

Running sigmas analysis of sampled molecular paraphyly in Pottiaceae (Bryophyta)

Richard H. ZANDER* 

Missouri Botanical Garden, 3444 Shaw Blvd., St. Louis, Missouri, USA

Received: 02.05.2021 • Accepted/Published Online: 14.07.2021 • Final Version: 30.09.2021

Abstract: A spreadsheet dealing with standard deviations allows statistical examination of continuous variables associated with paraphyly in systematics, with examples in the Pottiaceae (Bryophyta). Outlier values in taxonomic representation of distribution with the structure of $(w-)x-y(-z)$ of continuous variables can be empirically supported. The use of the geometric mean and intuitional estimation is reexamined. Published reports of the distance between maximally distant exemplars of individual paraphyletic species in molecular cladograms in several studies were reevaluated as metadata. Levels of standard deviations were assigned with the Running Sigmas spreadsheet. A spike in width of exemplars of a paraphyletic species of a moss in the family Pottiaceae, *Exobryum asperifolium*, was identified as an outlier. Techniques of macroevolutionary systematics, however, assured that no different evolutionary processes were involved that may have supported an explanation of taxonomic crypsis. Instead it was found that metadata heterogeneity was the problem in that the paraphyly width of 14 molecular cladogram nodes which is evolutionarily acceptable as a clear-cut progenitor-descendant lineage with no reversals.

Key words: Standard deviation, phylogenetics, paraphyly, metadata, spreadsheet, *Exobryum asperifolium*, Pottiaceae, macroevolutionary systematics

1. Introduction

This paper investigates the evolutionary significance of reports of molecular paraphyly in previously published systematic studies of groups of mosses in the Pottiaceae of the Bryophyta by other authors. Here, the term paraphyly includes short-distance polyphyly, based on demonstration (Zander, 2019a, 2019b) that branching molecular races of one progenitor species may separately generate descendant species creating a molecularly paraphyletic progenitor species. Statistical examination also demonstrated a clear congruence between use of morphological traits in macroevolutionary systematics and DNA traits in molecular studies, requiring only interpretation of results as species-to-species trait transformation as radiation within a genus. The genus concept of one progenitor species radiating two or more descendant species is called a dissilient genus. This concept is opposed to cladistic analysis which finds most parsimonious trait transformations between groups of species on a dichotomous tree. Only a clade is available as a possible genus. The present study involves a study of ranges of variable traits and ranges of species' traits within and among genera.

Evaluation of ranges of variable traits in alpha taxonomy is usually intuitional, with estimated common and expected values given in a range and more extreme values added in parentheses before and after the common range. In my experience, the stopping rule (Staley, 2004) in standard practice is mainly to stop quantifying continuous variables (leaf length, cell size, and similar measurements) after dimensions do not change significantly. The latter is in practice created as a running average, in taxonomy usually expressed as a range of common measurements with extreme measurements in parentheses, i.e. the familiar $(w-)x-y(-z)$ format.

Running Sigmas¹ is a spreadsheet now freely available on the Web that allows one to do running averages and running standard deviations as part of taxonomic analysis. Available standard deviation calculators are largely focused on business applications regarding profit and loss risk management (Ragsdale, 2004). As measurements are added to the Running Sigmas spreadsheet, the average, or mean, is sequentially calculated for the full range. Additionally, standard deviations (SD or less formally as σ , sigma) are calculated providing a measure of dispersion

¹ Zander RH (2020). Running Sigmas Calculator. A spreadsheet providing running average, absolute deviation, and one to four sigma running standard deviations. Res Botanica Technical Report 2020-08-2 [online] Website <http://www.mobot.org/plantscience/ResBot/Repr/1Reprints.htm> [accessed 26 September 2021]

* Correspondence: richard.zander@mobot.org

of the data, including deviation from the mean. Plotting data points in a spreadsheet as lines or bar graphs also allows estimates of variation in range, allows a check for bi- or multimodal distributions, and identification of extreme values. Standard deviation calculation in Running Sigmas works for small samples (the SD formula using N minus one as divisor) or all data and large samples (using N divisor). It provides ranges for first, second, third and fourth standard deviations using the standard formula for a sample of a larger distribution, and allows estimated probability for outliers being outside the estimated population distribution. Determining criteria for sample sizes is problematic, see discussions of Snedecor and Cochran (1967) and the general treatment by Thompson (2012), but sample sizes are sufficient when the values of the SD, as calculated with the formula for a population (σ), nearly match those as calculated for a sample (s). This comparison is provided in the spreadsheet and may be as few as 10 to 30 samples. It has been pointed out that once sample size is established (e.g., by convergence of population and sample SDs), increasing sample size is not effective (Wallis and Roberts, 1956).

