TempNet: An R script to produce temporal statistical parsimony networks for heterochronous DNA data in Epidemiology and Ancient DNA research.

Stefan Prost[1,2*], Christian Anderson[3]

1 Allan Wilson Centre for Molecular Ecology and Evolution, Department of Anatomy and Structural Biology, University of Otago, Dunedin, New Zealand
2 Department of Integrative Biology, University of California, Berkeley, United States of America
3
∗ **E-mail: stefan.prost@anatomy.otago.ac.nz**

Summary

1. The use of heterochronous data to study demographic changes in epidemiology and ancient DNA studies has revolutionized our understanding of complex evolutionary processes such as invasions, migrations or responses to drugs or climate change. While there are sophisticated applications based on MCMC or ABC to study this processes through time, no convenient way to display relationships within populations through time is available at the moment.

2. TempNet is a user-friendly R script that creates temporal statistical parsimony networks based on general genetic input-files such as fasta files.

Introduction

Heterochronous DNA data consists of sequences of different age. Such data is often used in epidemiology and ancient DNA to study demographic changes over time. In epidemiology it can provide substantial insights into the spread of infectious diseases and drug treatment and therefore help to control possible future outbreaks [1]. In ancient DNA research temporal DNA data is often used to investigate mammal populations responses to climate -, environmental change or the spread of modern humans (e.g. [2–5]). A break through in the study of heterochronous data was the development of sophisticated Markov Chain Monte Carlo approaches to study demography over time (such as the Skyline Plot [6]) and the development of Approximate Bayesian Computation [7]. Though, the plotting of haplotype relationships within populations still relies on 2 dimensional network reconstruction. 2 dimensional plotting of temporal data is often not a sophisticated way to explore and understand temporal coherence. The first 3 dimensional network based on individual 2 dimensional statistical parsimony networks constructed using the TCS software [8] and then combined by hand was published recently in Prost et al. 2010 [4]. However, constructing 3 dimensional - based on 2 dimensional networks by hand is far from being convenient and also error-prone. Here we present an R script to automatically produce 3 dimensional statistical parsimony networks and thus substantially alleviating its construction.

Methods

TempNet is written for the open-source project R (http://www.r-project.org/) and is based on the freely available R packages "ape" and "pegas".
It is an open-source script available at http://www.stanford.edu/group/hadlylab/tempnet/ . It uses statistical parsimony [9] to construct the haplotype network. Statistical parsimony estimates the maximum

number of differences (substitutions) among haplotypes with a 95% confidence interval [9, 10]. In the following steps first sequences with one difference, then with two and so on will be connected until the parsimony limit (confidence interval from the first step) will be reached [9,10].

TempNet uses the read.dna() function from the "ape" package to easily import DNA sequences from standard formats such as "fasta" and the haploNet() function from the "pegas" package to construct a statistical parsimony network object. The user has only to specify the location of the fasta file and the layer assignment. Additionally, it allows to change the size for the circles corresponding to the mutations and the size of the ellipses corresponding to the unsampled haplotypes within a certain layer. There is also the possibility to assign labels to the network layers. There are different ways to specify, which sequences belong to which layer of the network. The first and probably easiest way is to provide the assignment as part of the sequence name in the "fasta" file, simply by appending the symbol "$" and the specific layer number subsequently to the name. When different layer assignments will be tested TempNet also allows to provide the layer assignments as a simple vector in the order of the sequence appearance in the "fasta" file. The script will produce a 2 dimensional statistical parsimony network if neither labels or a vector is provided.

Network appearance
Haplotypes are represented by ellipses, for convenience the size of the ellipses is scaled according to the number of sequences that show the same haplotype. If a haplotype is found in the entire network, but not in the particular layer it will appear as a white ellipse. Two existing halpotypes are connected by a solid line, whereas two unsampled haplotypes or one sampled and one unsampled haplotypes will be connected by dotted lines. Two haplotypes are connected by a line if they are separated by one mutation, each additional mutation will be represented by a small black circle. If a haplotype is present in more than 1 layer the corresponding ellipses at the time layers will be connected by vertical lines. Once upon construction the appearance of the network can be reshuffled by hand for a better data representation.

Examples

We present 3 examples to illustrate the use and the strength of our R script and temporal network representations in general. First we use the data from ... as an example of a standard ancient DNA data set. In example two we use ancient DNA from Caramelli et al. 2007 [13] as well as a hugh modern day sampling from GenBank to illustrate the capability of our approach to illustrate hugh data-sets. In the last example we use the data from Bennett et al. 2003 [11] to show its applicability in epidemiological research. Virus are rapidly evolving pathogens that often exhibit much higher mutation rates than mammals [12] .

