SpecRat 1.0, a small computer program to study shifts in speciation rates

Klaus Bernhard von HAGEN

Abstract: HAGEN, K.B. v. 2003: SpecRat 1.0, a small computer program to study shifts in speciation rates. Schlechtendalia **10**: 67–63.

Shifts in diversification rates along a phylogenetic tree can be evaluated with a number of different sometimes rather complex methods. Here, the features of SpecRat 1.0, a small new computer program, which incorporates a likelihood method described in the literature some time ago are explained. It is written in C and should work with most standard compilers. The idea of the method, its implementation in C, the control features and an output option which can be useful for graphical presentation are outlined. It is also explained how the resulting parameters are used to manually calculate final probabilities applying a likelihood ratio test.

Zusammenfassung: HAGEN, K.B. v. 2003: SpecRat 1.0, a small computer program to study shifts in speciation rates. Schlechtendalia **10**: 67–63.

Um Änderungen der Diversifizierungsrate in einem Stammbaum zu untersuchen, sind mehrere zum Teil komplizierte Verfahren in der Literatur beschrieben worden. In dieser Arbeit wird SpecRat 1.0, ein kleines neues Computerprogramm, beschrieben, das auf einer schon länger bekannten Likelihood-Methode beruht. Es ist in C geschrieben und kann leicht für alle Betriebssysteme kompiliert werden. Die Methode, das Programm, Kontrollparameter und graphische Präsentationsmöglichkeiten werden erläutert. Am Ende wird die Anwendung des Likelihood-Verhältnistests beschrieben, der mit den erhaltenen Daten manuell ausgeführt werden muss.

Introduction

In a recent study on the phylogeny of *Halenia* Borkh. (Gentianaceae) we found several important events (e.g., evolution of spurs as a potential key innovation and long range dispersal to new habitats) which could have influenced speciation rates (HAGEN & KADEREIT, in press). We, therefore, were looking for methods in the literature which analyse potential rate changes along a phylogenetic tree with statistical certainty. Many such methods (with differing prerequisites) are reviewed in SANDERSON & DONOGHUE (1996) and others are described elsewhere (PURVIS et al. 1995, PARADIS 1998, STRIMMER & PYBUS 2001). However, most suitable seemed to us still another method described by SANDERSON & WOJCIECHOWSKI (1996) because it was well suitable for molecular data from extant species and not all species needed to be sampled in the phylogeny. It is a likelihood method and too complex to calculate it manually. Therefore, the computer program SpecRat 1.0 was written (the actual source code was written by G. Quast, Karlsruhe) and its use and some properties of the method are described in the following.

The Sanderson and Wojciechowski method is based on a Yule-model of diversification with Markov properties to describe the branching pattern of a phylogeny. In this model it is equally probable at any time for each branch to split (speciate) and the

probability follows a Poisson-distribution. Such a model does not correct for ending lineages (= extinction). Therefore, diversification equals speciation in this method. This could be a drawback under some circumstances, e.g., presence of heavy non-random extinction but no better method has yet been described. The prerequisites of this method are:

1) A well supported phylogeny of extant species must be available.
2) All branch lengths are determined and put under molecular clock constraint.
3) The standing diversity of all major clades must be known.

These conditions are not easily met, in our observation especially the assumption of a global molecular clock throughout a tree is often violated. The properties of using this method without a working clock hypothesis are unknown and it is potentially misleading to do so. However, a valid but yet underexplored alternative to the molecular clock constraint might be the application of non-parametric-rate-smoothing (NPRS; SANDERSON 1997).

When all named parameters are known such a tree can be described with the branching parameter lambda. This is more or less the average branching density of a tree. The method uses a likelihood formulation for this (number of clades = T; clade diversity = N1...NT; branch length of terminal clades = d1...dT; internal branch length = dT+1...dT+B):

$$L(\lambda) = \prod_{k=1}^{T} e^{-\lambda d_k} (1 - e^{-\lambda d_k})^{N_k - 1} \prod_{j=T+1}^{T+B} \lambda e^{-\lambda d_j}$$

