**PGEToolbox: a MATLAB Toolbox for Population Genetics and Evolution**

<table>
<thead>
<tr>
<th><strong>Journal:</strong></th>
<th>Bioinformatics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manuscript ID:</strong></td>
<td>BIOINF-2007-0386</td>
</tr>
<tr>
<td><strong>Category:</strong></td>
<td>Applications Note</td>
</tr>
<tr>
<td><strong>Date Submitted by the Author:</strong></td>
<td>12-Mar-2007</td>
</tr>
<tr>
<td><strong>Complete List of Authors:</strong></td>
<td>Cai, James; Stanford University, Biological Sciences</td>
</tr>
<tr>
<td><strong>Keywords:</strong></td>
<td>MATLAB Toolbox, Population Genetics, Evolution, Polymorphism, Neutrality Test</td>
</tr>
</tbody>
</table>
Applications Note

PGEToolbox: a Matlab Toolbox for Population Genetics and Evolution

James J. Cai
Department of Biological Sciences, Stanford University, 371 Serra Mall, Stanford, CA 94305, USA.

ABSTRACT
Summary: PGEToolbox is a Matlab-based software package for analysis of polymorphism and divergence data for population genetics and evolution. It estimates several basic statistics of DNA sequence variation and carries out statistical tests of selective neutrality under the infinite alleles model, such as Tajima's D test, Fu & Li's tests and Fay & Wu's H test. The significance of tests is determined from the distribution of the statistics obtained by coalescent simulation. The toolbox performs McDonald-Kreitman test (and several extensions). PGEToolbox also contains functions for handling SNP (Single Nucleotide Polymorphism) genotype data. PGEToolbox is open-sourced, can be easily extended or tailored for specific tasks, and scaled up for large data sets.
Availability: For academic uses, PGEToolbox is available free of charge at http://bioinformatics.org/pgetoolbox
Contact: jamescai@stanford.edu

1 INTRODUCTION

Assessing genetic diversity within populations is vital for understanding the nature of the evolutionary processes at the molecular level. Over many years, powerful methods have been developed to analyze genetic data to elucidate the influence of mutation, random genetic drift, migration and natural selection, on genetic diversity. Dedicated computer programs implementing these methods become essential for extracting embedded information. This is especially the case when large amounts of data are being produced nowadays in the accelerating rate.

PGEToolbox (from Population Genetics and Evolution) is a software package for data analysis in molecular population genetics under Matlab, a powerful technical computing environment. It assists population geneticists in many ways from manipulating data to performing statistical tests. It contains functions to manipulate polymorphism and/or divergence data and compute many population genetic statistics. It allows users to test the departure from selective neutrality with a number of established tests. The significance of results can be evaluated via coalescent simulations. It also provides tests based on comparison of polymorphism and divergence between species, which has been an effective strategy for testing population genetic hypotheses on the causes of variation. In addition, PGEToolbox includes a SNP analysis tool called snptool to provide the most frequently used functions related to SNP analysis. Open-source and sophisticated graphic function are major advantages of PGEToolbox over similar packages like DnaSP (Rozas et al., 2003) and libsequence (Thornton, 2003).

2 METHODS

Data Type and File Format: PGEToolbox supports two types of data - DNA sequences and SNP genotype data. For DNA sequences, it can read/write alignments in FASTA or Phylip formats. For SNP, it recognizes genotype information from the HapMap and Perlegen projects. The functions handling DNA sequences data and SNP data are separate, allowing the user to carry out the many same types of analyses irrespective of the data types. Functions working with SNP genotype data are named with the prefix 'snp_' for the user to carry out the many same types of analyses irrespective of the data types. Functions working with SNP genotype data are named with the prefix 'snp_'

System and Implementation: PGEToolbox is developed and tested in Matlab version 6.5 (R13) under Microsoft Windows. It is then deployed into versions for UNIX platforms (GNU/Linux and Solaris) and Macintosh platforms. PGEToolbox has been designed with several considerations in mind: batch-ability or scalability, extendibility and usability. As a result, PGEToolbox can be easily set up as scripts (calling one function after another) to perform an entire job in an unattended "batch mode". It is straightforward to apply to large data set. In many cases, implementing functions in the Matlab framework greatly reduced the complexity of the original implementation in other languages and allows users to debug or add new functions much more easily than before. The software is under an open source license which allows others to extend and re-use components, allows inter-operation via an open and published interfaces, and can reduce duplication of effort within the community. Graphic user interfaces (GUIs) are useful in hiding the complexity of the computations from the user. PGEToolbox contains many simple yet efficient menu- or dialog-driven GUIs

