Journal of Advanced Research 18 (2019) 9–18

Contents lists available at ScienceDirect

Journal of Advanced Research

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

Original article

Advanced DNA fingerprint genotyping based on a model developed from real chip electrophoresis data

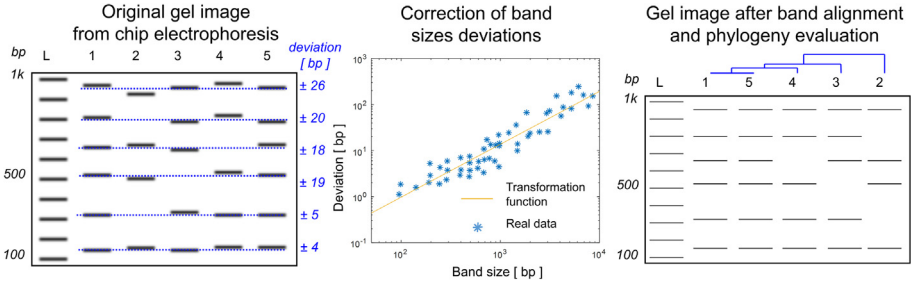
Helena Skutkova [a,*], Martin Vitek [a], Matej Bezdicek [b], Eva Brhelova [b], Martina Lengerova [b]

[a] Department of Biomedical Engineering, Brno University of Technology, Technicka 12, 616 00 Brno, Czech Republic
[b] Department of Internal Medicine, Hematology and Oncology, Masaryk University and University Hospital Brno, Cernopolni 212/9, 662 63 Brno, Czech Republic

HIGHLIGHTS

- Mapping chip electrophoresis distortion based on real data measurement.
- Determining the transformation function for the adaptive correction of band size deviation.
- Improving the ability to distinguish closely related DNA fingerprints.
- Using hierarchical clustering to adjust the global band position.
- Genotyping all DNA fingerprints from multiple runs at once.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 19 October 2018
Revised 6 January 2019
Accepted 10 January 2019
Available online 25 January 2019

Keywords:
DNA fingerprinting
Automated chip capillary electrophoresis
Genotyping
Band matching
Gel sample distortion
Pattern recognition

ABSTRACT

Large-scale comparative studies of DNA fingerprints prefer automated chip capillary electrophoresis over conventional gel planar electrophoresis due to the higher precision of the digitalization process. However, the determination of band sizes is still limited by the device resolution and sizing accuracy. Band matching, therefore, remains the key step in DNA fingerprint analysis. Most current methods evaluate only the pairwise similarity of the samples, using heuristically determined constant thresholds to evaluate the maximum allowed band size deviation; unfortunately, that approach significantly reduces the ability to distinguish between closely related samples. This study presents a new approach based on global multiple alignments of bands of all samples, with an adaptive threshold derived from the detailed migration analysis of a large number of real samples. The proposed approach allows the accurate automated analysis of DNA fingerprint similarities for extensive epidemiological studies of bacterial strains, thereby helping to prevent the spread of dangerous microbial infections.

Introduction

DNA fingerprinting methods are commonly used for typing bacterial strains, and electrophoretic separation methods are used for visualizing and evaluating the amplification results. Although standard planar electrophoresis (on an agarose gel) is still more commonly used than its automated equivalents, the popularity of modern automated chip electrophoresis is increasing, especially in the case of extensive comparative studies [1–4]. The main advantages are the elimination of the gel image digitalization process, the absence of sample distortion caused by the non-

Abbreviations: DBSCAN, density-based spatial clustering of applications with noise; DTW, dynamic time warping; ESBL, extended spectrum beta-lactamases; KLPN, *Klebsiella pneumonia*; MALDI-TOF, matrix assisted laser desorption ionization – time of flight; rep-PCR, repetitive element palindromic polymerase chain reaction; RMSE, root mean squared error; R-square, ratio of the sum of squares; SD, standard deviation; SLINK, single linkage; SSE, sum of squares due to error; UPGMA, unweighted pair group method with arithmetic mean.

Peer review under responsibility of Cairo University.

* Corresponding author.
E-mail address: skutkova@vutbr.cz (H. Skutkova).

https://doi.org/10.1016/j.jare.2019.01.005
2090-1232/

homogeneity of the electromagnetic field (smile effect), the simple adaptation of sample ranges from multiple electrophoretic runs, and the increased speed of the electrophoretic runs. Thus, the size of the DNA fragments can be obtained directly by using objective software analysis, in contrast to subjective estimates of the size from a low-quality image by a human operator. However, even automated chip electrophoresis has limited accuracy. For example, the Agilent 2100 Bioanalyzer System provides catalogue values of ±10 or ±15% sizing accuracy, depending on the kits and reagents used. The sizing resolution is also limited and dependent on the sizing range; for the DNA 7500 Kit from Agilent, the resolution is 5% in the 100–1,000 bp range and 15% in the 1,000–7,500 bp range. Thus, the resulting fragment size values are not completely accurate, and their deviation is not constant over the measured range. Although the deviation is smaller than that obtained in the subjective estimation of size from standard planar electrophoresis gel images, its existence and inconsistency still complicate subsequent comparative analyses, such as phylogeny reconstruction. The basis of these methods is the evaluation of the similarity between two sample lines (fingerprint patterns), depending on the presence/absence of bands of the same size. It is difficult to assess whether two bands are the same or belong to two different bands corresponding to various lengths of DNA fragments due to the inaccuracy in measurements. This problem has not been addressed, as evidenced by the lack of information in the literature.

The first reason is that planar electrophoresis is more commonly used than chip electrophoresis because the former is less expensive. Thus, DNA fingerprint gel images are still being analysed using tools, such as PyElph [5], GelClust [6], and GelJ [7], that focus primarily on image preprocessing tasks [8,9]. The similarity of two bands is evaluated trivially. Most often, the bands are identified as the same size if their deviation does not exceed the permitted constant threshold. The identification of bands of the same size or their alignment is generally performed using pairwise alignment. A more advanced solution can be found in the software GELect [10], where a density-based clustering method (DBSCAN) is used to identify band cluster centroids from all samples; however, it still uses a heuristically set constant threshold. Moreover, another decision parameter, the minimum number of samples containing bands, causes incorrect classification of unique samples. Another way to adapt band positions in gel images obtained from classic planar electrophoresis is the use of the dynamic time warping (DTW) method, which adaptively re-samples 1D signal representations of particular lines [11]. This method does not use a constant threshold for band position correction but requires a complete signal representation from raw data.

