Genetic diversity analysis with molecular marker data: Learning module

Software programs for analysing genetic diversity

Contents

Main characteristics

- Similar tasks
- Key differences
  - User interface
  - Type of data input and output
  - Platform
  - Capabilities
- Choice based on individual preferences

Numerous software programs are available for assessing genetic diversity. Most are freely available through Internet. Many perform similar tasks, with the main differences being in the user interface, type of data input and output, and platform. Thus, choosing which to use depends heavily on individual preferences. In this section, we describe some of the programs available, noting specific options that users may find preferable.

A summary table

Feature	Program					
	TFPGA[1]	Arlequin[1]*	GDA[1]	GENEPOP[1]*	GeneStrut	POPGENE*+
Diversity						
Heterozygosity (observed)	x	x	x		x	x
Expected heterozygosity	x	x	x		x	x
No alleles per locus		x	x		x	x
Effective alleles (no.)		x			x	x
Percentage polymorphic loci	x	x	x			x
Shannon-Weaver						
Population structure						
F statistics	x	x	x	x	x	x
G-statistics					x	x
ANOVA		x	x			
Rho-statistics		x		x		
Homogeneity	x			x		x
Migration		x		x		x
Isolation-by-distance				x		
Equilibrium						
Hardy-Weinberg	x	x	x	x	x	x
Two-locus		x	x	x	x	x
Multilocus			x			
U-test				x		
Genetic distance						
Nei's	x	x	x		x	x
Rogers'	x				x	
Pairwise Fsr	x	x	x			
Clustering						
Neighbour-joining			x			
UPGMA	x		x		x	x
Neutrality test		x				x

[1] Performs exact tests for significance
* Program can accommodate a null allele in the data
+User can specify an inbreeding coefficient to estimate the frequency of a null allele

After Labate (2000)

Joanne Labate (2000) wrote an excellent review of six programs: TFPGA (Miller, 1997), Arlequin (Schneider et al., 1997), GDA (Lewis and Zaykin, 1999), GENEPOP (Raymond and Rousset, 1995), GeneStrut (Constantine et al., 1994), and POPGENE (Yeh et al., 1997). Her review includes the particular options of each program, a table of functions available in each, and Web sites where they can be downloaded. To avoid redundancy, we have included only the Arlequin, which is possibly the most widely used program of the six.

For full references to these six programs and selected others, see Appendix 9.

Reference

Labate, J.A. 2000. Software for population genetic analyses of molecular marker data. Crop Sci. 40:1521-1528.

Some software programs, their authors, Web sites, …

Name	Author	Available from:
Arlequin[a]	Laurent Excoffier	http://lgb.unige.ch/arlequin
DnaSP	Julio and Ricardo Rozas	http://www.ub.es/dnasp
PowerMarker	Kejun (Jack) Liu	http://www.powermarker.net/
MEGA2	S. Kumar and others	http:www.megasoftware.net
PAUP*	David Swofford	http://paup.csit.fsu.edu/
TFPGA[a]	Mark Miller	http://bioweb.usu.edu/mpmbio/index.htm
GDA[a]	Paul Lewis, Dmitri Zaykin	http://lewis.eeb.uconn.edu/lewishome/software.html
GENEPOP[a]	Michel Raymond, Francois Rousset	ftp://ftp.cefe.cnrs-mop.fr/pub/PC/MSDOS/GENEPOP/Genepop.zip also at <http://wbiomed.curtin.edu.au/genepop/>
NTSYSpc	F.J. Rohlf	http://www.exetersoftware.com/cat/ntsyspc/ntsyspc.html
structure	Jonathan K. Pritchard	http://pritch.bsd.uchicago.edu/
GeneStrut[a]	Constantine, Hobbs & Lymbery	http://wwwvet.murdoch.edu.au/vetschl/imgad/GenStrut.htm
POPGENE[a]	F.C. Yeh, R.-C. Yang, T. Boyle	http://www.ualberta.ca/~fyeh/index.htm
MacClade	David R. & Wayne P. Maddison	http://phylogeny.arizona.edu/macclade
PHYLIP	Joe Felsentein	http://evolution.genetics.washington.edu/phylip.html
SITES	Jody Hey	http://lifesci.rutgers.edu/~heylab/ProgramsandData/Programs/SITES/SITES_Documentation.htm#Contents
CLUSTAL W	Thompson, Higgins & Gibson	http://www.ebi.ac.uk/clustalw
MALIGN	D. Janies and W.C. Wheeler	http://research.amnh.org/users/djanies/

