A Survey of Tools for Analysing DNA Fingerprints

Authors:

Jónathan Heras

Jónathan Heras has a PhD in Computer Science from the Universidad de La Rioja, Spain. He is currently working as a post-doc researcher at the Department of Mathematics and Computer Science, Universidad de La Rioja, Spain.

César Domínguez

César Domínguez has a PhD in Mathematics from the Universidad de La Rioja, Spain. He is currently working as a senior lecturer at the Department of Mathematics and Computer Science, Universidad de La Rioja, Spain.

Eloy Mata

Eloy Mata has a PhD in Computer Science from the Universidad de La Rioja, Spain. He is currently working as a senior lecturer at the Department of Mathematics and Computer Science, Universidad de La Rioja, Spain.

Vico Pascual

Vico Pascual has a PhD in Mathematics from the Universidad de La Rioja, Spain. She is currently working as a senior lecturer at the Department of Mathematics and Computer Science, Universidad de La Rioja, Spain.

Carmen Lozano

Carmen Lozano has a PhD in Pharmacy from the University of La Rioja, Spain. She has a contract associated with Project SAF2012-35474. She has a wide experience in different molecular typing methods among them PFGE.

Carmen Torres

Carmen Torres has a PhD in Pharmacy from the University Complutense of Madrid, Spain. She is Professor of Biochemistry and Molecular Biology of the University of La

Rioja, Spain and coordinates a research group on antimicrobial resistance and molecular bacteria epidemiology.

Myriam Zarazaga

Myriam Zarazaga has a PhD in Veterinary Sciences by the University of Zaragoza. She is Associate Professor of Biochemistry and Molecular Biology at Universidad de La Rioja, Spain. She has experience in bacterial characterization by molecular typing methods.

Corresponding author:

Jónathan Heras. Department of Mathematics and Computer Science, Universidad de La Rioja, Ed. Vives, C/ Luis de Ulloa 2, 26004, Logroño, La Rioja, Spain. Tel: +34-941299461; Fax: +34-941299460; E-mail: jonathan.heras@unirioja.es

Keywords. DNA fingerprinting tools; PFGE analysis; Gel images tools.

A Survey of Tools for Analysing DNA Fingerprints

Abstract. DNA fingerprinting is a genetic typing technique that allows the analysis of the genomic relatedness between samples, and the comparison of DNA patterns. This technique has multiple applications in different fields (medical diagnosis, forensic science, parentage testing, food industry, agriculture, and many others). An important task in molecular epidemiology of infectious diseases is the analysis and comparison of pulsed-field gel electrophoresis (PFGE) patterns. This is applied to determine the clonal diversity of bacteria in the follow-up of outbreaks or for tracking specific clones of special relevance. The resulting images produced by DNA fingerprinting are sometimes difficult to interpret, and multiple tools have been developed to simplify this task. In this paper, we present a survey of tools for analysing DNA fingerprints. In particular, we compare **33** tools using a set of predefined criteria. The comparison was carried out by hands-on experiences – whenever possible – and inspecting the documentation of the tools. Since no system is preferred in all the possible scenarios, we have created a spreadsheet that can be customised by researchers to determine the best system for their needs.

1. Introduction

Bioinformatics is fundamental to analyse, process, and understand the huge amount of biological data obtained with the use and development of the new technologies of molecular biology. This new multidisciplinary field gathers knowledge of different areas, such as biology, computer science, genetic, physics, and mathematics among others. There are many informatics tools designed to facilitate the study and annotation of genomes and the analysis of their expressions [1, 2], as well as to predict and identify particular sites, such as promoters [3], or splicing sites [4]; some of them developed as web-servers [5—7]. On the other hand, there are other tools for the analysis and processing of images which have also suffered an important advance, and have many applications in the biological field among others. In this sense, one technique of high utility is DNA fingerprinting: a genetic typing technique that allows the analysis of the genomic relatedness between samples, and the comparison of DNA patterns. This technique has multiple applications in different fields (medical diagnosis, forensic science, parentage testing, food industry and agriculture, just to name a few). An important task in molecular epidemiology of infectious diseases consists in analysing the genomic relatedness between bacterial clinical isolates. For this purpose, there are different molecular methods such as plasmid fingerprinting, ribotyping, amplified fragment length polymorphism (AFLP), random amplified polymorphic DNA (RAPD), restriction fragment length polymorphism (RFLP), rep-PCR, simple sequence repeats (SSR), or pulsed-field gel electrophoresis (PFGE). Regarding the latter, it enables the separation of large DNA molecules in an agarose gel matrix by applying an electric field that periodically changes direction. Due to its high discriminatory power, it is most useful for differentiating closely linked strains [8], and to determine the clonal diversity of bacteria in the follow-up of outbreaks, or for tracking specific clones of special relevance [9, 10]. **This technique is considered as the gold-standard approach for molecular epidemiological investigations.**

The interpretation of banding patterns by visual observation can be sometimes complicated, especially when comparing patterns that are distant, and it can be highly dependent on the researcher who reads them. There are multiple software tools which can help to simplify this task as well as to eliminate the possible suggestibility derived of the human eye. Namely, we can find several systems that allow the researcher to analyse lanes with a high amount of bands and to represent the results as dendograms.

In spite of the importance of DNA fingerprinting, and the considerable amount of tools that are currently available; there is not, at least as far as we are aware, a thorough comparison of the features included in each tool. Two small surveys were presented in [11] and [12] comparing, respectively, 3 and 2 tools for a particular case-study – from the tools studied in these surveys, only one tool is currently maintained. Additionally, a small comparison considering 4 criteria for 6 tools was presented in [13].