The branching paramter lambda is the only free parameter of this formula and it needs to be numerically optimized to maximize the likelihood on the left side of the formula (Fig. 1). At first, the maximum likelihood value for a complete phylogeny is evaluated with a single lambda. Next, the phylogeny is split in two parts which are then treated independently. A tree is usually separated at that branch where a switch in speciation rates is suspected. For each of the two smaller partial trees a separate optimization procedure for lambda has to be started. The added value of the two resulting likelihoods is then compared with the single original likelihood value of the complete tree using a likelihood ratio test. Therefore, the difference of the logarithms of the likelihood values is multiplied by minus two and the result is compared with a chi-square distribution with one degree of freedom. In other words, the Sanderson and Wojciechowski method tests whether a phylogeny is significantly worse described assuming only a single average branching density (speciation rate) across the whole tree or assuming two different branching densities. For more details about the rational of the method see the original description in SANDERSON & WOJCIECHOWSKI (1996).

The program

SpecRat is written in C and was successfully compiled with Leonardo IDE 3.4.1 (DEMETRESCU & FINOCCHI 1999) on different Macintosh computers but it also worked with the GNU GCC compiler under LINUX (http://gcc.gnu.org/). A compiler must have access to the stdio.h, math.h, and nrutil.h standard C libraries. When using

Leonardo the ansi.c, nrutil.c, SpecRat.c, and visualizer.c files need to be opened in one project and compiled with Leonardos rocket button. The source code is available on request from the author and soon from the FTP server of the University of Halle. SpecRat uses the Brent algorithm as implemented in PRESS (1997) to optimize lambda and the variables are calculated using double precision (64 bit) to avoid rounding artefacts. Apart from only calculating the optimal likelihoods SpecRat performs an additional control function. The lambda can be increased by a fixed increment and the resulting likelihood values can be provided in the first part of the output. These values can easily be read in other programs (using copy and paste), e.g., in Excel (Microsoft Corporation, Redmond, Washington) and this allows a graphical representation of the optimization procedure (Fig. 1). Such a graph may serve as a control, e.g., if there is reason to suspect that the optimum could be rather broad.

How to run SpecRat

Two text-files, control.dat and in.dat, provided by the user are necessary to run SpecRat and they must be placed in the same folder as SpecRat. The file control.dat determines the limits between which SpecRat tries to find the optimal value for lambda. It also allows to set the increment for the test output which later may be used to visualize the result in a graph. For most cases the control.xmp file that comes along with SpecRat should be fine and must just be renamed into control.dat before running. The file format for control.dat is: lower limit of lambda, starting value where the approximisation of lambda starts, upper limit of lambda, increment of lambda for the output. A standard example looks like:

1.0 300.0 900.0 20.0

With this file SpecRat would search for an optimal lambda between 1 and 900. These values should be suitable because they are well out of the range of lambda values we encountered yet in practical examples. The approximisation would start with lambda = 300. The increment for the test output here is 20. The test output is not necessary and may be supressed by using 0.0 as the increment.

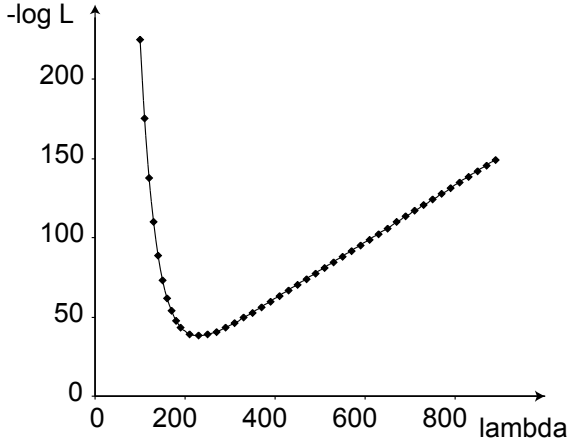


Fig. 1: Dependency of the log Likelihood from the branching parameter lambda. This graph was produced using the control feature of SpecRat on the example in-file detailed in the text.

The second text-file, in.dat, contains the data. The file format is:

number of groups compared
branch length group 1 number of OTUs group1
branch length group 2 number of OTUs group2
...
number of internal branches
length of internal branch 1
length of internal branch 2
...