The second reason for the insufficient examination of the band alignment in chip electrophoresis is that the processing of chip electrophoresis DNA fingerprinting data is almost exclusively realized through complex and expensive software platforms, such as BioNumerics (Fingerprint Data module or DiversiLab genotyping application distributed by Applied Maths NV, BioMérieux, France). These tools are copyrighted, and the principle of the methods used is not publicly available. According to the technical documentation from the company's website (http://www.applied-maths.com), the fingerprint data module uses a combination of nonlinear shift with fixed edges and global shift with linear stretch/compression for band position correction. Although the procedure is not described in detail, the shift correction is based on finding the highest correlation between samples. Since correlation describes the degree of linear dependence, correlation is expected between the deviation and band size. However, it can be assumed that the character of the dependence is not linear, because the sample mobility on the gel is not linearly dependent on band size.

In this study, a new method for the global alignment of the band positions using an adaptive threshold is presented. For this purpose, a large number of DNA weight markers were measured to confirm that the dependence between band size deviation (shift) and band size (band position) is not constant or linear. Based on these measurements, an empirical model of band size deviation was derived, which serves as a transformation function that adapts band size deviation to an approximately constant value across the measured range. It enables the use of hierarchical cluster analysis with one fixed threshold to identify bands of the same size in all samples without a pre-defined number of clusters or of objects in the clusters. The identification accuracy of the same bands was also verified on DNA weight markers, where the correct band size values are known. The designed method was finally tested on the study of the repetitive element palindromic polymerase chain reaction (rep-PCR) genotyping of 60 bacterial strains and comparison with the standard professional tool, the fingerprint data module in BioNumerics.

Material and methods

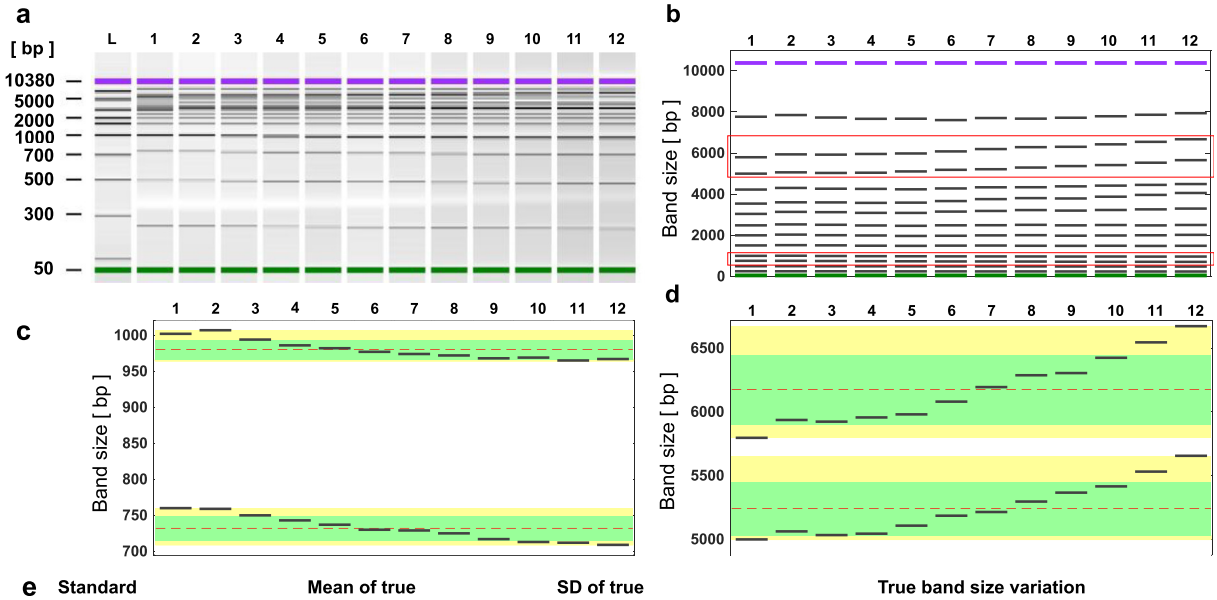
Problem description

The principle of the method for the global detection of the same size bands in all gel samples is composed of two key steps. The first step is the removal of the nonlinear dependence of band size deviations on the band size range. Samples with known DNA fragment sizes were used to describe true accuracy in band size determination. DNA weight markers (ladders) appeared to be appropriate for that purpose. However, during the first measurement of one ladder type (12 samples of GeneRuler 1 kb DNA Ladder) in one run, considerable variation was observed in sizes corresponding to the same size band (Fig. 1). A regular user may not be aware of this variance, because it is not highly noticeable in an artificial gel image with a logarithmic scale (Fig. 1a) as produced by the software supplied to the chip electrophoresis device (2100 Bioanalyzer Expert Software distributed by Agilent Technology, Inc., Santa Clara, California, USA). An illustration of the band positions in a graph with a linear scale band size axis (Fig. 1b) more clearly shows the variability of the same size bands. Detailed images of the four different band size levels (Fig. 1c, d) and their statistical evaluation (Fig. 1e) prove that the variance in band size is not constant across the whole sample range and varies even between individual samples. The measurements were performed with different ladder types (different size ranges) and with variable distributions of samples across several runs to reveal the maximum degree of band size variability.

The second step of the proposed method is global identification of the same size bands on the whole gel at once, instead of by individual local pairwise sample comparison. This step also allows us to obtain a corrected gel image (graphic representation of band sizes), where the "correct" band position is determined as the median size of the bands identified as the same. This process of positional adaptation of the same size bands in multiple samples is comparable to multiple sequence alignment [12,13], known for its application to symbolic DNA representations of protein sequences or genomic signals [14]. It is a necessary step preceding the subsequent phylogenetic analysis of biological sequences [15–17]. Therefore, global multiple alignments of band positions are a suitable step preceding the comparative analysis of gel samples, such as the genotyping of bacterial rep-PCR profiles.

