W166–W169 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh428

PDA: a pipeline to explore and estimate polymorphism in large DNA databases

Sònia Casillas and Antonio Barbadilla*

Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

Received February 15, 2004; Revised and Accepted April 13, 2004

ABSTRACT

Polymorphism studies are one of the main research areas of this genomic era. To date, however, no available web server or software package has been designed to automate the process of exploring and estimating nucleotide polymorphism in large DNA databases. Here, we introduce a novel software, PDA, *Pipeline Diversity Analysis*, that automatically can (i) search for polymorphic sequences in large databases, and (ii) estimate their genetic diversity. PDA is a collection of modules, mainly written in Perl, which works sequentially as follows: unaligned sequence retrieved from a DNA database are automatically classified by organism and gene, and aligned using the ClustalW algorithm. Sequence sets are regrouped depending on their similarity scores. Main diversity parameters, including polymorphism, synonymous and non-synonymous substitutions, linkage disequilibrium and codon bias are estimated both for the full length of the sequences and for specific functional regions. Program output includes a database with all sequences and estimations, and HTML pages with summary statistics, the performed alignments and a histogram maker tool. PDA is an essential tool to explore polymorphism in large DNA databases for sequences from different genes, populations or species. It has already been successfully applied to create a secondary database. PDA is available on the web at http://pda.uab.es/.

INTRODUCTION

Molecular data is growing dramatically and the need to develop efficient large-scale software to deal with this huge amount of information has become a high priority in this genomic era (1). Polymorphic studies are one of the main focuses of genomic research because of their promise to unveil the genetic basis of phenotypic diversity, with all their potential implications in basic biology, health and society. So far, several software programs have been developed that successfully analyze local data in terms of nucleotide variability [DnaSP (2), Arlequin http://lgb.unige.ch/arlequin/, SITES http://lifesci.rutgers.edu/~heylab/ProgramsandData/Programs/WH/WH_Documentation.htm], but they usually require that input sequences are previously aligned, which assumes that sequences are known to be polymorphic. None of these programs include a first step that permits to explore for potential polymorphic sequences from a large source of heterogeneous DNA, and then to extract and sort them out by gene, species and extent of similarity. Finally, for each group of two or more sequences already aligned, the main diversity parameters can be estimated.

With this prospect in mind we have developed PDA, *Pipeline Diversity Analysis*, a web-based tool which retrieves information from large DNA databases and provides a consistent (3), user-friendly interface to explore and estimate nucleotide polymorphisms. PDA can deal with large sets of unaligned sequences, which can be retrieved automatically from DNA databases given a list of organisms, genes or accession numbers. Even though it is web based, the source code can also be downloaded and installed locally.

A typical user of this site is a researcher who wants to know how many polymorphic sequences are available in Genbank (4) for one or several species of interest and how much variation there is in such sequences. Then, the researcher addresses to the PDA main page, writes the species names and chooses Genbank as the data to search for. Additionally, the user defines some parameter values such as the minimum ClustalW pairwise similarity score from which the sequences or the different gene regions to be analyzed will be grouped. The researcher will receive as output a database containing all the sequences and measures of DNA diversity, as well as HTML pages with summary statistics, the performed alignments and a histogram maker tool for graphical display of the results.

PDA has already been successfully used to explore the amount of polymorphism in the *Drosophila* genus and to create the DNA secondary database DPDB, *Drosophila Polymorphism Database* (http://dpdb.uab.es). This is the first

*To whom correspondence should be addressed. Tel: +34 935 812 730; Fax: +34 935 812 387; Email: Antonio.Barbadilla@uab.es

database that allows the search of DNA sequences by genes, species, chromosome, etc., according to different parameter values of nucleotide diversity. PDA is available on the web at http://pda.uab.es.

PROGRAM OVERVIEW

PDA is a *pipeline* made of multiple programs written in Perl (http://www.perl.com). This language was chosen for the development of PDA because of its initial orientation to the search, extraction and formatting of sequence strings, its support for object-oriented programming, the existence of a public repository of reusable Perl modules [the Bioperl project, http://www.bioperl.org (5)], and the ease of Perl commands to control and execute external programs in other languages (6).

