W632–W634 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl080

PDA v.2: improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA

Sònia Casillas and Antonio Barbadilla*

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

Received February 14, 2006; Revised and Accepted March 3, 2006

ABSTRACT

Pipeline Diversity Analysis (PDA) is an open-source, web-based tool that allows the exploration of polymorphism in large datasets of heterogeneous DNA sequences, and can be used to create secondary polymorphism databases for different taxonomic groups, such as the *Drosophila* Polymorphism Database (DPDB). A new version of the pipeline presented here, PDA v.2, incorporates substantial improvements, including new methods for data mining and grouping sequences, new criteria for data quality assessment and a better user interface. PDA is a powerful tool to obtain and synthesize existing empirical evidence on genetic diversity in any species or species group. PDA v.2 is available on the web at http://pda.uab.es/.

INTRODUCTION

The first version of Pipeline Diversity Analysis (PDA), was announced in the Web Server Issue of this journal (1) as a web-based tool that allowed the exploration for polymorphism in large datasets of heterogeneous DNA sequences. The pipeline automatically extracts a set of sequences from a DNA database given a list of organisms, genes or accession numbers, and sorts them by gene, species and extent of similarity. Then it aligns the homologous sequences and calculates the standard population genetic diversity parameters on the generated alignments. PDA is not aimed to provide exhaustive measures of DNA diversity (2), but rather to be an exploratory tool to transform the huge amounts of sequences available in public databases into information that can be analyzed from a population genetic perspective. PDA gives an overview of the empirical evidence on genetic diversity in any species or group of species.

PDA has already been used successfully to explore the amount of polymorphism in the *Drosophila* genus and to create the secondary database DPDB, *Drosophila* Polymorphism Database (http://dpdb.uab.es) (3). This is the first database that allows the search of DNA sequences and polymorphic alignments by diversity values, in addition to filter the results by organism, gene region or data quality criteria. At present, PDA is being used to create a database for mammalian sequences (MamPol, http://pda.uab.es/mampol/) of nuclear and mitochondrial genes that will include new features with respect to DPDB. A modified version of PDA is also being developed to extend the DPDB database to include sequences from non-coding regions.

In this paper we introduce a new version of the pipeline, PDA v.2, which incorporates novel features and substantial improvements with respect to the original version, including new methods for data mining and grouping, new criteria for data quality assessment and a much better interface usability.

NEW METHODS FOR DATA GROUPING AND ANALYSIS

The input raw data for PDA are polymorphic sets formed by groups of orthologous sequences (alleles or haplotypes) for a given species and DNA region. Sequences belonging to a polymorphic set can come either: (i) from previous polymorphism studies, or (ii) from independent studies of the same gene and species, possibly not primarily focused on polymorphism. This second subset of sequences increases significantly the amount of polymorphic sets, although it raises the question whether the estimations are reliable. Due to the heterogeneous origin of the source sequences, PDA can mix together fragmented sequences coming from different regions of the same gene that do not align together, paralogous sequences or sequences coming from different populations or arrangements that have very distinct haplotypes. These cases were already resolved in PDA v.1 using a minimum similarity score for each

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

pair of sequences in the alignment that is customizable by the user. The default score is 95%, so sequences differing in more than the 5% of the sequence (excluding gaps) are split into separate alignments. PDA v.2 includes new features to handle the heterogeneity of the source sequences and to improve the quality of the alignments.

Algorithm for maximization of the number of informative sites

Although sequences from a given alignment are usually very similar in terms of sequence identity, they can vary widely in length. Because estimates of genetic diversity usually exclude gapped sites, a significant amount of information can be lost if large and short sequences are aligned together, since only the sites included in the shortest sequences will be used in the analyses. To maximize the amount of information that can be used in such estimates, we have implemented an algorithm that works as follows (Figure 1). First, sequences from an alignment are grouped according to their length, so that sequences in a group cannot differ in more than 20% of their length. After that, the amount of informative sites in each accumulative group of sequences is calculated, starting with the group of the longest sequences (group 1) and adding in each step the next group of sequences ordered by their length (groups 1 + 2, groups 1 + 2 + 3, etc.). By informative sites we mean the number of non-gapped positions multiplied by the number of sequences in the set (note that this differs from the definition of 'informative site' typically used in phylogenetics). Finally, PDA v.2 uses the set of sequences which offers the largest number of informative sites, in some cases discarding the shortest sequences. This algorithm can be used optionally in PDA v.2.

