Computer Methods and Programs in Biomedicine 140 (2017) 69–76

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Extending GelJ for interoperability: Filling the gap in the bioinformatics resources for population genetics analysis with dominant markers

César Domínguez[a], Jónathan Heras[a,*], Eloy Mata[a], Vico Pascual[a], Maria Soledad Vázquez-Garcidueñas[b], Gerardo Vázquez-Marrufo[c,*]

[a] Department of Mathematics and Computer Science, University of La Rioja, Logroño, Spain
[b] Division of Postgraduate Studies, Faculty of Medical and Biological Sciences "Dr. Ignacio Chávez", Universidad Michoacana de San Nicolás de Hidalgo, Mexico
[c] Multidisciplinary Center of Biotechnology Studies (CMEB), Faculty of Veterinary Medicine, Universidad Michoacana de San Nicolás de Hidalgo, Mexico

ARTICLE INFO

Article history:
Received 26 May 2016
Revised 14 October 2016
Accepted 5 December 2016

Keywords:
Interoperability
GelJ
Dominant markers
Population genetics
Phylogenetic trees
PopXML

ABSTRACT

Background and objective: The manual transformation of DNA fingerprints of dominant markers into the input of tools for population genetics analysis is a time-consuming and error-prone task; especially when the researcher deals with a large number of samples. In addition, when the researcher needs to use several tools for population genetics analysis, the situation worsens due to the incompatibility of data-formats across tools. The goal of this work consists in automating, from banding patterns of gel images, the input-generation for the great diversity of tools devoted to population genetics analysis.

Methods: After a thorough analysis of tools for population genetics analysis with dominant markers, and tools for working with phylogenetic trees; we have detected the input requirements of those systems. In the case of programs devoted to phylogenetic trees, the Newick and Nexus formats are widely employed; whereas, each population genetics analysis tool uses its own specific format. In order to handle such a diversity of formats in the latter case, we have developed a new XML format, called PopXML, that takes into account the variety of information required by each population genetics analysis tool. Moreover, the acquired knowledge has been incorporated into the pipeline of the GelJ system – a tool for analysing DNA fingerprint gel images – to reach our automatisation goal.

Results: We have implemented, in the GelJ system, a pipeline that automatically generates, from gel banding patterns, the input of tools for population genetics analysis and phylogenetic trees. Such a pipeline has been employed to successfully generate, from thousands of banding patterns, the input of 29 population genetics analysis tools and 32 tools for managing phylogenetic trees.

Conclusions: GelJ has become the first tool that fills the gap between gel image processing software and population genetics analysis with dominant markers, phylogenetic reconstruction, and tree editing software. This has been achieved by automating the process of generating the input for the latter software from gel banding patterns processed by GelJ.

1. Introduction

Dominant markers including AFLPs, ISSRs, rep-PCR (BOX, ERIC, and REP), and RAPDs are useful tools for population genetic analysis – several applications of these markers are listed in Supplementary material S1. Some of the advantages of dominant markers are that they do not require previous knowledge about the genome of the studied species, and that they allow detection of intraspecific differences across the whole genome at different ploidy levels [1,2]. In addition, except for AFLP assays based on fluorescent detection in capillary electrophoresis, techniques for fingerprinting with dominant markers are simple and can be made in agarose or polyacrylamide gels using basic molecular laboratory equipment; making them a cheap solution for population genetic analysis of

* Corresponding authors.
E-mail addresses: cesar.dominguez@unirioja.es (C. Domínguez), jonathan.heras@unirioja.es, joheras@gmail.com (J. Heras), vico.pascual@unirioja.es (V. Pascual), marisolvaz@yahoo.com (M.S. Vázquez-Garcidueñas), gvazquezmarrufo@yahoo.com.mx (G. Vázquez-Marrufo).

http://dx.doi.org/10.1016/j.cmpb.2016.12.001
0169-2607/

plants, animals, and microorganisms – both prokaryotic and eukaryotic. In spite of the existence of some disadvantages of dominant markers [2], those drawbacks can be overcome by a combined use with other dominant or codominant markers and by a suitable bioinformatics analysis [3], thus becoming useful for analysing inter and intra-population genetic differentiation and structure, dispersion, migration, genotype-environment associations, and gene flow, among others.

Nowadays, there is a wide variety of software tools for population genetics analysis with dominant markers featuring, among other functionalities, the computation of genetic diversity indices and F-statistics, and the visualisation and edition of phylogenetic trees – see [4] and Supplementary materials S2 and S3. In order to employ software tools for population genetics analysis, it is necessary to transform dominant marker fingerprints, that consist of complex gel banding patterns, into either band presence (1) – absence (0) binary matrices or phylogenetic trees that will be used as input of those systems. Unfortunately, the generation of such an input from banding patterns of gel images might be a challenge.