A little statistical background is necessary for most systematists. The standard deviation is the square root of the variance (Wallis and Roberts, 1956, Winkler and Hays, 1975). The variance is the average of the squared differences from the mean. We can expect about 0.68 of sampled randomly distributed values to fall within plus or minus 1 SD in any standard normal distribution. That is, 0.34 on either side of a normal distribution of data (bell curve). Doubling the SD allows 0.95 of data to fall within 2 SD of the mean on both sides of the distribution; and tripling the SD makes about 99.7 of data fall between 3 SD of mean on both sides. The three-sigma rule of thumb implies that nearly all values outside 3 SD of the mean have a one minus 0.997 chance (as Bayesian posterior probability) of correctly being included in the distribution. That is, a low chance of being generated by the same natural processes. In particle physics five sigmas of accuracy (odds of having occurred by chance alone are less than 1 in 3.5 million) are necessary for a discovery and six sigmas for manufacture of critical aircraft parts (1 in 500 million). In less structured sciences like taxonomy, 2 sigmas allow 0.95 BPP to comprise the expected range of values, which is taken as adequate for a basis for continued research, and certainly 3 sigmas or 0.997 BPP is decisive in noncritical, corrigible studies. If there are many data points outside a 3-sigma range or if skewness (imbalance in the bell curve) is not negligible, one may question the normality of the data. For further information on multiple sigmas and the 68–95–99.7 rule see account by Moore and Notz (2006).

Therefore, if some taxonomic data are not within the 2 SD or 0.95 range of the distribution of values, there is a better than 1 in 20 chance they are outliers caused by other natural processes, including mistakes. This translates to 0.95 Bayesian posterior probability that the values should be mistrusted and require additional study. Because outliers can be merely rare, however, extreme values may not be wrong. As part of scientific reporting in taxonomic descriptions, they may be placed in parentheses. The use of the Running Sigma spreadsheet may be used for empirically calculating a descriptive range of values for variable traits in taxonomy, and constructing the usual formula of $(w-x)-y(-z)$ for detailing ranges of continuous variables like leaf lengths or cell sizes. A study of such ranges as used in taxonomic practice with mosses (Zander, 2013) demonstrated that intuitional estimation organizes the central range around the geometric mean of the two extreme values if the range is near zero.

Using the number of samples (N) to calculate the average of the squared differences from the mean is appropriate when you have complete data on a population to work with. A modification (Bessel correction) of the SD formula corrects for the increased inaccuracy of only having a sample of the population. This is used for a sample of a larger population, replacing N with $N - 1$ to provide an unbiased estimator for calculating the average. The sample standard deviation (s) is the estimator of the population SD (σ or sigma). The two values for standard deviation are quite similar when N is greater than 10 (or in some references, 30) samples.

The formula for calculating standard deviation for total population (σ) is:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

where σ = population SD; N = size of population; x_i is each data value; μ = population mean. The formula for calculating SD of a sample (s) is simply replacing N with $N - 1$, which is the number of degrees of freedom. One explanation² for $N - 1$ is that the sum of deviations of n observations from their sample mean must be zero. This means that if $N - 1$ of the deviations are known, they completely determine the n th deviation. It is the squared deviations from the mean that are used to construct the sample variance and hence we say that the sample variance has $N - 1$ degrees of freedom.

The Running Sigma spreadsheet is used here to evaluate the dispersion of values given in a metadata study (Zander, 2019b) of maximum distance of exemplars of paraphyletic moss species of the family Pottiaceae on a molecular cladogram. The 23 paraphyletic taxa contributing to the

² Williams S (2015). Statistical concept for degrees of freedom. Mathematics Stack Exchange [online]. Website <https://math.stackexchange.com/questions/1218076/intuitive-explanation-for-dividing-by-n-1-when-calculating-sample-variance> [accessed 10 July 2020].

metadata in that paper were *Anoetangium aestivum*, *Barbula gregaria*, *B. indica*, *Chionoloma bombayense*, *Didymodon ancerinocapitatus*, *D. asperifolius*, *D. gaochenii*, *D. topiaceus*, *Oxystegus daldinianus*, *O. recurvifolius*, *O. tenuirostris*, *Pseudosymblypharis angustata*, *Streblotrichum convolutum*, *Tortella arctica*, *T. flavovirens*, *T. fragilis*, *T. tortuosa*, *Trichostomum duidense*, *T. leptocylindricum*, *Weissia condensa*, *W. controversa*, and *W. jamaicensis*. There were 27 instances of paraphyly, with some species occurring in different studies.