Pipeline design

PDA runs sequentially several modules in a pipeline process as illustrated in Figure 1. Initially, sequences and their annotations are extracted from the input source defined by the user in the PDA home page. Input sources include DNA databases such as Genbank (4) (http://www.ncbi.nlm.nih.gov/Genbank/index.html), EMBL-Bank (http://www.ebi.ac.uk/embl/index.html) or the DPDB database (http://dpdb.uab.es). Low quality sequences coming from large-scale sequencing projects (i.e. *working draft*) are excluded from the analysis. Searches to these databases are done according to a list of accession numbers, organisms and/or genes. Alternatively, sequences can be introduced manually in Fasta or Genbank formats. All the retrieved sequences are introduced into a database (Figure 1: 1a) and passed to the next module (Figure 1: 1b). The second module organizes the sequences by organism and gene and filters these groups according to a minimum number of sequences per group set by the user (Figure 1: 2). Then, every group is aligned using the ClustalW algorithm (7) (Figure 1: 3). Default values have been fitted for the optimal alignments obtained in DPDB, but they can be alternatively defined by the user. The percentage of similarity between each pair of sequences (*ClustalW score*) is taken into account to group again the sequences in subgroups having a higher score than the minimum defined (Figure 1: 4). The value of this score can also be defined by the user and is set to 90% by default. Later on, the alignments are input into the Diversity Analysis module (Figure 1: 5–6), where the main nucleotide diversity, linkage disequilibrium and codon bias measures can be estimated. Finally, the results of the analyses are presented in four formats: a complete output database (in MySQL or MS-Access format) which can be downloaded as a compressed *.gz* file, a web-based output with summary statistics and the estimators, all the performed alignments, and a histogram maker tool for graphic display (Figure 1: 7).

Different gene regions can be analyzed separately. In this case, some additional steps are taken before presenting the results (Figure 1: 8–10). First, a module reads the annotations of the gene corresponding to the sequences on each alignment resulting from previous analyses. The fragments of the sequences from every gene region to analyze (e.g. exon, intron, etc., defined by the user) are extracted from the initial sequence according to the annotations and reversed-complemented if needed. Finally, the resulting sequences fragments are aligned and analyzed as before (Figure 1: 3–7).

Limitations

The heterogeneous nature of the source sequences is intrinsically problematic because the grouping module can lump

