Chapter 6

CoalFace: a graphical user interface program for the simulation of coalescence

"I've never had a conflict between teaching and research as some people do because when I'm teaching, I'm doing research"

Raymond Smullyan

CoalFace: a graphical user interface program for the simulation of coalescence

Wayne Delport

Molecular Ecology and Evolution Programme, Department of Genetics, University of Pretoria, Pretoria, 0002, South Africa
wdelport@postino.up.ac.za

Abstract

In this manuscript I describe a computer program that simulates the coalescent process and provides visual outputs of coalescent genealogies to the screen. *CoalFace* is a user-friendly program for teaching the principles of coalescence to both undergraduate and postgraduates in population genetics. In addition, *CoalFace* can generate data for distributions of the time to the most recent common ancestor, number of segregating sites and common diversity indices, from multiple simulations. Windows and Linux (Intel) executables are available at http://www.up.ac.za/academic/genetics/staff/Bloomer/Research/software.htm.

keywords: Coalescence, simulation, genealogies

Accepted for publication in *Molecular Ecology Notes*

The principle of coalescence has gained much recent attention within the population genetics and phylogeography literature recently (Fu & Li 1999, Emerson, Paradis & Thébaud 2001, Rosenberg & Nordborg 2002, Nordborg 2003). This increase is largely attributed to several authors who have worked at developing both coalescent theory (Hudson 1991, Donnelly & Tavaré 1995, Tavaré *et al.* 1997, Bahlo & Griffiths 2000, Wakeley & Aliacar 2001, Excoffier 2004), and analytical software based on the coalescent (Beerli & Felsenstein 2001, Kuhner, Yamato & Felsenstein 1998, Nielsen & Wakeley 2001). Although there is this increase in studies involving the coalescent, few phylogeographic studies take the randomness of the coalescent into account, with notable exceptions (Fleischer *et al.* 1998, Schneider *et al.* 1998, Edwards & Beerli 2000, Irwin 2002). Indeed Knowles (2004) has highlighted the need for coalescent-based modeling in phylogeographic studies. I believe the lack of reference to the coalescent among typical phylogeographical studies, in South Africa at least, is the result of a lack of understanding of how data can be interpreted in light of coalescent models. Furthermore, even though there are several good reviews that explain the difference between gene trees and phylogenetic trees (Smouse 1998, Nichols 2001, Posada & Crandall 2001), and the inappropriateness of the latter in phylogeographical studies, I still find that South African postgraduates and researchers involved in phylogeography misunderstand the implications of coalescent theory for phylogeographical analyses. To this end I have developed a software program, *CoalFace*, which simulates the coalescent process, with the aim of teaching coalescence to both undergraduate students and workshop participants.

CoalFace is a user-friendly software program, written in the Borland Delphi/Kylix programming language, which simulates the coalescent process. The software is written as a teaching aid is therefore, strongly visually orientated, with the ability to draw coalescent genealogies on screen (Figure 1). *CoalFace* can simulate coalescent genealogies, with or without mutations. Mutations are implemented in *CoalFace*, given a mean mutation rate per site per generation. *CoalFace* can simulate mutations under either an infinite or finite sites/alleles model. In the former, the number of segregating sites is simply output as the number of mutations that have occurred on the genealogy. In addition, *CoalFace* is capable of specifying either a finite sites sequence or finite alleles microsatellite mutation model. In the sequence mutation model, mutations can accumulate according to a JC69 (Jukes & Cantor 1969), F81 (Felsenstein 1981), Kimura 2-parameter (Kimura 1980) or HKY85 (Hasegawa,