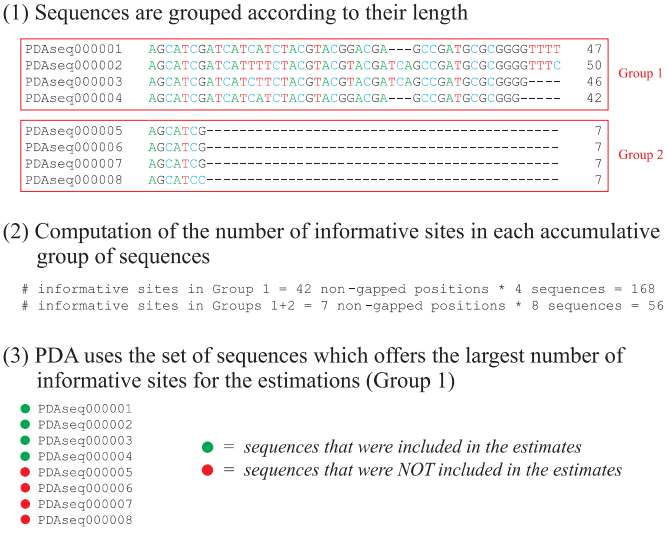

Figure 1. Example showing the new algorithm for maximizing the number of informative sites. (1) Input sequences are grouped according to their length, so that sequences in a group cannot differ in more than the 20% of their length. In this example, the eight input sequences are split into two different groups (group 1 and group 2). (2) Assuming that an 'informative site' is the number of non-gapped positions multiplied by the number of sequences in the set (note that this differs from the definition of 'informative site' typically used in phylogenetics), PDA v.2 calculates the amount of informative sites in each accumulative group of sequences, starting with the group of the longest sequences (group 1 = 168 informative sites) and adding in each step the next group of sequences ordered by their length (groups 1 + 2 = 56 informative sites). (3) Finally, PDA v.2 shows the alignment with all the sequences, but uses the set of sequences which offer the largest number of informative sites for the estimations, in some cases discarding the shortest sequences. In this case, PDA v.2 would use only the four longest sequences for the estimations (group 1). To distinguish which sequences were used in the analyses from those which were discarded, PDA v.2 uses a color code: green for sequences that were included in the estimates, and red for sequences that were not included.

Filtering raw sequences for well annotated genes

PDA v.1 analyzed raw sequences directly from GenBank regardless of the annotation quality or the number of genes included in the sequence. So, large genomic fragments including more than one gene could be aligned together with sequences of single genes. To avoid these noisy data, only well annotated sequences for the different functional regions of the genes (genes, CDSs, exons, introns, UTRs, promoters, etc.), as defined in the Features section of the GenBank format files, are now analyzed in PDA v.2. Note that sequences lacking these annotations, even coming from polymorphic studies, will not be included in the analyses. Thus, in PDA v.2 raw data is more appropriately pre-processed by functional category, and the main unit for storing information in the database is not the raw sequence coming from GenBank but the corresponding polymorphic sets for each organism and gene region [see Figure 1 in (3)].

Additional alignment programs

We have incorporated two new programs within PDA that can be used to align the polymorphic sequences in addition to ClustalW (4,5): Muscle (6) and T-Coffee (7). These programs have been shown to achieve better accuracy than the commonly used ClustalW for sequences with a high proportion of gaps, such as non-coding sequences (see the Help section of the Web site). We suggest using these alternative programs when analyzing non-coding regions (introns, promoters, UTRs, etc.).

DATA QUALITY ASSESSMENT

In PDA v.2 we provide several measures concerning the quality of each dataset so that the user can assess the confidence on the data source and the estimations. A quick guide is also supplied explaining how to use these quality measures and how to easily reanalyze the data.