In this paper, we have surveyed the functionalities supported by several tools to analyse DNA fingerprint images (from now on gel-images). The workflow to process gel-images is summarised in Figure 1; and, the stages of this procedure have been the basis to define the criteria evaluated in our survey. Let us briefly explain each stage.

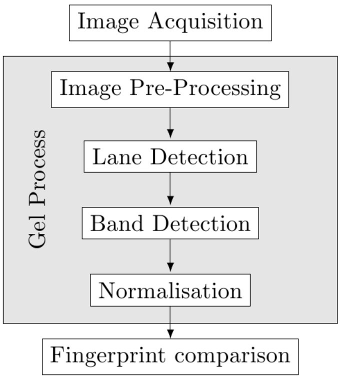

Figure 1. Workflow to analyse DNA fingerprint images.

After the acquisition of the gel-image, such an image is pre-processed. As a general principle, gel-images should remain as close as possible to the original acquired data. However, there are some attributes (e.g. the brightness and contrast of the image) that can be changed to increase the quality of the image, and to facilitate its analysis (see Step (1) of Figure 2).

In the second stage, the *lanes* (or *gelstrips*) of the image are detected (see Step (2) of Figure 2). In the literature, we can find a vast number of methods for automatic lane detection (see, for instance, [14—23]). The common idea of these methods is the construction of a *vertical densitometric-curve* (or *histogram*) averaging the pixel values on the same vertical line. In the densitometric curve, the local minima correspond to the gap between the lanes, and this fact is used to detect the lanes of the image.

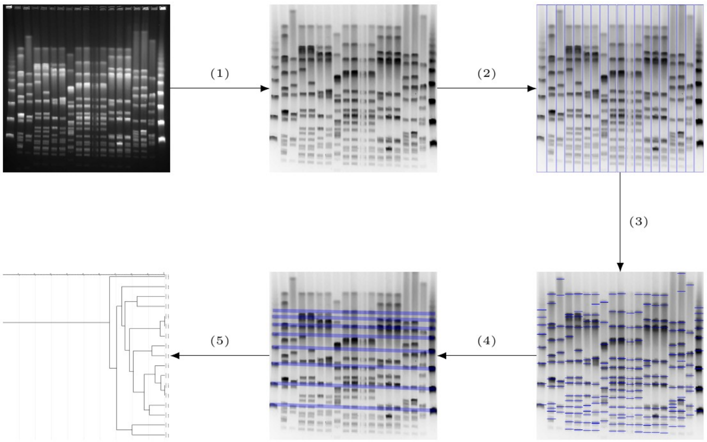

Figure 2. Example of the workflow to analyse DNA fingerprints. (1) The original image is rotated, cropped, and the colours are inverted. (2) Lane-Detection. (3) Band-Detection. (4) Normalisation using the first and last lane as reference lanes **(markers could be located in other positions of the gel)**. (5) Dendrogram generation.

Once the lanes of the gel-image have been defined, the third stage of the procedure consists in finding the bands in each lane (see Step (3) of Figure 2). The process to locate bands is almost analogous to the detection of lanes: a *horizontal densitometric-curve* is computed from each lane, and the local maxima in that curve indicate the position of the bands. Different variations of this method have been studied in the literature [14, 16, 17, 19, 24—26].

The next stage is the normalisation phase (see Step (4) of Figure 2). This step is required to compare banding patterns within the same gel – the band-positions of a lane are influenced by experimental conditions – and to compare banding patterns from different gels. Normalisation is achieved thanks to the use of reference lanes in the same gel in which the different strains are running. The utilisation of known reference positions of the reference lanes allows the researcher to normalise the rest of the lanes in the gel. A detailed description of the normalisation process can be found in [27].

The last step is the comparison of the similarity among the different lanes. Different methods exist to compute such a similarity using either densitometric curves [27—29] or band positions [30—32]. From the similarity among the lanes, a similarity matrix is constructed, and in turn, such a matrix is used to graphically represent the relatedness among lanes using a dendrogram [33] (see Step (5) of Figure 2).

These five stages (image pre-processing, lane detection, band detection, normalisation and fingerprint comparison) are implemented in most of the tools for DNA fingerprint analysis; hence, they have been the basis to define the evaluation-criteria for the survey presented in this paper.

Outline. The rest of the paper is organised as follows. In Section 2, we present the evaluation criteria and our evaluation method. A description of the obtained results is provided in Section 3, followed by a discussion in Section 4. The paper ends with the conclusions and the bibliography.

2. Materials and methods

2.1. Selection of tools

We screened PubMed Central and Google Scholar looking for corpora publications, and used the Google search-engine to create a list of tools specialised in analysing DNA fingerprints – the search-strategy that we have followed is described in Appendix A.1. **This search produced 33 tools (see Table 1). We have evaluated these tools using the criteria described as follows.**

