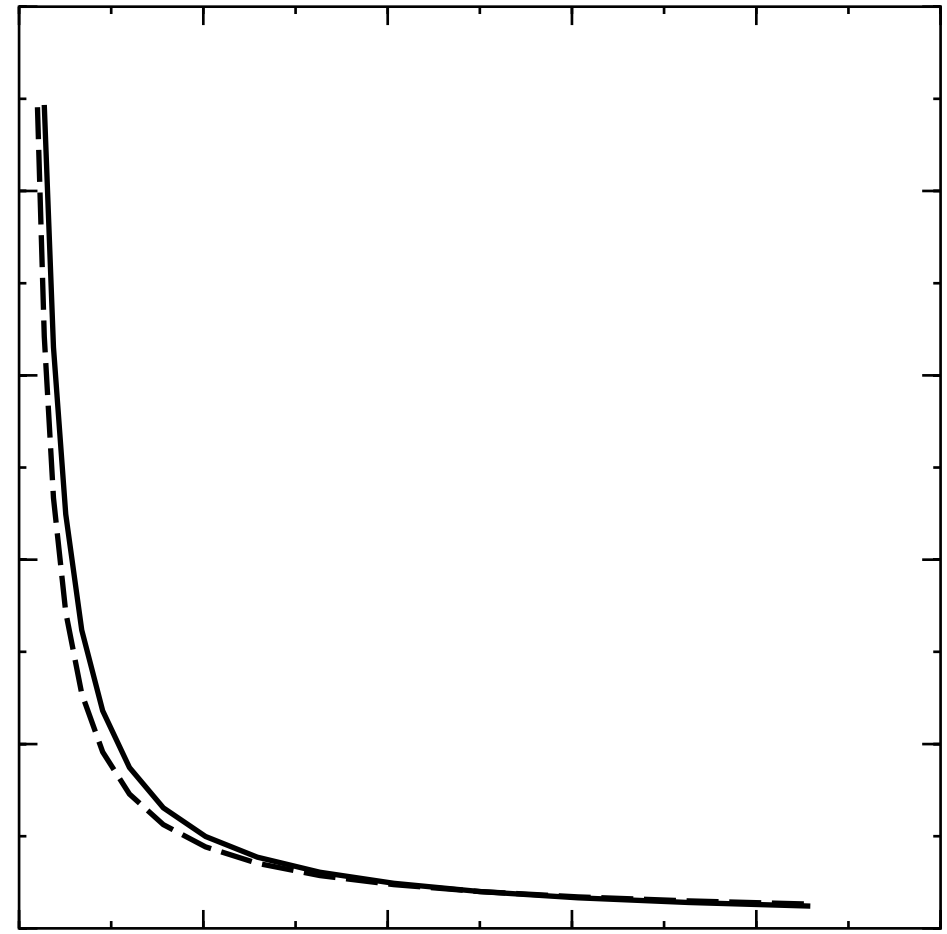

Figure 5