BMC Bioinformatics

BioMed Central

Methodology article

Comparison of mode estimation methods and application in molecular clock analysis

S Blair Hedges* and Prachi Shah

Address: NASA Astrobiology Institute and Department of Biology, Pennsylvania State University, 208 Mueller Laboratory, University Park, PA 16802-5301, U.S.A

Email: S Blair Hedges* - sbh1@psu.edu; Prachi Shah - pss11@psu.edu

* Corresponding author

Published: 31 July 2003

BMC Bioinformatics 2003, **4**:31

Received: 26 May 2003
Accepted: 31 July 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/31

Abstract

Background: Distributions of time estimates in molecular clock studies are sometimes skewed or contain outliers. In those cases, the mode is a better estimator of the overall time of divergence than the mean or median. However, different methods are available for estimating the mode. We compared these methods in simulations to determine their strengths and weaknesses and further assessed their performance when applied to real data sets from a molecular clock study.

Results: We found that the half-range mode and robust parametric mode methods have a lower bias than other mode methods under a diversity of conditions. However, the half-range mode suffers from a relatively high variance and the robust parametric mode is more susceptible to bias by outliers. We determined that bootstrapping reduces the variance of both mode estimators. Application of the different methods to real data sets yielded results that were concordant with the simulations.

Conclusion: Because the half-range mode is a simple and fast method, and produced less bias overall in our simulations, we recommend the bootstrapped version of it as a general-purpose mode estimator and suggest a bootstrap method for obtaining the standard error and 95% confidence interval of the mode.

Background

It is not uncommon in many fields to encounter data distributions that are skewed or contain outliers. In those cases, the arithmetic mean may not be an appropriate statistic to represent the center of location of the data. Alternative statistics with less bias are the median and the mode. The median is the value of the variable, in an ordered array, which has an equal number of data points on either side, whereas the mode is the value of the peak of the distribution (Figure 1). The mode is biased the least by outliers and contaminants [1–3] and is used commonly in astronomy [4,5] and occasionally in other fields, including biology [6–10]. However, calculation of the mode is more difficult than the mean or median and this has limited its widespread application.

For discrete data involving a relatively small number of possible values and a large number of data points, the mode is easily calculated as the most frequent value. Otherwise, the most common mode estimation method for discrete or continuous data involves construction of a histogram. The value of the bin with the greatest number of data points is the mode, and this value can be fine-tuned by simple interpolation with adjacent bins [11]. The major drawback of the histogram method is that different modes can be obtained using different bin sizes, although

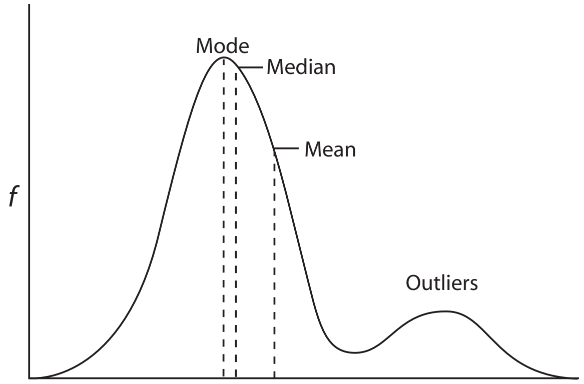

Figure 1
A normal distribution with outliers, showing the relative positions of the mean, median, and mode. In this case, the outliers (contaminants) are normally distributed and centered at twice the distance between the true mode and the 99th percentile of original normal distribution and account for 20% of the total data points.

some stability can be gained by using the mean of modes obtained from different bin sizes. For continuous data, two other simple methods have been proposed. The Dalenius method [1] is calculated by selecting the interval with the maximum number of data points and using the mean of that interval as the mode. The Grenander method [12] uses two parameters; one defines an interval by limiting the number of data points in the interval and, the other defines the weight exponent applied as penalty to the range of an interval. However, the former method is sensitive to the size of the interval selected, and the latter method is sensitive to outliers [13].

Several new mode estimation methods have been proposed in recent years. Two related ones are the Half-Sample Mode (HSM) and Half-Range Mode (HRM) [2]. The HSM uses shortest half samples, in an iterative fashion, to estimate the mode. Similarly, the HRM uses densest half samples to estimate the mode. Two substantially different methods are the Standard Parametric Mode (SPM) and the Robust Parametric Mode (RPM) [13]. These methods transform the data distribution to an approximate normal and use the probability density function of the approximated distribution to estimate the mode. The SPM uses the mean and standard deviation, whereas the RPM uses the median and the median absolute deviation as the average and variance parameters. Each of these methods has been tested in simulations with normally and asymmetrically distributed data, with and without contamination [13]. The results showed that different estimators perform better under different conditions, but that the RPM method is perhaps more versatile than the others.

Non-normal distributions of divergence time estimates are often obtained in molecular clock studies involving many genes, or with genomic data [7,8,14]. Typically those distributions are right skewed, or have noticeable outliers on the right (older) side. The reason for such non-normal distributions is unclear, but they are consistent with contamination from paralogous comparisons (those involving gene duplications), although other sources of bias have been proposed [15]. (The right-skewed distributions seen in some plots of evolutionary distances are different, because they have not been normalized by a calibration time). It is sometimes difficult to distinguish a paralogous comparison from an orthologous comparison (speciation event) when comparing different species, because of gene loss and incomplete taxon sampling. For these reasons, the mode – as estimated using the histogram method – has been used to avoid the bias of such outliers [7].