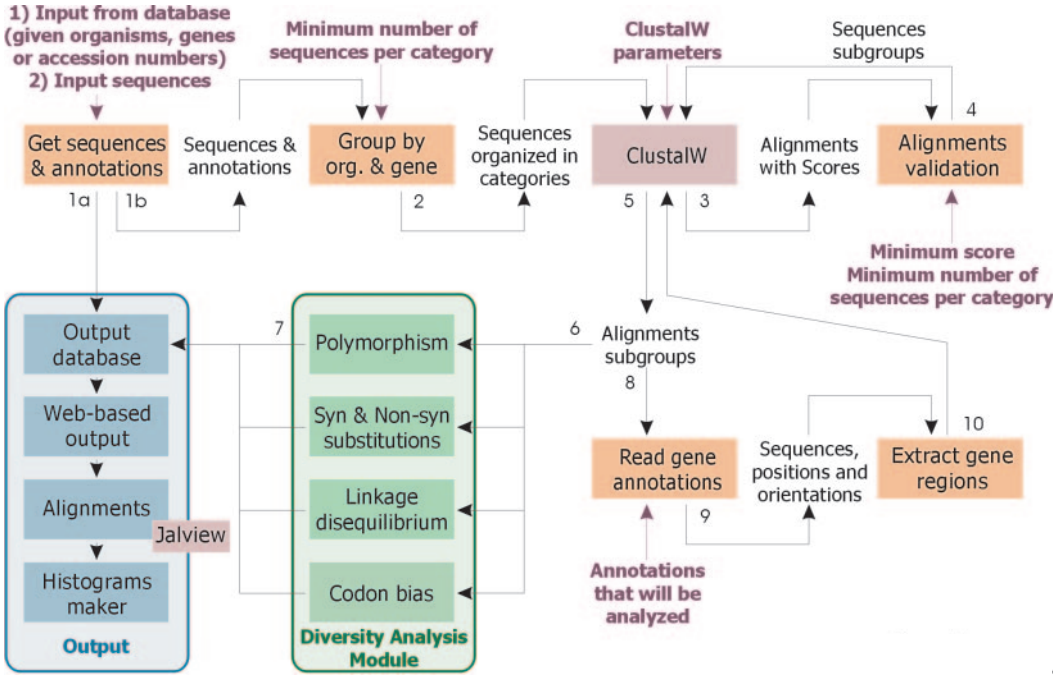

Figure 1. PDA program design and data flow. Independent Perl modules are represented by color boxes, and data flow by arrows and numbers. Lettering in purple corresponds to user-defined parameters. Meanings of color boxes: orange, sequences manipulations; green, nucleotide diversity analysis; blue, output; purple, external programs implemented in PDA. See text for details.

together sequences that are fragmented, or paralogous, or coming from different populations or arrangements, or simply incorrectly annotated, among other reasons. This can distort, to different degrees, the estimated diversity values and therefore, a first analysis must be seen as preliminary. To minimize this problem it would be useful to define an appropriate similarity score between each pair of sequences (*ClustalW score*) or to repeat the analysis with different values. High values of this score would make more restrictive the grouping of sequences. Nevertheless, after a first analysis it is always advisable to inspect visually the alignments, mainly those that yield extreme values, that have a high proportion of gaps or ambiguous bases, or whose sequence lengths vary widely. Two parameters, the percentage of excluded sites due to gaps or ambiguous bases within the aligned sequences and the relative and absolute differences between the longest and shortest sequences are estimated. A warning message appears in the output when the percentage of excluded sites is >30%. In addition, sequences with lengths <100 nt are excluded from the analysis. Both values are set by default and can be modified by the user. Since every sequence from an alignment is linked to its annotation, it is easy to trace the origin of the sequence and to assess its suitability to be included in the analysis. After this inspection, dubious, incorrect or unequal sequences can be manually eliminated via the *Jalview* editor (8), implemented in the *Alignments* section of the output and a reanalysis performed.

PDA has been optimized in terms of speed analysis. However, the process is intensive by nature and the analysis is run in a batch queue. We are putting our effort into parallelizing different instances of PDA using a large cluster of computers through the Condor batch queues specialized management system (http://www.cs.wisc.edu/condor/). However, we encourage users aiming to conduct large and frequent analyses to download, install and use PDA locally in their computers.

DIVERSITY PARAMETERS ESTIMATED

PDA provides a wide range of polymorphic estimations (with their respective variances and SD measures) and statistical tests for polymorphism, codon bias and linkage disequilibrium analyses. Table 1 lists all estimated parameters that have been implemented. All the algorithms have been checked with specific examples or by comparing the results with other available software such as DnaSP (2). Future improvements of the program will include the implementation of typical measures of divergence between different species and the reconstruction of phylogenetic trees. In this way, PDA should be seen as a general tool for large-scale DNA diversity analysis, both for within and among species gene variation.

Table 1. List of estimators implemented in PDA for DNA polymorphism, codon bias and linkage disequilibrium analysis

Nucleotide polymorphism	
Number of segregating sites (S, s)	Nei (9)
Minimum number of mutations (H, η)	Tajima (10)
Nucleotide diversity (π) (with and without Jukes and Cantor correction)	Nei (9); Jukes and Cantor (11)
Theta (θ) per DNA sequence from S	Tajima (12)
Theta (θ) per site from S	Nei (9)
Theta (θ) per site from Eta (η)	Tajima (10)
Theta (θ) per site from π, from S and from η under the Finite Sites Model	Tajima (10)
Average number of nucleotide differences (k)	Tajima (13)
Tajima statistic test (D)	Tajima (14)
Total number of synonymous and non-synonymous sites	Nei and Gojobori (15)
Number of non-synonymous substitutions per non-synonymous site (Ka) and number of synonymous substitutions per synonymous site (Ks)	Nei and Gojobori (15)
Codon bias	
Relative Synonymous Codon Usage (RSCU)	Sharp (16)
Effective Number of Codons (ENC)	Wright (17)
Codon Adaptation Index (CAI)	Sharp and Li (18)
Scaled Chi Square	Shields (19)
G + C content in second, third and total positions	Wright (17)
Linkage disequilibrium	
Nucleotide distance (Dist) between a pair of polymorphic sites	
D	Lewontin and Kojima (20)
D′	Lewontin (21)
R and R^2	Hill and Robertson (22)
ZnS statistic	Kelly (23)
Chi-square test	
Fisher's exact test	