Tool	Free	Operating system	Year (last release)	Fingerprint comparison
Advanced Quantifier [34]	No	Win/Mac	2012	Yes
BioDocAnalyze [35]	No	Win	2011	Yes
Dolphin 1D [36]	No	Win	2006	No
EzQuant [37]	No	Win	2005	No
Gel plugin ImageJ [38]	Yes	All	2014	No
Gel-Pro Analyzer [39]	No	Win	2011	No
Gel-Quant [40]	No	Win	2014	No
GelAnalyzer [41]	Yes	All	2010	No
Gelclust [13]	Yes	Win	2013	Yes
GelComparII [42]	No	Win	2013	Yes
GelQuant Pro [43]	No	Win	2013	Yes
GelQuant.Net [44]	Yes	Win	2011	No
gelQuest [45]	No	Win	2010	Yes
GelScan [46]	No	Win	2007	Yes
GeneTools [47]	No	Win/Mac	2013	Yes
Image [48]	Yes	Linux	2000	No
ImageLab [49]	No	Win/Mac	2014	No
ImageQuant [50]	No	Win	2011	Yes
ImageStudio [51]	No	Win/Mac	2014	No
Intelligent Quantifier [52]	No	Win/Mac	2011	No
Jelmarker [53]	No	Win/Mac	2010	Yes
LabImage [54]	No	All	2014	No
Laneruler [22]	Yes	All	2007	No
Logger Pro [55]	No	All	2014	No
Molecular Imaging Software [56]	No	Win/Mac	2012	No
myImageAnalysis [57]	No	Win	2012	No
Phoretix 1D Pro [58]	No	Win	2013	Yes
PyElph [19]	Yes	All	2013	Yes
Quantity One [59]	No	Win/Mac	2008	Yes
TotalLab [60]	No	Win	2013	Yes
Ultraquant [61]	No	Win	2014	No
Un-Scan-it [62]	No	Win/Mac	2013	No
VisionWorks [63]	No	Win	2014	Yes

Table 1. Surveyed tools and some of their features.

2.2. Evaluation Criteria

Based on the workflow to process gel images depicted in Figure 1, and on discussions with experts in the subject; we have split the evaluation criteria into 5 categories:

C.1. *Image pre-processing*. In this category, we review the available options to edit (e.g. crop, rotate, or flip), and enhance the quality (e.g. adjust the contrast and brightness, or perform gamma correction) of gel-images.

C.2. *Lane detection*. The criteria in this category are related to the options that wrap the **automatic**-detection of lanes in a gel-image. **For instance, if it is possible to add and delete lanes manually, modify the detected lanes (e.g. adjust their thickness and position), or whether the detected lanes can be curved.**

C.3. *Band detection*. Analogously to the criteria in Category **C.2**, we are interested in the options offered to locate bands, and not in surveying the algorithms employed for automatic band-detection.

C.4. *Normalisation*. This category gathers the functionality featured for normalising gel-images.

C.5. *Fingerprint comparison*. In this category, we investigate the methods supported by the different systems for comparing fingerprints; namely, the computation of similarity matrices, and the construction of dendrograms. Additionally, we study how the dendrograms are presented to the user.

We have also included other 2 categories: **C.0.** *General features* (basic information about the tools; for instance, year of last release or whether the software is free), and **C.6.** *Additional features* (functionality that is not necessary for the processing of gel images; for instance, database storage or the generation of reports). For these 7 categories, we have fixed a total of **44** criteria. The list of those criteria is provided in Table 2 – a more detailed description of each criterion can be seen in Appendix B.

Category	Criteria	Code
General	Free	C.0.1
	Demo	C.0.2
	Operating System	C.0.3
	Year	C.0.4
	Format support	C.0.5
	Hardware requirements	C.0.6
	PubMed Central cites	C.0.7
	Google Scholar cites	C.0.8
Pre-processing	Image Editing	C.1.1
	Contrast/Brightness	C.1.2
	Image Histogram	C.1.3
	Gamma correction	C.1.4
	Background subtraction	C.1.5
	Filtering	C.1.6
Lane detection	Lane creation	C.2.1
	Number of lanes	C.2.2
	Add/Delete lanes	C.2.3
	Lane edition	C.2.4
	Curved lanes	C.2.5
	Different thickness	C.2.6
	Background subtraction	C.2.7
Band detection	Band creation	C.3.1
	Add/Delete bands	C.3.2
	Threshold	C.3.3
	Histogram display	C.3.4
Normalisation	Band matching	C.4.1
	Band matching several images	C.4.2
	rf-lines	C.4.3
	Loading standards	C.4.4
	Methods	C.4.5
	Tolerance	C.4.6
Fingerprint Comparison	Comparison support	C.5.1
	Similarity methods	C.5.2
	Dendrogram methods	C.5.3
	Dendrogram output	C.5.4
	Similarity matrices	C.5.5
Additional features	Database	C.6.1
	Reports	C.6.2
	Save	C.6.3
	Export	C.6.4
	Smiling	C.6.5
	3D	C.6.6
	Annotation	C.6.7
	GLP/CFR 21 Part11 compliance	C.6.8

Table 2. List of criteria used for the evaluation of tools.

3. Results

In this section, we present an overview of the results obtained for the different criteria presented in the previous section – the complete evaluation for each category is given in Tables 12—20 of Appendix C.

3.1. General features

Most of the systems included in this survey are commercial tools **(26 out of 33)**, but they usually provide a demo version that is limited to either a number of usage-days or fixed images – only 5 of the commercial tools do not offer a demo version. **All but one of the tools are** available for the Windows operating system, **15** for Mac, and **7** for Linux. Additionally, we can notice that there is an interest in developing software tools for DNA fingerprint analysis – most of the tools (**26 out of 33**) were released in the last 5 years.

Another important criterion is the format of the images that can be processed by the different systems. All but one tool work with images in a standard format (e.g. tiff or jpeg). The tiff format – a widely used format for gel-images, and, in general, biological images – is accepted by the **87**% of the tools.

In this category, we also consider as criterion the minimum hardware requirements that the surveyed tools need to work correctly. In general, these tools can be run without problems in any basic computer.

Finally, the last criterion considered in this category is the number of citations in PubMed Central and Google Scholar. For this criterion, the commercial tools clearly overcome the free tools; namely, the systems with the highest number of citations are: ImageQuant (10595 cites in Pubmed Central and 25500 in Google Scholar), GelComparII (**10307** cites in Pubmed Central and 11120 in Google Scholar), and Quantity One (8300 cites in Pubmed Central and 23400 in Google Scholar) – **the searches on Pubmed Central and Google Scholar were carried out on 17th December 2014**. In the case of free tools, GelAnalyzer (19 cites in Pubmed Central and 107 in Google Scholar) is the most cited tool (but far from the number of citations of commercial systems).