In this study, we extend the analyses of Bickel [13] to assess the accuracy and precision of mode estimation methods for continuous data using simulations. We evaluate additional types of distributions, the effect of coarse data (sub peaks), and additional levels and distributions of contamination. Finally, we applied these different estimators to five published data sets from a molecular clock study. Our immediate goal was to determine the optimal method for use with molecular clock data (divergence times) involving large numbers of genes, although the results are generally applicable.

Results

General Patterns

The mode estimators were not biased by randomly distributed contamination, and the cauchy-distributed contamination produced similar results (albeit at higher levels of bias) to the normally distributed contamination. For these reasons, we confine our discussion to results using the normally distributed contamination. Also, simulations using contaminant data located close to the true mode (67th percentile) showed that mode estimators performed poorly under those conditions, as noted elsewhere [2], especially with the normal and coarse original distributions. In those cases, all estimators (mean and modes) showed similar bias, being equally misled by the contamination. On the other hand, most mode estimators performed better (lower bias) than the mean or median in simulations using contamination located at the 99th and twice the 99th percentiles. Because we wished to compare the efficiencies of the various mode estimators, we further confine the discussion to results using

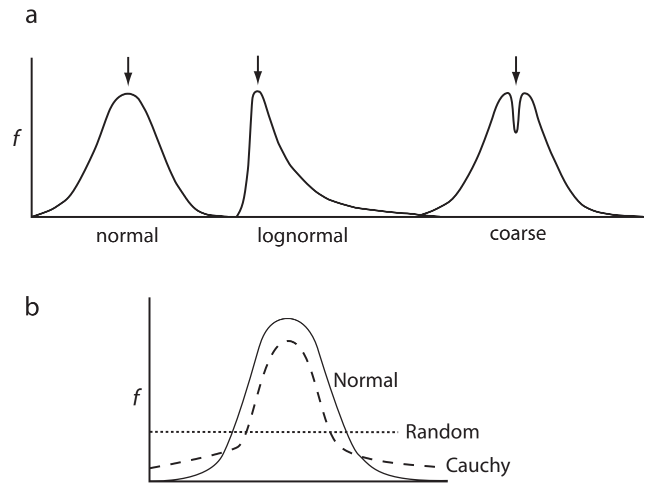

Figure 2
Data distributions used in the study. (a) Original distributions tested: normal, lognormal, and coarse. Arrows indicate the location of the true mode. (b) Contaminant distributions tested: normal (solid), cauchy (dashed), and random (dotted).

contamination centered at twice the 99th percentile (Figs. 1, 4, 5; Table 1).

The normal and coarse distributions produced similar results for the different mode estimators, although HRM showed the least bias as contamination level increased (Figure 4e,4i). RPM and SPM showed an increase in bias in a negative direction (away from contamination) as contamination level increased. However, those two estimators did not show such negative bias with the lognormal distribution (Figure 4e).

All estimators of central location (mean, median, mode) performed well (had little or no bias) when contamination was absent, at least with the normal and coarse original distributions. Bias increased as the level of contamination increased, with RPM and SPM exhibiting negative bias at higher levels in the normal and coarse original distributions (as noted previously). In almost all cases (at different levels of contamination), the mean exhibited the greatest bias, followed by the median and then various mode estimates. The HRM method showed almost no bias at any level of contamination and original distribution type (Figure 4). However, the highest levels of contamination (40%) produced a spike in the bias and variance of several mode estimators (especially HRM), probably because the contamination was a significant peak in the distribution, competing with the true mode. The SPM method generally produced the highest bias of the mode estimators, in some cases even performing more poorly than the median. These patterns also are reflected in rankings of the methods (Table 1).

Bootstrapping

Molecular Clock Analyses

Analysis of the published molecular clock data for fungi and plants (Figure 6, Table 2) showed that the mean was higher than the median in most cases, indicating asymmetric distributions and supporting the use of the mode. Although the true modes are not known in these cases, some patterns were evident. Among the mode estimates, those using SPM often were the lowest and appeared to visually underestimate the center of the distribution. Of the remaining mode estimators, HRM and RPM were 10.1% different from each other, on average, across the five sample data sets (Table 2). HRM-BMO (mode of the bootstrapped modes) and RPM-BMO averaged 7.2% different, and HRM-BME (mean of the bootstrapped modes) and RPM-BME were only 1.4% different. A greater difference was observed between HRM and its bootstrapped estimate (HRM-BME) than between RPM and RPM-BME. All of these results from analyses of real data were consistent with the simulation results (Figure 5).

The use of bootstrapping with mode estimation had little or no affect on bias, except that bias increased slightly for the lognormal original distribution (Figure 4f,4g). On the other hand, bootstrapping lowered the variance of all mode estimators, with HRM showing the greatest improvement and RPM (lognormal) the least improvement (Figure 5b,5c,5d,5f,5g,5h,5j,5k,5l). In effect, bootstrapping eliminated the disparity between HRM and RPM in terms of variance, with bootstrapped versions of those estimators having similar low levels of variance.

The results of all simulations, and software for estimating the mode and its standard error, are available as Supplementary Data http://www.evogenomics.org/publications/data/mode/index.htm.

