Genetics and population analysis

Processing and population genetic analysis of multigenic datasets with ProSeq3 software

Dmitry A. Filatov

Department of Plant Sciences, University of Oxford, South Parks Rd, Oxford OX1 3RB, UK

ABSTRACT

Motivation: The current tendency in molecular population genetics is to use increasing numbers of genes in the analysis. Here I describe a program for handling and population genetic analysis of DNA polymorphism data collected from multiple genes. The program includes a sequence/alignment editor and an internal relational database that simplify the preparation and manipulation of multigenic DNA polymorphism datasets. The most commonly used DNA polymorphism analyses are implemented in ProSeq3, facilitating population genetic analysis of large multigenic datasets. Extensive input/output options make ProSeq3 a convenient hub for sequence data processing and analysis.

Availability: The program is available free of charge from http://dps.plants.ox.ac.uk/sequencing/proseq.htm

Contact: dmitry.filatov@plants.ox.ac.uk

1 INTRODUCTION

With ever decreasing costs of DNA sequencing and increasingly sophisticated analyses, the number of loci used in population genetic, phylogeographic and phylogenetic studies increases steadily. Only a few years ago it was normal to base the conclusions of experimental population genetic studies on the analysis of a single gene (Filatov and Charlesworth, 1999), while these days it is not uncommon to use hundreds of loci (or more) in a single study (Begun *et al.*, 2007; Foxe *et al.*, 2008). With the advent of high throughput sequencing the use of hundreds of loci will become the norm even for non-model organisms within a few years.

Many population genetic programs, such as IMa (Hey and Nielsen, 2007), Structure (Pritchard *et al.*, 2000) or Compute (Thornton, 2003) use multiple genes for analysis, however, preparation of such datasets, even with sequences in hand, is far from straightforward. Although there are ways to manipulate multigenic datasets using scripts, this requires programming skills, and in practice experimental population geneticists often do that manually. Here I report a program, ProSeq3, with a convenient graphic user interface that simplifies the preparation and basic population genetic analysis of multigenic datasets. It has been tested and fine-tuned for several years in our laboratory and its use leads to significant time savings at the dataset preparation and analysis stages.

2 FEATURES

ProSeq was originally developed as a Windows-based sequence editor with some DNA polymorphism analysis capability for single

gene datasets (Filatov, 2002). The new version is now available for both Windows and Linux and can handle large datasets with thousands of genes. The size of the datasets is limited by memory and by the maximal value of 32-bit signed long integers (2 147 483 647) used for internal indexing. The program can be used for sequence editing, annotation of sequence features, handling of output from high throughput sequencers, or from BLAST searches, as well as for various population genetic analyses. ProSeq3 supports and facilitates all steps of DNA sequencing workflow from sequence chromatogram editing to DNA polymorphism analysis of multigenic data.

2.1 DNA sequence editing, alignment and annotation

To help with the processing of raw sequence data ProSeq3 allows users to open and visualize sequence chromatograms, edit the sequence and assemble sequence contigs. Integration with popular phred and phrap programs (de la Bastide and McCombie, 2007; Ewing and Green, 1998) makes it possible to automatically assess chromatogram quality and assemble contigs. Raw sequences with or without associated chromatogram and base quality information can be further edited and annotated in ProSeq3 to obtain finished sequences.

ProSeq3 supports and facilitates the functional annotation of individual sequences in the dataset with several handy functions, such as selection and assignment of a functional (e.g. coding) region in the editor window, and the ability to copy assigned regions from another sequence in the dataset. All annotations are preserved if the dataset is saved in the data file (*.df) 'native' for ProSeq3.

Multiple sequence alignment can be done within ProSeq3, which includes Clustal (Higgins *et al.*, 1996). Alternatively alignment can be done manually using the ProSeq3 editor or an external program. In the latter case alignment information (position and length of gaps) can be imported back into the annotated dataset in ProSeq3. Following automated alignment, it is usually necessary to check, correct and trim the alignment manually, and check sequence differences between individual sequences, which is easily done in the sequence editor included in ProSeq3. The editor is fairly flexible and includes three viewing/editing modes, allowing the user to see/edit the sequence, polymorphisms in the alignment and the functional regions assigned to the sequence. Using these modes the user can scroll along the sequence, zoom in to see a region of the sequence or zoom out to visualize the entire sequence with annotation shown in a graphical form.