[a] Discussed in Labate (2000)

We review other programs, selecting them for their wide use and giving priority to those that are available for no charge (except for PAUP*). Web sites listing and linking many other available programs are also given in this and following slides. We include information on authors, costs, platform specificities and Web sites. Note that, while these programs are sometimes made specifically for only one platform (usually Windows or Macintosh), with the recent advent of 'emulators' (such as SoftWindows, VirtualPC), most programs can be run on any computer, regardless of platform. Although we note the cases where a program has been successfully used with one of these emulators, we do not say that all listed programs can be used across platforms; simply that we know for sure that these have been successfully used. Sometimes, using emulators can cause the program to run more slowly or create other problems. Where possible, prefer using the platform for which the program was designed.

Programs listed in **bold face** in the slide are discussed in this paper. Web sites were available as of 28 February, 2003.

… and other features

Name	Platform			Cost (US$)
	Windows	Macintosh	Other	
Arlequin	X	X	X	0
DnaSP	X	c		0
PowerMarker	X			0
MEGA2	X	c	c	0
Arlequin	X	X	X	100
TFPGA	X			0
GDA	X		X	0
GENEPOP			X	0
NTSYSpc	X			230-300
structure	X		X	0
GeneStrut		X		0
POPGENE	X	c		0
MacClade		X		125
PHYLIP	X	X	X	0
SITES	X	X		0
CLUSTAL W	X	X	X	0
MALIGN			X	0

c = program runs well with an emulator such as SoftWindows or VirtualPC.

To save space, references to the programs discussed are given in Appendix 9.

Five software programs in detail

- Arlequin
- PowerMarker
- DnaSP
- PAUP*
- MEGA

Five software programs were selected to show detail. The choice was made after informally surveying users on the programs they use most and their opinions as to the most useful or representative. Users included graduate students, postdoctorates and research associates, as well as faculty. A list was compiled of the most-mentioned programs. In cases of doubt, those that were freely available or seemed more widely used were chosen. To be more representative, an additional criterion was to choose those programs that did different things.

Arlequin

- <http://lgb.unige.ch/arlequin>
- Windows, Mac, and Linux versions available, all free of charge
- Can use different types of data (but not yet dominant markers)
- Missing or ambiguous data can be included
- Data can be entered as DNA sequences, RFLP haplotypes, microsatellite profiles or multilocus haplotypes
- Data can be imported from files created for other programs

(continued on next slide)

Released in 1997, Arlequin (current version 2.001) is still very popular. It is 'an exploratory population genetics software environment able to handle large samples of molecular data (RFLPs, DNA sequences, microsatellites), while retaining the capacity of analysing conventional genetic data (standard multi-locus data or mere allele frequency data).'

Arlequin can use many different types of data, such as molecular data and genotype or haplotype frequencies, including codominant or recessive data but not yet dominant data. Molecular data can be entered as DNA sequences, RFLP haplotypes, microsatellite profiles, or multilocus haplotypes. The data format is specified in an input file. The user can create a data file from scratch, using a text editor and appropriate keywords, or use the 'Project Outline Wizard'. Data can be imported from files created for other programs, including MEGA, BIOSYS, GENEPOP, and PHYLIP. Missing or ambiguous data can be included. A very detailed user manual is available, which includes a large amount of theoretical information, formulae, and references. A large number of data can be analysed, and a Batch Files option is available.