Datasets

All data used in this article were obtained by chip capillary electrophoresis using the 2100 Bioanalyzer platform. All reactions were performed using the Agilent DNA 7500 kit (Agilent



e

Standard band size [bp]	Mean of true band size [bp]	SD of true band size [bp]	True band size variation [bp]	
6k	Ø 6,174.75	± 275.47	max - min 6,672 – 5,796	difference » 876
5k	Ø 5,242.25	± 212.96	max - min 5,655 – 4,999	difference » 656
1k	Ø 980.25	± 14.19	max - min 1,007 – 965	difference » 42
700	Ø 732	± 17.99	max - min 760 – 709	difference» 51

Fig. 1. Visualization of band size variance within 12 samples of GeneRuler 1 kb DNA Ladder. (**a**) Original gel image from 2100 BioAnalyzer software. (**b**) A graphical representation of band positions with a linear scale band size axis (red rectangles are enlarged for detailed analysis in images **c** and **d**). (**c**) Details of the size variance in the 750 bp and 1 k bp bands. (**d**) Details of the size variance in the 5 k bp and 6 k bp bands (the red dashed line is the mean of the same band sizes; the green area is the standard deviation (SD); and the yellow area is the maximum-minimum range). (**e**) Statistical description of band size variance from detailed images **c** and **d**.

Technology, Santa Clara, California, USA) with the manufacturer's default settings. The results were analysed using the 2100 Expert software. The input data for the proposed method are the sizes of the bands in each sample, determined by the device-supplied software with the default settings.

The DNA weight markers were measured 120 times in ten runs for the set up and validation of the proposed method. Four different types of DNA ladders were used to evaluate the band size deviation variability across the whole band size range of the Agilent DNA 7500 kit. The measurements were carried out by two different operators across five days. The samples of each ladder type were separated into multiple runs and randomly combined within one run. The samples were measured at two concentration levels, 12.5 and 25 ng/µl, to ensure the maximum possible variability in the standardized measurement and to enable the determination of the real-time measurement error in the whole range. The ladder types used and the measurement parameters are summarized in Table 1.

For the validation of the proposed method, 60 isolates from 12 extended-spectrum betalactamase-producing *Klebsiella pneumonia* (ESBL KLPN) strains (one to ten isolates per strain) were collected at the Department of Clinical Microbiology, University Hospital Brno and identified using matrix assisted laser desorption ionization – time of flight (MALDI-TOF). DNA was extracted using an UltraClean Microbial DNA Isolation Kit (MO BIO Laboratories, USA).

DNA fingerprints of the mentioned bacterial strains were evaluated by rep-PCR, which was performed using the primers and protocol described by Versalovic et al. [18]. The rep-PCR products were then analysed by chip capillary electrophoresis as described above.

The original records of chip electrophoresis for both datasets are available on the deposition site (https://figshare.com/s/6e1ebc0c396756597ecf).

Variance analysis of band size deviation

The aim of the variance analysis of band size deviation is to derive a transformation function for correcting band size deviation from a set of DNA molecular weight markers. The principle is described in the block diagram in Fig. 2. The input data consist of 1,566 bands with known DNA fragment sizes. The first step is the division of all bands into 52 band levels based on the consistency of their sizes. The SD was calculated for each of the 52 band levels (2nd block in Fig. 2). During the measurement, different types of ladders were found to have different variability for equally sized

Table 1
DNA weight markers and their measurement parameters used for band size error description.

Ladder type	Range			Samples	Bands in sample	Bands	Divided into runs
GeneRuler 1 kb DNA Ladder	250 bp	–	10 kbp	39	13	507	4
GeneRuler 100 bp Plus DNA Ladder	100 bp	–	3 kbp	27	12	324	3
GeneRuler 50 bp DNA Ladder	50 bp	–	1 kbp	33	14	462	4
O'GeneRuler 1 kb DNA Ladder	250 bp	–	10 kbp	21	13	273	2
Sum:	50 bp	–	10 kbp	120	52 band levels	1566	10 runs in total*

* One run contains 12 samples. Samples from one run can be composed of several types of ladders in a different arrangement.

Fig. 2. The principle of the derivation of the transformation function for eliminating the trends in band size deviation.

DNA fragments. Therefore, although some of the DNA fragments for the chosen ladder types were of the same size, which could lead to a reduced number of band levels, they need to be assessed separately. The SD values were determined from the real measured data, not as a deviation from the declared sizes, because the real measured band size levels were significantly different from the expected values specified in the ladder composition. In particular, in the case of the O'GeneRuler 1 kb DNA Ladder, different chemical compositions of the sample buffer caused considerable differences in sample mobility against the GeneRuler 1 kb DNA Ladder with the same sizes of DNA fragments. The complete results are shown in supplement S1.

The 3rd block in Fig. 2 represents the evaluation of the graphical dependency between the SD of the band levels and the arithmetic mean of their sizes. The best fitting analysis (MATLAB 2017a, with The Curve Fitting Toolbox™ distributed by The MathWorks, Inc., Natick, Massachusetts, USA) was implemented to estimate the dependence trend between band size SD and band size (5th block in Fig. 2). Although a logarithmic or exponential trend could be expected due to the logarithmic character of sample mobility across the gel range, none of these trends could approximate the measured data faithfully enough. Therefore, the logarithmic trend of sample mobility was compensated for by the logarithmic expression of both the assessed parameters (4th block in Fig. 2) before the fitting process; thus, the $\times$ and y axes both have a logarithmic scale (Fig. 3). The linear polynomial function was then determined to be the most accurate in approximating the characteristics of the measured data. Fig. 3 shows the results of the best fitting analysis, with the provided function equation and statistical evaluation of the fitting correctness. This transformation function was consequently used for detrending all the measured data. This step ensured that the band size deviation would be almost constant across the gel range.

Six samples of the same ladder (GeneRuler 1 kb DNA Ladder) were chosen for the demonstration of the detrending, as shown in Fig. 4. Panel **a** presents the original band position distribution, and the SD of the chosen levels is highlighted. The SD values significantly differ across the gel range. Panel **b** shows the variation in the same size bands after detrending. The SD values are almost constant at a value of approximately 0.25 and do not depend on the position in the gel.

This empirical trend model is valid for Agilent DNA 7500 Kits with standard reagents. An estimate of a trend specific for other chip electrophoresis devices can be obtained using the approach described above. The same approach can also be applied to band sizes (positions) obtained from planar electrophoresis gel images after digitalization.