3.2. Pre-processing

Before analysing gel-images, researchers tend to apply some safe image-transformations to simplify their work. The most common transformations are cropping (to select the region of interest of the image), flipping and rotating (to adjust the position of the gel), and inverting the colours (depending on user's preference to work with images with light- or dark-background). These transformations can be applied using general imaging software (e.g. Photoshop [64] or GIMP [65]); however, from the user point of view, it is simpler if that functionality is integrated in the tool used for analysing the image. From the **22** tools that allow those image transformations, all of them can either rotate or flip the image, **75**% allow cropping, and 33% can invert the colours of the image.

In general, it is recommended to optimise the contrast and brightness of the image [27]. This task can be carried out either manually, or automatically using different optimisation algorithms (e.g. linear or logarithmic) – **21** tools offer the functionality to adjust manually the contrast and brightness, and **13** of them can perform this task automatically. In addition, the user can enhance the quality of the images applying gamma correction – a technique employed to adjust the lightness of the image [66] – in **10** of the surveyed tools.

Finally, gel-images might contain noise produced during the acquisition stage. This noise can be removed using background subtraction techniques (e.g. the rolling ball mechanism [66]), or using filtering methods (e.g. median or average filters [66]). Background subtraction methods correct local background differences and are available in **8** of the surveyed tools. Several filtering methods are provided by **16** of the tools in order to remove “salt-and-pepper” noise or sharpen the bands.

The most complete systems regarding this category are Molecular Imaging Software, and **the gel plugin of ImageJ** which offer all the pre-processing options previously explained.

3.3. Lane detection

The majority of the surveyed tools (**84**%) can either automatically or semi-automatically detect the lanes of a gel-image – some of them (5 to be more precise) require as input the number of lanes in the image to obtain a more accurate result. Since the precision of the lane-detection step influences the rest of the process, the functionality that serves to manually adjust the detected-lanes is essential. The basic functionality to edit lanes is related to the addition and removal of lanes (supported by **27 out of the 33 surveyed** tools), and the modification of thickness and position (this functionality is available in **24 out of the 33 tools**).

The lanes of gel-images do not usually run completely straight; therefore, it is an important issue whether the detected-lanes can be curved, and whether they can have different thickness. If these options are not supported, the detected lanes might either include irrelevant information, or lose some relevant part of the lane. The functionality to manage lanes with different thickness is implemented in most of the tools (**84**%); on the contrary, less than half of the tools (36%) can work with curved lanes.

Once that the lanes have been detected, and before continuing with the analysis of the gel-image, it might be useful to enhance the quality of the lanes subtracting their background. As this operation is less computing-intensive than subtracting the background of the whole image, tools usually provide this functionality – 22 tools can subtract the background from lanes, 7 of them also support the background subtraction from the whole image, and **only the gel plugin of ImageJ provides the option of subtracting the background from the whole image but not from the lanes**.

There are several outstanding software tools for the lane-detection task; namely, GelComparII, Gel-Pro Analyzer, TotalLab, Phoretix 1D Pro, Quantity One, ImageQuant, GelQuant Pro, VisionWorks, Molecular Imaging Software, and LabImage provide all the functionality evaluated in this category.

3.4. Band detection

Analogously to the detection of lanes, the majority of the tools (**84**%) automatically locate the bands in a gel-image – note that automatic band detection is more relevant than automatic lane detection since a gel-image contains just a few lanes, but it might contain dozens or even hundreds of bands.

Roughly speaking, the procedure to locate the bands of a lane consists in finding the local peaks of the densitometric curve associated with such a lane (see Figure 3). Some of the local peaks come from noise and are excluded by the algorithm using a height criterion (see the dotted square in Figure 3); however this threshold can also exclude low-intensity bands (see the non-dotted square in Figure 3). The optimum threshold-height varies from image to image, and the user can take advantage of tools that allow her to modify this parameter (a functionality provided by **69**% of the tools).

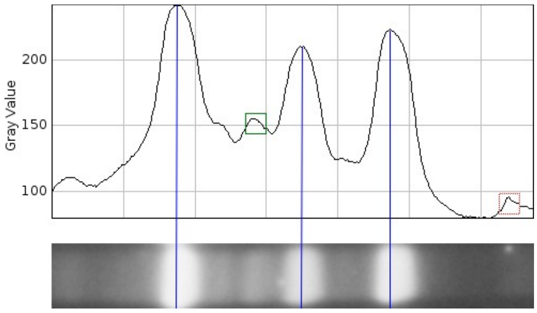

Figure 3. Densitometric curve from a lane of a gel-image. The vertical lanes indicate the bands located from the peaks, the dotted square is a local peak coming from noise, and the non-dotted square is a peak that comes from an uncertain band.

Even if the user can modify the threshold-height, it is usually necessary to add and remove bands manually (an instrumental functionality included in **29 out of the 33** surveyed tools). Some uncertainties might arise during the manual picking of bands; in those situations, the user can inspect the densitometric curve to decide about the inclusion of concrete bands – the densitometric curves are shown by **82**% of the tools, and only 3 tools that support band picking do not include this “hint” for the user.

In this category, 23 out of 33 tools support all the studied options. This illustrates that the detection of bands is the most important step in the analysis of gel-images; and, therefore, software tools try to simplify this task as much as possible.