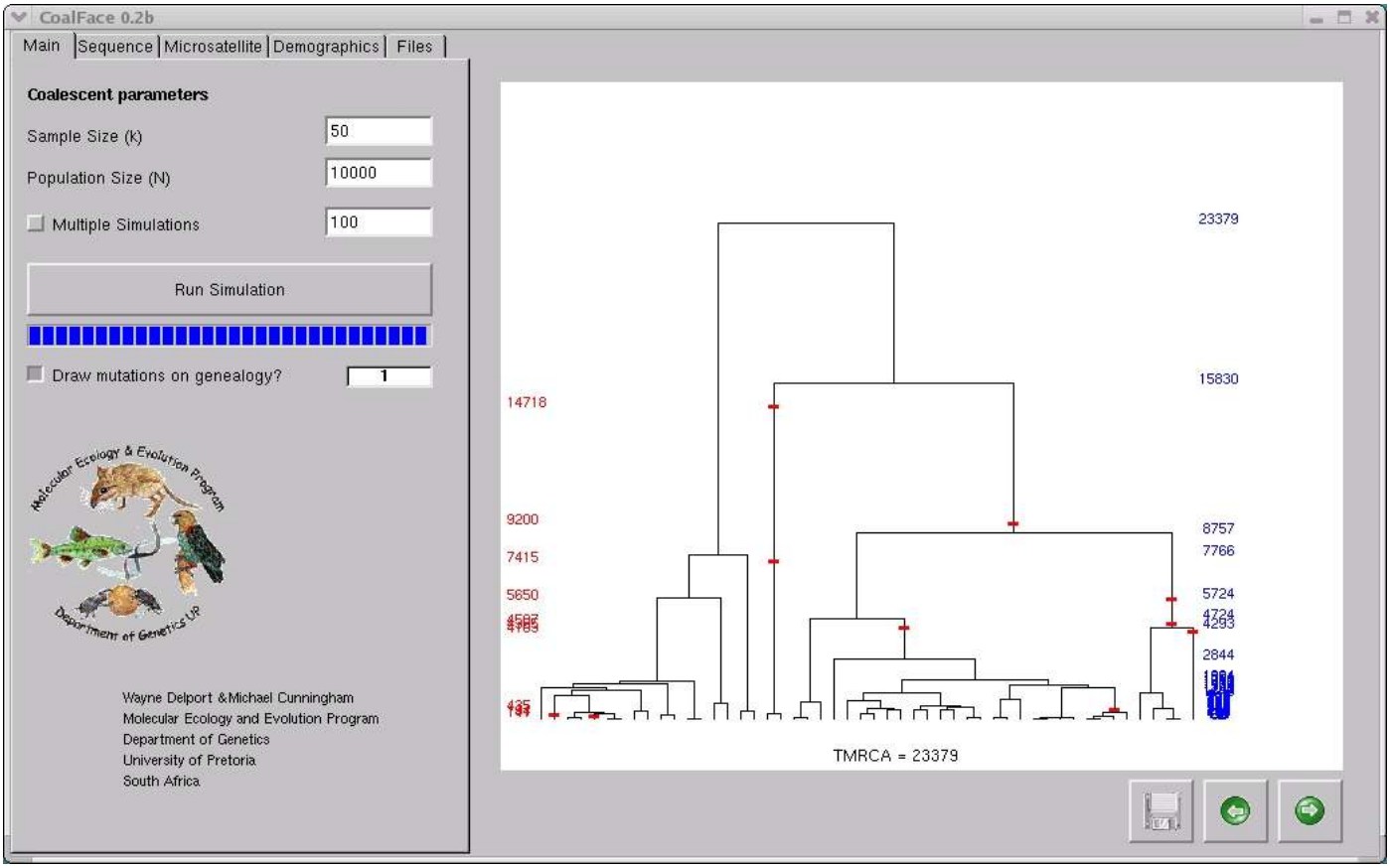

Figure 1: Screenshot of CoalFace showing the visually orientated representation of coalescent genealogies.

Kishino & Yano 1985) substitution model. Among-site rate variation can be specified according to the alpha parameter of a gamma distribution (Yang 1996), as can the proportion of invariable sites. I have found the deviation of the number of segregating sites in an infinite sites model from the number of alleles in a finite sites model especially useful to represent the relationship between mutation rate and the incidence of homoplasy. In addition, the relationship between population size, mutation rate and theta can be investigated, such that students can begin to understand that a large population size with a low mutation rate, is equivalent to a small population, with high mutation rate, in terms of coalescence and population genetic theory. This understanding aids the interpretation of much population genetic literature, which is largely theta-based. In the finite alleles microsatellite model, coalescent simulations are performed independently for each locus, according to a stepwise mutation model (Ohta & Kimura 1973) or a random allele model. Again a comparison of the number of alleles derived from the infinite alleles model, and the finite alleles models, at different mutation rates aids in the understanding of the incidence of homoplasy.

