Evolutionary Bioinformatics

SHORT REPORT

Structurama: Bayesian Inference of Population Structure

John P. Huelsenbeck[1], Peter Andolfatto[2] and Edna T. Huelsenbeck[1]

[1]Department of Integrative Biology, University of California, Berkeley, CA 94720, USA. [2]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA. Corresponding author email: johnh@berkeley.edu

Abstract: Structurama is a program for inferring population structure. Specifically, the program calculates the posterior probability of assigning individuals to different populations. The program takes as input a file containing the allelic information at some number of loci sampled from a collection of individuals. After reading a data file into computer memory, Structurama uses a Gibbs algorithm to sample assignments of individuals to populations. The program implements four different models: The number of populations can be considered fixed or a random variable with a Dirichlet process prior; moreover, the genotypes of the individuals in the analysis can be considered to come from a single population (no admixture) or as coming from several different populations (admixture). The output is a file of partitions of individuals to populations that were sampled by the Markov chain Monte Carlo algorithm. The partitions are sampled in proportion to their posterior probabilities. The program implements a number of ways to summarize the sampled partitions, including calculation of the 'mean' partition—a partition of the individuals to populations that minimizes the squared distance to the sampled partitions.

Keywords: Bayesian estimaion, Dirichlet Process Prior, Markov chain Monte Carlo, population structure

Evolutionary Bioinformatics 2011:7 55–59

doi: 10.4137/EBO.S6761

This article is available from http://www.la-press.com.

Introduction

Natural populations of organisms often exhibit some degree of population subdivision. Identifying this population structure is important for several reasons. Practically speaking, undetected population structure can adversely affect statistical tests for the presence of natural selection[1] or of genetic association.[2] Population structure is also known to affect the evolutionary dynamics of alleles in populations; understanding patterns of population subdivision, then, is often a first step in learning about the evolutionary forces affecting a species.

Identifying population structure is a difficult problem that has motivated a variety of statistical and computational approaches. Here, we focus only on Bayesian methods for inferring population structure.[3–8] Pritchard et al[8] proposed a widely-used Bayesian method for inferring population structure. The simplest variant of the Pritchard et al[8] method assumes a fixed number of populations, K, and a Dirichlet prior probability distribution on the allele frequencies for each population. The method allows one to assign individuals to populations by calculating the posterior probability that an individual is assigned to each of the K populations.

Like many of the other methods that have been proposed, the Bayesian one proposed by Pritchard et al[8] assumes a fixed number of populations. Determining the correct number of populations for a particular set of observations is itself a difficult problem. With some reluctance, Pritchard et al[8] suggested determining the number of populations by approximating the marginal likelihoods of the data when the number of populations varies. In short, one performs repeated analyses with different numbers of populations; the number of populations that results in the maximum marginal likelihood for the data is chosen as the optimal value for the analysis. In a simulation study, Evanno et al[9] found that the method based upon marginal likelihoods performs poorly. (The poor performance may be related to the instability of the harmonic mean estimator of the marginal likelihood; see.[10])

More recently, Pella and Masuda[7] proposed a Bayesian method for determining population structure in which the number of populations is a random variable. Structurama implements the methods of Pritchard et al[8] and Pella and Masuda;[7] also see.[11] The program also implements a hierarchical variant of the Dirichlet Process prior model.[12] We use this model to account for admixture when the number of populations is considered a random variable. Structurama also implements a novel method for summarizing the results of a Bayesian analysis of population structure using the mean partition.

Approach

Pella and Masuda[7] assume that the assignment of individuals to populations and the number of populations follow a Dirichlet process prior.[13,14] Like Pritchard et al[8] Pella and Masuda[7] assume Hardy-Weinberg equilibirum of allele frequencies within a population, linkage equilibrium of the loci, and a Dirichlet prior probability distribution for the allele frequencies within a population. Their application of a Dirichlet process prior to the problem, however, is original. The Dirichlet process prior has been described extensively elsewhere,[13–15] and effective Markov chain Monte Carlo methods for sampling under this model have been described by Neal.[15] Here, we will provide an intuitive explanation of the Dirichlet process prior, which is sometimes referred to as the 'Chinese Restaurant Table Process'.[16,17] One imagines a (presumably very large) Chinese restaurant with a countably infinite number of tables. Patrons enter the restaurant one at a time (there are a total of n patrons that will enter the restuarant). The first patron enters and sits at some table (this with probability one). The number of occupied tables is now $K = 1$. The next patron can either sit at the same table as the first or sit at a new table. This patron sits at the same table as the first person with probability $1/(1 + \alpha)$ or at an unoccupied table with probability $\alpha/(1 + \alpha)$. If the patron sits at an unoccupied table, the number of occupied tables will increase by one, and $K = 2$. The process continues, with the kth patron that enters the restaurant sitting at table i, which is already occupied by η_i people, with probability $\eta_i/(k + \alpha)$ or at an unoccupied table with probability $\alpha/(k + \alpha)$.