Discussion

General performance of the mode

There were few cases in this study where mode estimates performed worse than the mean in determining the center of location of a data set, suggesting a more general application for the mode. For example, in situations where centers of location are needed but distributions are not predictably normal, the mode might be used as a routine statistic. In those cases it would produce a similar result to the mean for normal distributions and would estimate the center of location with less bias than the mean for non-normal and contaminated distributions. However, our simulations showed that contamination located close to the true mode will mislead any estimator (mean or

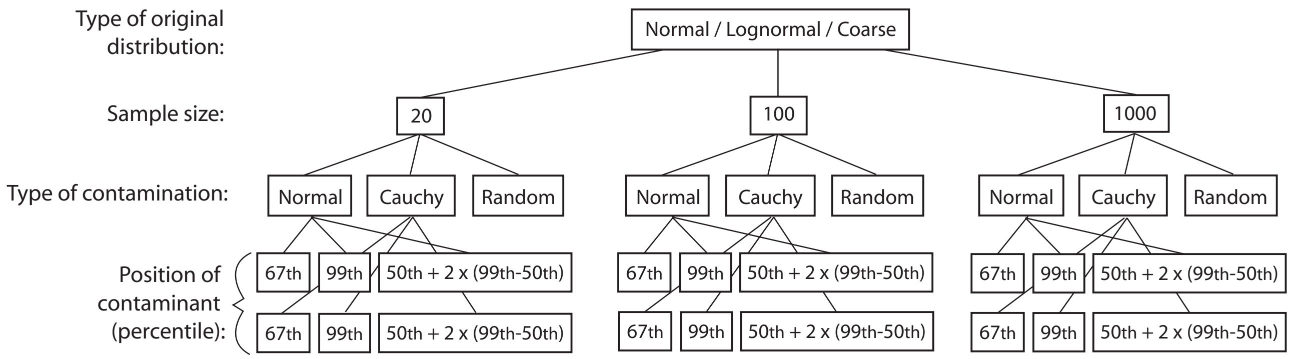

Figure 3
Design of the simulations. Each path from the root to a leaf defines the parameter values for an individual simulation run.

Table 1: Ranking of different estimators of location based on the results of the simulations. Rankings shown here are those involving estimator bias, each with normal contamination centered at twice the 99th percentile and with a sample size 100. The first group of three columns shows the average rank for all different levels of contamination. The other two groups of columns show rankings for 10% and 30% levels of contamination. Smaller rank numbers indicate lower bias.

Rank	Average			10% contamination			30% contamination		
	N = 20	N = 100	N = 1000	N = 20	N = 100	N = 1000	N = 20	N = 100	N = 1000
				Normal original distribution					
RPM	3	2	2	3	2	2	3	2	3
HRM	1	1	1	1	1	1	1	1	1
SPM	2	4	4	2	4	3	2	4	4
Median	4	3	3	4	3	4	4	3	2
Mean	5	5	5	5	5	5	5	5	5
				Lognormal original distribution					
RPM	1	2	2	1	1	2	1	1	1
HRM	3	3	3	3	3	3	2	3	3
SPM	2	1	1	2	2	1	3	2	2
Median	4	4	4	4	4	4	4	4	4
Mean	5	5	5	5	5	5	5	5	5

mode). In the future, it would be useful to explore the performance of the various mode estimators under a greater diversity of "coarse" data distributions.

Negative Bias of some Mode Estimators

We found that RPM and SPM, the two estimators that use a data transformation method, exhibited increasing negative bias (bias away from the position of contamination) as the contamination level increased. Because the contamination was applied to the right side of the distribution, the bias resulted in underestimates of the true mode. This may be attributed to the fact that both estimators are based on the transformation of sample data to conform to a single normal distribution, not to additional secondary distributions often encountered with outliers. The HRM estimator would appear to be less affected by such secondary distributions and therefore is at an advantage. Compared with the normal original distribution and same location of contamination, the right tail of the lognormal distribution integrates with the outliers to a greater extent and thus obscures more of the secondary peaks. This may explain the lack of a negative bias by RPM and SPM in the lognormal distribution (Figure 4e).