Quality assessment of the alignments

To assess the quality of an alignment we use three criteria: (i) the number of sequences included in the alignment; (ii) the percentage of gaps or ambiguous bases within the alignment; and (iii) the percent difference between the shortest and the longest sequences. Three qualitative categories are defined for each criterion: high, medium and low quality, which are shown in the main output table to quickly visualize the confidence on the results (further details are given in the Help section of the Web site).

Quality assessment of the data sources

According to the data source, we use four criteria to determine if the sequences from a polymorphic set come from a

population study: (i) one or more sequences from the alignment are stored in the PopSet database; (ii) all the sequences have consecutive GenBank accession numbers; (iii) all the sequences share at least one reference; and (iv) one or more references are from journals that typically publish polymorphism studies (*Genetics*, *Molecular Biology and Evolution*, *Journal of Molecular Evolution*, *Molecular Phylogenetics and Evolution* or *Molecular Ecology*). This information is shown in the main output table by means of a confirmatory tick where the dataset satisfies the corresponding criterion.

Origin of the sequences

PDA v.2 reports the origin of each sequence (country, strain and population variant) when this information is available in the GenBank annotations. This allows the user to trace the origin of the source sequences and to assess the suitability of each sequence to be included in the dataset.

INTERFACE AND NEW UTILITIES

Important improvements in the text and graphic interface and other new features make PDA a much more useful tool.

Completely renewed interface

PDA v.2 offers a more intuitive and visually improved interface for both data input and output. For example, the page for job submission is designed in layers, which substantially facilitates the understanding of the available options. The output is more clearly displayed, and is based on the design of the DPDB database (3).

Management of previous analyses

On submitting a job, PDA v.2 can optionally store user information to allow them enter the 'Previous IDs' section and manage their previous analyses, either to revisit or to delete them. This new feature extends the previous 'Request by ID' option of PDA v.1, which is still available.

Improved database structure

The database has been extended to store the new data gathered by PDA, e.g. the storage of polymorphism datasets by functional categories (see above). Moreover, existing tables have been redefined, improving the performance of the search responses.

Tools for extraction and representation of polymorphic sites

A new module for extraction of SNPs from the aligned sequences has been incorporated. It lists the position of each SNP in the alignment and the frequency of the different alleles. Moreover, the data can be directly submitted to the SNPs-Graphic tool of the DPDB database to perform sliding windows and graphs for detailed analyses of polymorphism.

Improved sections of the web site

We have extended the Help section of the Web to provide a more complete and detailed description of PDA and to explain the new features of PDA v.2. We have also included links to the polymorphic databases created with this software.

AVAILABILITY

PDA v.2 can be accessed on the web at http://pda.uab.es/, together with examples and documentation. Jobs are run in a batch queue. Although at present the number of sequences that can be analyzed on the Web is limited to 500, we are working to have ready a parallel version of PDA to extend the number of sequences that can be analyzed. In addition, the source code of PDA is distributed under the GNU General Public License (GPL) as a package of Perl programs to be downloaded and run locally without limitations (http://pda.uab.es/pda2/pda_download.asp).

ACKNOWLEDGEMENTS

The authors would like to thank Casey Bergman for valuable discussions and guidance in implementing the alignment programs, and for his critical reading of the manuscript. The authors also thank Francesco Catania for helpful comments on this manuscript. S.C. was supported by the Ministerio de Ciencia y Tecnología (Grant BES-2003-0416) and a Marie Curie fellowship from the European Commission (HPMT-GH-01-00285-13). The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Casillas,S. and Barbadilla,A. (2004) PDA: a pipeline to explore and estimate polymorphism in large DNA databases. *Nucleic Acids Res*., **32**, W166–W169.
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. Casillas,S., Petit,N. and Barbadilla,A. (2005) DPDB: a database for the storage, representation and analysis of polymorphism in the *Drosophila* genus. *Bioinformatics*, **21**, ii26–ii30.
4. 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.
5. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
6. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.
7. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*., **302**, 205–217.