OUTPUT

The results of PDA are stored in the PDA server and can be accessed through an HTML page using a unique ID that is assigned to every job. The output includes: (i) a MySQL or MS-Access 2002 database with all the retrieved sequences and the results of the analyses, which can be downloaded as a compressed *.gz* file or searched directly through the PDA server in the case of MySQL; (ii) a set of HTML pages with most of the contents of the database and summary statistics both for the whole gene length and for gene regions; (iii) the performed alignments in Fasta and Clustal formats, and the alignments visualization java applet Jalview (8); and (iv) a histogram maker tool for graphic display of personalized histograms and frequency representations of all the estimations. A sample output can be seen at http://pda.uab.es/pda/pda_example.asp.

PDA has already been used on all the sequences of the *Drosophila* genus. The results have been introduced in a relational database which is integrated in the web bioinformatics platform DPDB (http://dpdb.uab.es). Using the DPDB interface, these estimations and the original sequences analyzed can be searched and retrieved according to different parameter values of nucleotide diversity, and many tools can be used online with the users input, including the PDA itself.

AVAILABILITY

PDA can be accessed on the web at site http://pda.uab.es together with examples and documentation. In addition, the source code to PDA is distributed as a package of programs to

be downloaded and run locally (http://pda.uab.es/pda/pda_download.asp) under the GNU General Public License (GPL).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR online.

ACKNOWLEDGEMENTS

The authors would like to thank Jordi Pijoan and Helena Norén for help in developing the software, and Rosemary Thwaite, Marta Puig, Natalia Petit and Alfredo Ruiz for their critical reading of the manuscript. This work was funded by the Ministerio de Ciencia y Tecnología (Grant PB98-0900-C02-02). S.C. was supported in part by the bioinformatics company e-Biointel and the Ministerio de Ciencia y Tecnología (Grant BES-2003-0416).

REFERENCES

1. Collins,F.S., Green,E.D., Guttmacher,A.E. and Guyer,M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
2. Rozas,J., Sanchez-DelBarrio,J.C., Messeguer,X. and Rozas,R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.
3. Stein,L. (2002) Creating a bioinformatics nation. *Nature*, **417**, 119–120.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res*., **32**, D23–D26.
5. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*., **12**, 1611–1618.
6. Stein,L.D. (2001) Using Perl to facilitate biological analysis. In Ouellette,B.F.F. (ed.), Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Wiley-Liss, Inc., pp. 413–449.
7. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*., **31**, 3497–3500.
8. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.
9. Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.
10. Tajima,F. (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics*, **143**, 1457–1465.
11. Jukes,T.H. and Cantor,C.R. (1969) Mammalian protein metabolism. In Munro,H.N. (ed.), *Evolution of Protein Molecules*. Academic Press, New York, pp. 21–132.
12. Tajima,F. (1993) Mechanisms of molecular evolution. In Takahata,N. and Clark,A.G. (eds), *Mesurement of DNA Polymorphism*. Sinauer Associates Inc., Suderland, Massachusetts.
13. Tajima,F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.
14. Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
15. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol*., **3**, 418–426.
16. Sharp,P.M., Tuohy,T.M. and Mosurski,K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res*., **14**, 5125–5143.
17. Wright,F. (1990) The 'effective number of codons' used in a gene. *Gene*, **87**, 23–29.
18. Sharp,P.M. and Li,W.H. (1987) The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*., **15**, 1281–1295.
19. Shields,D.C., Sharp,P.M., Higgins,D.G. and Wright,F. (1988) 'Silent' sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol*., **5**, 704–716.
20. Lewontin,R.C. and Kojima,K. (1960) The evolutionary dynamics of complex polymorphisms. *Evolution*, **14**, 458–472.
21. Lewontin,R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, **49**, 49–67.
22. Hill,W.G. and Robertson,A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet*., **38**, 226–231.
23. Kelly,J.K. (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.