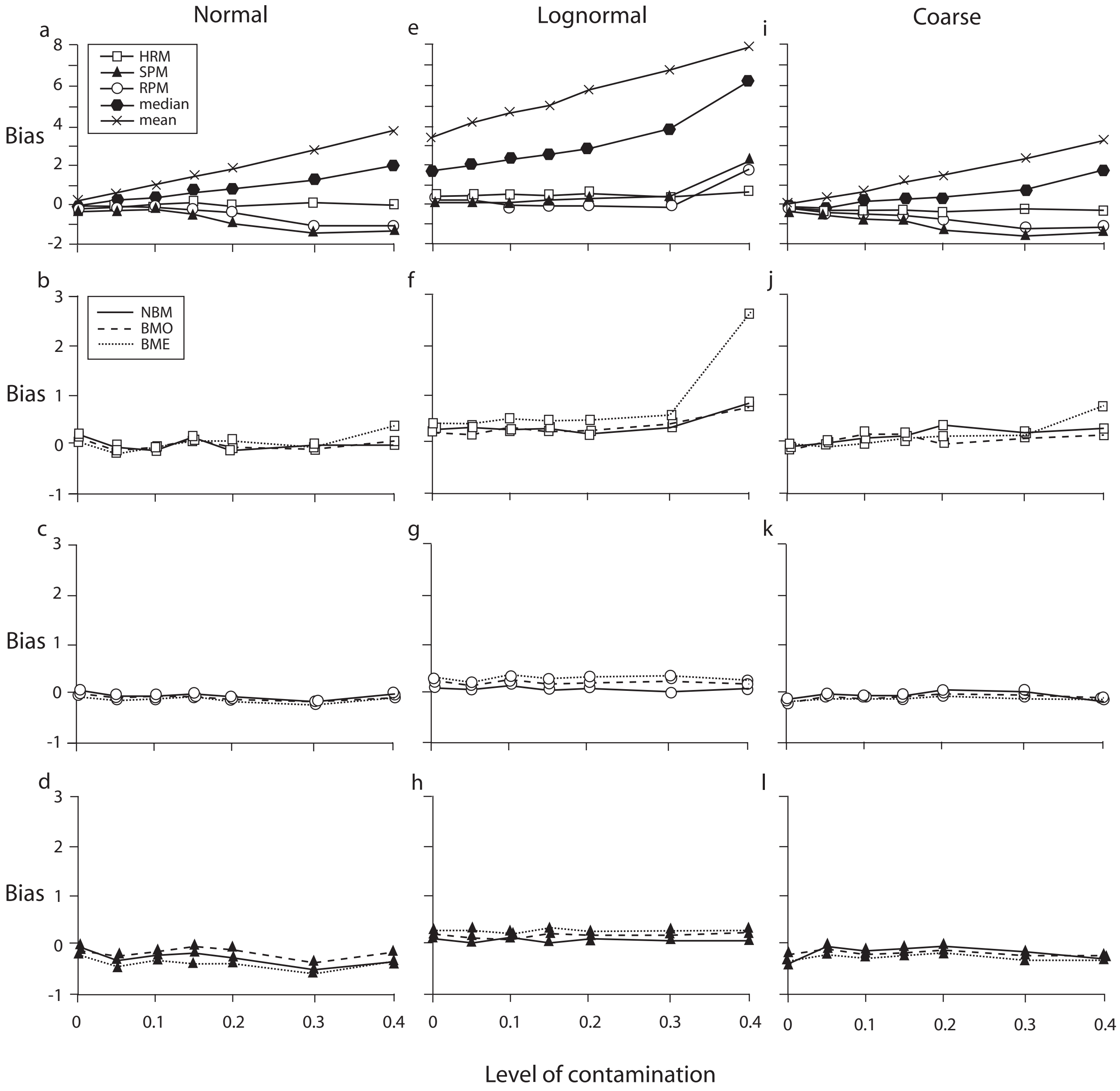

Figure 4

Selected results of the simulations for bias of the estimators. Graphs shown here are for intermediate sample size (n = 100), and normally distributed contamination located at twice the distance between the true mode and the 99th percentile; see Supplementary Data for full results. Each column of graphs represents simulations for different original distributions: normal (left, a-d), lognormal (middle, e-h), and coarse (right, i-l). The top row (a, e, i) shows bias results for a comparison of three mode estimators with mean and median, all using 1000 replications. The lower three rows show bias results for comparisons between mode estimators with no bootstrapping (NBM, solid line), mode of 100 bootstrapped modes (BMO, dashed line), and mean of 100 bootstrapped modes (BME, dotted line), for HRM (b, g, l), RPM (c, h, n), and SPM (d, i, n); all using 100 simulation replications (100 bootstrap replications per simulation replication). In each case, the level of contamination applied to the original distribution ranged from zero to 40% (shown on X-axis). Bias is indicated in absolute units (Y-axis).