Algorithm for band alignment

The principle is described in the block diagram in Fig. 5. The key step of the presented approach is the identification of bands of the same size by cluster analysis (2nd block in Fig. 5). The unassigned vector of all band size values from all samples is hierarchically linked to a dendrogram. Then, the constant threshold subdivides the dendrogram into partial clusters. The correct threshold value ensures that each cluster contains only bands of the same size and that all bands of the same size are in one cluster. This goal is achieved by the nearest neighbour hierarchical clustering method (single linkage, SLINK), with Euclidean distance as the similarity metric. The SLINK clustering approach has been recommended for strongly interconnected and distinct data [19]. The advantage of hierarchical clustering utilization is that it does not require prior knowledge about the number or the size of the clusters. However, a constant value of the threshold for subdividing data into individual clusters is required. Therefore, detrending

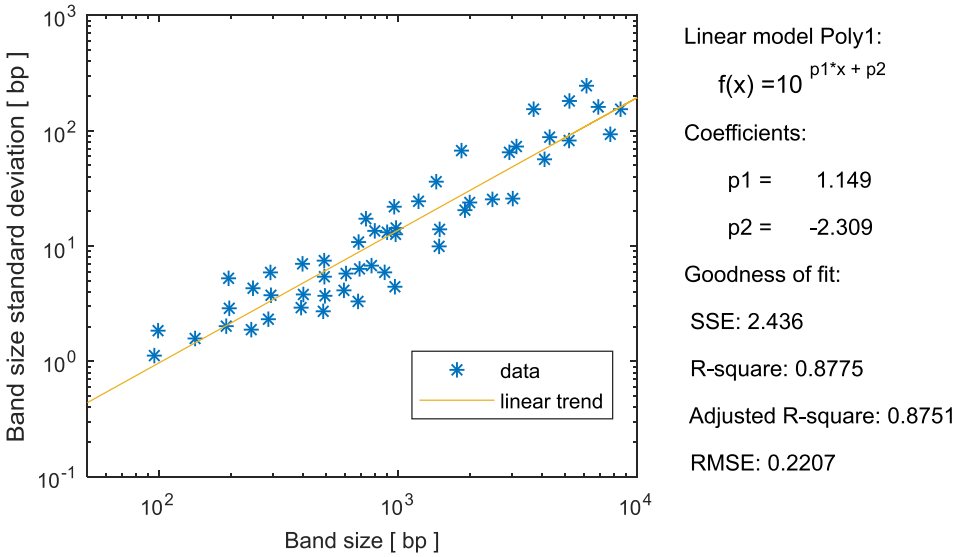


Fig. 3. The result of the best fitting analysis of the dependence between the band size standard deviation and band size. The statistical evaluation of the fitting process is given on the right-hand side using the following parameters: the sum of squares due to error (SSE), the root mean squared error (RMSE) and the ratio of the sum of squares of the regression to the total sum of squares (R-square).

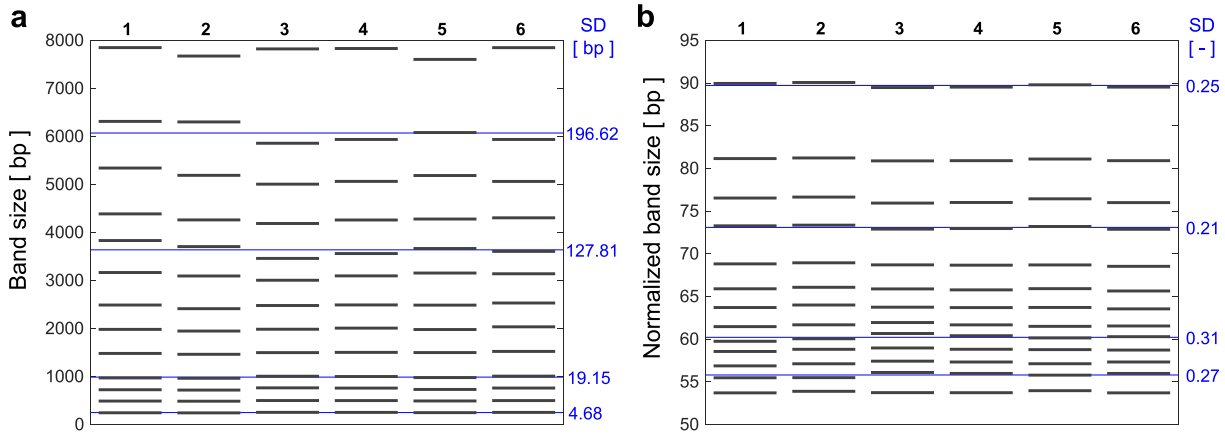


Fig. 4. The visualization of the band positions of six ladder samples of the same type (GeneRuler 1 kb DNA Ladder) (**a**) before and (**b**) after detrending by the empirical model of band size deviation. Blue lines mark the mean values of the selected band levels, and blue values represent their SDs.

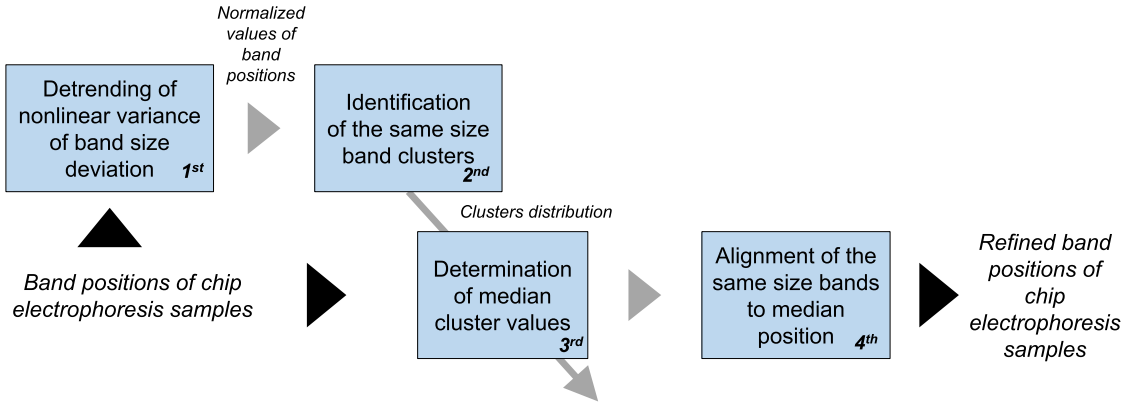


Fig. 5. The principle of the band alignment algorithm.