3.5. Normalisation

Normalisation among gels is achieved by introducing reference lanes that contain known DNA fingerprint patterns (*reference markers*). A reference marker consists of a set of band positions together with a physical property (mainly, the molecular weight) of each band of such a set. For example, in PFGE, these reference lanes can consist in commercial molecular markers (such as Lambda Ladder PFG Marker, Middle Range PFG Marker or Low Range PFG Marker) or reference strains (e.g. Salmonella enterica Braenderup H9812). From the reference marker, the molecular weight of each band in the gel can be computed. This computation requires two interpolation stages: (1) a vertical interpolation within a reference lane serves to derive a migration model; and, (2) a horizontal interpolation is carried out to calculate the shift in each position of the non-reference lanes that fall between the reference lanes.

Matching bands within the same gel using a reference marker is a feature included in most of the surveyed tools (**25 out of 33**); however, matching bands across multiple gels is only available in **7** tools – note that the later feature requires database-storage support. All but one of the tools supporting band-matching (i.e. **24** tools) provide the functionality to load and save reference markers for further use – this reduces the burden of introducing the molecular weight of the bands manually each time that the normalisation step is required.

Several interpolation methods, both linear [67] and non-linear [27], can be applied in the two-stages of the normalisation process. The most common migration models are linear, logarithmic, and cubic spline; and, the user is in charge of choosing the most suited model for her concrete problem.

Usually, two bands are matched even if their molecular weights are not exactly the same, but they are close enough. This “closeness” value is obtained from a tolerance that is either fixed or can be modified by the user. In the latter case, the user has more control over the results – this functionality is provided by **16 out of the 25** tools that support band matching.

The three most complete tools in this category are: GelComparII, Gel-Pro Analyzer, and Phoretix1D-Pro.

3.6. Fingerprint comparison

Not all the tools surveyed in this paper can be used to compare fingerprints; namely, **15** tools provide this functionality (see the column “Fingerprint comparison” in Table 1). The process to compare fingerprints consists of two steps: the computation of similarity matrices, and the construction of dendrograms [27].

Given a list of *n* lanes *L*, the similarity matrix of *L* is an $n \times n$ matrix where the element of row *i* and column *j* encodes the distance between the *i*-th and *j*-th lane of *L*. There are two approaches to calculate the similarity between lanes: band-based and curve-based [27] – a search in PubMed (see Appendix A.3) shows that both approaches are equally used in the literature. In the former approach, the similarity between two lanes is calculated as a coefficient based on the number of matching and non-matching bands. In the latter approach, the similarity is determined using a correlation coefficient computed from the densitometric curves of the lanes. In both cases, different coefficients can be used. The most-common band-based coefficients used in the literature are: DICE [30] (72%), Jaccard [31] (10%), Ochiai [32] (8%), and Band difference (8%); and, the most-used curve-based coefficients are: Pearson coefficient [28] (75%), Euclidean distance [29] (18%), and cosine correlation [27] (6%). All the surveyed tools that allow the researcher to compare DNA fingerprints can compute similarity matrices using, at least, a band-based coefficient; but, only 7 of them work with curve-based coefficients. The two most-used coefficients (DICE and Pearson) are available in all but one of the tools that work, respectively, with band-based and curve-based coefficients.

The similarity matrices are fed as input to hierarchical clustering algorithms [68]. These algorithms are employed to visualise the relations among fingerprints using either a *dendrogram* or a *tree*. The main algorithms employed in the literature are UPGMA (26%), single linkage (18%), neighbour joining (16%), complete linkage (11%), Ward (8%), maximum linkage (7%), and minimum linkage (6%) – the parameters used for this literature-search can be found in Appendix A.3. The **15** tools support several methods for cluster analysis, and the UPGMA algorithm is implemented in all of them.

The generated dendrograms can include additional information like the images of each lane (supported by **8 out of 15 tools**), the band positions (available in **3 tools**) or an overlapping of images and bands (supported by **3 tools**). Additionally, **8** of the tools that generate dendrograms can also display the similarity matrices. On the contrary to dendrograms that provide an overview of the relatedness among the studied lanes, the similarity matrices can be used to inspect the concrete relation (a numerical value) between two lanes.

In this category, the “best” system is GelComparII since it offers the most used methods both for the computation of distance matrices and for the construction of dendrograms. If we focus only on the computation of distance matrices, gelQuest is the most complete tool, offering a wide variety of methods. In the case of dendrogram construction, GelComparII, Quantity One and VisionWorks are the systems that support the most common methods applied in the literature. Finally, from the output point of view, GelComparII is the only system that includes all the evaluated criteria.

3.7. Additional features

The features included in this category are not strictly necessary to analyse gel-images, but they improve the user experience. We include some figures about that functionality:

- **78**% of the tools can load unfinished studies previously saved, instead of starting from-scratch every time that the user wants to analyse an image.
- **78**% of the tools can export the results (similarity matrices, molecular weights, and so on) to a spreadsheet format (e.g. Excel) for further analysis.
- **66**% of the tools automatically generate reports.
- **54**% of the tools can perform smiling correction – this has the disadvantage that the images are altered, and this goes against principle of staying as close to the original data as possible.
- **48**% of the tools can be used to annotate the images.

- **39**% of the tools generate 3D models of the gels. This functionality, as the display of the densitometric curves, might help the user to decide whether to include a band in the band-detection phase.
- **33**% of the tools are compliant with the GLP/CFR 21 part 11 regulation, that ensures the integrity and quality of data.
- 24% of the tools include a database to store and compare several gels.

There are only 2 systems that include all these features: Gel-Pro Analyzer and Phoretix 1D Pro.