The reported range was from two to 14 continuous, directly connected cladogram nodes between maximally distant exemplars (the original paper, Zander, 2019a, reported 15 as maximum value, a miscount for 14). The single data point of 14 nodes was outside of the otherwise continuous range of 2 to 9 nodes and deprecated (ignored) as anomalous. The percent of species with demonstrable molecular races (46) that were also paraphyletic was 0.41. The average maximum distance between paraphyletic exemplars was 4.52 continuous nodes. This is the reason why the number of nodes between the most distant nodes of a set of exemplars of a paraphyletic species are used for calculation of standard deviations; it is the entire set of internal branching lines of races that generates apophyletic (descendant) species and expectation is for the full width. The average number of apophyletic (embedded) species was 3.6 per paraphyly.

What was the statistical basis for intuitively deprecating the data point of 14 cladogram nodes, when other paraphyletic ranges were between 2 and 9 nodes? Since data for multiple exemplars of each species are not common in the literature, the metadata study of Zander (2019b) is significant as it may apply to other taxonomic groups in which molecularly established but otherwise nearly cryptic species, genera and families have been established. I use the term nearly cryptic because any polythetic genus can be split into subsets each somewhat supported by unbalanced polythetic morphology. Because different data are used, support for molecular studies is possible only from morphological studies that can stand entirely on their own.

A similar study of Pottiaceae genera with largely different species was provided by Zander (2019a). This also reported that about half the species sufficiently sampled to demonstrate multiple internal races also were paraphyletic. There was also an average number of 4.5 continuous nodes between maximally distant exemplars of paraphyletic species.

A quasi-phylogenetic study by Aubert (2017) using macroevolutionary evaluation of cladogram nodes, resulted in a caulogram (tree of serial speciation). This study agreeably concluded that *Didymodon asperifolius* deserved inclusion in a new genus.

2. Materials and methods

The large moss family Pottiaceae is the subject of this study because the author has studied its taxonomy for five decades (e.g., Zander, 1993, 2019a, 2019b), and because other researchers have begun to analyze its evolutionary relationships with molecular (DNA) techniques.

Fundamental to this study is the expectation that apophyletic species (those embedded or nested in a molecular paraphyly) are descendants of the paraphyletic species. This may be direct or indirect if extinction of intermediates has happened. The extant paraphyletic species is equivalent to the inferred progenitor if the morphology is essentially the same (see Zander, 2013, for discussion). A Microsoft Excel spreadsheet of 13 columns was constructed with the following columns starting with row 4 of spreadsheet and giving a formula for row 6:

- (1) Numbers from 1 to 50;
- (2) Original data, the data values in order of observation, essentially random;
- (3) Sorted data, the same data sorted from lowest value to highest;
- (4) Running average, calculating average for each row from data top to present row, using =AVERAGE(\$C\$4:C6), where \$C\$4 is unchanging value of first data entry;
- (5) Absolute deviation, for each row, using =ABS(C6-D6);
- (6) Running median =MEDIAN(\$C\$4:C6);
- (7) Running SD for a sample, s calculated with =STDEV.S(\$C\$4:C6);
- (8) Skewness, running average minus running median divided by running SD of a sample, =((D5-F5)/G5);
- (9) Running SD assuming data on a complete population, with σ calculated with =STDEV.P(\$C\$4:C6);
- (10) 1st SD range, adding running average and running SD for a sample for that row;
- (11) 2nd SD range, adding running average and two times SD for a sample;
- (12) 3rd SD range, adding running average and 3 times SD for a sample; and
- (13) 4th SD adding running average and 4 times SD for a sample.

Adding the dollar signs (\$C\$4) to the cell location in the Excel formula anchored that cell so its pointer does not change when copying. One then may simply copy rows to extend the spreadsheet to however many data points are available.

The above description should allow anyone to create his or her own Running Sigmas spreadsheet. In addition, a sample spreadsheet with proper formulas and loaded with sample data is available online. Data from the Zander (2019a) study was inserted into the spreadsheet, then sorted from least up to largest value. Examination was made for

possible multimodal distribution with a bar graph. The running SD from the sample and from the sample treated as a population were compared, the standard deviation for the full sample was noted, and ranges calculated for first through fourth SD Outliers were noted and evaluated as to just how much they differed as absolute deviations from common range of the metadata. Degree of skewness (Snedecor and Cochran, 1967) was noted.