the band size deviation is an essential first step (1st block in Fig. 5). The subsequent band alignment is realized by redefining the positions of the bands within each cluster to their median cluster value (4th block in Fig. 5). The normalized values of band sizes obtained by detrending (output from the 1st block in Fig. 5) serve only to identify the same size clusters. The median is determined (3rd block in Fig. 5) from the original band size values identified by the cluster distribution (output from 2nd block in Fig. 5). Between 2nd and 3rd block there is no direct data transfer, but one block controls the other. The more samples there are that contain bands of the same size, the more precise is the estimation of the resulting band positions. Incorrect cluster subdivision can cause a split of the same-sized bands into several band size levels or the fusion of different bands. If bands of the same size are identified in only two samples, the arithmetic mean redefines them. The occurrence of a unique band in only one sample is preserved unchanged. The result of the alignment process (output from the 4th block in Fig. 5) is a set of refined band positions (sizes) in the original units [bp].

The result of the cluster analysis used for the identification of the bands in the same six samples in Fig. 4 is shown in Fig. 6. The upper dendrogram (Fig. 4a) illustrates clusters subdivided by a constant threshold applied to normalized band size distances. The normalization was performed by detrending with an empirical model of band size deviation. The agglomeration process rapidly links the same size bands to one cluster compared to the linking of two clusters containing bands of different sizes, which allows a wide range of values for the threshold. Specifically, in this case, the maximal distance of the same size bands is 0.42, whereas the minimal distance between different bands is 0.77 (these values are dimensionless after the transformation and normalization). Thus, the threshold value can be set anywhere within this range without producing any error. For comparison, the same procedure of hierarchical clustering was performed without the proposed detrending. The bottom dendrogram (Fig. 4b) shows the result. In this case, the setting of a constant threshold for correct cluster subdivision was not possible. The best value of the threshold selected for the demonstration was 222 bp. However, the selected setting caused the merging of three clusters with different band sizes into one (the grey cluster). Decreasing the threshold to the value subdividing these three clusters would lead to splitting the cluster containing 6 kbp size bands into two different clusters. The consequence of this setting (using original band size distances) is demonstrated in Fig. 7c and d, where the first image shows the colour differentiation of the bands according to the colour of the individual clusters, and the second image shows the result of alignment, where the bands with values of approximately 500, 700, and 1000 bp are merged (highlighted in red). The correct result, according to the upper dendrogram in Fig. 6, is shown in Fig. 7a and b. The first image is colour coded according to the cluster colours, and the second image illustrates the final band alignment.

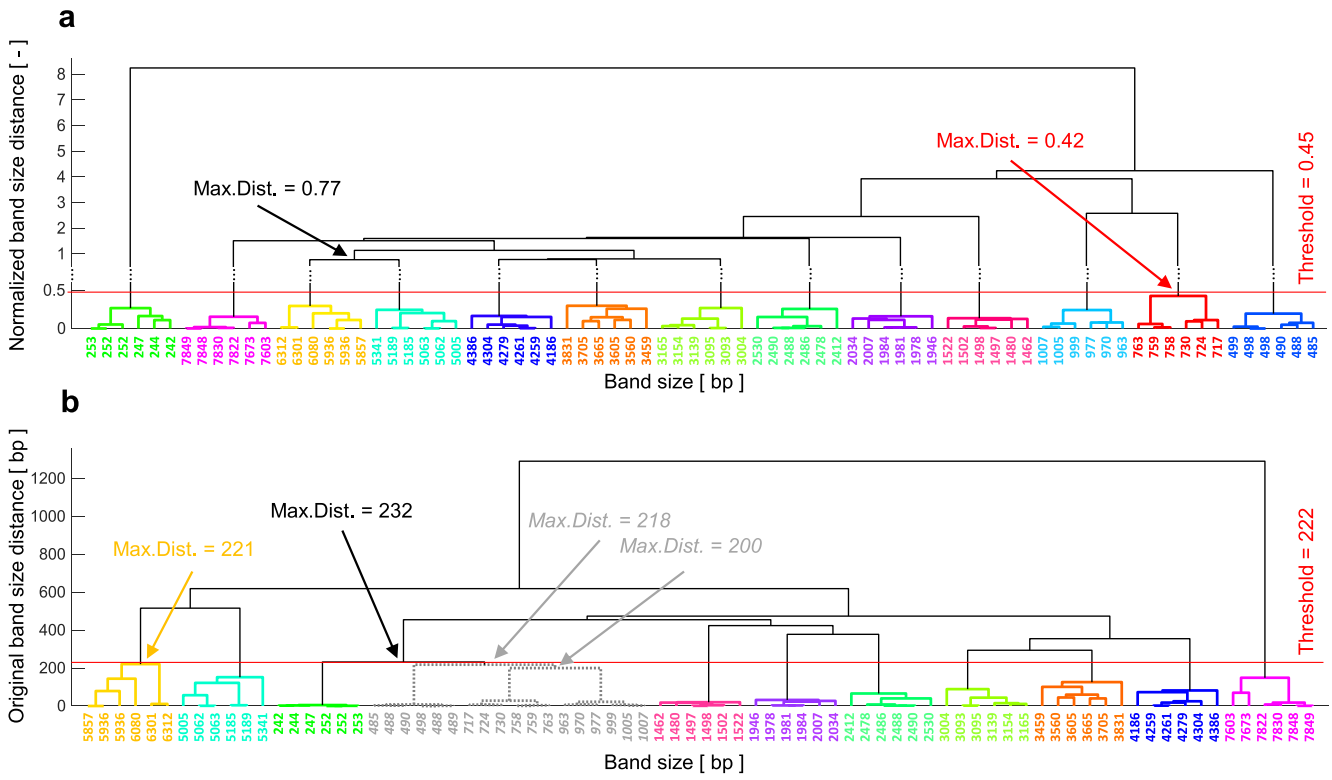


Fig. 6. Identification of the same band sizes in six samples (GeneRuler 1 kb DNA Ladder) using cluster analysis (band size values correspond to gel image in Fig. 4). The result of cluster analysis with a constant threshold for common band level identification (**a**) after detrending by the empirical model of band size deviation and (**b**) without detrending. The Y axis of the dendrogram in **a** has a double scale for better readability.