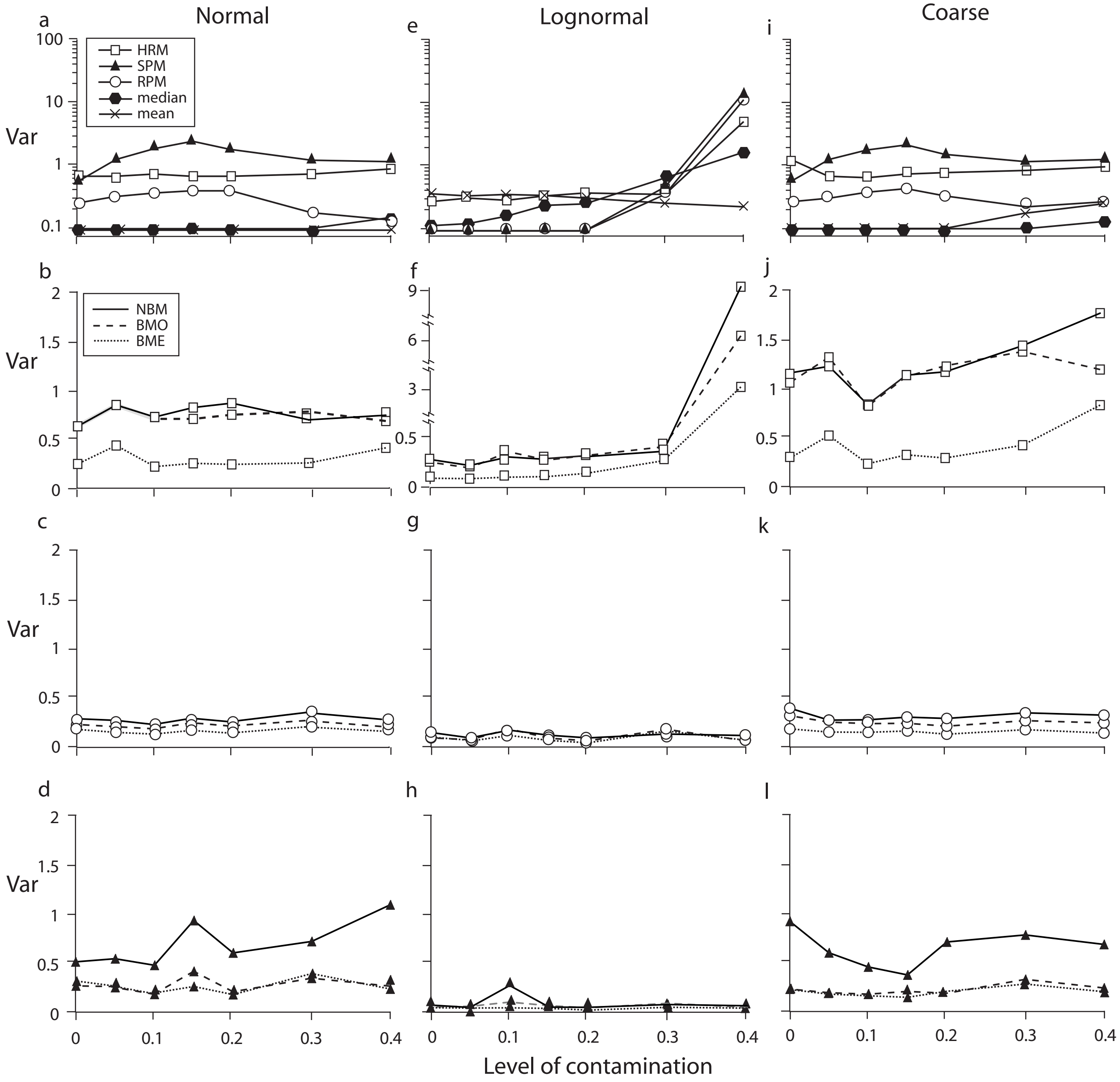

Figure 5
Selected results of the simulations for variance of the estimators. See legend to Figure 4 for description of the simulation conditions.

Increased Performance of the Bootstrapped Mode Estimators

Bootstrapping clearly improved the variance of the mode estimators. Of the two bootstrap methods tested, the bootstrap mean of the modes (BME) is preferred because it performed substantially better than the bootstrap mode of the modes (BMO) in terms of lowering variance. The better performance of BME may result from the lower variance usually associated with the mean. The bootstrapping probably acts to "smooth" the coarseness and irregularity that otherwise might cause inaccurate estimation of the center of location. This is important for a sta-

Table 2: Application of different estimators of location to data sets from a molecular clock study [8]. Each column in the table corresponds to a different evolutionary branch point, between various groups of fungi and plants, being timed; data from the first two columns are shown in Figure 6. Individual data points are divergence time estimates (millions of years ago) from constant rate proteins (number of proteins in parentheses). Time estimates in bold are those obtained with the two preferred estimators.

	Archiascomycetes vs. other Ascomycota (N = 70)	Hemiascomycetes vs. filamentous Ascomycota (N = 48)	Basidiomycota vs. Ascomycota (N = 37)	Green algae vs. plants (N = 41)	Mosses vs. vascular plants (N = 50)
Estimates of central tendency					
Mean	1283	1181	1354	1217	769
Median	1181	1027	1249	1109	765
Histogram	1139	1050	1244	1111	715
RPM	1076	929	1123	1002	703
RPM-BME	**1135**	**948**	**1208**	**1037**	**705**
RPM-BMO	1150	943	1184	1014	744
HRM	878	934	1249	969	858
HRM-BME	**1122**	**942**	**1216**	**1061**	**688**
HRM-BMO	1315	934	1248	1053	862
SPM	991	744	966	724	692
SPM-BME	1204	831	1257	984	707
SPM-BMO	1044	759	1066	786	707
Difference between estimates					
HRM vs. RPM	18.4%	0.5%	10.1%	3.3%	18.1%
HRM-BMO vs. RPM-BMO	12.5%	1.0%	5.1%	3.7%	13.7%
HRM-BME vs. RPM-BME	1.1%	0.6%	0.7%	2.3%	2.4%

tistic (the mode) that relies on the overall shape of a distribution or peak in density.

Without bootstrapping, the RPM estimator has a lower variance than the HRM (Figure 5), in part leading to the recommendation of RPM as a general purpose estimator [13]. However, bootstrapping not only lowers the variance of both estimators but it reduces that disparity, with RPM-BME showing only a small improvement in variance over HRM-BME. The same pattern was seen in the analysis of divergence times (Figure 6), where RPM and HRM mode estimates exhibited a large difference (10%) without bootstrapping and a much smaller difference (1%) with bootstrapping.

Standard Error and Confidence Interval of the Mode

Methods for direct calculation of the standard error of the mode have been suggested [16] but are rather complex. Another approach would be to trim the outliers from the distribution and estimate the standard error of the mean of the trimmed data set, assuming that the outliers can be identified and the underlying distribution is normal. However, a simpler method, and one that we recommend, is to estimate the bootstrap 95% confidence interval (range) and/or bootstrap standard error of the mode. Although bootstrapping already is used to calculate the mode (M_{HRMB} or M_{RPMB}), that must be considered separately. The error estimation requires that each resampled data set be used to calculate the mode; in this case, each such data set is used as a starting point for further resampling to calculate M_{HRMB} or M_{RPMB}. Although a bootstrap standard error can be calculated, it would not be appropriate for skewed distributions where the error is distributed asymmetrically around the mode. In those cases, or more generally, we recommend calculating the bootstrap 95% confidence interval (range).