Authors: Laurent Excoffier, Stefan Schneider and David Roessli, University of Geneva, Switzerland.

Reference

Schneider, S., D. Roessli and L. Excoffier. 2000. Arlequin: A Software for Population Genetics Data Analysis, Version 2.000. Genetics and Biometry Laboratory, Dept. of Anthropology, University of Geneva, Switzerland.

Arlequin (continued)

- Advantages:
  - Good support provided through a detailed manual, which includes large amounts of theoretical information, formulae, and references; and a well-organized Web site with such features as 'Frequently Asked Questions'
  - The graphic interface is very user-friendly
- Disadvantages:
  - Numerous features and options to learn
  - Setting up a data file can be complex and must be formatted correctly (but good examples are given in the manual)

PowerMarker

- <http://www.powermarker.net>
- Designed for use with SSR/SNP data in population genetics analyses
- Available options include summary statistics, consensus trees, population structure, Mantel's test, triangle plotting and visualization of linkage disequilibria results

(continued on next slide)

PowerMarker is a new program, with the first official version released in January 2004. It was designed specifically for the use of SSR/SNP data in population genetics analyses. Data can be imported from Excel or other formats, making data set-up very easy. Data can also be exported to NEXUS and Arlequin formats. It includes a '2D viewer' for linkage disequilibrium visualization. The user can edit graphics within PowerMarker or export them for publication. The program has been tested extensively for accuracy and efficieny. Full documentation is included. Several new modules for association study are included in the package. Several demonstration datasets available to get started. The program is free, but requires having PHYLIP, TreeView and the Microsoft.net framework system (all freely available) and Excel 2000 (not free). Another disadvantage is that it is available only for Windows 98 and above (not for Macintosh or other systems). Email support (registered user only): powermarker@hotmail.com

Author: Kejun (Jack) Liu, North Carolina State University.

Reference

Liu, K. 2003. PowerMarker: New Genetic Data Analysis Software, Version 3.0. Free program distributed by the author over Internet at <http://www.powermarker.net>

PowerMarker (continued)

- Advantages:
  - Allows importing data from Excel, which makes data management very easy
  - The graphical interfaces are very user-friendly
  - Graphics are 'guaranteed' to have publication quality
- Disadvantages:
  - Very new program with possible bugs still to be ironed out
  - Not a stand-alone program, needs downloading of several other software programs and buying Excel

DnaSP

- <http://www.ub.es/dnasp>
- It uses DNA sequence data to perform population genetics analyses
- It performs a very large number of analyses, including measures of polymorphism, divergence between populations (including a measure of gene flow), synonymous and non-synonymous substitutions, linkage disequilibrium, recombination, and many statistical tests (Hudson, Kreitman and Aguade's, Fu and Li's, Tajima's, and the McDonald and Kreitman test)

(continued on next slide)

DnaSP, for DNA Sequence Polymorphism, uses DNA sequence data. This program is widely used for sequence analysis because it does all the necessary analyses and at the same time is easy to use. It was written exclusively for the Windows operating system, but can be run on a Macintosh using SoftWindows or VirtualPC software emulators. DnaSP can import and export several types of data formats, including FASTA and NEXUS, which is very convenient, and can handle large numbers of long sequences, depending on your computer's memory. The authors are currently working on version 4. It is freely available, downloadable from the Web site. Although no manual is available, a Help file is incorporated into the program. In addition, the Web site includes much explanatory material, as well as many references. The authors have several publications about the program (e.g. see citations below).

Authors: Julio and Ricardo Rozas

References

Rozas, J. and R. Rozas. 1995. DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. Comput. Appl. Biosci. 11:621-625.

Rozas, J. and R. Rozas. 1997. DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Appl. Biosci. 13:307-311.

Rozas, J. and R. Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174-175.