3 MAIN FEATURES

3.1 Polymorphism Statistics

The calculation of polymorphism statistics is a routine task in molecular population genetics. PGEToolbox calculates several polymorphism statistics as a routine task in molecular population genetics, such as polymorphic sites, number of segregating sites, site-frequency spectrum, nucleotide diversity and its sampling variance, Fay's $\theta_1$ statistic (Fay and Wu, 2000), and linkage disequilibrium (LD) like D, $D'$, or $r^2$ from sequences. For the population mutation parameter, $\theta = 4N_\mu$ (where $N$ is the effective population size and $\mu$ is the per-locus mutation rate per generation), PGE-Toolbox computes several common estimates of $\theta$, including the number of segregating sites, $\theta_K$ (Watterson, 1975), the mean pairwise difference between nucleotide sequences, $\theta_S$ (Nei, 1987), Fay's $\theta_H$ (2000).
3.2 Neutrality Tests

PGEToolbox conducts several statistical tests for detecting departures from neutrality. These tests include Ewens-Watterson's homozygosity test, Tajima's D test (1989), Fu & Li's D*- F*- tests (1993), Strobeck's S statistic (1987), Wall's B- & Q-tests (1999), Fay & Wu's H-test (2000), Watterson's homozygosity test of neutrality (1978), and Kelly's ZνS test (1997). PGEToolbox also offers tests for detecting population growth, Fu's Fs test (1997) and the R2 test (Ramos-Onsins and Rozas, 2002). For most important tests, like Tajima's D test, three functions are evaluated to complete the whole test. The three functions are tajima89d_test, tajima89d_simu and tajima89d. The first function tajima89d_test takes sequences as input to evaluate parameters: \( \theta_1 \) and \( \theta_2 \), required for calculating the statistics D. The second function tajima89d_simu generates multiple samples using coalescent simulation (see below) so that the significance of test can be evaluated. Both functions finally call a common procedure deployed in the third function tajima89d, which takes merely the necessary parameters \( \theta_1 \) and \( \theta_2 \) to compute D directly. Such a design allows user to start with sequence data, simulated data or direct parameter to compute a statistic.

3.3 Coalescent Simulations

Testing of the significance of computed statistics like Tajima’s D requires generate parametric bootstrap samples from a wide variety of neutral models using a coalescent approach and an infinitely many sites model of mutation. Hudson (2002)’s program ms has been incorporated as a Matlab MEX-function (the C interface to Matlab) to do coalescent simulation, giving PGEToolbox extensive capabilities in coalescent-based analyses. Simulations can be conducted for different parameter combinations. A dialog interface for coalescent simulation called coalsimdlg was developed to assist users in setting up those parameters.

3.4 McDonald-Kreitman Test and Derivatives

PGEToolbox provides methods to analyze patterns of genetic diversity within and between population samples – the McDonald-Kreitman (MK) test (1991) and extensions. Functions are available to count the numbers of synonymous (Ds) and nonsynonymous (Dn) divergences, and the numbers of synonymous (Ps) and nonsynonymous (Pn) polymorphisms. The MK test can be initiated from the command-line function mktestcmd or from another function called mktestgui, which invokes a pop-up dialog of 2x2 contingency table. The function, sewfw, estimates the average proportion of amino-acid substitutions driven by positive selection by using the method of Fay, Wyckoff & Wu (2001) and the method of Smith & Eyre-Walker’s (2002).

3.5 SNP Tool

PGEToolbox provides functions to explore the frequency and distribution of SNPs. The interface of those SNP-related functions is snptool. snptool not only opens genotype data file in HapMap, Perlegen, or pedigree format, but can also retrieve genotype data from the HapMap and Perlegen databases over the Internet. snptool computes several statistics from genotype data, including observed/predicted heterozygosity, minor allele frequency (MAF), p-value for the Hardy-Weinberg equilibrium test, allele frequency and genotype frequency. It calculates composite likelihood, Tajima’s D (1989) and Fay & Wu’s H (2000) statistics for SNPs with frequencies. Display and interpretation of large genotype data sets can be simplified by using snptool’s graphical display, such as the pie chart of allele/genotype frequencies among populations for a given SNP, plot for relative position of SNPs on chromosome, and the visual genotype (VG) view, which presents complete raw datasets of individuals’ genotype data. Finally, snptool uses the expectation-maximization (EM) algorithm to estimate probabilities of haplotypes and calculates LD statistics, such as, D D’ and R between pairs of SNPs.

ACKNOWLEDGEMENTS

JJC thanks Dmitri Petrov and Mike Macpherson at the department of Biological Sciences, Stanford University, Gavin Huttley, Peter Maxwell, Ray Sammut and Helen Lindsay of the Centre for Bioinformation Science (CBiS) at the Australian National University for valuable technical discussions.

REFERENCES