Results and discussion

The quality test results of the proposed algorithm can be divided into two separate parts. The first test was focused on the accurate identification of the same bands. For this purpose, samples containing DNA fragments of known sizes are needed. The dataset of ladders was used. The second testing process was performed on a real dataset of bacterial strain fingerprints without prior knowledge of the band distribution in the samples. Although the corresponding bands in real samples cannot be evaluated because the exact sizes of their DNA fragments are unknown, analysis of the influence of the correct alignment on bacterial genotyping is possible. All analyses were performed on a regular desktop PC (Intel Core i7-3770K CPU @ 3.50*GHz*, 16GB DDR3 RAM). The program codes for both innovative steps of presented method (derivation of the transformation function and band alignment algorithm) are available on the deposition site (https://doi.org/10.6084/m9.figshare.7464452.v2).

Accuracy of the same size band identification

The same size band identification in samples with known molecular weights was evaluated in two stages. The first quality assessment evaluated each of the four ladder types separately. In this case, only one band in only one ladder type (from 1,566 bands) was incorrectly assigned to a higher band size level. The second stage of quality assessment was performed on all 120 ladder samples immediately. In an ideal case, the 1,566 bands should be divided into 22 different band size levels. This reduction from the original 52 band size clusters (used to derive the transformation function) is caused by the occurrence of equal band size fragments in different ladder types. However, 10 bands were classified to a lower band size level, one (the same as in the previous case) was shifted to a higher level, and two bands created their own class. As a result, 13 bands were not identified correctly, which contributed less than one percent of all bands. The detailed results are provided in Table 2. The processing time of the 120 ladder repprofiles averaged 8.75 s.

All mentioned errors occur only in the GeneRuler 1 kb Ladder type. This ladder has the largest band size variation among all the ladders used (see Supplement S1). The increase in error rate in the combined analysis is caused by a large deviation of band sizes compared to the standard O'GeneRuler 1 kb Ladder samples. This ladder contains bands of the same sizes, but different compositions of its loading buffer cause different mobilities. The hierarchical clustering process had a tendency to assign similar bands from the GeneRuler 1 kb Ladder to O'GeneRuler 1 kb. These errors can be compensated for by addition of logic to the algorithm, which would consider sample indices instead of blind analysis, as was used in this case. On the other hand, the difference between the maximal and minimal size for one band in the upper part of the ladder range (for the 6 kbp band level in the case of GeneRuler 1 kb Ladder in Fig. 1) is more than two-thirds of the distance between the two different neighbouring band sizes. In the analysis of real samples, this difference could be even higher than the distance between neighbouring bands, thus reducing the possibility of correct band size determination.

Similarity analysis of aligned samples

The previous quality testing of the identification of the same bands in the ladder samples showed that the proposed algorithm could compensate for device error (sizing accuracy + resolution) to a great extent. However, its effect on a subsequent biological analysis should be determined. The most common usage is the similarity analysis of DNA fingerprints, which is the comparison

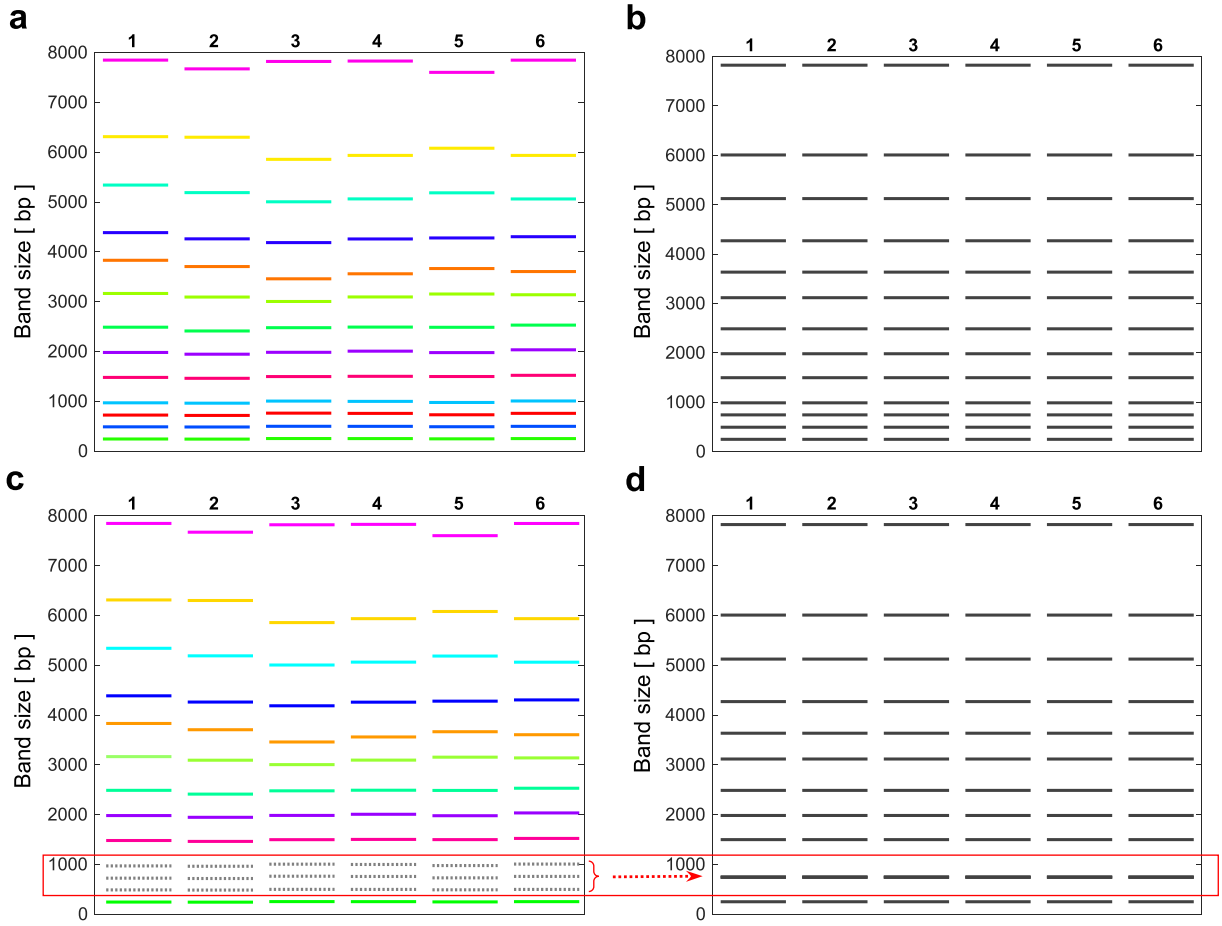


Fig. 7. Graph visualization for the identification of the same bands and multiple alignments. (**a**) Colour coding of the bands according to the results of cluster analysis (corresponding with Fig. 4). (**b**) The results of aligning the same band size to the median line after detrending. The results without detrending are shown in **c** and **d,** respectively. The merging of the three band levels into one is highlighted in red.