Conclusions

The HRM and RPM mode methods [2,13] both performed reasonably well under a diversity of conditions. However, bootstrapping is recommended for both methods because it reduces the variance. Suggested conventions for indicating modes that are estimated using these methods are M_{HRM} and M_{RPM} for the non-bootstrapped versions and M_{HRMB} and M_{RPMB} for the bootstrapped versions. The bootstrapping should be accomplished by taking the mean of the bootstrapped mode estimates and by using a relatively large number of replications, such as 1,000 or more [17]. In our implementation of the two methods, the M_{RPMB} estimate takes 16 times longer to compute. In cases where time permits, it might be informative to use and compare both estimators (M_{HRMB} and M_{RPMB}) and to construct a histogram for visual inspection of the distribution. The RPM method might have a slight advantage in cases where the data form a single, asymmetric (right- or left-skewed) distribution and a disadvantage in some other cases (Figs 4, 5). However, because M_{HRMB} is simpler, faster to estimate, and does not exhibit negative bias in cases of outlier contamination, we recommend it as a general-purpose mode estimator, along with a bootstrap standard error or 95% confidence interval (range).

Methods

Mode Estimation Methods

The HSM method [2] iteratively divides the data set into samples of half the size as the original set and uses the half-sample with the minimum range, where range is defined as the difference between the maximum and the minimum value of the sample. This method terminates when the half-sample is less than three data points. An average of these three or fewer values is the mode. The HRM method [2] is similar but uses the sub-sample with the densest half-range, where range is defined as the absolute difference between the maximum and the minimum values in a sample. Of these two methods, only the HRM was used in this study because HRM has been shown to have lower bias with increasing contamination and asymmetry [2].

The parametric methods [13] are based on the idea of transforming the data set to an approximate normal distribution by raising it to a real power. Different values for the exponent are tested and a correlation coefficient between the transformed data and an ideal normal distribution is calculated. The exponent with the maximum correlation coefficient is used to transform the data and approximate it to a normal distribution. For a normal distribution, the mode is the value that maximizes the probability density function. Thus, by equating the first derivative of the probability density function to zero to find the maxima, the mode can be estimated.

Thus,

$$M = \left[\frac{1}{2} \left(\bar{y} + \sqrt{\bar{y}^2 + \frac{4\sigma^2(\alpha - 1)}{\alpha}} \right) \right]^{1/\alpha} \qquad (1)$$

where

α = transforming exponent

$\bar{y}$ = mean parameter

σ = standard deviation parameter

In the SPM method [13], the sample mean and sample standard deviation of the transformed data are used as the mean and standard deviation parameters ($\bar{y}$ and σ) for mode estimation. The correlation coefficient used to estimate the best value of the exponent α in Equation (1) is expressed as:

$$r(\alpha) = \frac{s_+^2(\alpha) - s_-^2(\alpha)}{s_+^2(\alpha) + s_-^2(\alpha)}, \qquad \text{where}$$

$$s_\pm(\alpha) = \delta\left(\frac{y_i(\alpha)}{\delta y_i(\alpha)} \pm \frac{z_i}{\delta z_i} \right)$$

with the operator δ giving the sample standard deviation.

In the RPM method [13], the sample median and the sample standardized median absolute deviation (MAD) of the transformed data are used as the mean and standard deviation parameters ($\bar{y}$ and σ) for the mode estimation. The MAD is defined as the average difference of each data point from the median. The correlation coefficient used to estimate the best value of the exponent α in (1) is expressed as:

$$r(\alpha) = \frac{s_+^2(\alpha) - s_-^2(\alpha)}{s_+^2(\alpha) + s_-^2(\alpha)}, \qquad \text{where}$$

$$s_\pm(\alpha) = \Delta\left(\frac{y_i(\alpha)}{\Delta y_i(\alpha)} \pm \frac{z_i}{\Delta z_i} \right)$$

with the operator Δ giving the median absolute deviation.

Evaluation of the Methods

We designed 63 simulated data sets to compare the three mode estimation methods (HRM, SPM, and RPM) with the mean and median. Each data set was defined using combinations of the following six parameters, modeled in part after the study of Bickel [13] but with slightly larger variance (standard deviation parameter = 2) to more closely approximate real biological data sets [7,9], additional levels, distributions, and locations of contamination, and other features: (1) Type of original distribution: normal (mode = 6), lognormal (mode = 1, median = 2.72), and coarse (mode = 6.25) distributions (Figure 2a). (2) Type of contamination: normal, cauchy, and random (Figure 2b). (3) Level of contamination: 0%, 5%, 10%, 15%, 20%, 30%, and 40%. (4) Location of the contaminant: 67th percentile of the original distribution, 99th percentile of the original distribution, and "twice the 99th percentile" (true mode plus twice the distance between the 99th percentile of original distribution and the true mode). (5) Spread of the contaminant: standard deviation = 2.0. (6) Sample size: N = 20, 100, 1000. For simulations testing the non-bootstrap mode methods, 1000 replications were used. Additionally, 100 replicates were used for simulations testing the use of bootstrapping, with 100 bootstrap iterations performed during each replication. An overview of the simulation design is shown in Figure 3.