The stochastic nature of the coalescent however, cannot be understood from single simulations of the coalescent. Typically, one should derive distributions of statistics of interest (Figure 2) to gain an understanding of the potential level of stochasticity given a population size and mutation rate. Therefore, I have implemented multiple simulations in *CoalFace*, where the results of each simulation are output to various files. A common statistic of interest is the distribution of times to the most recent common ancestor, and I have used this to teach students the effects of drawing from a given distribution. Furthermore, from an understanding of the distribution of time to the most recent common ancestor, students gain an understanding of why it is notoriously difficult to estimate divergence times from population genetic data. Clearly, the time to the most recent common ancestor is not the only statistic of interest, and *CoalFace* can calculate common diversity indices from sequences generated in the simulation. In additon, I have added a procedure that allows one to output both sequence (nexus, fasta, mrbayes, clustal formats) and microsatellite data to data files. These can then be imported into other software packages for analyses. *CoalFace* can also assemble Arlequin (Schneider *et al.* 2000) files and Arlequin batch files, from multiple simulations. When running multiple simulations, only the last genealogy is drawn, yet the genealogies of previous simulations can be rapidly flipped

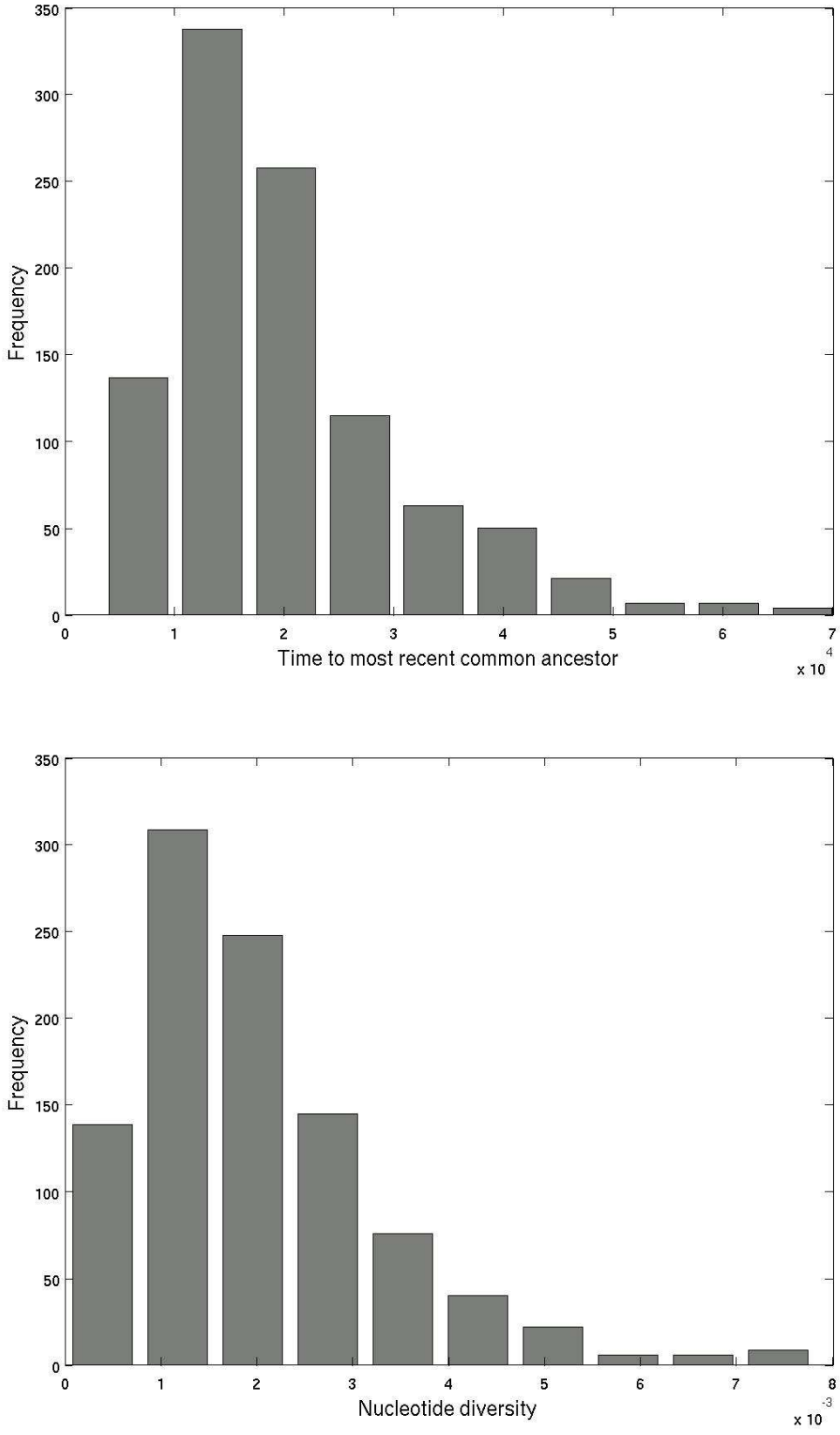

Figure 2: Distributions of (a) time to the most recent common ancestor, and (b) nucleotide diversity derived from 1000 simulations of the coalescent process (k = 50, N = 10000).

over in *CoalFace*. The genealogies are also output to a newick tree file and can be viewed in Rod Page's TreeView (Page 1996).

Finally, I have implemented some basic demographic scenarios in *CoalFace*, such that students can gain an understanding of how increases or decreases in population size, or variance in reproductive success can influence the time to coalescence of a population, and consequently the occurrence of mutations. These demographic simulations are simple and should not be used to test hypotheses of population expansion and contraction; there are far better and faster programs available for these purposes (Excoffier Novembre & Schneider 2000, Grassly, Harvey & Holmes 1999). *CoalFace* is available free from http://www.up.ac.za/academic/genetics/staff/Bloomer/Research/software.htm. Executables for both Windows (NT, 2000, XP) and Linux (intel) are available, as is a simple manual describing