BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 9 2000
Pages 847–848

LIAN 3.0: detecting linkage disequilibrium in multilocus data

Bernhard Haubold[1,*] and Richard R. Hudson[2]

[1]Max-Planck-Institut für Chemische Ökologie, Carl-Zeiss-Promenade 10, D-07745 Jena, Germany and [2]Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St., Chicago, IL 60637, USA

Received on January 11, 2000; revised on March 13, 2000; accepted on May 12, 2000

Abstract

Summary: *LIAN is a program to test the null hypothesis of linkage equilibrium for multilocus data. LIAN incorporates both a Monte Carlo method as well as a novel algebraic method to carry out the hypothesis test. The program further returns the genetic diversity of the sample and the pairwise distances between its members.*

Availability: *LIAN can be accessed at http://soft.ice.mpg.de/lian*

Contact: *haubold@ice.mpg.de*

Introduction

Linkage equilibrium is characterized by statistical independence of alleles at all loci. LIAN (from **LI**nkage **AN**alysis) tests for this independent assortment by first computing the number of loci at which each pair of haplotypes differs. From the distribution of mismatch values a variance, V_D, is calculated, which is compared to the variance expected for linkage equilibrium, V_e. This approach was poineered by Brown *et al.* (1980), who applied it to wild barley (*Hordeum spontaneum*), but the method can be applied to any set of haplotype data. The null hypothesis $H_0 : V_D = V_e$ is tested by Monte Carlo simulation and by a new parametric test (Haubold *et al.*, 1998). LIAN is written in standard Fortran90 and is accessible through a web interface.

Testing the null hypothesis of linkage equilibrium

Monte Carlo simulation

The input data set is scrambled by resampling the loci without replacement and computing a V_D-value for each resampled data set. The significance of any difference between V_D and V_e is the frequency with which a V_D-value greater or equal to the original V_D-value is returned from the randomization procedure. Similarly, a 5% critical value can be calculated by ordering the V_D-values and removing the top 5% from the distribution (one-tailed test; Souza *et al.*, 1992). The simulation approach may be slow for large data sets.

Parametric method

Instead of simulating the null distribution of V_D by the Monte Carlo method outlined above, one can assume that this distribution is normal (Haubold *et al.*, 1998). Then a 5% critical value, L, for V_D is simply computed as

$$L = V_e + 1.654\sqrt{\mathrm{Var}(V_D)},$$

where $\mathrm{Var}(V_D)$ is the variance of V_D. A formula for $\mathrm{Var}(V_D)$ was given by Brown *et al.* (1980) and popularized by Maynard Smith *et al.* (1993). Haubold *et al.* (1998) showed that this formula was incorrect, leading to faulty conclusions about population structure. A correct formula for $\mathrm{Var}(V_D)$ was derived by Haubold *et al.* (1998) and is now implemented in LIAN.

Input and output files

Input file

LIAN takes haplotype data as input. This might have been obtained by one of a number of experimental methods, including allozyme, microsatellite and DNA sequencing. The input file essentially follows the PHYLIP format (Felsenstein, 1993). The first line in a plain text file states the number of haplotypes, followed by the number of loci. Each line of the subsequent data matrix looks like this:

```
name 1 3 2.5 4
```

that is, a name, of which the first 10 characters are stored and which must not contain blanks, followed by numerical allele designations separated by spaces. An example data file can be obtained together with the program. There are no intrinsic limitations on the number of haplotypes, loci or Monte Carlo iterations the program can handle.

Results output file

The main output file quotes V_D- and V_e-values as well as a '*standardized* I_A'. The latter is a measure of linkage.

*To whom correspondence should be addressed.