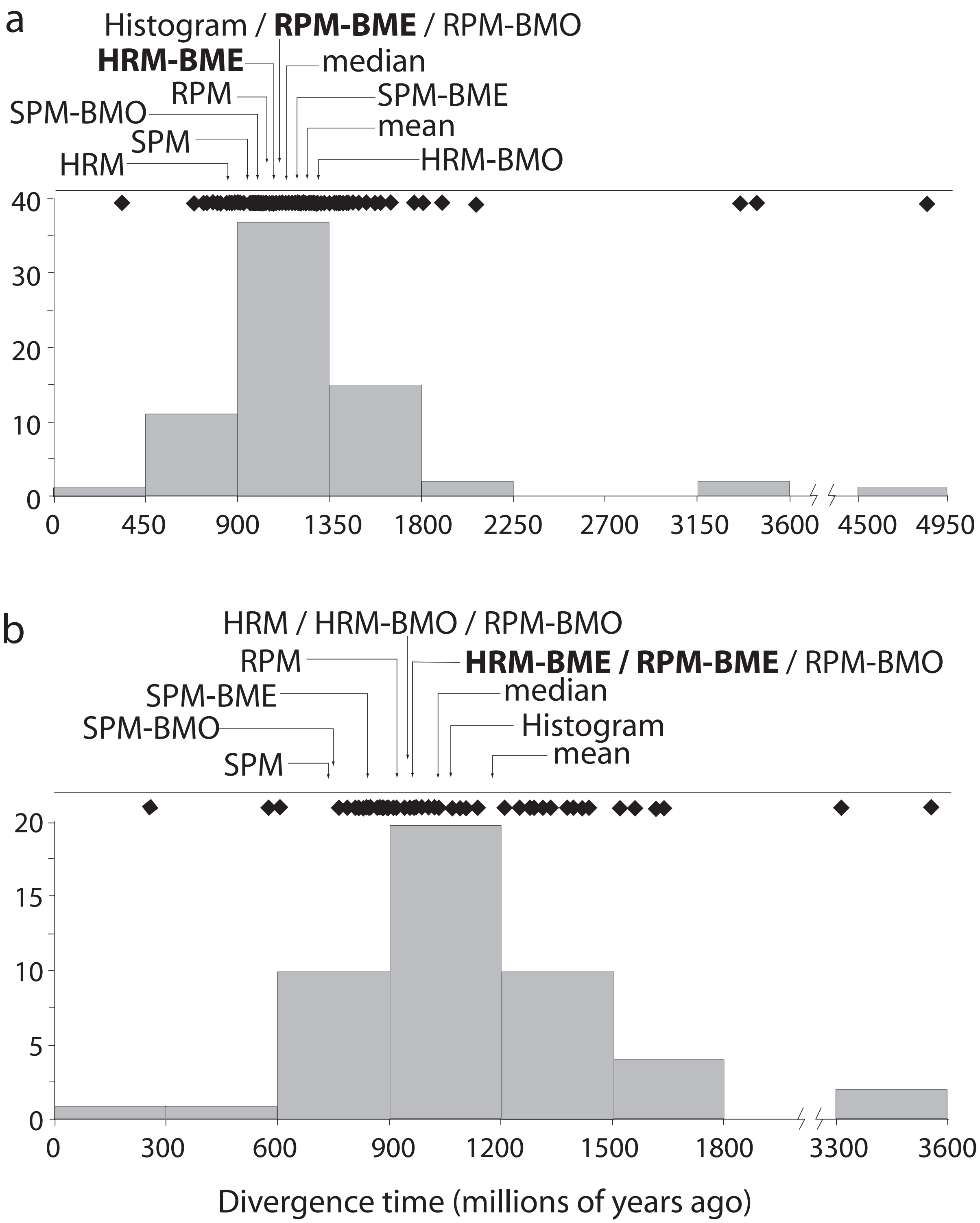

Figure 6
Application of mode estimation methods to published data sets. The data are divergence time estimates (millions of years ago) from a molecular clock study of fungi and plants [8]. Both graphs include the histogram distribution, the actual data points plotted in a horizontal line, and positions of the various estimates of central tendency (Table 2). The two recommended mode estimators are highlighted in bold. (a) Archiascomycetes versus other Ascomycota (n = 70 constant rate proteins), (b) Hemiascomycetes versus filamentous Ascomycetes (n = 48).

The coarse original distribution was intended to model the class of real data where there is a central tendency, but there are multiple subpeaks within an otherwise single peak. This may arise from limited sampling, minor clumping of data, or other factors. For the purpose of these simulations, we have modeled the simplest case involving two subpeaks of a single distribution; it is not intended to be a bimodal distribution. To construct this coarse distribution, we used two normal distributions, with one distribution centered at the 60th percentile of the other distribution. Thus, the latter distribution is partly to the left of the first one. We sampled 33% of the data points from the first two quartiles of the distribution on the left, 33% of the data points from the last two quartiles of the distribution on the right, and 33% of the data points from the remaining half portions of both the distributions (Figure 2a). There are many other possible types of coarse distributions, with multiple subpeaks, but we considered only this simple case. The cauchy distribution was used as one of the contaminants to model erratic contamination and a high number of outliers (Figure 2b).

In previous simulations, contamination was applied only at the 99.99th percentile, and with a narrow standard deviation equal to 1/100th of the interquartile range of the main distribution divided by the interquartile range of the standard normal distribution [13]. This resulted in a sharp spike (peak) in the tail of the distribution. In real data sets involving distributions of molecular time estimates, outliers more frequently appear removed from the main distribution and spread broadly rather than focused at one point (e.g., Figure 1). Therefore, we used contamination locations that were further removed (twice the 99th percentile) and more broadly distributed (standard deviation = 2). For comparison, we also simulated contamination at the 99th percentile and 67th percentile. We predicted that the mode estimators would perform poorly with contaminants located close to the true mode.