A simple example looks like:

4
0.0265 529
0.0265 2496
0.0415 30
0.0580 86
2
0.0140
0.0165

Using this data file SpecRat would search for an optimal lambda on a phylogeny where four major clades are present. Two clades (sistergroups in this case) have originated 0.0265 distance units from the present and contain 529 and 2469 OTUs (= species in most cases). A third clade contains 30 species and originated 0.0415 distance units from the present and the fourth clade contains 86 species and originated 0.0580 distance units from the present. There are two internal branches which connect the four terminal clades. One has the length of 0.0140, the other 0.0165. Its best to use patristic likelihood distances under molecular clock constraint which can be received easily from PAUP 4.0beta (SWOFFORD 2001). Look in HAGEN & KADEREIT (in press) or other literature for how to define a suitable likelihood model for sequence evolution and whether a molecular clock constraint is applicable or not.

The within order of the clades or branch length is not important for SpecRat. That means an identical result is achieved when using the rearranged example from above as the in-file:

4
0.0580 86
0.0265 2496
0.0415 30
0.0265 529
2
0.0165
0.0140

The output of SpecRat

In the output first, the datafile is repeated to control for a correct input. Next, the

optional test output with a fixed increment of lambda is provided. In the last line the essential output of SpecRat is given and looks like:

min. - Log = 38.0331 at 231.4

The first number is the negative logarithm of the likelihood. The second number is the optimal lambda value and is additionally given for information but it is not necessary for the later steps of the procedure. That these values are clearly optimal may be controlled in Fig. 1.

It is obviously not necessary to have a completely bifurcating tree. It also seems not necessary to use the best resolution available. When we defined as many smaller clades and internal branches as possible for the in-file in our *Halenia* data set we qualitatively always got the same results as using few major clades only. Nevertheless, this should be tested for each data set. For the analysis of single clades there may be no internal branches. This has to be indicated by a 0 in the in-file as in the following example:

```
1
0.0347  54
0
```

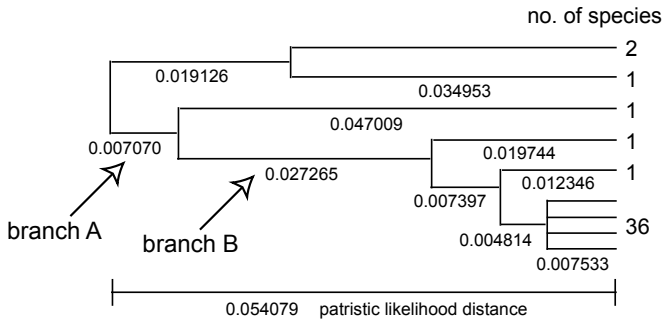
Manual steps after the likelihood calculation

After the log likelihoods have been calculated by SpecRat (three different values are necessary) a likelihood ratio test has to be performed. The procedure is shown in detail for a significant and a non-significant branch in Fig. 2. The final value must be positive so it could be necessary to multiply the result by minus one. The value must be compared with a criticial table for the chi-square distribution with one degree of freedom or the following values: P = 0.10, 2.71; P = 0.05, 3.84; P = 0.025, 5.02; P = 0.01, 6.63; P = 0.005, 7.88; P = 0.001, 10.83 (taken from http://www.ncat.edu/~warrack/chisquaretable.pdf). If the final value is lower than found for a particular level of significance it is statistically not justified to assume that a switch in speciation rate happended on the branch tested. If the value is higher a switch in speciation rate could have happened along the branch.

Further methodological advice

This method was not tested on many data sets yet but we feel that it is overly sensitive and shows a potential switch in speciation rates to often. On Fig. 2, for example, it is shown that no switch happened on branch A, but for all branches above and including branch B a switch in speciation rate was shown (HAGEN & KADEREIT, in press). It seems not likely that so many switches happened consecutively. Rather, the more inclusive clades seem not completely independent from nested clades with this method. That means on the one hand that non-significant results are very well supported (which can be very useful) but on the other hand that significant results should be treated with some caution. In the *Halenia* case, for example, we also used other tests (a lineage-through-plot and a sistergroup method) to further analyse the exact position of a switch in speciation rates and we found that a true rate switch probably happened on the highest branch of Fig. 2 only and not on the lower branches which were shown

a)



b)

