CABIOS

Vol. 11 no 6 1995
Pages 621–625

DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data

J.Rozas[1] and R.Rozas

Abstract

DnaSP, DNA sequence polymorphism, is an interactive computer program for the analysis of DNA polymorphism from nucleotide sequence data. The program, addressed to molecular population geneticists, calculates several measures of DNA sequence variation within and between populations, linkage disequilibrium parameters and Tajima's D statistic. The program, which is written in Visual Basic v. 3.0 and runs on an IBM-compatible PC under Windows, can handle a large number of sequences of up to thousands of nucleotides each.

Introduction

Population genetics is a branch of evolutionary biology that tries to determine the level and distribution of genetic polymorphism in natural populations and also to detect the evolutionary forces that could determine the pattern of genetic variation observed in natural populations. Ideally, the best way to quantify genetic variation in natural populations is by comparison of DNA sequences (Kreitman, 1983). However, until 1990 the use of DNA sequence data had had little impact on population genetics. This is because the effort (in terms of both money and time) required to obtain DNA sequence data from a relatively large number of alleles was substantial.

The introduction of the polymerase chain reaction (PCR) (Saiki *et al.*, 1985, 1988), which allows direct sequencing of PCR products and avoids, therefore, their cloning, has changed the situation. Undoubtedly this has produced a revolutionary change in population genetics. Although, at present, population studies at the DNA sequence level are still scarce and primarily carried out in *Drosophila* (e.g. McDonald and Kreitman, 1991; Schaeffer and Miller, 1992; Rozas and Aguadé, 1994), they will certainly increase in the future.

Here we present a program, DnaSP, addressed to molecular population geneticists, that uses this new molecular data. The program computes several measures of DNA sequence variation within and between populations, linkage disiquilibrium parameters and Tajima's *D* statistic. DnaSP takes advantage of Microsoft Windows capabilities, so that it can handle a large number of sequences of thousands of nucleotides each on a microcomputer. Furthemore, DnaSP can easily exchange data with other programs, e.g. programs for performing multiple sequence alignments, phylogenetic tree analysis or statistical analysis. Additionally, the program provides a standard Microsoft Windows Help file.

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Diagonal 645, 08071 Barcelona, Spain

[1]*To whom correspondence should be addressed. Email julio@porthos-bio.ub.es*

System and methods

The program DnaSP is written in Visual Basic v. 3.0 (Microsoft) and runs on an IBM-compatible PC under Windows. The program has been developed and tested on a microcomputer with a 66 MHz Intel 80486 DX2 processor (with maths co-processor) with 8 Mbyte of main memory and with MS-DOS operating system v. 6.2 running under Microsoft Windows v. 3.1. The minimum hardware requirements for the program are an Intel 80386 processor, 2 Mbytes of RAM memory, a mouse and a hard disk. DnaSP also requires the MS-DOS operating system (version 3.2 or later) and Microsoft Windows (version 3.0 or later). For large data sets an 80486 processor and 4 Mbytes of RAM are highly recommended.

Algorithm and implementation

DnaSP user interface

DnaSP has a standard Microsoft Windows user interface, including the menu bar, pull-down menus, dialog boxes and windows with scroll bars (Figure 1). The DnaSP menu bar displays five pull-down menu titles: File, Analysis, Display, Window and Help. Each menu contains a set of related commands, which are displayed when the menu is pulled down.

Input and output

The number and length of the sequences that can be handled by DnaSP depends mainly on the available memory. However, DnaSP is able to use all RAM memory available in a computer, both conventional and extended memory. DnaSP can also use virtual memory (it can use the hard disk space as memory, although in this case the