DnaSP (continued)

- Advantages:
  - The Windows user-interface makes the program very easy to use
  - A very large number of analyses are possible
- Disadvantages:
  - Currently only available for Windows users (and Macintosh with the use of a Windows emulator)
  - No manual or tutorials are yet available
  - Support is, apparently, not readily available

PAUP*

- <http://www.lms.si.edu/PAUP/about.html>
- Used for inferring and interpreting evolutionary trees
- Versions available for different computer platforms
- Includes: parsimony, distance matrix, invariants, maximum likelihood methods, and many indices and statistical analyses

(continued on next slide)

PAUP* is widely used for inferring and interpreting evolutionary trees. It originally meant Phylogenetic Analysis Using Parsimony, but now has many other options. PAUP* is available from Sinauer Associates, Sunderland, MA, at <http://www.sinauer.com/detail.php?id=8060>. Although not free, it is relatively inexpensive (US$100 at writing). A new version, 4.0 beta, has been released as a provisional version. Macintosh, PowerMac, Windows and Unix/OpenVMS versions are available; the Mac version has some extra features. PAUP* is closely compatible with MacClade (another program available from Sinauer), since they use a common data format (NEXUS, Maddison et al. 1997).

Author: David Swofford, Laboratory of Molecular Systematics, National Museum of Natural History, Smithsonian Institution, Washington, DC.

Reference

Swofford, D.L. 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4. Sinauer Associates, Sunderland, MA.

PAUP* (continued)

- Advantage:

  Good support available, both online (including a long list of 'Frequently Asked Questions') and via e-mail

- Disadvantage:

  It is not free

MEGA

- <http://www.megasoftware.net/>
- Uses DNA sequence, protein sequence, evolutionary distance or phylogenetic tree data
- The methods of analyses possible are numerous, including distance calculations, tree formations, and many ways to view data

(continued on next slide)

MEGA (Molecular Evolutionary Genetics Analysis) software has been widely used since its creation in 1993; MEGA2 has since come out. It uses DNA sequence, protein sequence, evolutionary distance or phylogenetic tree data. The authors' goal was to take advantage of advances in computer power and graphic user interfaces to make available a 'flexible and easy-to-use genetic data analysis workbench'. Although it was designed for the Windows platform, it runs well on Macintosh with a Windows emulator, Sun workstation (with SoftWindows95) or Linux (with Windows by VMWare). The newest version, 2.1, has many important additions, such as the ability to import data from NEXUS or CLUSTAL W, unlimited dataset sizes, and many others.

A book by the software authors Nei and Kumar (2000) includes theoretical information about statistical analyses and how to interpret results from both their software and other software programs. Online, a thorough manual is available (although the format is not easy to page through), together with a bulletin board for users to interact with each other.

Authors: Sudhir Kumar, Koichiro Tamura, Ingrid Jakobsen and Masatoshi Nei.

References

Nei, M. and S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, NY.

Sudhir, K., T. Koichiro, I.B. Jakobsen and M. Nei. 2001. MEGA2: Molecular Evolutionary Genetics Analysis software. Bioinformatics 12(17):1244-1245.

MEGA (continued)

- Advantages:
  - As this software has been available since 1993, it's likely that most of the bugs have already been discovered
  - There is good support available, including an online manual and a published book
- Disadvantage:

  It is not the easiest program for beginners to learn

Internet resources

- The European Molecular Biology Laboratory–European Bioinformatics Institute (EBI) site <http://www.ebi.ac.uk/services/>
- Biological software page from the Institut Pasteur in France <http://bioweb.pasteur.fr/intro-uk.html>
- Phylogeny Programs (listed by Joe Felsenstein at the University of Washington) <http://evolution.genetics.washington.edu/phylip/software.html>

(continued on next slide)

In this and the next slide, we give a sample list of Internet resources that you may find useful for locating information related to, for example, genetic diversity analysis, population genetics, other software available, and links to useful extra information. For each resource presented, we briefly describe their contents in the notes below.

- The European Molecular Biology Laboratory–European Bioinformatics Institute (EBI) site: not only does it contain links to many useful programs and other sites but it also describes what they do and is therefore a good source of general information as well.