Example 1

Example 2
Ancient DNA data sets often include by far more modern DNA sequences than ancient ones, as modern DNA sequencing is way cheaper and faster. We used the ancient DNA data set from Caramelli et al. 2007 [13] along with 234 modern Sardinian DNA sequences from GenBank (DQ067827-DQ067877, DQ081420 -DQ081607 and DQ081669-DQ081715) to show that temporal network reconstruction proof to be suitable for bigger data sets and represent easy ways to summarize relationships within the data (see figure 2).

Example 3
Heterochronous sampling is a common and powerful practice in epidemiological studies. Virus show low

generation times and high mutation rates, which makes them perfect model organism to study temporal demographic changes. We used the data from Bennett et al. 2003 [11], which consists of 75 sequences from the dengue virus (DENV-4 isolates) randomly sampled in the years 1982 (n=14), 1986/1987 (n=19), 1992 (n=15), 1994 (n 5=14), and 1998 (n=13) to show the applicability of our approach in epidemiological studies (see figure 3).

Discussion

We show that temporal networks are suitable to display and summarize relationships within heterochronous data common in ancient DNA or epidemiological research. The presented R script is user-friendly and will most likely be of great use to illustrate complex relationships in research areas dealing with temporal DNA data. Using 3 examples we showed that TempNet can deal with standard as well as hugh data-sets. It proofs useful even if big data-sets of fast evolving organisms such as virus are used.