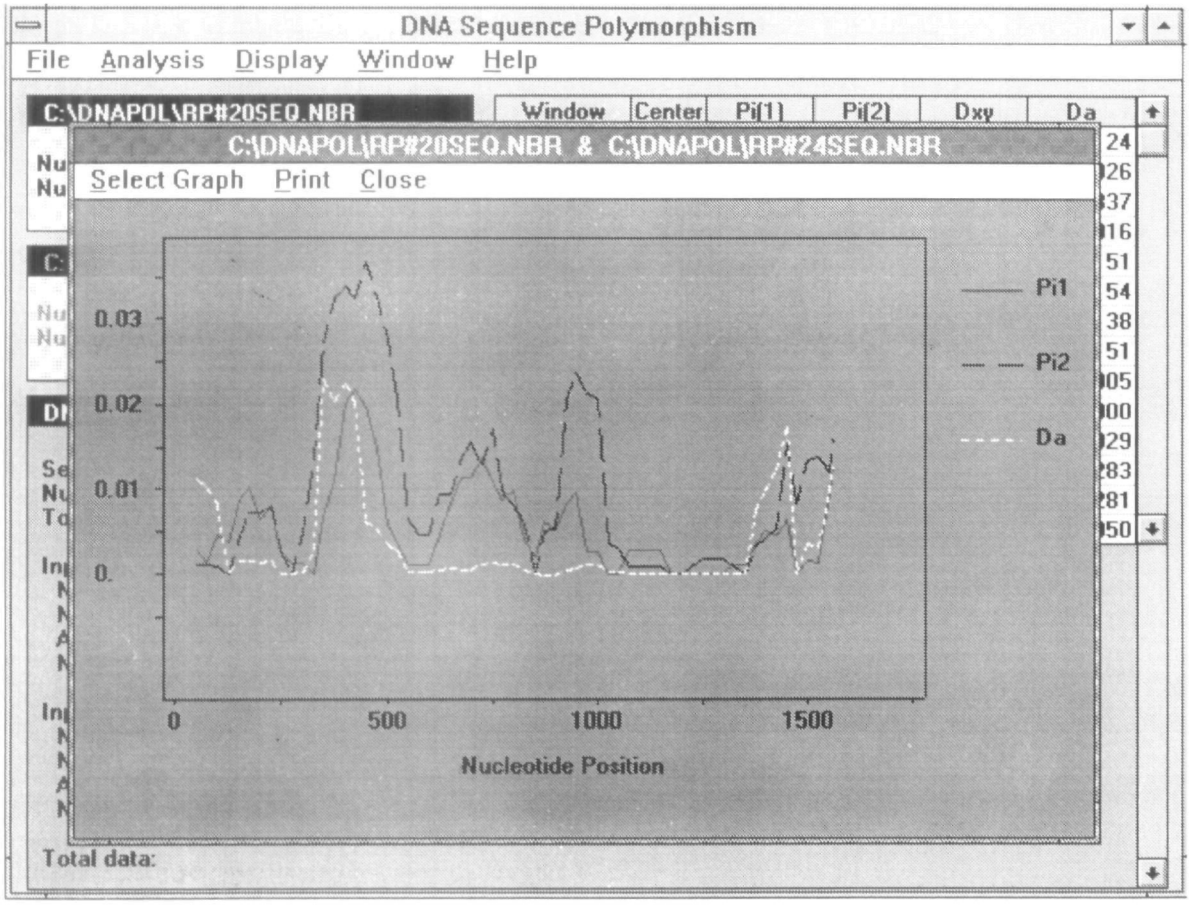

Fig. 1. The DnaSP user interface Example of a graph produced by DnaSP in the analysis of DNA divergence between populations

computation speed will be much lower than when using RAM). Thus the program can handle large numbers of sequences of up to thousands of nucleotides each. For example, a huge input data file (160 sequences of 7620 nucleotides each; i.e. 1 219 200 nucleotides) has been analysed in a machine with an 80486 processor.

DnaSP can automatically read two types of formats: NBRF/PIR (Barker *et al.*, 1991) and MEGA (non-interleaved format) (Kumar *et al.*, 1994). In both cases two or more homologous and aligned (i.e. the sequences must have the same length) nucleotide sequences should be included in one (ASCII) file. The nucleotide data can be displayed in a window (View Data command in the Display menu).

To date there are several available programs for performing multiple sequence alignments. For this reason we have not included any sequence alignment option in DnaSP. Sequence alignments can be performed, for example, using CLUSTAL V (Higgins *et al.*, 1992), which can produce an output (NBRF/PIR format) that can be used by DnaSP as input.

The NBRF/PIR and MEGA formats can also be converted by DnaSP; this feature can be useful to perform additional evolutionary analysis (estimating evolutionary distances, reconstructing phylogenetic trees, etc.) using other programs, e.g. MacClade (Maddison and Maddison, 1992), MEGA (Kumar *et al.*, 1994), PAUP (Swofford, 1991) and PHYLIP (Felsenstein, 1993).