Under the Dirichlet process prior, one can calculate the probability of a particular configuration of patrons at tables, and importantly, this probability does not depend upon the order in which the patrons enter

the restaurant. The joint probability of the assignment of individuals to tables and the number tables is

$$f(z,K \mid \alpha,n)=\alpha^{k}\frac{\prod_{i=1}^{K}(\eta_i-1)!}{\prod_{i=1}^{n}(\alpha+i-1)} \quad (1)$$

The parameter α determines the tendency of patrons to sit at the same table. If α is small, then patrons are more likely to sit at the same table. In fact, the probability that patron i and patron j find themselves sitting at the same table is

$$f(z_i=z_j)=\frac{1}{1+\alpha} \quad (2)$$

Finally, the probability that a total of K tables are occupied by patrons is

$$f(K|\alpha,n)=\frac{{}_{n}a_{K}\,\alpha^{K}}{\prod_{i=1}^{n}(\alpha+i-1)} \quad (3)$$

where ${}_{n}a_{K}$ is the absolute value of the Stirling number of the first kind.

In the context of determining population structure, populations are equivalent to the 'tables' of the Chinese restaurant example. Moreover, all of the individuals in a particular population share a common set of allele frequencies. These allele frequencies are drawn from a flat Dirichlet prior probability distribution. (It is unfortunate that 'Dirichlet' is used to name two very different probability distributions: the Dirichlet probability distribution on allele frequencies and the Dirichlet process prior describing how individuals are grouped into populations.)

We also implemented a hierarchical version of the Dirichlet process prior model[12] that allows us to accommodate admixture while treating the number of populations as a random variable. The hierarchical Dirichlet process prior has not been applied to the problem of assigning individuals to populations. Under the hierarchical Dirichlet process prior, there are n restaurants—one for each individual—and the alleles for the ith individual are only seated at tables in the ith restaurant. The tables are then assigned to different populations, which themselves have an independent DPP model. [Teh et al[12] describe this as the 'Chinese restaurant franchise', with the franchise allowing the sharing of data elements across groups.] An individual with no admixture would have the alleles assigned to only one table in its restaurant, whereas an admixed individual would have its alleles assigned to more than one restaurant table, and these tables would be assigned to multiple populations in the franchise.

We consider the assignment of individuals to populations to be a partition, where a partition is a division of a set into nonempty and disjoint sets which completely cover the set. Structurama implements Algorithm 3 of Neal[15] to sample partitions using a Gibbs sampling method when there is no admixture. We use a similar algorithm, described by Teh et al[12] for performing MCMC under the hierarchical Dirichlet procoess model (for a model with admixture). Partitions of individuals among populations are sampled in proportion to their posterior probabilities. The end result is a file with sampled partitions. Part of a sample of n = 10 individuals among populations might look like the following:

MCMC cycle	Individuals
	1 2 3 4 5 6 7 8 9 10
1	1 1 2 1 3 3 1 2 2 2
2	1 1 2 1 3 3 1 1 2 2
3	1 2 1 1 3 3 1 2 2 2
4	1 1 2 1 1 1 1 1 1 2
5	1 2 2 1 1 1 1 2 1 2

where partitions are labeled according to the restricted growth function notation of Stanton and White.[18] The first sample taken from the Markov chain has three populations, with individuals 1, 2, 4, and 7 grouped together into one population, individuals 3, 8, 9, and 10 grouped together into a second population, and individuals 5 and 6 grouped together into a third population.

Although the meaning of any single partition is unambiguous, it can be difficult to describe features in common among a set of partitions. How can one summarize features in common for a collection of partitions? One approach is simply to assign each population an index in computer memory. Instead of reporting the restricted growth function notation for a partition, one simply reports the index for each individual. The problem with this approach is that if the MCMC works properly, the labels should switch. That is, the MCMC algorithm

should visit the following two partitions equally often (they imply equivalent groupings of individuals into populations): (1,1,1,1,1,2,2,2,2,2) and (2,2,2,2,2,1,1,1,1,1). When the number of populations is fixed, this problem is more theoretical than practical because MCMC fails to visit equivalent labelings of the partitions. However, when the number of populations is a random variable, it is no longer suficient to use an arbitrary index for populations; the meaning of an index can change over the course of the Markov chain. Structurama summarizes the results on partitions using a number of methods. Perhaps the most notable is the use of the mean partition. We define the mean partition as the partition of individuals to populations which minimizes the sum of the squared distances to the sampled partitions. We use Gusfield's[19] distance on partitions. The partition distance is the minimum number of individuals that need to be removed from two partitions to make the induced partitions identical. Structurama also calculates the posterior probability of grouping each of the $\binom{n}{2}$ pairs of individuals together into the same population.

Program Details

Structurama takes as input a text file containing the allelic information for the sampled loci for each individual. The file format is a structured one, in the style of the Nexus format used by many phylogeny programs.[20] The following illustrates the file format for a study of impala.[21,22]

Note that this input file has $n = 216$ individuals each of which has $L = 8$ loci. The allele labels are arbitrary.