3. Results

Table below presents a summary to the evaluation of the distribution of metadata presented by Zander (2019a)

on maximum distances between exemplars of species exhibiting paraphyly in species of various genera of Pottiaceae. The running average of small-to-large valued sorted data, the absolute deviation, and the running standard deviation for samples all showed (Table) that the number of nodes between exemplars of *Didymodon asperifolius* was unusually large in sharp distinction (Figure 1) from the range of the other data points. When skewness is large, the SD can be grossly inflated and should be doubted when similar to the mean. In the present study, skewness was minimal and positive (distribution curve leaning to the left, tail longer to the right). The data point

Table. Running Sigma spreadsheet with calculated data on running average, absolute deviation from running average, absolute deviation, running median, running standard deviation for samples, running skewness, running S.D. for total population, and first, second, third and fourth running standard deviation ranges as calculated for samples, including respectively 0.68, 0.95, 0.997, and 0.999 of total data values.

	Original data	Data sorted	Run. ave.	Abs. dev.	Run. -median	s - Run. SD sample	Skewness: (mean - median) / s	σ - Run. SD population	s - 1st SD range	s - 2nd SD range	s - 3rd SD range	s - 4th SD range
1	2	2	2.00	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
2	6	2	2.00	0.00	2.00	0.00	N.A.	0.00	2.00	2.00	2.00	2.00
3	6	2	2.00	0.00	2.00	0.00	N.A.	0.00	2.00	2.00	2.00	2.00
4	5	2	2.00	0.00	2.00	0.00	N.A.	0.00	2.00	2.00	2.00	2.00
5	2	2	2.00	0.00	2.00	0.00	N.A.	0.00	2.00	2.00	2.00	2.00
6	3	2	2.00	0.00	2.00	0.00	N.A.	0.00	2.00	2.00	2.00	2.00
7	14	2	2.00	0.00	2.00	0.00	N.A.	0.00	2.00	2.00	2.00	2.00
8	6	3	2.13	0.88	2.00	0.35	0.35	0.33	2.48	2.83	3.19	3.54
9	2	3	2.22	0.78	2.00	0.44	0.50	0.42	2.66	3.10	3.55	3.99
10	3	3	2.30	0.70	2.00	0.48	0.62	0.46	2.78	3.27	3.75	4.23
11	7	3	2.36	0.64	2.00	0.50	0.72	0.48	2.87	3.37	3.88	4.38
12	7	4	2.50	1.50	2.00	0.67	0.74	0.65	3.17	3.85	4.52	5.20
13	5	4	2.62	1.38	2.00	0.77	0.80	0.74	3.38	4.15	4.92	5.69
14	8	4	2.71	1.29	2.50	0.83	0.26	0.80	3.54	4.37	5.19	6.02
15	7	5	2.87	2.13	3.00	0.99	-0.13	0.96	3.86	4.85	5.84	6.83
16	4	5	3.00	2.00	3.00	1.10	0.00	1.06	4.10	5.19	6.29	7.38
17	7	6	3.18	2.82	3.00	1.29	0.14	1.25	4.46	5.75	7.04	8.32
18	3	6	3.33	2.67	3.00	1.41	0.24	1.37	4.75	6.16	7.58	8.99
19	2	6	3.47	2.53	3.00	1.50	0.31	1.46	4.98	6.48	7.99	9.49
20	2	7	3.65	3.35	3.00	1.66	0.39	1.62	5.31	6.98	8.64	10.30
21	4	7	3.81	3.19	3.00	1.78	0.46	1.74	5.59	7.37	9.14	10.92
22	2	7	3.95	3.05	3.50	1.86	0.24	1.82	5.82	7.68	9.55	11.41
23	3	7	4.09	2.91	4.00	1.93	0.05	1.89	6.02	7.94	9.87	11.80
24	9	7	4.21	2.79	4.00	1.98	0.11	1.94	6.19	8.16	10.14	12.12
25	4	8	4.36	3.64	4.00	2.08	0.17	2.04	6.44	8.52	10.60	12.68
26	7	9	4.54	4.46	4.00	2.23	0.24	2.19	6.77	9.00	11.23	13.46
27	2	14	4.89	9.11	4.00	2.85	0.31	2.79	7.74	10.58	13.43	16.27