In spite of the existence of several programs for dealing with banding patterns of gel images of dominant markers [5], these systems have not been designed to interact with software for population genetics analysis. In the case of presence/absence matrices, software for gel banding patterns sometimes construct those matrices internally; but, in general, they do not support their exportation, and when they do, the format of the exported matrices is not compatible with the input format of software tools for population genetics analysis. In the case of phylogenetic trees, several tools for gel banding patterns feature the generation of such trees; however, the generated trees can only be saved as images; and, hence, they cannot be fed as input to the tools for population genetics analysis. Therefore, the task of creating the input for population genetic analysis software must be carried out manually. This is a laborious, subjective, time-consuming, error-prone, and unreproducible task, which might produce unreliable results. Moreover, the risk of generating unreliable inputs is increased when using a large number of individuals and loci, which are needed to obtain reliable population genetics data.

In addition to the drawback of generating the input of software tools for population genetics analysis, there is another challenge in this context: interoperability among tools. Researchers normally need to analyse the same data with several programs; unfortunately, most of the programs that take presence/absence matrices as input use specific data-file formats [4]. Then, it is necessary to either manually transform the data across formats or use conversion tools. Neither approach is fully satisfactory, the former since is tedious, error-prone and not suitable when dealing with a large number of individuals; and the latter because conversion tools do not cover all the possible systems, depend on the version of the programs and, in general, cannot be easily adapted to handle new file formats that might arise with new systems. The situation is much better in the case of phylogenetic trees [6], since there are two standard formats widely employed by the majority of the systems: Newick [7] and Nexus [8]. Hence, the same input can be employed by several systems.

In this paper, we present how we have tackled the aforementioned problems to achieve our goal: automatic input generation for the wide variety of systems devoted to population genetics analysis and phylogenetic-trees editing from banding patterns of gel images. The first step to reach that aim has been a thorough review of software for population genetics analysis and phylogenetic trees to identify the characteristics of the input of such tools. From that review, we have checked that the formats Newick and Nexus are widely employed by software for phylogenetic-trees editing; and, in addition, we have defined a new format, called PopXML, that puts together all the information needed by the diversity of formats employed to encode presence/absence matrices in software for population genetic analysis with dominant markers. Finally, we have expanded the functionality of the GelJ system [9], an open-source and free tool for analysing DNA fingerprint gel images, to generate presence/absence matrices and phylogenetic trees that can be employed not only by the tools surveyed in our review, but also by new tools that might appear in the future. As a result, we have created the first existing tool that fills the bioinformatics gap between gel image processing software and population genetics analysis, phylogenetic reconstruction, and tree editing software.

2. A review of software for population genetics analysis and phylogenetic trees

In this section, we survey software tools for population genetic analysis with dominant markers that take presence/absence matrices as input, and tools for managing phylogenetic trees that work with the standard formats Newick and Nexus. The final aim of this survey is threefold: find the tools available for population genetics analysis with dominant markers; identify the characteristics of the input files of those tools; and, check whether tools that use either the Newick or the Nexus format to encode phylogenetic trees provide the necessary functionality to handle this kind of trees.

We screened PubMed Central and Google Scholar looking for corpora publications, and used the Google search-engine to create two lists of software tools. This search produced 29 tools for population genetic analysis with dominant markers, and 31 tools for managing phylogenetic trees with the Newick or the Nexus format.

2.1. Software for population genetics analysis with dominant markers

In the last 20 years, a great diversity of software tools for population genetics analysis has been developed with different aims and handling different kinds of data. A survey of 25 of those tools was provided in [4]. Such a survey included tools that support data types like DNA sequences, dominant markers, or multi-allelic markers. In our case, we are focused on the tools that work with dominant markers and take as input presence/absence matrices. The programs that have been included in our survey are listed in Table 1.

As can be seen in Table 1, our main interest was not to perform a thorough analysis of the features of each system. On the contrary, we were interested in spotting how the presence/absence matrices are represented in the input files for each system, and what is the Supplementary information needed in those input files. As we will explain in Sections 3 and 4, this knowledge has been employed to, first define a new format that takes into account the requirements of the variety of systems; and, then to allow the connection of GelJ with all the tools of Table 1.

In most tools, the presence/absence matrices are encoded using a 1 to indicate the presence of a band and a 0 to indicate its absence, but there are some systems that use a different representation (e.g. Mcheza and NewHybrids); in fact, two systems (ABC4F and Bayescan) do not work directly with the presence/absence matrix but with a frequency matrix. The additional information varies from system to system, and might include some of the following information: number of loci, number of populations, number of individuals, number of individuals per population, names of loci, names of populations, names of individuals, and individuals of each population. In addition to the differences among systems presented in Table 1, the input files of each system have their own peculiarities that are not related with the data (e.g. keywords, order of the data, characters employed to separate data, and so on). This variety of formats shows the diversity of the field and the difficulty of manually transforming data across formats.