The output can be displayed in three kinds of windows: text, grid (the output data are laid out in rows and columns like in a spreadsheet application) and graphic (scatter graph and line chart). All commands produce an output text window at the bottom of the screen, moreover, some of them also produce a grid. The data in the grid can be used to create a graph (Graphs command in the Display menu). The data generated from DnaSP can be saved as an ASCII text file. The grid output data file can easily be used in other applications, such as spreadsheet, statistical or graphics applications, by simply removing the header.

Analysis menu

This menu has the following five commands: Polymorphic Sites, DNA Polymorphism, DNA Divergence Between Populations, Linkage Disequilibrium and Tajima's Test.

1. Polymorphic Sites. This command displays some

```
DNA Polymorphism
----------------

Input Data File: C:\DNAPOL\RP#44SEQ.NBR
Number of sequences: 44
Selected region, 1-1599     Number of sites: 1599
Number of polymorphic sites: 104
Total sites (excluding sites with alignment gaps). 1437

Nucleotide diversity, Pi: 0 00973
   Sampling variance of Pi: 0.00000
      Standard deviation of Pi: 0 00052

Theta (per nucleotide)  0 01664
   Variance of theta (no recombination)  0.00002
      Standard deviation of theta (no recombination)  0 00493
   Variance of theta (free recombination)  0 00000
      Standard deviation of theta (free recombination): 0.00163

Average number of nucleotide differences, k: 13.983
   Stochastic variance of k (no recombination), Vst(k): 39 024
   Sampling variance of k (no recombination), Vs(k)  1 870
      Total variance of k (no recombination), V(k): 40 894
   Stochastic variance of k (free recombination), Vst(k): 4 661
   Sampling variance of k (free recombination), Vs(k): 0 217
      Total variance of k (free recombination), V(k)  4.878

M = 4Nv (per sequence). 23.908
   Variance of M (no recombination). 50.189
   Variance of M (free recombination)  5.496
```

Fig. 2. Example of the DNA polymorphism output file (partial) for 44 sequences (1599 nucleotides each) of the *rp49* gene region of *Drosophila subobscura* (Rozas and Aguadé, 1993, 1994).

general information about the polymorphisms in the data file: the number of sites with alignment gaps, the number of monomorphic sites and the number of polymorphic sites segregating for two, three or four nucleotides. For sites segregating for two nucleotides the total number of parsimony-informative sites is also indicated.

2. DNA polymorphism. This command computes several measures of the extent of DNA polymorphism and their variances (Tajima, 1983, equation A3; Nei, 1987, equations 10.3, 10.5 and 10.7; Tajima, 1993, equations 3, 4, 8 and 13–18) (Figure 2).

For the analysis sites with alignment gaps are not used (these sites are completely excluded).

DnaSP can also calculate the nucleotide diversity π (Nei, 1987, equation 10.5) by the sliding window method. In this method a window (segment of DNA) is moved along the sequences in steps. The nucleotide diversity is calculated in each window and the value is assigned to the nucleotide at the mid-point of the window. Both the window length and the step size default values can be

