The χ^2 test for compositional homogeneity, improved

Peter G. Foster *

13 April, 2003

The χ^2 test for compositional homogeneity uses a contingency table (or $R \times C$ table) of the compositions of the various taxa against the mean composition to determine if the composition of the data as a whole is homogeneous. The underlying question we are asking here is: Does the composition of the data fit a homogeneous composition model with empirical composition? (The same question but with ML optimized composition, and the same question but with heterogeneous composition models, are different, although the former is not very different.)

When this test is done in PAUP, the output is like this:

```
paup> basefreq

Base frequencies:

Taxon              A         C         G         T     # sites
-------------------------------------------------------------
A            0.09300   0.20800   0.29300   0.40600        1000
B            0.11600   0.16100   0.33600   0.38700        1000
C            0.10300   0.19100   0.30900   0.39700        1000
D            0.09787   0.21277   0.30426   0.38511         940
-------------------------------------------------------------
Mean         0.10254   0.19289   0.31066   0.39391      985.00

Chi-square test of homogeneity of base frequencies across taxa:

Taxon                  A         C         G         T
-------------------------------------------------------
A           O         93       208       293       406
            E     102.54    192.89    310.66    393.91
B           O        116       161       336       387
            E     102.54    192.89    310.66    393.91
C           O        103       191       309       397
            E     102.54    192.89    310.66    393.91
D           O         92       200       286       362
            E      96.39    181.32    292.02    370.27

Chi-square = 15.161065 (df=9), P = 0.08660575
Warning: This test ignores correlation due to phylogenetic structure.
```

I will call the statistic X^2, in the sense used by Sokal and Rohlf in Biometry. They, and I, distinguish the statistic X^2 from χ^2, the distribution. Here $X^2 = 15.2 = \sum (obs - exp)^2/exp$, where obs is observed and exp is expected from the

*Department of Zoology, The Natural History Museum. email: p.foster@nhm.ac.uk

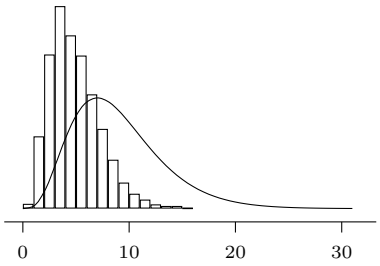

Figure 1: A histogram of 2000 X^2 statistics collected from simulations on the tree and model. The line is the χ^2 curve with 9 degrees of freedom.

data. The χ^2 distribution, with degrees of freedom $dof = 9 = (R-1) \times (C-1)$, is used as the null distribution to determine the P value and whether X^2 is significant. In the example above, $P = 0.087$ and so X^2 is not significant. So we conclude that the data are homogeneous and reassure ourselves that if we use a homogeneous model using empirical composition then the composition of the data will not be in violation of the stationarity assumption of the model. If the composition fails this test then we usually plow ahead and use a homogeneous model anyway, keeping in mind that the data are in violation.

The output to the χ^2 test above warns us that the test "ignores correlation due to phylogenetic structure". The problem is that sequences related on a tree with realistically short branch lengths have correlated composition. This means that this test, when applied to phylogenetic data, suffers from Type II error. This means that data that are truly not homogeneous may not be rejected by this test. The source of the problem is that the χ^2 null distribution is inappropriate.

A workaround is to use a null distribution tailored to the specific problem at hand. We can use the same statistic, X^2. We can simulate true null hypothesis data based on the specific tree and model at hand, and collect X^2 statistics from those simulations to be the null distribution. If we use that null distribution rather than the χ^2 distribution then the test will not suffer from Type II error due to correlated compositions. I first heard this articulated by Dave Swofford and Jack Sullivan, although I do not know of it in a publication.

For example, we use the data above, and optimize the tree ((A, B), C, D) with the Jukes-Cantor model. We use that tree and model to generate data sets of the same size as the original data set, collecting X^2 as we go. These are plotted in the histogram in the Figure. The statistic for the original data is the same as before, 15.2, but instead of using the χ^2 curve to determine significance we use the histogram, and get $P \approx 0.0005$. Only one of the 2000 simulation X^2 values was greater than that from the original data. With that, we can reject homogeneity of the data, wheras using the χ^2 curve we cannot.

The tree and model based composition fit test that I have concocted is different. It asks whether the possibly heterogeneous model fits the possibly heterogeneous model. It uses a similar X^2_m statistic, where the expected composition comes from the model, not from the data.