	complete phylogeny	rate change tested on branch A: lower partial tree	rate change tested on branch A: upper partial tree	rate change tested on branch B: lower partial tree	rate change tested on branch B: upper partial tree
infiles for SpecRat	5 0.034953 3 0.047009 1 0.019744 1 0.012346 1 0.007533 36 5 0.019126 0.007070 0.027265 0.007397 0.004814	1 0.034953 3 1 0.019126	4 0.047009 1 0.019744 1 0.012346 1 0.007533 36 3 0.027265	2 0.034953 3 0.054079 1 1 0.019126	3 0.019744 1 0.012346 1 0.007533 36 2 0.007397

c)

optimal lambda:	134.9	36.7	164.2
- log likelihood:	16.4841	-0.968647	17.4045
		16.4841 - 16.435853	
		0.048247 x 2 = 0.096494	
		This is lower than 2.71 which is not significant on a P = 0.1 level	
		Conclusion: A rate change did not happen along branch A.	

optimal lambda:	134.9	21.62	257.9
- log likelihood:	16.4841	0.533342	7.67652
		16.4841 - 8.209862	
		8.274238 x 2 = 16.548476	
		This is higher than 10.83 which is significant on a P = 0.001 level.	
		Conclusion: A rate change could have happened on branch B because the tree is significantly better described using two branching parameters than just one.	

Fig. 2: Complete schematic procedure of the Sanderson and Wojciechowski method using SpecRat. a) Simplified, molecular clock constrained phylogeny including clade diversity (original data from HAGEN & KADEREIT, in press). b) Five different in-files for SpecRat derived from the above phylogeny. c) Application of the likelihood ratio test using the SpecRat output for the data sets directly above.

to be significant (HAGEN & KADEREIT, in press). Therefore, I would suggest to use $P < 0.001$ as the level of significance which reduces the overall sensitivity of the test. We currently test a number of other methods dealing statistically with switches in speciation rates on practical data sets, e.g., PURVIS et al. (1995), PARADIS (1998), and STRIMMER & PYBUS (2001). We have no final results yet but, similar with the Sanderson and Wojciechowski method incorporated in SpecRat and described here, all of which seem to have the one or the other drawback.

Literature:

DEMETRESCU, C. & FINOCCHI, I. 1999: Leonardo 3.4.1 Power PC. Computer program available via http://www.dis.uniroma1.it/~demetres/leonardo/.

HAGEN, K.B. von & KADEREIT, J.W. in press: The diversification of *Halenia* (Gentianaceae): Ecological opportunity versus key innovation. Evolution.

PARADIS, E. 1998: Detecting shifts in diversification rates without fossils. American Naturalist **152**: 176–187.

PRESS, W.H. 1997: Numerical recipes: the art of scientific computing. Cambridge University Press, Cambridge.

PURVIS, A., NEE, S. & HARVEY, P.H. 1995: Macroevolutionary inferences from primate phylogeny. Proceedings of the Royal Society London, Series B **260**: 329–333.

SANDERSON, M.J. 1997: A nonparametric approach to estimating divergence times in the absence of rate constancy. Molecular Biology and Evolution **14**: 1218–1231.

SANDERSON, M.J. & DONOGHUE, M.J. 1996: Reconstructing shifts in diversification rates on phylogenetic trees. Trends in Ecology and Evolution **11**: 15–20.

SANDERSON, M.J. & WOJCIECHOWSKI, M.F. 1996: Diversification rates in a temperate legume clade: are there "so many species" of *Astragalus* (Fabaceae). American Journal of Botany **83**: 1488–1502.

STRIMMER, K. & PYBUS, O.G. 2001: Exploring the demographic history of DNA sequences using the generalised skyline plot. Molecular Biology and Evolution **18**: 2298–2305.

SWOFFORD, D.L. 2001: PAUP* Phylogenetic analysis using parsimony (* and other methods) Version 4. Sinauer Associates, Sunderland.

Address of the author:

Dr. K.B. v. Hagen, Martin-Luther-Universität Halle-Wittenberg, Institut für Geobotanik und Botanischer Garten, Neuwerk 21, D-06099 Halle/Saale, Germany.
(email: vonhagen@botanik.uni-halle.de)