2.2 Handling data with a relational database

Tracking what sequence in a dataset comes from which individual becomes problematic when the number of sequenced genes is large. ProSeq3 resolves this problem by storing all the data in an internal relational database where the sequences are linked to individuals and individuals can be combined into groups (populations). This data structure makes it trivial to manipulate multiple datasets in the project; e.g. exclusion of one individual from analysis can be done with a couple of mouse clicks, which results in automatic exclusion of all sequences linked to that individual. Similarly, individual sequences or parts of sequences can be excluded from the analysis. Grouping sequences into populations is also done at the level of individuals: if an individual is assigned to the particular population, all the sequences across multiple datasets in the project that are linked to that individual are automatically assigned to that population. The assignment of sequences to individuals and individuals to groups can be done by a simple drag and drop approach. Relational information of the database is preserved if the project is saved in the native (*.df) ProSeq3 file format.

2.3 DNA polymorphism analysis

Once the alignments for several genes are complete and ready for analysis, they are usually analysed one by one using such programs as MEGA (Tamura *et al.*, 2007) or DnaSP (Librado and Rozas, 2009). This process is relatively quick when there are only a few genes, but it becomes prohibitively time-consuming with larger numbers of genes. ProSeq3 solves this problem by allowing the user to run all the datasets in the project through the particular analysis in one go. Several most commonly used population genetic analyses are implemented in ProSeq3: visualisation and analysis of single nucleotide polymorphisms, common statistics for DNA polymorphism (π, θ; Nei and Kumar, 2000), various neutrality tests such as Tajima's D (Tajima, 1989), and analysis of population subdivision/divergence. The distribution of DNA polymorphism or neutrality statistics along the length of a gene can be visualised with a sliding window option.

Although ProSeq3 was developed for population genetic analyses it also includes a tool for basic phylogenetic analysis that can construct and visualise neighbor-joining trees (Nei and Kumar, 2000). A combination of a sequence editor and tree visualisation tool in one program is particularly handy at the stage of preliminary evaluation and checking of the datasets, as oddities in the data, such as misalignment or sequencing errors make a sequence appear more diverged, which is easily identifiable from the inspection of a gene tree and can be quickly fixed within ProSeq3.

Other analysis options include the tool for creating bootstrap replicates of a dataset, and a tool for coalescent simulations (Hein *et al.*, 2005) with or without recombination in panmictic or subdivided populations.

2.4 Input/output options

ProSeq3 supports 25 different file formats. It can create input files for such popular programs as DnaSP (Librado and Rozas, 2009), MEGA (Tamura *et al.*, 2007), PAML (Yang, 2007), Arlequin (Excoffier *et al.*, 2005), Structure (Pritchard *et al.*, 2000) and IMa (Hey and Nielsen, 2007). The multitude of supported file formats and flexible data structure of ProSeq3 make it a convenient hub for sequence data processing and analysis.

3 IMPLEMENTATION

ProSeq3 has been developed in Delphi7 with the CLX library and it can be compiled for Windows and Linux operation systems.

REFERENCES

Begun,D.J. *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biol.*, **5**, e310.

de la Bastide,M. and McCombie,W.R. (2007) Assembling genomic DNA sequences with PHRAP. *Curr. Protoc. Bioinformatics*, **Chapter 11**, Unit 11 14.

Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.

Excoffier,L. *et al.* (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evol. Bioinform. Online*, **1**, 47–50.

Filatov,D.A. (2002) PROSEQ: A software for preparation and evolutionary analysis of DNA sequence data sets.. *Mol. Ecol. Notes*, **2**, 621–624.

Filatov,D.A. and Charlesworth,D. (1999) DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia PgiC locus. *Genetics*, **153**, 1423–1434.

Foxe,J.P. *et al.* (2008) Selection on amino acid substitutions in Arabidopsis. *Mol. Biol. Evol.*, **25**, 1375–1383.

Hein,J. *et al.* (2005) *Gene Genealogies. Variation and Evolution.* Oxford University Press, Oxford, UK.

Hey,J. and Nielsen,R. (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl Acad. Sci. USA*, **104**, 2785–2790.

Higgins,D.G. *et al.* (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.*, **266**, 383–402.

Librado,P. and Rozas,J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Nei,M. and Kumar,S. (2000) *Molecular Evolution and Phylogenetics.* Oxford University Press, New York.

Pritchard,J.K., Stephens,M. and Donnelly,P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

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

Tamura,K. *et al.* (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.*, **24**, 1596–1599.

Thornton,K. (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*, **19**, 2325–2327.

Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.