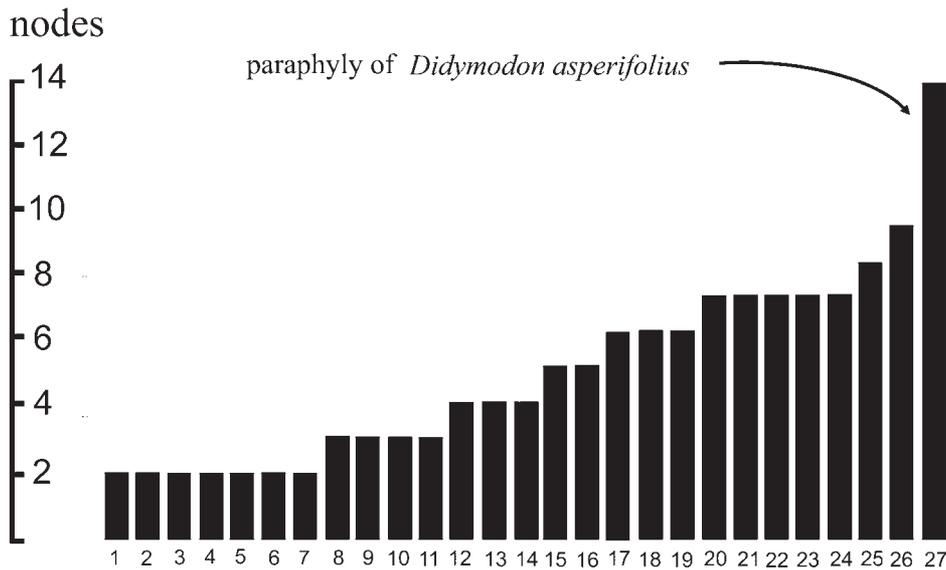


Figure 1. Twenty-seven instances of paraphyly from a metadata study of phylogenies of several genera of Pottiaceae (Zander, 2019a). Width of each paraphyly is given as number of nodes (sorted, low to high) between extreme exemplars of paraphyletic species. The paraphyly of *Didymodon asperifolius* is shown as a clear outlier on the right.

of 14 nodes for *Didymodon asperifolius* is a clear outlier (Figure 1) and occurred within the fourth SD when the data was included and outside the fourth SD when deprecated (by deletion or ignoring the last row in the spreadsheet). This means that the data on *D. asperifolius* is only 0.03 BPP of participating in the evolutionary processes generating the remainder of the data. A bar chart (Figure 1) clarified expectation that the distribution (outside of the 14 value) was unimodal or at least not bimodal.

The ranges of standard deviations obtained from the full metadata study and applied just to species of *Didymodon* sect. *Rufiduli* in the study of Kučera and Ignatov (2015) are given in Figure 2. Kučera and Ignatov gave three cladograms, of ITS, rps4 and trnM-trnV, and a combined data set. The ITS cladogram was chosen because it gave the most paraphyletic examples. This is important in that the empirically molecularly demonstrated number of immediate descendant species from an ancestral species in genera of the Pottiaceae studied is about 3.5 (Zander, 2019a, 2019b) and optimization on a dichotomous tree is misleading. ITS may also have been the most sensitive sequence or it may be biased, but in either case, the reader is encouraged to view this study as an example of how to interpret molecular cladograms with true evolutionary model (the dissilient or radiative genus). The ranges are shown as starting with one end of the ITS molecular cladogram, although either end of the *D. asperifolius* paraphyly is acceptable. First SD for the full data set is a range of 7.7 nodes, second SD is 10.58 nodes, third is 13.43 nodes, and fourth SD is 16.27 nodes. The four

ranges account for an expected 0.68, 0.95, 0.997 and 0.999, respectively, of total data values for samples of paraphyletic species in several genera of Pottiaceae (presented by Zander (2019a)) and here applied just to the outlier paraphyly (that of *D. asperifolius*). The paraphyletic range of 14 contiguous nodes across *Didymodon asperifolius* is very large and is expected to be very rare.

The bold-faced letters in Figure 2 represent exemplars of *Didymodon asperifolius* and the light face other, apophyletic (descendant) species. These are A–E: *Didymodon asperifolius* 1 exemplar each; F–H: *D. johansenii* 1 exemplar each; I: *D. johansenii* 6 exemplars; J: *D. zanderi* 3 exemplars; K: *D. asperifolius* 4 exemplars; L: *D. gaochenii* 3 exemplars, and *D. fragilicuspis* 3 exemplars; M: *D. gaochenii* 1 exemplar; N: *D. hedysarifformis* 11 exemplars and *D. gaochenii* × *hedysarifformis* 1 exemplar; and O: *D. asperifolius* 1 exemplar. Exemplars A through E well represent molecular races of *D. asperifolium*, as does any cladogram multifurcation with several exemplars of one species with internal nodes and branches.