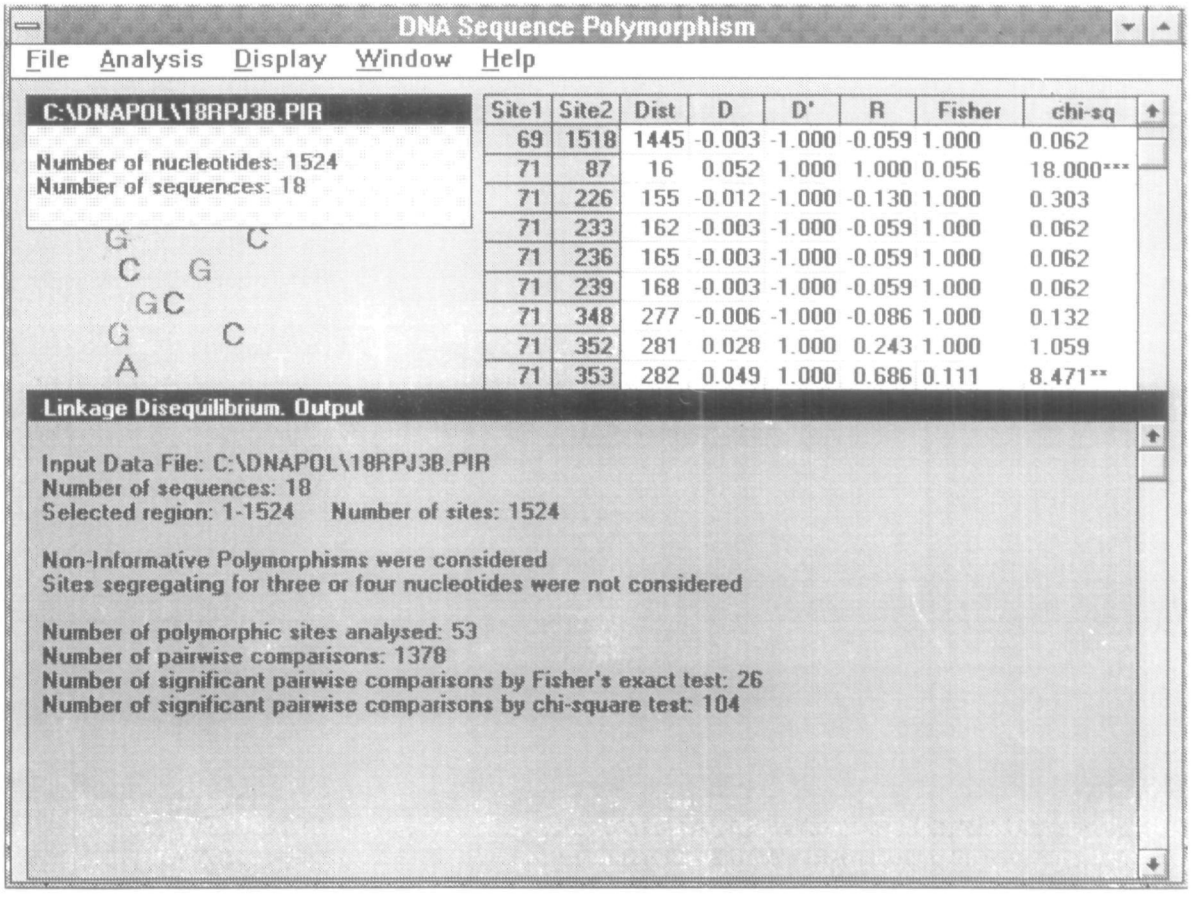

Fig. 3. Example of the linkage disequilibrium output produced by DnaSP in the analysis of 18 sequences of the *rp49* gene region (O_{3+4} chromosomes) of *Drosophila subobscura* (Rozas and Agvadé, 1994).

changed by the user. This option can also be used to analyse nucleotide diversity in non-overlapping windows.

The output of the sliding window analysis is given in a grid. The results can also be presented graphically (as a line chart).

3. DNA Divergence Between Populations. This command computes some measures of the extent of DNA divergence between populations, taking into account the effect of DNA polymorphism (d_{XY} and d_A; Nei, 1987, equations 10.20 and 10.21 respectively).

Two data files (one for each population) are needed for the analysis. Both data files must contain aligned sequences all of equal length. Sites containing alignment gaps in any population are completely excluded.

The program can also compute the nucleotide diversity for populations 1, 2, d_{XY} and d_A by the sliding window method. The output of the analysis is given in a grid. The results can also be presented graphically (as a line chart).

4. Linkage disequilibrium. This command calculates the degree of association between nucleotide variants at different polymorphic sites (Figure 3). Sites containing alignment gaps or polymorphic sites segregating for three or four nucleotides are completely excluded from the analysis. The analysis can be performed with all polymorphic sites in the data or with only parsimony-informative sites (sites that segregate for only two nucleotides that are present at least twice). The degree of linkage disequilibrium (or non-random association between variants of different polymorphic sites) is estimated by the following parameters: *D* (Lewontin and Kojima, 1960), *D'* (Lewontin, 1964) and *R* (Hill and Robertson, 1968). For the analysis we consider as coupling gametes those with the most or the least common variants (Langley *et al.*, 1974). Both Fisher's two-tailed exact test and the χ^2 test are computed to determine whether the associations between polymorphic sites are or are not significant.