how to use the software. Typical run times, on a Linux based Intel 2.0 Ghz Pentium 4 with 384MB Ram, for a single simulation of 50 individuals in a population size of 10000 are on the order of 10-15 seconds. Multiple simulations (1000), with the calculation of diversity indices and exporting of sequence data files take approximately 90 minutes, whereas multiple simulations (1000) with no diversity index calculations take approximately 15-30 minutes. These test runs were performed by sampling 50 individuals from a population size of 10000, and reducing the population size can increase the speed of the runs.

Acknowledgements

This work was partially funded from a Mellon Foundation grant to Wayne Delport and Prof. J. Willem H. Ferguson, and we would like to thank Wayne Delport's PhD supervisors; Prof Paulette Bloomer and Prof J. Willem H. Ferguson, for affording him the time to develop this software program. In addition, we would like to thank participants at the 2003 Phylogeography workshop, presented by the Systematics Society of Southern Africa, for test-driving an earlier version of this software. Finally, we would like to thank Joe Felsenstein for answering some of our coalescence related questions. Also thanks to Michael Cunningham for useful and thoughtful discussion.

References

Bahlo M, Griffiths RC (2000) Inference from gene trees in a subdivided population. *Theoretical Population Biology*, **57**, 79-95.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences, USA*, **98**, 4563-4568.

Donnelly P, Tavaré S (1995) Coalescents and the genealogical structure under neutrality. *Annual Review of Genetics*, **29**, 401-421.

Edwards SV, Beerli P (2000) Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, **54**, 1839-1854.

Emerson BC, Paradis E, Thébaud C (2001) Revealing the demographic histories of species using DNA sequences. *Trends in Ecology and Evolution*, **16**, 707-716.

Excoffier L (2004) Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Molecular Ecology*, **13**, 853-864.

Excoffier L, Novembre J & Schneider S (2000) SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbritary demography. *Journal of Heredity*, **91**, 506-510.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368-376.

Fleischer RC, McIntosh CE, Tarr CL (1998) Evolution on a volcanic conveyor belt: using phylogeographic reconstructions and K-Ar-based ages of the Hawaiian Islands to estimate molecular evolutionary rates. *Molecular Ecology*, **7**, 533-545.

Fu Y-X, Li W-H (1999) Coalesceing into the 21st Century: An overview and prospects of Coalescent theory. *Theoretical Population Biology*, **56**, 1-10.