After the data has been read into computer memory (using the execute command), the user specifies the details of the model using the model command. Here the user has four choices:

model numpops=<number> admixture=no
model numpops=<number> admixture=yes
model numpops=rv admixture=no
model numpops=rv admixture=yes

The first two commands specify the models described by Pritchard et al.[8] The third model specifies the Dirichlet process prior model without admixture described by Pella and Masuda[7] and Huelsenbeck and Andolfatto.[11] The final model is unique to Structurama and specifies the hierarchical Dirichlet process prior model.[12] The user also can specify a hierarchical prior for the parameters of the Dirichlet process model.

Once the user has specified the model, the Markov chain Monte Carlo analysis is performed using the mcmc command. The MCMC algorithm samples partitions in proportion to their posterior probabilities. The sampled partitions are saved to a file in the restricted growth function notation for partitions.[18] The user can summarize the results of the MCMC analysis using one of several commands. The posterior probabilities of individuals being assigned to the same population are obtained using

```
#NEXUS
begin data;
dimensions nind=216 nloci=8;
info
ka117 (177,183) (122,124) (71,72) (61,61) (77,78) (54,62) (148,150) (105,107),
ka118 (181,185) (124,126) (72,72) (61,61) (77,78) (58,62) (146,146) (105,105),
ka119 (181,181) (126,128) (72,72) (60,61) (79,79) (57,62) (148,162) (105,105),
.
.
.
sa359 (179,187) (128,128) (73,73) (61,61) (80,80) (59,59) (140,140) (106,107),
sa360 (?, ?) (128,132) (72,73) (61,61) (?, ?) (57,60) (140,140) (108,108),
sa1077 (179,187) (?, ?) (72,73) (61,61) (80,81) (56,57) (140,146) (107,107)
;
end;
```

the showtogetherness command. The posterior probability distribution for the number of populations is obtained using the shownumpops command. Finally, the mean partition is obtained using the showmeanpart command.

Program Availability

Structurama is available for download from www.structurama.org.

Acknowledgements

The development of this program was supported by NSF and NIH grants DEB-0445453 and GM-069801, respectively, awarded to J.P.H. The authors would also like to thank Youn-Ho Lee and the Kordi South Sea Institute for their support during the development of this program.

Disclosure

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Nielsen R. Statistical tests of selective neutrality in the age of genomics. *Heredity*. 2001;86:641–7.
2. Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. *American Journal of Human Genetics*. 1995;57:455–64.
3. Corander JP, Waldmann MJ, Sillanpää. Bayesian analysis of genetic differentiation between populations. *Genetics*. 2003;163:367–74.
4. Corander JP, Waldmann P, Marttinen MJ, Sillanpää. BAPS2: Enhanced possiblities for the analysis of population structure. *Bioinformatics*. 2004;20:2363–9.
5. Falush DM, Stephens JK, Pritchard. Inference of population structure using multi-locus genotype data: Linked loci and correlated allele frequencies. *Genetics*. 2003;164:1567–87.
6. Holsinger KE, Lewis PO, Dey DK. A Bayesian approach to inferring population structure from dominant markers. *Molecular Ecology*. 2002;11:1157–64.
7. Pella J, Masuda M. The Gibbs and split-merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences*. 2006;63:576–96.
8. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155:945–59.
9. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software Structure: a simulation study. *Molecular Ecology*. 2005;14:2611–20.
10. Raftery AE, Newton MA, Satagopan JM, Krivitsky PN. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*. 2007;8:1–45.
11. Huelsenbeck JP, Andolfatto P. Inference of population structure under a Dirichlet process model. *Genetics*. 2007;175:1787–802.
12. Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*. 2006;101:1566–581.
13. Antoniak CE. Mixtures of Dirichlet processes with applications to non-parametric problems. *Annals of Statistics*. 1974;2:1152–74.
14. Ferguson TS. A Bayesian analysis of some nonparametric problems. *Annals of Statistic*s. 1973;1:209–30.
15. Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*. 2000;9:249–65.
16. Aldous D. Exchangeability ad related topics. In École d'Eté de Probabilités de Saint-Flour XIII-1983, Springer, Berlin; 1985:1–198.
17. Pitman J. *Combinatorial stochastic processes*. Technical Report 621, University of California at Berkeley, lecture notes for St. Flour Summer School; 2002.
18. Stanton D, White D. Constructive Combinatorics. Springer-Verlag, New York, 1986.
19. Gusfield D. Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters*. 2002;82:159–64.
20. Maddison DR, Swofford DL, Maddison WP. NEXUS: an extensible file format for systematic information. *Systematic Biology*. 1997;46:590–621.
21. Lorenzen ED, Arctander P, Siegismund HR. Regional genetic structuring and evolutionary history of the impala Aepyceros melampus. *Journal of Heredity*. 2006;97:119–32.
22. Lorenzen ED, Siegismund HR. No suggestion of hybridization between the vulnerable black-faced impala (Aepyceros melampus petersi) and the common impala (A.m. melampus) in Etosha NP, Namibia. *Molecular Ecology*. 2004;13:3007–19.