Table 2
Quality assessment of the same size band identification.

		Analysis of separate ladder types		Analysis of all ladder types together	
Ladder type	Bands	Error bands	Accuracy [%]	Error bands	Accuracy [%]
GeneRuler 1 kb	507	1	99.80	13	97.44
GeneRuler 100 bp Plus	324	0	100.00	0	100.00
GeneRuler 50 bp	462	0	100.00	0	100.00
O'GeneRuler 1 kb	273	0	100.00	0	100.00
All ladders together	1,566	1	99.94	13	99.17

of fragment length polymorphisms of samples from certain DNA amplification or restriction techniques, including restriction fragment length polymorphism, amplified fragment length polymorphism, and rep-PCR. Comparative analysis does not differ among these methods. The main principle is the evaluation of the sample distance by the Jaccard index and the subsequent construction of a similarity tree (or dendrogram in general) by unweighted pair group method with arithmetic mean (UPGMA) clustering methods. The quality of the similarity analysis is not the subject of this paper, so the commonly used methods have been used for a general comparison [9,20]. An important step of similarity analysis is band detection. The default settings of the detection process provided by the 2100 Bioanalyzer Expert Software tool, supplied with the chip electrophoresis device, were used to assess the quality of the proposed algorithm.

A blind comparison test of 60 rep-PCR samples of 12 ESBL KLPNs with an unequal distribution of the individual strains (from one to ten samples per strain) was performed. The dataset was obtained in five runs (12 samples in each run) (Fig. 8a). The resulting dendrogram (Fig. 8b), describing the relationship of the chosen strains, is obtained by the procedure described above from rep-profiles aligned by the proposed algorithm. The same datasets were analysed by the fingerprint data module from BioNumerics software (with default settings), and the resulting dendrogram is shown in Fig. 8c. Both dendrograms were modified (for better clarity) to use the same colour coding for clusters (branches) representing the same strain (the original result from BioNumerics software is in online supplement S2). The classification quality assessment of both methods was performed according to the following scheme: the number of correctly classified samples is equal

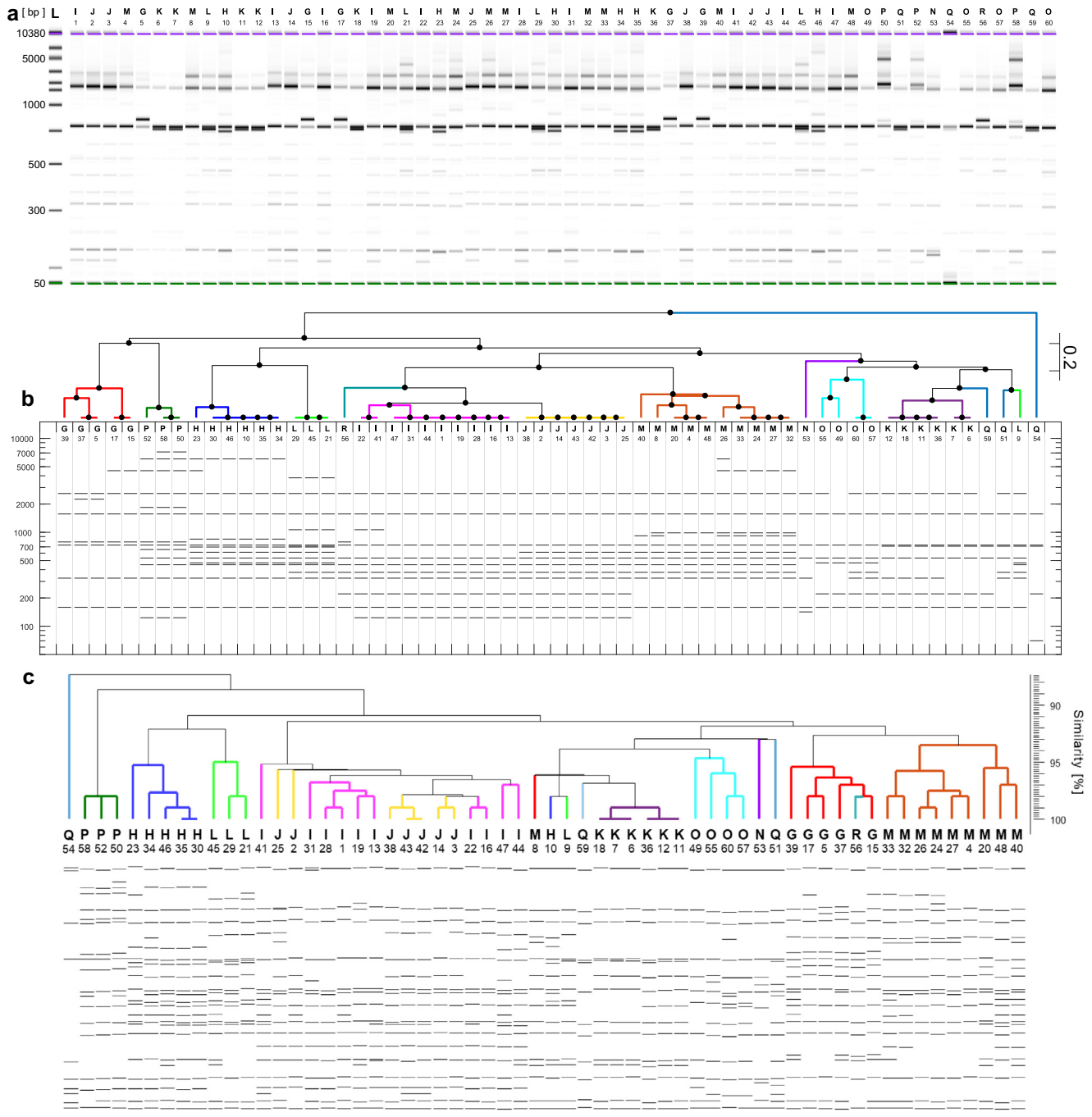


Fig. 8. The results of similarity analysis of 12 different bacterial strains by rep-PCR. (**a**) Original data: 60 samples from five runs with variable positions of 12 bacterial strains; capital letters represent strains; Arabic numerals represent sample number. (**b**) The resulting dendrogram with assigned samples after the proposed alignment procedure; the default band detection process was performed by using the BioAnalyzer tool. (**c**) The resulting dendrogram obtained by BioNumerics software with the default settings. Colour coding of strain types (clusters) is the same for **b** and **c** images.