We evaluated the bias (the difference between the estimated value and the true mode) and variance of the estimators for each simulation. A ranking system was used to make comparisons of the different methods. Ranking was done individually (each simulation) and in groups, to determine a consensus order. Grouped ranking for a particular simulation run was calculated as the average rank of each estimator for all levels of contamination. The grouping was done by collecting the nodes at the third and fourth levels of the tree in Figure 3. In other words, the ranks for the different types and positions of the contaminants were averaged for a unique combination of the original distribution type and sample size. This was done because the original distribution type and sample size are characteristics that are more easily determined for real data sets, and were therefore of greater interest in this study. Average and individual ranks for particular levels of contamination shown in Table 1 are for two simulation sets (normal and lognormal), each using normal contamination located at twice the 99th percentile and with a sample size of 100. This combination of intermediate sample size and contamination location most closely corresponded to real data sets involving distributions of time estimates [8].

In addition to the non-bootstrapped mode estimate (NBM, using original data) for each method, we calculated a bootstrap estimate using the mean (BME) and the mode (BMO) of bootstrapped modes. This permitted us to compare biases and variances associated with bootstrapped and non-bootstrapped versions of each method for each data set. We predicted that bootstrapping [18] might improve mode estimates because of the smoothing effect of the resampling, emphasizing (by chance) different subpeaks and thus generating different modes centered around a single (overall) peak in the distribution. Therefore, the mean (or mode) of those multiple subpeaks might better represent the central tendency of the overall data set.

Finally, we applied these mode estimators to published data sets of divergence time estimates of fungi and plants [8]. The objective was to observe how the different estimators performed with real data and to compare their performance with simulation results to assist in formulating recommendations for mode estimation.

Authors' contributions

SBH conceived the study; SBH and PS designed the simulations; PS carried out the simulations; SBH and PS drafted the manuscript and both authors approved the final manuscript.

Acknowledgements

We thank David Hunter and Sudhir Kumar for comments on the manuscript and Jennifer Hines for drafting the figures. This work was supported by grants to SBH from the National Science Foundation (DBI-0112670) and the National Aeronautics and Space Administration (NCC2-1057).

References

1. Dalenius T: **The mode - a neglected statistical parameter** *Journal of Royal Statistical Society, Serial A* 1965, **128:**110-117.
2. Bickel DR: **Robust estimators of the mode and skewness of continuous data** *Computational Statistics and Data Analysis* 2002, **39:**153-163.
3. Sokal RR and Rohlf FJ: **Biometry: the principles and practice of statistics in biological research** 3rdth edition. *New York, W.H. Freeman and Co.*; 1995.
4. Markov H, Valtchev T, Borissova J and Golev V: **An algorithm to "clean" close stellar companions** *Astronomy and Astrophysics Supplement Series* 1997, **122:**193-199.
5. Stetson PB: **DAPHOT: A computer program for crowded-field stellar photometry** *Publications of the Astronomical Society of the Pacific* 1987, **99:**191-222.
6. Hanes DP and Schall JD: **Neuronal control of voluntary movement initiation** *Science* 1996, **274:**427-430.

7. Kumar S and Hedges SB: **A molecular timescale for vertebrate evolution** *Nature* 1998, **392:**917-920.
8. Heckman DS, Geiser DM, Eidell BR, Stauffer RL, Kardos NL and Hedges SB: **Molecular evidence for the early colonization of land by fungi and plants** *Science* 2001, **293:**1129-1133.
9. Wang Daniel Y.-C., Kumar Sudhir and Hedges S. Blair: **Divergence time estimates for the early history of animal phyla and the origin of plants, animals, and fungi** *Proceedings of the Royal Society of London B* 1999, **266:**163-171.
10. Huberman BA, Pirolli PL, Pitkow JE and Lukose RM: **Strong regularities in World Wide Web surfing** *Science* 1998, **280:**95-97.
11. Agarwal BL: **Basic Statistics** *New Dehli, Wiley Eastern Limited*; 1991.
12. Grenander U: **Some direct estimates of the mode** *Annals of Mathematical Statistics* 1965, **36:**131-138.
13. Bickel DR: **Robust and efficient estimation of the mode of continuous data: The mode as a viable measure of central tendency** *Journal of Statistical Computation and Simulation (in press); preprint: http://interstat.stat.vt.edu/interstat/articles/2001/abstracts/n01001.html-ssi* 2001:1-22.
14. Hedges SB and Kumar S: **Genomic clocks and evolutionary timescales** *Trends in Genetics* 2003, **19:**200-206.
15. Nei M, Xu P and Glazko G: **Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms** *Proceedings of the National Academy of Sciences (U.S.A.)* 2001, **98:**2497-2502.
16. Yasukawa K: **On the probable error of the mode of skew frequency distributions** *Biometrika* 1926, **18:**263-292.
17. Hedges SB: **The number of replications needed for accurate estimation of the bootstrap p-value in phylogenetic analysis** *Molecular Biology and Evolution* 1992, **9:**366-369.
18. Efron Bradley and Tibshirani Robert: **An Introduction to the Bootstrap** *Monographs on Statistics and Applied Probability 57 New York, Chapman & Hall*; 1993.