SimCo: a program to automate the comparison of multiple Structure runs

Clare A Rebbeck[1], Owen R. Jones[1*], Isheng J. Tsai[1]

[1] Division of Biology, Imperial College London, Silwood Park Campus, Ascot, Berks. SL5 7PY, UK

[*] Corresponding author: Email: owen.jones@imperial.ac.uk

ABSTRACT

SUMMARY: The computer program "Structure" has become an increasingly popular tool for scientists seeking to assign individuals to discrete populations using molecular data. Typically, it is necessary to compare several runs of Structure outputs in order to evaluate the consistency of assignments. A commonly used metric for comparing these assignments is the similarity coefficient. Here we present a computer program, SimCo, which rigorously calculates the similarity coefficient between multiple Structure runs. We provide versions that can be implemented in PERL (SimCo-PERL) and R (SimCo-R) and make these publicly available.

AVAILABILITY: The programs are available under the GNU General Public License and are provided as supplementary information available from the journal's website. Simco for R is available from http://cran.r-project.org/.

CONTACT: owen.jones@imperial.ac.uk

SUPPLEMENTARY INFORMATION: Available on Bioinformatics website

INTRODUCTION

Population geneticists often need to assign individuals to distinct populations. For example, analysis on the genetic structure of domestic dog breeds (Parker et al. 2004) and studies on human evolution (Adeyomo et al. 2005) and human migration (Falush et al. 2003) focus on the evolutionary relationships of modern populations and the individuals within populations. The program Structure (Pritchard et al. 2000; Falush et al. 2003) implements a Bayesian model-based clustering algorithm which attempts to infer population structure and assign individuals to populations probabalistically based on patterns of allele frequencies [multilocus genotype data (e.g. SNPs, RFLPs and microsatellites etc.]. A drawback of the clustering method implemented in Structure is that it assumes K populations, where K is defined by the user. If the value of K chosen is inappropriate, individuals can potentially form weak, inconsistent associations with their assigned clusters, which can influence any inferences drawn from an analysis (Pritchard et al. 2000). This, together with the inherent uncertainty in the execution of Bayesian programs using the Markov Chain Monte Carlo simulation (Pearse & Crandall 2004) make it advisable to perform multiple runs with the same data set.

Quantifying the similarity of outputs, by calculating the similarity coefficient is often used as a measure of confidence for the populations (Parker et al. 2004, Rosenberg et al. 2002). These calculations are, however, time consuming and complicated to carry out, especially with increasing numbers of simulations, where the number of possible combinations increases exponentially with the number of runs. To resolve this problem we have written programs, which we call "SimCo", implementable in R (R Development Core Team, 2005) and PERL, to take the probabilistic assignments that Structure gives and carry out the necessary permutations and similarity coefficient calculations automatically. We are releasing these programs under

the GNU General Public License. The programs are available as text files in the supplementary information or from the authors.

System and Methods

Structure output files contain data on the probability of assignment to one of K populations. SimCo takes these output files and uses the assignment probabilities given therein to calculate similarity coefficients as previously described by Rosenberg et al. (2002). It does this for all possible pairwise combinations of n Structure runs (n!/(2!*(n-2)!) and then outputs the summary statistics of the distribution of the calculated similarity coefficients (range, moments, mean).

Because the population identities do not correspond with any particular output column number, any two matrices cannot be directly compared. Therefore, the similarity coefficient calculated within SimCo permutes the columns of the second matrix for every possible combination of column ordering (K! combinations), and calculates the absolute difference between the first and second matrix for each permutation. The column-permutation with the smallest difference is deemed to be the best fit for a direct comparison between the two matrices.

Algorithm

The similarity coefficient for every pair of runs is calculated using equation 1;

$$C(M1, M2) = 1 - ||M1-M2||F \ / \ ||M1-1/K||F$$

Equation 1

where C(M1, M2) is the similarity coefficient for the comparison between stochastic matrices M1 and M2. The matrices have dimensions I x K (I = number of individuals; K= number of supposed clusters that the individuals are assigned to) and are the outputs (i.e. the probability of assignment to a particular cluster) from two separate Structure runs using the same data. The notation ||x|| F is the Frobenius matrix norm (Golub and Van Loan 1996) and 1/K is the I x K matrix with all the entries equal to 1/ K.

IMPLEMENTATION

Both the Perl and R versions of the program can be used without the need to reformat the Structure output file. Simply by editing the SimCo input command to include the file pathways of simulations you wish to compare, any number of Structure runs can be analyzed (memory dependent) simultaneously. Further details for running both the R and Perl versions are given in PDF documents in the supplementary information. In addition there are help files associated with the R version of simco (the 'simco' package). R is available for Mac, Windows PC and Linux from http://www.R-project.org.

DISCUSSION

When implemented on a test data set with 3 runs, a K of 3 and 74 samples, both SimCo-R and SimCo-PERL produced the same similarity coefficient value. Both programs were fast with the test dataset completing the analysis in less than 1 second (on a Mac G5 with 4 x 2.5 GHz PowerPC processors and 2GB SDRAM). Tests indicate that, with datasets where a larger number of runs and more assumed clusters are used, the program will take significantly longer to run. We provide some test files as text documents in the supplementary information.

CONCLUSION

SimCo provides significant assistance to users of Structure, allowing the rigorous comparison of the generated output files without the need to manipulate them. We are aware that the need to carry out such matrix comparisons is widespread in biology (and the other sciences). We are, therefore, making SimCo freely available on a license that allows free use, distribution and alteration as long as we are acknowledged for our original work.

ACKNOWLEDGEMENTS

We thank Elaine Ostrander for providing us with a test dataset to work with. CAR and IJT were supported by Wellcome Trust Studentships.

CONFLICT OF INTEREST: none declared

REFERENCES

Adeyomo, A. A. et al. (2005) Genetic structure in four West African population groups. *BMC Genet*. 6:38, 1471-2156

Falush, D. et al. (2003) Traces of human migrations in Helicobacter pylori populations. *Science*. 299, 1582-1585.

Falush, D.et al. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated frequencies. *Genetics*. 164, 1567-1587.

Golub, G. H., and Van Loan, C. F. (1996) *Matrix computations*. Johns Hopkins University Press, Baltimore.

Parker, G. H. et al. (2004) Genetic structure of the purebred domestic dog. *Science*. 304, 1160-1164.

1 Pearse, D. E. & Crandall, K. A. (2004) Beyond FST: Analysis of population genetic data for
2 conservation. *Conserv. Genet.* 5, 585-602.

3 Pritchard, J. K. et al. (2000) Inference of population structure using multilocus genotype data.
4 *Genetics*. 155, 945-959.

5 R Development Core Team (2005) R: A language and environment for statistical computing. R
6 Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL
7 http://www.R-project.org.

8 Rosenberg, N. A. et al. (2002) The genetic structure of human populations. *Science*. 298, 2381-
9 2385.