4. Discussion

In the previous section, we have performed an objective study of several tools for DNA fingerprint data analysis. This study might help researchers to decide the best tool for their needs. Such a decision usually depends on several factors (e.g. the quality of the acquired images, experience of the researchers using software tools, or their current budget); and, hence, there is not a preferred system for all the possible scenarios. In order to facilitate the decision process, we have created a **set of tables summarising the advantages and disadvantages of each tool (see Appendix D), and a** customisable spreadsheet (see the supplementary materials) that allows researchers to adjust the weight of each criterion to their needs. In this section, we use that spreadsheet to determine the best tool for 4 different scenarios. We finish this section with a comparison between commercial and free systems.

4.1. Case study 1: The most complete tool

As a first case study, we are interested in discovering the most-complete tool included in our survey. In particular, if we only study the “yes/no” criteria included in Table 2 (and detailed in the tables of the appendices), we discover that GelComparII is the most-complete tool.

Considering the most complete tools for each category, the most complete programs for image pre-processing (Category C.1.) are Molecular Imaging Software and the gel plugin of ImageJ that offer more options (e.g. filtering, background subtraction or gamma correction) than the rest of the systems. There are several outstanding software tools for the lane-detection and band-detection tasks (Categories C.2. and C.3.); namely, 10 out of the 33 surveyed tools offer all the analysed options for lane-detection, and 23 out of the 33 inspected tools provide all the features for band-detection. The three most optimised systems for normalisation (Category C.4.) are GelComparII, Gel-Pro Analyzer, and Phoretix 1D-Pro. GelComparII is also the most complete system in Category C.5. offering several methods to compute distance matrices and construct dendrograms. Finally, Phoretix 1D Pro and Gel-Pro Analyzer excel at Category C.6. supporting all the surveyed advanced features.

4.2. Case study 2: The most automatic tool

In this second scenario, we suppose that researchers work with “perfect” images – i.e. high-quality images without noise, with straight lanes, and well-differentiated bands. In this situation, the most suitable tool will be the system that requires less user-intervention. Therefore, the marking-scheme of this case-study rewards the systems that offer automatic processing.

There are 5 tools that are more automatic than the rest: GelComparII, GelQuant Pro, ImageQuant, Phoretix 1D Pro and TotalLab. All those tools are commercial systems;

however, all of them offer a fully-functional demo version. Hence, we can evaluate them counting the number of “clicks” that are necessary to process a “perfect” image (i.e. complete the workflow presented in Figure 1). After the hands-on evaluation, we can conclude that GelQuant Pro is the most-automatic tool. In addition to the automatic options for each step that are available in the other 4 tools, GelQuant Pro offers an option to automatically analyse gel-images based on pre-defined protocols.

4.3. Case study 3: The best tool for low-quality images

In general, the quality of gel-images varies from experiment to experiment, and, low-quality images sometimes arise. For those images, the most suited system is not a fully-automatic tool, but a tool that implements several image-editing options, is highly customisable (allowing the user to adjust several parameters), and helps the user to take decisions.

Taking those parameters into account, the best two systems for processing low-quality images are: Molecular Imaging Software and GelComparII. The former supports more options for pre-processing images, and the latter allows a better adjustment in the normalisation phase – they are equally good for handling lanes and bands. The disadvantage of Molecular Imaging Software is that it does not generate dendrograms.

4.4. Case study 4: The best tool for PFGE analysis

In our last case study, we consider a scenario where the researcher wants to determine the best tool for PFGE analysis. A good system for PFGE analysis should detect accurately both lanes and bands (this might require some user-intervention like the selection of missing bands), be precise in the normalisation process (supporting several options), and offer the most-used algorithms for computing similarity matrices and constructing dendrograms. Additionally, the best tools should compare not only fingerprints from one gel-image, but from several images. GelComparII and Phoretix 1D Pro fully satisfy the above requirements and can be considered the best tools for PFGE analysis.

4.5. Commercial vs. free tools

In the above case studies, the best tools are always commercial systems. In fact, in the classifications for those case studies, free-tools appear either at the bottom of the third quartile, or in the last quartile (PyElph is the best free-system in Case studies 1, 3—4; and, GelAnalyzer is the best free-tool in Case Study 2). This does not mean that free tools are not useful for fingerprint analysis, but that they offer less functionality than commercial systems.

In general, free-tools implement the basic functionality for analysing gel-images. However, they lack features to obtain more accurate results (e.g. they do not handle curved lanes), offer less options (for instance, there is not any free-tool working with curve-based similarity-matrices), and do not include advanced features (e.g. database support or 3D visualisation).

5. Conclusions

In this paper, we have surveyed different tools for analysing DNA fingerprint data using several criteria. The requirements for the analysis of gel-images vary from researcher to researcher; and, there is not a best tool for all the possible scenarios. Therefore, our survey does not pick a tool, but offers an overview of the available systems and their features to researchers.

As a by-product of this work, we have created a dynamic survey (in the form of a customisable spreadsheet) that can be adjusted by researchers to determine the most suited tool for their actual needs.

Funding

This work was supported by the Agencia de Desarrollo Económico de La Rioja (ADER) [2013-I-IDD-00123].

Key points

- DNA fingerprinting is a genetic typing technique applied in a wide variety of contexts.
- Analysis of DNA fingerprint just by visual observation is a complex and subjective task.
- Several commercial and freely-available tools have been developed to deal with the analysis of DNA fingerprints.
- Freely-available tools provide just the basic functionality, commercial tools enhance that basic functionality with features that improve the precision of studies and the user-experience.
- There is not a best tool for all the possible scenarios, since this decision depends on several factors.

References

1. Yu U, Lee S-H, Kim Y-J et al. Bioinformatics in the post-genome era. J. Biochem. Mol Biol. 2004; 37(1):75—82.

2. Ekblom R and Wolf J-B. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 2014; 7(9):1026—1042.

3. Lin H, Deng E-Z, and Ding H. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 2014; 42:12961—12972.

4. Chen W, Feng P-M, and Lin H. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed Research International 2014; 2014:623149.