- Biological software page from the Institut Pasteur in France: although some pages are in French only, it provides a very comprehensive list of software programs available online, including links, and is current (last updated December 2002). It also contains links to many programs developed at the Institut Pasteur.

- Phylogeny Programs: this is the longest list of phylogeny programs we have seen, counting 194. The author adds the caveat that he has not tried to assess their quality or cost. Nor has the list been updated since 2001, but it contains so many links to programs, and sorts them in various ways (e.g. by methods, system used) that it is still very useful.

Internet resources (continued)

- Dr. Ed Buckler's Maize Genetics site <http://www.maizegenetics.net/bioinformatics/index.htm>
- Kent Holsinger's site at the University of Connecticut <http://darwin.eeb.uconn.edu/archives.html>
- Claire Constantine's site at Murdoch University <http://wwwvet.murdoch.edu.au/vetschl/imgad/GSLinks.htm>
- Software page of the Institute of Forest Genetics and Forest Tree Breeding, University of Göttingen, Germany <http://www.uni-forst.gwdg.de/forst/fg/index.htm>

- Dr. Ed Buckler's Maize Genetics site: contains freely available software programs developed by his laboratory group. Although some information is specific to maize, the site also contains useful information about genomics, links to many journals, and PDF versions of Dr. Buckler's publications.

- Kent Holsinger's site at the University of Connecticut: has links to many software programs, including ones for biology, programming and statistics. The author does not post things he does not use regularly, so some guarantee of good quality exists. At writing, this was also the most recently updated.

- Claire Constantine's site at Murdoch University: although not updated recently, this site contains links to the most used programs for population genetics analyses. Even more useful, it includes a comparison table of the kinds of statistics available in each of 7 commonly used programs (Arlequin, GENEPOP, POPGENE, GDA, GeneStrut, DnaSP and SITES).

- Software page of the Institute of Forest Genetics and Forest Tree Breeding, University of Göttingen, Germany: this site contains just 4 software programs, developed in-house, but they are freely available, include good descriptions, and the page is regularly updated.

Appendix 9

Appendix 9. References to software programs

In summary

- Many computer programs exist for analysing molecular data for genetic diversity
- Most programs perform similar tasks and their main differences should be evaluated, depending on resources available and/or individual preferences
- Nowadays, in addition to freely available computer programs, plenty of resources are also found on Internet to help us obtain both basic and more specialized information on methods

By now you should …

- Be familiar with:
  - The contrasting features of available software for genetic diversity analysis
  - The advantages and disadvantages of some of the major computer programs
- Have become acquainted with some Internet resources that may assist you with the task of studying genetic diversity

References

References

Labate, J.A. 2000. Software for population genetic analyses of molecular marker data. Crop Sci. 40:1521-1528.

Liu, K. 2003. PowerMarker: New Genetic Data Analysis Software, Version 1.0. Free program distributed by the author over Internet at <http://www.powermarker.net>

Maddison, D. R., D.L. Swofford and W.P. Maddison. 1997. NEXUS: an extensible file format for systematic information. Syst. Biol. 46:590–621.

Nei, M. and S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, NY.

Rozas, J. and R. Rozas. 1995. DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. Comput. Appl. Biosci. 11:621-625.

Rozas, J. and R. Rozas. 1997. DnaSP, version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Appl. Biosci. 13:307-311.

Rozas, J. and R. Rozas. 1999. DnaSP, version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174-175.

Schneider, S., D. Roessli and L. Excoffier. 2000. Arlequin: A Software for Population Genetics Data Analysis, Version 2.000. Genetics and Biometry Laboratory, Dept. of Anthropology, University of Geneva, Switzerland.

Sudhir, K., T. Koichiro, I.B. Jakobsen and M. Nei. 2001. MEGA2: Molecular Evolutionary Genetics Analysis software. Bioinformatics 12(17):1244-1245.

Swofford, D.L. 2002. PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4. Sinauer Associates, Sunderland, MA.

- Glossary