How does one examine whether or not the extreme rarity of a continuous variable is merely rare or is not causally involved in the same processes as the other values? Here we use the analytic key of Zander (2013) as used later by Zander (2019a) in a study of the large pottiaceous genus *Didymodon*. The apophyletic taxa (those distal to the paraphyly) in the molecular ITS study of *Didymodon* sect. *Rufiduli* by Kučera and Ignatov (2015) were *Didymodon hedysarifformis*, *D. fragilicuspis* (= *D. murrayae*), *D. zanderi* and *D. johansenii*. The paraphyletic

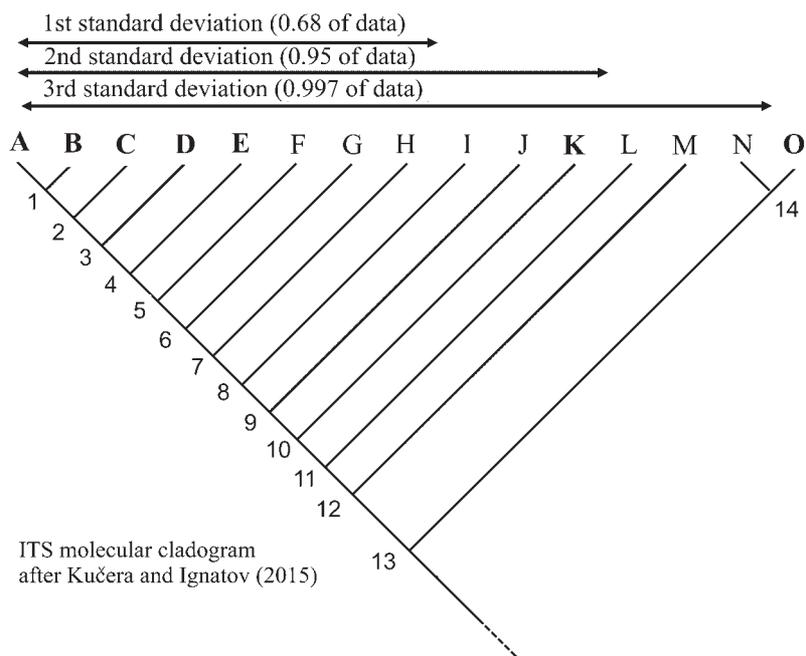


Figure 2. ITS molecular cladogram extracted from that of Kučera and Ignatov (2015). Nodes are numbered, there being 14 nodes between maximally distant exemplars or subclades of *Didymodon asperifolius*. This cladogram is evaluated with standard deviations calculated from all 27 instances of paraphyly of this and other species. See text for exemplar species represented by letters. Exemplars of *D. asperifolius* are in boldface.

taxon *Didymodon asperifolius* and the above apophyletic taxa continued across 14 nodes of the Kučera and Ignatov molecular cladogram between the two most extreme distant exemplars of *D. asperifolius*, and enclosed medially other exemplars (Figure 2) of that species.

Both the paraphyletic taxon and all the apophyletic taxa were placed in the segregate genus *Exobryum* R. H. Zander by Zander (2019a). Based on morphological evaluation (analytic key) of radiative clusters (dissiliient genera), all apophyletic species in the Kučera and Ignatov (2015) study were direct descendants of *Exobryum rufidulum*, except that *E. johansenii* was considered a direct descendant of *E. asperifolius*. The macroevolutionary formula for the caulogram (Zander, 2019a) of *Exobryum* is (putative progenitors boldfaced):

Exobryum rufidulum > (*E. fragilicuspis*, *E. hedysariforme*, *E. zanderi*, (***E. asperifolium*** > *E. johansenii*))

Comparing this formula (illustrated in Figure 3) with the cladogram in Figure 2, one can see that the inferred evolutionary relationships are rather similar. In the case of the caulogram of Zander (2019a) (see Figure 3), *Exobryum rufidulum*, not included in the Kučera and Ignatov (2015) study, is the progenitor of most species, and *E. asperifolius* has only *E. johansenii* in *Exobryum* as descendant. The Zander (2019a) study calculated high Bayesian posterior

probabilities in support of the morphological evolutionary relationships using Shannon information theory (traits treated as informational bits) and Turing sequential Bayesian analysis (adding bits and translating to posterior probabilities).