5. Chen W, Lei T-Y, and Jin D-C. PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Anal. Biochem. 2014; 456:53—60.

6. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/ (17 December 2014, date last accessed).

7. The European Bioinformatics Institute. http://www.ebi.ac.uk/services (17 December 2014, date last accessed).

8. Goering R-V. Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease. Infection, Genetics and Evolution 2010; 10(7):866—875.

9. Ruiz E, Rojo-Bezares B, Sáenz Y et al. Outbreak caused by a multiresistant Klebsiella pneumoniae strain of new sequence type ST341 carrying new genetic environments of aac(6')-Ib-cr and qnrS1 genes in a neonatal intensive care unit in Spain. International Journal of Medical Microbiology 2010; 300(7):464—469.

10. Lópen M, Rezusta A, Seral C et al. Detection and characterization of a ST6 clone of vanB2-Enterococcus faecalis from three different hospitals in Spain. European Journal of Clinical Microbiology & Infectious Diseases 2012; 31(3):257—260.

11. Gerner-Smidt P, Graves L-M, Hunter S et al. Computerized Analysis of Restriction Fragment Length Polymorphism Patterns: Comparative Evaluation of Two Commercial Software Packages. Journal of Clinical Microbiology 1998;36(5):1318—1323.

12. Rementeria A, Gallego L, Quindós G et al. Comparative evaluation of three commercial software packages for analysis of DNA polymorphism patterns. Clinical Microbiology and Infection 2001; 7(6):331—336.

13. Khakabimamaghani S, Najafi A, Ranjbar R et al. GelClust: A software tool for gel electrophoresis images analysis and dendrogram generation. Computer Methods and Programs in Biomedicine 2013; 111(2):512—518.

14. Bailey D-G and Christie B-C. Processing of DNA and Protein Electrophoresis Gels by Image Analysis. In: 2nd New Zealand Conference on Image and Vision Computing 1994; pages 2.2.1—2.2.8.

15. Labyed Y, Kaabouch N, Schultz R-R et al. An Improved 1-D Gel Electrophoresis Image Analysis System. In: Advances in Computational Biology, volume 680 of Advances in Experimental Medicine and Biology 2010; pages 609—617.

16. Labyed Y, Kaabouch N, Schultz R-R et al. Automatic segmentation and band detection of protein images based on the standard deviation profile and its derivative. In: IEEE International Conference on Electro/Information Technology 2007; pages 577—582.

17. Wang D, Keller J, and Carson C. Pulsed-Field Gel Electrophoresis Pattern Recognition of Bacterial DNA: A Systemic Approach. Pattern Analysis & Applications 2001; 4:244—255.

18. Chan Y-K, Guo S-W, Cheng H-M et al. Automatic band detection of 1D-gel images. In: International Conference on Electronics, Communications and Control 2011; pages 3586—3589.

19. Pavel A and Vasile C. PyElph – a software tool for gel images analysis and phylogenetics. BMC Bioinformatics 2012; 13(9).

20. Cheol Park S and Lee S. Lanes detection in pcr gel electrophoresis images. In: 11th IEEE International Conference on Computer and Information Technology 2011; pages 306—313.

21. Machado A, Campos M, Siqueira A et al. An Iterative Algorithm for Segmenting Lanes in Gel Electrophoresis Images. In: 10th Brazilian Symposium of Computer Graphics and Image Processing 1997; pages 140—146.

22. Wong R, Flibotte S, Corbett R et al. LaneRuler: automated lane tracking for DNA electrophoresis gel images. IEEE Transactions on Automation Science and Engineering 2010; 7:706—708.

23. Moreira B, Sousa A, Mendonça A et al. Automatic Lane Segmentation in TLC Images Using the Continuous Wavelet Transform. Computational and Mathematical Methods in Medicine 2013:id=218415.

24. Skutkova H, Vitek M, Krizkova S et al. Preprocessing and Classification of Electrophoresis Gel Images Using Dynamic Time Warping. International Journal of Electrochemical Science 2013; 8:1609—1622.

25. Lee J, Huang C, Wang N et al. Automatic DNA sequencing for electrophoresis gels using image processing algorithms. Journal of Biomedical Science and Engineering 2011; 4:523—528.

26. Bajla I, Holländer I, and Burg K. Improvement of Electrophoretic Gel Image Analysis. Measurement Science Review 2001;1(1):5—10.

27. Vauterin L and Vauterin P. Molecular Identification, Systematics, and Population Structure of Prokaryotes, chapter Integrated Databasing and Analysis. Springer-Verlag 2006.

28. Pearson K. Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 1895; 58:240—242.

29. Fullaondo A, Vicario A, Aguirre A et al. Quantitative analysis of two-dimensional gel electrophoresis protein patterns: a method for studying genetic relationships among Globodera pallida populations. Heredity 2001; 87:266—272.

30. Dice L-R. Measures of the Amount of Ecologic Association Between Species. Ecology 1945; 26(3):297—302.

31. Jaccard P. Nouvelles recherches sur la distribution orale. Bulletin de la Société vaudoise des sciences naturelles 1908; 44:223—270.

32. Ochiai A. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring region. Bulletin of the Japanese Society for the Science of Fish 1957; 22:526—530.

33. Anderberg M-R. Cluster Analysis for Applications. Academic Press, 1973.

34. Bio Image Systems, Inc. Advanced quantifier version 4.3.4, 2012. http://bioimage.net/AQ.htm (17 December 2014, date last accessed).