Table 1. Analysis of the multilocus data set published by Holmes *et al.* (1999) using LIAN. The simulation results were obtained from 100 000 random resamplings

Quantity	Value
Haplotypes	30
Loci	12
Mean genetic diversity (H)	0.941 ± 0.009
Observed mismatch variance (V_D)	1.085
Expected mismatch variance (V_e)	0.659
Standardized index of association (I_A^s)	0.059
Simulated 5% critical value (L_{MC})	0.753
Calculated 5% critical value (L_{para})	0.730
Simulated significance (P_{MC})	0.00001
Calculated significance (P_{para})	0.00000

The 'classical' I_A was defined by Maynard Smith *et al.* (1993) as

$$I_A = \frac{V_D}{V_e} - 1.$$

The I_A is a function of the rate of recombination and is zero for linkage equilibrium. However, Hudson (1994) showed that this statistic scales with $l - 1$, where l is the number of loci analyzed. For this reason we define the '*standardized* I_A' as

$$I_A^s = \frac{1}{l-1}\left(\frac{V_D}{V_e} - 1\right).$$

This has the advantage of being comparable between studies as long as it can be assumed that the neutral mutation parameter $\theta = 2N_e\mu$ is constant (Hudson, 1994).

The measure of disequilibrium is followed by testing the null hypothesis of linkage equilibrium. P-values derived from the parametric as well as from the Monte Carlo method are quoted. In addition, the program returns the 5% critical value as determined by the Monte Carlo process (L_{MC}) and the 5% critical value calculated from the parametric approach (L_{para}). Finally, the mean genetic diversity $\pm$ standard error is printed, followed by the genetic diversity at each locus going from left to right along the columns of the data matrix.

Distance file

LIAN also generates pairwise distances between the sampled taxa. There is a choice between applying the *matching*, *dice*, *ochiai* or *jaccard's* index (cf. Manly, 1986, p. 86). The resulting distance file can be imported directly into one of the distance programs of the PHYLIP package Felsenstein (1993).

Application example

Maynard Smith *et al.* (1993) showed in a simulation study that if recombination is at least 20 times more frequent than mutation, most populations will be indistinguishable from linkage equilibrium. Hudson (1994) corroborated this result in a theoretical investigation. There are only a few estimates of the relative recombination rate from real data available, but Feil *et al.* (1999) reported that in the human pathogen *Neisseria meningitidis* recombination affects a locus 3.6 times more frequently than mutation. Since $3.6 < 20$, it would be expected that *N. meningitidis* is in linkage disequilibrium. However, based on the formula for $\mathrm{Var}(V_D)$ introduced into population biology by Brown *et al.* (1980), Holmes *et al.* (1999) claimed that this organism was in linkage equilibrium. Reanalysis of the data presented by Holmes *et al.* (1999) using LIAN showed that *N. meningitidis* has a highly significant linkage disequilibrium ($P < 10^{-4}$; Table 1). This conclusion fits with what is currently known about the population genetics of bacteria in general and of *N. meningitidis* in particular.

Acknowledgements

We thank Tim Anderson for testing an earlier version of LIAN and Steffi Gebauer-Jung for help with programming LIAN's web interface. We are also grateful to Thomas Wiehe for useful comments on the manuscript. B.H. thanks the Max-Planck-Society for financial support.

References

Brown,A.H.D., Feldman,M.W. and Nevo,E. (1980) Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics*, **96**, 523–536.

Felsenstein,J. (1993) PHYLIP (phylogeny inference package); version 3.4c: distributed by the author, University of Washington, Seattle.

Feil,J.E., Maiden,M.C.J., Achtman,M. and Spratt,B.G. (1999) The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol. Biol. Evol.*, **16**, 1496–1502.

Haubold,B., Travisano,M., Rainey,P.B. and Hudson,R.R. (1998) Detecting linkage disequilibrium in bacterial populations. *Genetics*, **150**, 1341–1348.

Holmes,E.C., Urwin,R. and Maiden,M.C.J. (1999) The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Mol. Biol. Evol.*, **16**, 741–749.

Hudson,R.R. (1994) Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation. *J. Evol. Biol.*, **7**, 535–548.

Manly,B.F.J. (1986) *Multivariate Statistical Methods; A primer*. Chapman and Hall, London.

Maynard Smith,J., Smith,N.H., Dowson,C.G. and Spratt,B.G. (1993) How clonal are bacteria? *Proc. Natl. Acad. Sci. USA*, **90**, 4384–4388.

Souza,V., Nguyen,T.T., Hudson,R.R., Piñero,D. and Lenski,R.E. (1992) Hierarchical analysis of linkage disequilibrium in *Rhizobium* populations: evidence for sex? *Proc. Natl. Acad. Sci. USA*, **89**, 8389–8393.