The present analysis indicates that the extreme range of nodes between maximally distant exemplars of *E. asperifolium* was simply a rare event. This is because *E. asperifolium* and its associated species of apparent descent form a coherent, monothetic, radiative genus that was in fact informative of descent in the internally evolutionarily coherent genus *Exobryum*. It is possible that if *E. rufidulum*, which is considered (Zander, 2019a) equivalent to the progenitor of *Exobryum*, were included in the molecular study, the nesting of descent as exposed by molecular paraphyly would be somewhat different or better resolved and the caulogram would perhaps approach homeomorphism with the cladogram. Translating between caulograms and cladograms is difficult because the shared ancestors of any two cladogram branches are unconstructed, meaning one may accept that they inferentially and theoretically exist but cannot be described as real entities. See Barrow (1992) for a thorough but accessible explanation of constructivism in science.

In addition, Bayes factors must be used to evaluate if one evolutionary tree is better than another when they

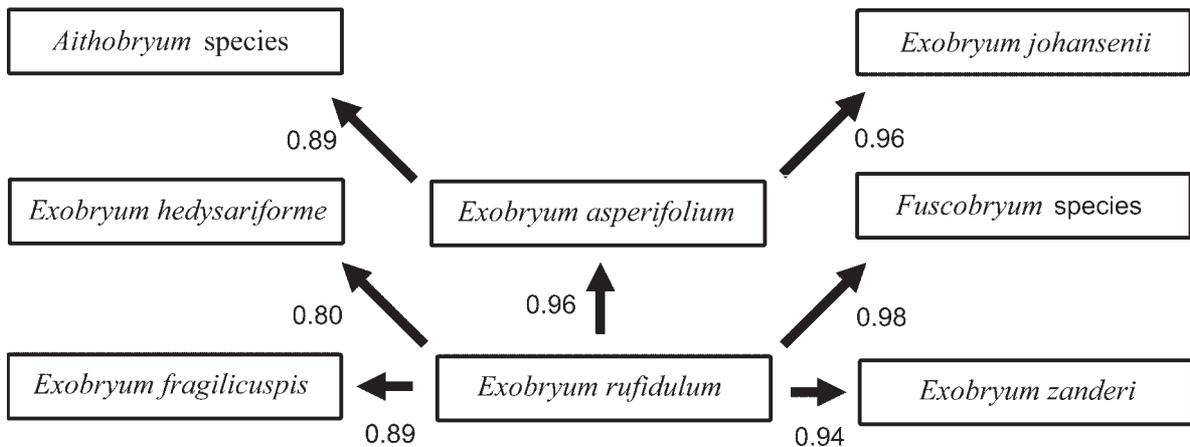


Figure 3. Caulogram from Zander (2019a) here limited to the *Exobryum* putative progenitor and descendant genera and species. The values are Bayesian posterior probabilities (BPP). The topology is actually quite similar to that of the cladogram of Figure 2, which lacks the putative progenitor, *E. rufidulum*. The morphologically based evolutionary relationships in the above caulogram are coherent and stepwise, thus the molecular cladogram supports the caulogram of this genus in large part.

are based on different data (Zander, 2013), and these simply consist of dividing one BPP by the other BPP. A Bayes factor of three (BPP of one tree three times as large as that of another) is needed for any helpful support from one of two alternatives. If two contrary morphologically based and molecularly based evolutionary trees both have BPP support for internal branches greater than 0.50, then both should be deemed refuted unless some third synthetic argument explains both. That third argument is the interpretation of molecular parphyly as implying evolutionary descent, and reexamination of evolutionary relationships via construction of morphological transformation series as in macroevolutionary systematics (Zander, 2013, 2019a, 2019b).

The geometric mean (nth root of the product of all n values) is important in taxonomic description (Zander, 2013). For the full data set, the geometric mean was 3.99. When the outlier of 14 nodes was deleted, the geometric mean was 4.18. The average for the full data set was 4.89, for the trimmed data set it was 4.54. The geometric mean of just the extreme values 2 and 14 is 5.29, of 2 and 9 is 4.24. If intuition played a larger role in evaluating the range, then the geometric mean between 2 and 14 would be higher than the average of the full data set, while the geometric mean of extremes of the trimmed data set, 2 and 9, would be below the average of the data set trimmed of the *Didymodon asperifolius* parphyly. The geometric mean is expected to be less than the average in statistical computation. A range of 2–9(–14) describes the range of variation in numbers of contiguous nodes between farthest exemplars of paraphyletic species in this metadata set involving several genera of the same family. Both empirically and intuitively, 2–9(–14) is a better representation of the range than 2–14.