Acknowledgements

Figures

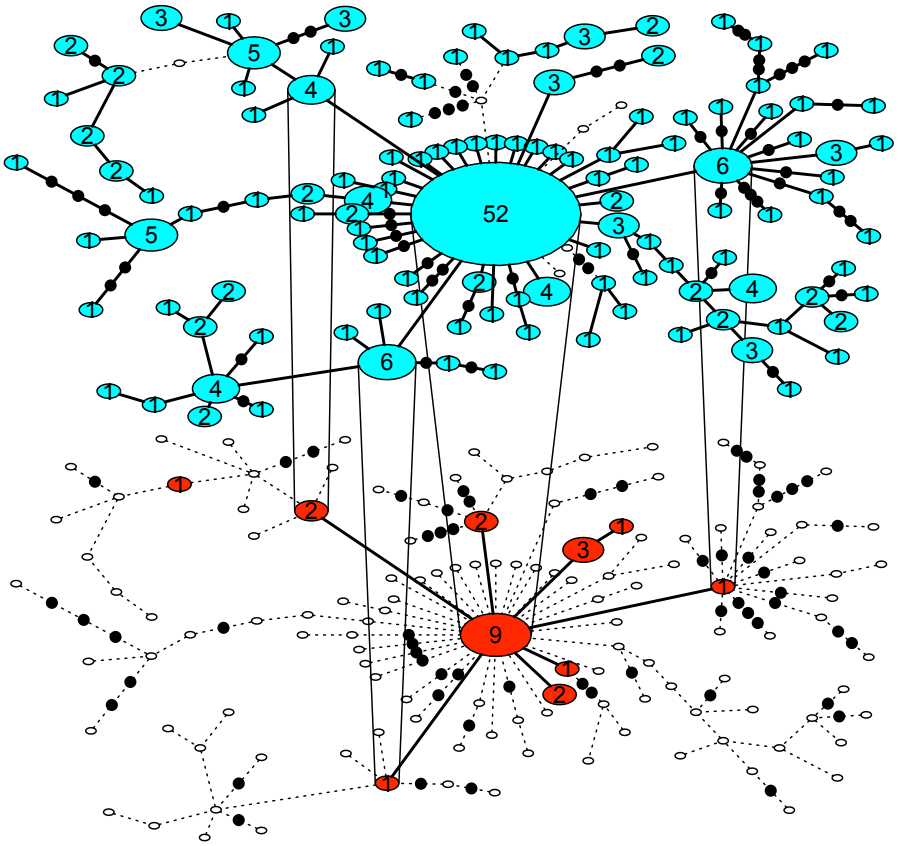

Figure 1. Scheme

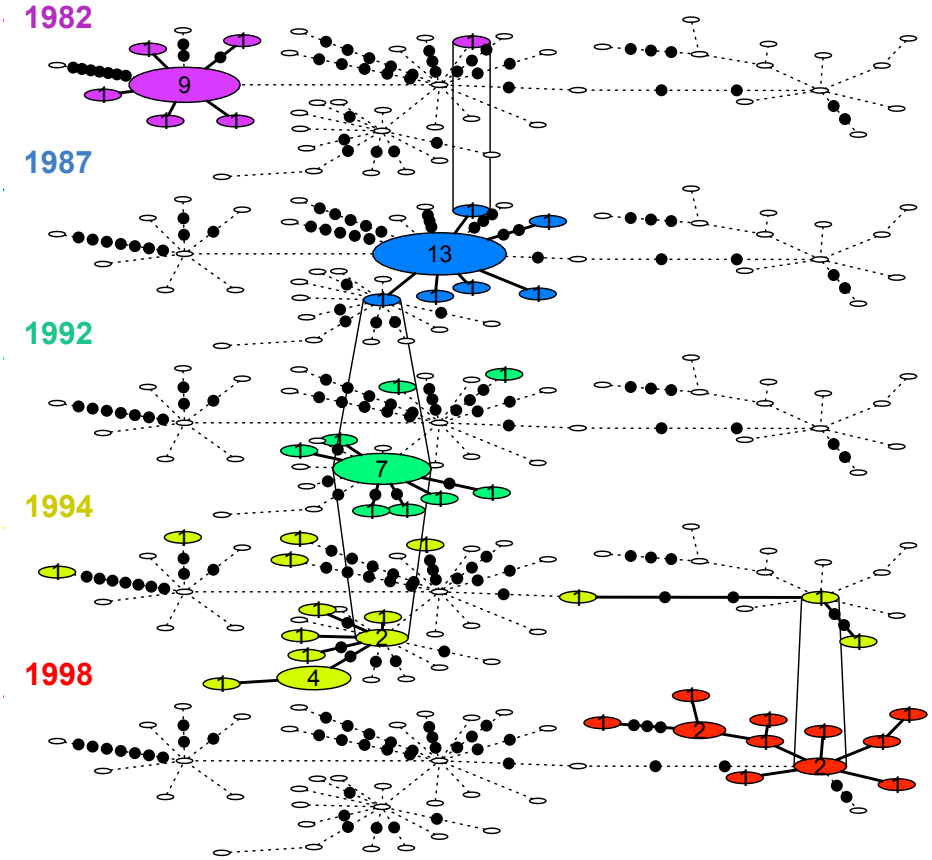

Figure 2. Scheme

References

1. Pybus O, Rambaut A (2009) Evolutionary analysis of the dynamics of viral infectious disease. Nature Reviews Genetics 10: 540–550.

2. Hadly E, Ramakrishnan U, Chan Y, Mv T, O'Keefe K, et al. (2004) Genetic response to climatic change: insights from ancient DNA and phylochronology. PLoS Biol 2: 1601-1609.

3. Shapiro B, Drummond A, Rambaut A, Wilson M, Matheus P, et al. (2004) Rise and fall of the Beringian steppe bison. Science 306: 1561.

4. Prost S, Smirnov N, Fedorov V, Sommer R, Stiller M, et al. (2010) Influence of Climate Warming on Arctic Mammals? New Insights from Ancient DNA Studies of the Collared Lemming Dicrostonyx torquatus. PloS one 5: e10447.

5. Campos P, Willerslev E, Sher A, Orlando L, Axelsson E, et al. (2010) Ancient DNA analyses exclude humans as the driving force behind late Pleistocene musk ox (Ovibos moschatus) population dynamics. Proceedings of the National Academy of Sciences 107: 5675.

6. Pybus O, Rambaut A, Harvey P (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155: 1429.

7. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian Computation in population genetics. Genetics 162: 2025-2035.

8. Clement M, Posada D, Crandall K (2000) TCS: a computer program to estimate gene genealogies. Molecular Ecology 9: 1657–1659.

9. Templeton A, Crandall K, Sing C (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics 132: 619.

10. Posada D, Crandall K (2001) Intraspecific gene genealogies: trees grafting into networks. Trends in Ecology & Evolution 16: 37–45.

11. Bennett S, Holmes E, Chirivella M, Rodriguez D, Beltran M, et al. (2003) Selection-driven evolution of emergent dengue virus. Molecular biology and evolution 20: 1650.

12. Duffy S, Shackelton L, Holmes E (2008) Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews Genetics 9: 267–276.

13. Caramelli D, Vernesi C, Sanna S, Sampietro L, Lari M, et al. (2007) Genetic variation in prehistoric Sardinia. Human genetics 122: 327–336.