Grassly NC, Harvey PH & Holmes EC (1999) Population dynamics of HIV-1 inferred from gene sequences. *Genetics*, **151**, 427-438.

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160-174.

Hudson R (1991) Gene genealogies and the coalescent process. In Oxford Surveys in Evolutionary Biology vol 7. Futuyma D, Antonovics J (eds). Oxford University Press, Oxford.

Irwin DE (2002) Phylogeographic breaks without geographic barriers to gene flow. *Evolution*, **56**, 2383-2394.

Jukes TH, Cantor C (1969) Evolution of protein molecules. In Mammalian Protein Metabolism. Munro MN (ed). Academic Press, New York.

Kimura M (1980) A simple model for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111-120.

Knowles LL (2004) The burgeoning field of statistical phylogeography. *Journal of Evolutionary Biology*, **17**, 1-10.

Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429-434.

Nichols R (2001) Gene trees and species trees are not the same. *Trends in Ecology and Evolution*, **16**, 358-364.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov Chain Monte Carlo approach. *Genetics*, **158**, 885-896.

Nordborg M (2003) Coalescent Theory. In Handbook of Statisical Genetics, 2nd edition. Balding DJ, Bishop M, Cannings C (eds). John Wiley and Sons, Ltd.

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, **22**, 201-204.

Page RDM (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, **12**, 357-358.

Posada D, Crandall KA (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology and Evolution*, **16**, 37-45.

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, **3**, 380-390.

Schneider CJ, Cunningham M, Moritz C (1998) Comparative phylogeography and the history of endemic vertebrates in the wet tropics rainforests of Australia. *Molecular Ecology*, **7**, 487-498.

Schneider S, Roessli D, Excoffier L (2000) Arlequin: a software for population genetics data analysis. Ver 2.000. Genetics and Biometry Lab, Department of Anthropology, University of Geneva.

Smouse PE (1998) To tree or not to tree. *Molecular Ecology*, **7**, 399-412.

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505-518.

Wakeley J & Aliacar N (2001) Gene genealogies in a metapopulation. *Genetics*, **159**, 893-905.

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution*, **11**, 367-372.

Chapter 7

Conclusions

" I love deadlines, I love the wooshing noise they make as they go by."

Douglas Adams

The purpose of the research conducted in this dissertation was to estimate demographic parameters, such as migration and population size changes, using population genetic data in the continuously distributed African Wild Silk Moth, *Gonometa postica*. The African Wild Silk Moth is a species that exhibits large inter-annual population size fluctuations. However, it has proven difficult to estimate these parameters using population genetic data as a result of these fluctuations. Previous theoretical work in this area has shown that the effect of population size fluctuations on spatial genetic structure in metapopulations is dependent on the number of individuals colonizing a deme versus the number of recurrent migrants between demes. In general, however most population size fluctuations generate an increase in spatial genetic structure as a result of local genetic drift. However, in the species considered here high levels of gene flow are inferred from microsatellite data. These results are in agreement with other population genetic studies of cyclical species. In this dissertation I used simulations to determine the dispersal levels at which spatial genetic structure would become panmictic as a result of complex population demographics in a continuously distributed species. Using simulations I could demonstrate the increase in population genetic structure as a result of size fluctuations given low dispersal distances. However, I was unable to show an overestimate of dispersal inferred from population genetic data, as a result of population size fluctuations. Rather, very low dispersal distances, less than 1% of the distribution of the species, resulted in population size fluctuations having no effect on spatial genetic structure versus that of simulations of constant population size. These results seem contradictory to the observed patterns in empirical studies of cyclical species. I believe the failure to show an overestimate in the inference of dispersal from cyclical species is the result of two factors not addressed in this dissertation.