4. Discussion

A metadata study of molecular systematics of many paraphyletic species of the moss family Pottiaceae (Zander, 2019a) was evaluated on an empirical, statistical basis. The discount of an outlier value associated with paraphyletic distance of *Didymodon asperifolius* by Zander (2019a) was demonstrated as unjustified, but the analytic key of that paper matched to a great extent the molecular study of Kučera and Ignatov (2015). It is possible that the parphyly of the outlier species is not unusual but that the more narrow parphyly of all other paraphyletic species was truncated by paucity of sampling.

There is no evidence, morphological or molecular, that any of the molecularly paraphyletic exemplars of *D. asperifolius* represented a cryptic taxon at the species or genus level. The coherence (close, stepwise trait transformations with no reversals) of the morphological relationships in the *Didymodon asperifolius* parphyly as demonstrated by Zander (2019a) goes against the criterion of minimal meaningful distance (Tabachnick and Fidell, 1989) for rejecting a null hypothesis (i.e. taxonomic crisis). The extreme width (14 nodes) of the *E. asperifolius* parphyly is probably due to heterogeneity in the metadata, among which the Kučera and Ignatov (2015) study of taxa now placed in *Exobryum* involved sufficient sampling to get an inkling of true molecular descent and its relationship with evolution of expressed traits. It might be expected that with sufficient sampling of other genera, the width of parphyly of related genera may also extend to about 14 nodes as two standard deviations, not four, and well within expected distribution, assuming survival of informative paraphyletic molecular races. Heterogeneity in metadata studies may well prove to be either a good measure of adequate sampling in taxonomy, or, in other

cases, the degree of extinction of molecular races in some more ancient species.

A Running Sigmas evaluation of the range of variation is useful for any continuous variable in taxonomy. A bar chart is easily developed in Excel to expose bimodal distributions revealing significant trait differences. In this study, samples larger than $N = 10$ would have been sufficient to converge sample S.D. and population S.D. I recommend, however, that more than 10 samples are

important for any required sampling of continuous variables. This includes data on molecular analyses of species, genera and families.

Acknowledgments

The author thanks the Missouri Botanical Garden for its continuing support of novel methods in bryophyte systematics. He also thanks two anonymous referees for their valuable suggestions.

References

- Aubert D (2017). A simple parsimony-based approach to assess ancestor-descendant relationships. *Ukrainian Botanical Journal* 74: 103–121.
- Barrow JD (1992). *Pi in the Sky: Counting, Thinking, and Being*. Oxford; Clarendon Press.
- Kučera J, Ignatov MS (2015). Revision of phylogenetic relationships of *Didymodon* sect. *Rufiduli* (Pottiaceae, Musci). *Arctoa* 24: 79–97.
- Ragsdale CT (2004). *Spreadsheet Modeling and Decision Analysis*. Fourth Edition. Mason, Ohio, U.S.A.; Thomson-Southwestern.
- Moore D S & Notz WI (2006). *Statistics: Concepts and Controversies*. Sixth Edition. New York; W. H. Freeman.
- Snedecor GW, Cochran WG (1967). *Statistical Methods*. Sixth Edition Ames, Iowa, U.S.A.; Iowa State University Press.
- Staley KW (2004). *Evidence for the Top Quark: Objectivity and Bias in Collaborative Experimentation*. Cambridge, U.K.; Cambridge University Press.
- Tabachnick BG & Fidell LS (1989). *Using Multivariate Statistics*. Second Edition. New York; Harper-Collins, Publishers.
- Thompson SK (2012). *Sampling*. Third Edition. Hoboken, New Jersey, U.S.A; John Wiley and Sons, Inc.
- Wallis WA, Roberrts HV (1956). *Statistics: A New Approach*. Glencoe, Illinois, U.S.A; Free Press.
- Winkler RL, Hays WL (1975). *Statistics: Probability, Inference, and Decision*. Second Edition. New York; Holt, Rinehart and Winston.
- Zander R.H (1993). Genera of the Pottiaceae: mosses of harsh environments. *Bulletin of the Buffalo Society of Natural Sciences* 32: i–vi, 1–378.
- Zander RH (2013). *Framework for Post-Phylogenetic Systematics*. St. Louis, U.S.A; Zetetic Publications.
- Zander RH (2019a). Macroevolutionary versus molecular analysis: Systematics of the *Didymodon* segregates *Aithobryum*, *Exobryum* and *Fuscobryum* (Pottiaceae). *Hattoria* 10: 1–38.
- Zander RH (2019b). Intraspecific molecular trees are associated with serial macroevolution in Pottiaceae (Bryophyta). *Ukrainian Botanical Journal* 76: 390–405.