DnaSP also computes the nucleotide distance (in base pairs) between a pair of polymorphic sites. This physical distance is calculated as the average number of nucleotides that separate two particular polymorphic sites (Schaeffer and Miller, 1993).

The output of the analysis is given as a grid. The results can also be presented graphically (as a scatter graph). In the graph *D*, *D'*, *R* and R^2 can be plotted against the nucleotide distance (*x* axis).

5. Tajima's Test. This command calculates the *D* statistic proposed by Tajima (1989) to test the neutrality hypothesis (Kimura, 1983). The confidence limits of *D* can be obtained from Table 2 of Tajima (1989).

Availability

Copies of DnaSP can be obtained freely for academic use. The program and the documentation are available by Email from the EMBL file server (for instructions send an Email with the words HELP and HELP SOFTWARE in the body of the message to netserv@ebi.ac.uk) and by anonymous ftp to: ftp.ebi.ac.uk (directory /pub/software/dos). Queries and suggestions may be addressed via Email to julio@porthos.bio.ub.es.

Acknowledgements

We would like to thank M.Aguadé for critical reading of this manuscript We would also like to thank the numerous people who tested the program for their comments and suggestions This work was supported by a grant from the Dirección General de Investigación Científica y Técnica, Spain (PB88-0196 and PB91-0245) to M Aguadé.

References

Felsenstein,J. (1993) Phylogeny inference package (PHYLIP), version 3 5. University of Washington, Seattle, WA.

Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) CLUSTAL V: improved software for multiple sequence alignment. *Comput. Applic. Biosci*, **8**, 189–191.

Hill,W.G. and Robertson,A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, **38**, 226–231.

Kimura,M. (1983) *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, MA.

Kreitman,M. (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster. Nature*, **304**, 412–417.

Kumar,S , Tamura,K. and Nei,M (1994) MEGA· molecular evolutionary genetics analysis software for microcomputers. *Comput Applic Biosci.*, **10**, 189–191.

Langley,C.H., Tobari,Y.N. and Kojima,K. (1974) Linkage disequilibrium in natural populations of *Drosophila melanogaster Genetics*, **78**, 921–936.

Lewontin,R.C. and Kojima,K. (1960) The evolutionary dynamics of complex polymorphisms. *Evolution*, **14**, 458–472.

Lewontin,R C. (1964) The interaction of selection and linkage. I General considerations· heterotic models. *Genetics*, **49**, 49–67.

Maddison,W.P. and Maddison,D R. (1992) MacClade: analysis of phylogeny and character evolution, version 3. Sinauer Associates, Sunderland, MA.

McDonald,J.H. and Kreitman,M. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila Nature*, **351**, 652–654.

Nei,M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York, NY.

Rozas,J. and Aguadé,M. (1993) Transfer of genetic information in the *rp49* region of *Drosophila subobscura* between different chromosomal gene arrangements *Proc. Natl Acad Sci. USA*, **90**, 8083–8087

Rozas,J. and Aguadé,M (1994) Gene conversion is involved in the transfer of genetic information between naturally occurring inversions of *Drosophila. Proc Natl Acad. Sci. USA*, **91**, 11517–11521.

Saiki,R.K., Scharf,S., Faloona,F., Mullis,K.B., Horn,G.T., Erlich,H.A. and Arnheim,N. (1985) Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, **230**, 1350–1354

Saiki,R K., Gelfand,D H , Stoffel,S , Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K B and Erlich,H.A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase *Science*, **239**, 487–491.

Schaeffer,S.W. and Miller,E L. (1993) Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura Genetics*, **135**, 541–552

Sidman,K.E. George,D.G., Barker,W.C. and Hunt,L T (1988) The protein identification resource (PIR). *Nucleic Acids Res.*, **16**, 1869–1871.

Swofford,D.L. (1991) PAUP: phylogenetic analysis using parsimony, version 3 0. Illinois Natural History Survey, Champaign, IL

Tajima,F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595

Tajima,F. (1993) Measurement of DNA polymorphism. In Takahata,N and Clark,A.G. (eds), *Mechanisms of Molecular Evolution*. Sinauer Associates, Sunderland, MA, pp. 37–59.

Received on May 5, 1995, accepted on September 7, 1995