Firstly, I only used Rousset's (2000) estimate of neighbourhood size to monitor the effect of population size fluctuations on spatial genetic pattern. This is the only appropriate statistic for continuously distributed species. However, observations of simulations conducted in this study suggest that neighbourhood size might not be a particularly good statistic to use to monitor the effects of population size fluctuations. Some simulations generated either large over-estimates or large under-estimates of neighbourhood size for some generations. This observation is not the result of an

incorrect implementation of the neighbourhood size calculation, since checking the procedure used in the simulations against calculations performed in Genepop (Raymond & Rousset 1995) using the same data, gave identical results. Leblois *et al.* (2003) have suggested that the incorrect scale of analysis might influence the accurate inference of neighbourhood size. This result was taken into the consideration for the calculation of neighbourhood size in the LatticeFlucIII code, where neighbourhood sizes were averaged over areas of 10σ (where σ is the mean parent-offspring axial dispersal distance) as suggested by Leblois *et al.* (2003). I am currently investigating the influence of outliers in the calculation of neighbourhood size and the suitability of neighbourhood size calculations for unstable populations. Furthermore, I have considered another statistic that may be useful to detect the effects of population size changes on spatial genetic structure in continuous populations. The statistic, A_{DI}, describes the geographic distribution of alleles summed across all loci, where

$$A_{DI} = \sum_{i=1}^{L} \sum_{j=1}^{N} \sum_{k=1}^{n-1} \sum_{l=2}^{n} \frac{D_{kl} S}{D_{kl}}$$

L = number of loci, N = number of alleles for locus i, n = number of individuals/samples, D_{kl} = geographic distance between individual k and l, S = allele sharing coefficient. The value of S is dependent on the degree of sharing for allele j between individual k and l. If both individuals are homozygous for allele j, $S = 1$. If one individual is homozygous for allele j and the other is heterozygous for allele j, $S = 0.75$. If both individuals are heterozygous for allele j then $S = 0.5$. Finally, if only one or neither individual has allele j, $S = 0$. In this way the average geographic distance over which alleles are shared are expressed as a proportion of the total geographic distance over which they could be shared. In a totally homozygous population that is fixed for a single allele, $A_{DI} = 1$, yet in a population where each individual is heterozygous and fixed for unique alleles, $A_{DI} = 0$.

The aforementioned statistic may be suitable for evaluating the effects of population size fluctuations on spatial genetic pattern in continuously distributed species. Preliminary simulation runs under a constant population size and low and high dispersal distances (Figure 1) indicated that the statistic is mostly influenced by allelic