35. Biometra. BioDocAnalyze (BDA) Software, 2011. http://www.biometra.de/1163.0.html (17 December 2014, date last accessed).

36. Wealtec. Dolphin-1D software version 2.4, 2006. http://www.wealtec.com/products/imaging/software/dolphin-1d-software.htm (17 December 2014, date last accessed).

37. EZQuant Biology Software Solutions. EZQuant-Gel version 2.2, 2005.

http://www.ezquant.com/en/products/ezquant-gel/ (17 December 2014, date last accessed).

38. ImageJ team. Gel Quantification Analysis for ImageJ, 2014. http://imagejdocu.tudor.lu/doku.php?id=video:analysis:gel_quantification_analysis (17 December 2014, date last accessed).

39. MediaCybernetics. Gel-Pro Analyzer, 2011. http://www.mediacy.com/index.aspx?page=GelPro (17 December 2014, date last accessed).

40. AMPL Software. Gel-Quant Electrophoresis Image Analysis Software, 2014. http://www.ampl.com.au/gelquant_home.htm (17 December 2014, date last accessed).

41. Lazar I. Gelanalyzer 2010a, 2010. http://www.gelanalyzer.com/ (17 December 2014, date last accessed).

42. Applied Maths NV. GelCompar II version 6.6.11, 2013. http://www.applied-maths.com (17 December 2014, date last accessed).

43. DNR Bio-Imaging Systems Ltd. GelQuant Pro version 13, 2013. http://www.dnr-is.com/Product.asp?Par=3.19&id=81 (17 December 2014, date last accessed).

44. BiochemLab Solutions. GelQuant.NET version 1.7.8, 2011. http://biochemlabsolutions.com/GelQuantNET.html (17 December 2014, date last accessed).

45. SequentiX. Gelquest version 3.2.1, 2010. http://www.sequentix.de/gelquest/index.php (17 December 2014, date last accessed).

46. BioSciTec GmbH. Gelscan 6 Standard, 2007. http://www.bioscitec.com/produkte-bioscitec/software/gelscan-6-0/ (17 December 2014, date last accessed).

47. Syngene | a division of Synoptics Ltd. GeneTools version 4.03.05, 2013. http://www.syngene.com/genetools/ (17 December 2014, date last accessed).

48. The Wellcome Trust Sanger Institute. Image: The fingerprint image analysis system version 3.10.b, 2000. http://www.sanger.ac.uk/resources/software/image (17 December 2014, date last accessed).

49. Bio-Rad Laboratories, Inc. Image Lab Software, 2014. http://www.bio-rad.com/en-us/product/image-lab-software (17 December 2014, date last accessed).

50. GE Healthcare Life Sciences. Imagequant tl 7.0, 2010. http://www.gelifesciences.com/ (17 December 2014, date last accessed).

51. LI-COR Biosciences. Image Studio Software version 4, 2014. http://www.licor.com/bio/products/software/image_studio/index.html (17 December 2014, date last accessed).

52. Bio Image Systems, Inc. Intelligent quantifier version 3.3.4, 2011. http://bioimage.net/IQ.htm (17 December 2014, date last accessed).

53. SoftGenetics. JelMarker version 1.4, 2010. http://www.softgenetics.com/jelMarker.html (17 December 2014, date last accessed).

54. Kapelan Bio-Imaging. LabImage Platform version L360, 2014. http://www.kapelan-bioimaging.com/ (17 December 2014, date last accessed).

55. Vernier Software & Technology. Logger Pro version 3.8.7, 2014. http://www.vernier.com/biology/biotechnology/gel-electrophoresis/ (17 December 2014, date last accessed).

56. Bruker Corporation. Molecular Imaging Software, 2012. http://www.bruker.com/service/support-upgrades/software-downloads/molecular-imaging.html (17 December 2014, date last accessed).

57. Thermo Scientific Pierce Protein Biology Products. myImageAnalysis Software v2.0, 2012. http://www.piercenet.com/product/myimageanalysis-software (17 December 2014, date last accessed).

58. TotalLab Limited. Phoretix 1d pro version 11.4, 2013. http://www.totallab.com/products/1dpro/ (17 December 2014, date last accessed).

59. Bio-Rad Laboratories, Inc. Quantity One 1-D Analysis Software version 4.6.9, 2008. http://www.bio-rad.com/en-us/product/quantity-one-1-d-analysis-software (17 December 2014, date last accessed).

60. TotalLab Limited. Totallab quant version 13, 2013. http://totallab.com/products/totallabquant/ (17 December 2014, date last accessed).

61. Aplegen, Inc. UltraQuant Analysis Software version 13.11.18, 2014. http://www.aplegen.com/software.php (17 December 2014, date last accessed).

62. Silk Scientific, Inc. Un-Scan-it gel version 6.1, 2013. http://www.silkscientific.com/gel-analysis.htm (17 December 2014, date last accessed).

63. Ultra-Violet Products Ltd. VisionWorks LS Analysis Software, 2014. http://uvp.com/visionworks.html (17 December 2014, date last accessed).

64. The Adobe team. Adobe Photoshop, 2014. http://www.adobe.com/products/photoshop.html (17 December 2014, date last accessed).

65. The GIMP team. GIMP | The GNU Image Manipulation Program, 1997—2014. http://www.gimp.org/ (17 December 2014, date last accessed).

66. Burger W and Burge M-J. Principles of Digital Image Processing: Fundamental Techniques. Springer, 2009.

67. Pot B, Gillis M, Hoste B et al. Intra- and intergeneric relationships of the genus Oceanospirillum. International Journal of Systematic and Evolutionary Microbiology 1989;39(1):23—34.

68. Xu R and Wunsch D C. Clustering. IEEE Computer Society Press, 2008.