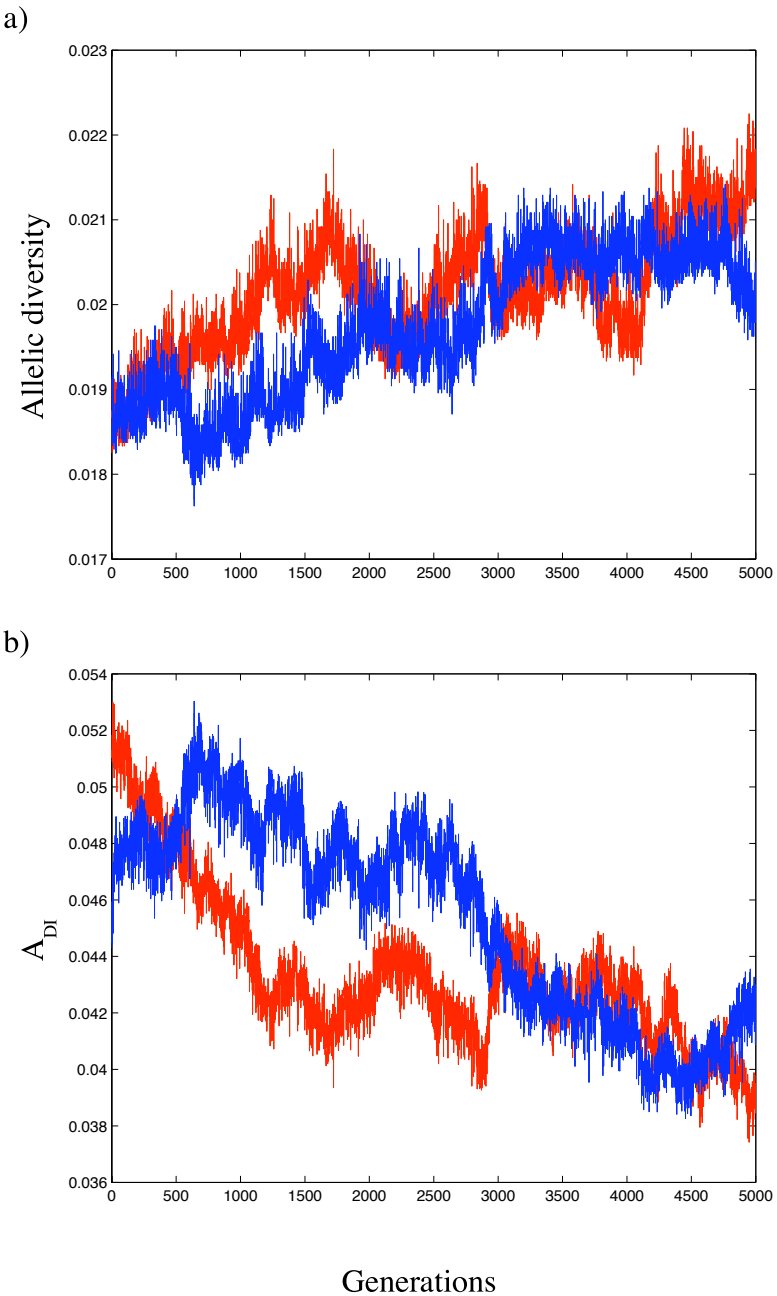

Figure 1: Per generation changes in a) allelic diversity and b) A_{DI} in a population of constant size (N = 125000) over 5000 generations with a microsatellite mutation rate of 2.5 x 10^{-4}. Red = high dispersal, blue = low dispersal.

diversity and not by dispersal distance. An increase in allelic diversity is closely tracked by a decrease in A_{DI}, and dispersal appears to have little effect (Figure 1). Since allelic diversity is typically reduced in populations that experience eruptions this statistic may be of some use. I am currently investigating the properties of this statistic and its utility in tracking the effects of population size fluctuations.

Secondly, the influence of the mutation rate, of the loci analysed, and its effect on inference of spatial genetic patterns from cyclical species has not been considered in this dissertation. The mutation rate of a locus has a direct influence on the observed levels of allelic diversity, and the potential for populations to recover from the effects of genetic drift when populations crash. Furthermore, given low allelic diversity the probability for the same alleles to become fixed as a result of genetic drift in different populations is greater than when allelic diversity is high. The effects of mutation rate of loci will be investigated with future simulations. Furthermore, the number of loci analysed is likely to determine the extent to which one can infer dispersal patterns in cyclical species. Since allelic diversity is generally reduced by population cycles, more loci may be required for such analysis, than in species with stable population dynamics.

From the point of view of understanding the population dynamics of *G. postica* in southern Africa several questions still need to be addressed. Apart from the theoretical considerations of population size fluctuations explored above a thorough understanding of the population biology of *G. postica* is required before recommendations regarding the sustainability of a Wild Silk Industry can be made. Firstly, the interaction between climatic factors and eruptions needs to be explored. In Chapter 2 I showed that eruptions, in general, were correlated with total rainfall. The steady decrease in total rainfall in the Kalahari over the past three years has resulted in substantially lower numbers of *G. postica* eruptions. However, the spatial heterogeneity in eruptions does not appear to be the result of spatial heterogeneity in rainfall, yet this should be explored further with suitable climatic modeling approaches. The interaction between *G. postica* larvae and their host plants is another aspect of the biology of this species that is not well understood. The larval requirements, in terms of foliage quantity and quality, for complete development and

pupation should be evaluated. The integration of this information with host-plant phenology, and the timing of adult moth emergence, is necessary to gain an understanding of the complex exogenous and endogenous factors that contribute to eruptions in this species. The work presented in this dissertation is the initiation of a population dynamics and genetics research programme that has the principle aim of constructing population models for this species. The results presented in this dissertation are paramount in the planning of future research, and in particular the planning of the scale at which to conduct population dynamics research. I look forward to contributing to the development of a long-term population dynamics programme which will attempt to entangle the complex population dynamics of this fascinating moth species.

References

Leblois R, Estoup A & Rousset F (2003) Influence of mutational and sampling factors on the estimation of demographic parameters in a “continous” population under isolation by distance. *Molecular Biology and Evolution*, **20**, 491-502.

Raymond M & Rousset (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248-249.

Rousset F (2000) Genetic differentiation between individuals. *Journal of Evolutionary Biology*, **13**, 58-62.