to the highest number of branches of one strain type within one cluster. The smallest value is one for a strain with each representative in a different cluster. In an ideal case, the number of correctly classified samples would be equal to the number of all samples of one strain. This ideal result occurred in 10 out of 12 cases. In the BioNumerics analysis, only five classified strains were completely correct. In total, the percentage success of sample classification using the proposed method was 95%, in comparison to <72% using BioNumerics. An overview of the results and description of the strains is provided in Table 3. The processing time of the 60 bacterial rep-profiles averaged 5.87 s.

The proposed classification approach did not correctly evaluate one sample from strain L, but a closer examination of the pseudo-gel image in Fig. 8b shows that this sample (id 9) is also different from the other samples of strain L. Similarly, three samples of strain Q were classified separately using the proposed approach, but their rep-profiles were also significantly different. These errors were most likely caused by the inaccuracy of rep-PCR, which occurs in large data sets [21], rather than by the classification approach and the proposed alignment technique.

The limitation of the proposed approach lies in the need to redefine the transformation function for other types or ranges of

Table 3
The description of 12 different strains of *Klebsiella pneumoniae* and the results of the similarity analysis of rep-PCR samples.

Bacterial strain	Sample ID	Number of samples	Correctly classified samples	
			Proposed method	BioNumerics
G	5, 15, 17, 37, 39	5	5	2
H	10, 23, 30, 34, 35, 46	6	6	5
I	1, 13, 16, 19, 22, 28, 31, 41, 44, 47	10	10	5
J	2, 3, 14, 25, 38, 42, 43	7	7	3
K	6, 7, 11, 12, 18, 36	6	6	6
L	9, 21, 29, 45	4	3	3
M	4, 8, 20, 24, 26, 27, 32, 33, 40, 48	10	10	9
N	53	1	1	1
O	49, 55, 57, 60	4	4	4
P	50, 52, 58	3	3	3
Q	51, 54, 59	3	1	1
R	56	1	1	1
	Sum:	60	57	43
	Percentage:		95.00%	71.67%

chip electrophoresis devices according to the principle of derivation of the transformation function for detrending of the band size deviation, as shown in Fig. 2. The accuracy of the similarity analysis can be improved only partially by the proposed band size correction because it depends on the correctness of the digitization electrophoretogram generated by the software supplied with the chip electrophoresis devices.

Conclusions

A key step in the similarity-based analysis of gel samples from chip electrophoresis is the reliable recognition of bands of equal size in different samples. This step is complicated by the influence of the device sizing accuracy. The recognition of the same bands, which is based only on this declared accuracy, would significantly reduce the ability to distinguish between samples. This study introduces a novel and unique technique to determine and compensate for band size error. The main benefit of the proposed approach is the creation of an empirical model of band size error determination across the whole gel range, based on real measurements of a large number of standardized samples. The measurements confirm that the band size deviation is not constant across the gel range and does not depend linearly on the band size value. The transformation function was derived from the empirical model to achieve a constant value for the band size deviation across the whole gel range. Another unique step of the proposed approach lies in the utilization of the hierarchical clustering method with a constant threshold to identify the same size bands in the samples. This utilization allows the identification of the same bands in all samples at once instead of a simple pairwise comparison, which is currently more commonly used. In contrast to other tools where the accuracy drops as the number of samples increases, in the proposed approach, a large number of samples leads to better results. With an increasing number of samples, precise estimation of the true position of the same size bands on the gel can be performed. A resulting accuracy of over 99% for the identification of the same size bands was achieved on 120 standardized samples containing 1,566 bands. However, the influence of the proposed processing pipeline in real applications should be confirmed. Only three of 60 bacterial rep-profiles were incorrectly assigned to different related strains in the classification process using the proposed processing pipeline. Thus, the results improved from 71.67%, achieved by the commonly used tool BioNumerics 7, to 95%, achieved using the proposed method. Although the proposed methodology has been designed and tested on only one type of chip electrophoresis technology, it could also be utilized for other devices.

Conflict of interest

The authors declare no competing interests.

Compliance with Ethics Requirements

This article does not contain any studies with human or animal subjects.

Acknowledgements

This work has been supported by the grant project GACR 17-01821S. The authors would like to thank the Department of Biology and Wildlife Diseases (Faculty of Veterinary Hygiene and Ecology, University of Veterinary and Pharmaceutical Sciences Brno) for the analysis of the clinical isolates using BioNumerics software.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2019.01.005.

References

[1] Serrano I, De Vos D, Santos JP, Bilocq F, Leitão A, Tavares L, et al. Antimicrobial resistance and genomic rep-PCR fingerprints of Pseudomonas aeruginosa strains from animals on the background of the global population structure. BMC Vet Res 2017;13(1):58.
[2] Hirzel C, Donà V, Guilarte YN, Furrer H, Marschall J, Endimiani A. Clonal analysis of Aerococcus urinae isolates by using the repetitive extragenic palindromic PCR (rep-PCR). J Infect 2016;72(2):262–5.
[3] Viau RA, Kiedrowski LM, Kreiswirth BN, Adams M, Perez F, Marchaim D, et al. A comparison of molecular typing methods applied to enterobacter cloacae complex: hsp60 Sequencing, Rep-PCR, and MLST. Pathog Immun 2017;2(1):23–33.
[4] Momeni SS, Whiddon J, Cheon K, Ghazal T, Moser SA, Childers NK. Genetic diversity and evidence for transmission of Streptococcus mutans by DiversiLab rep-PCR. J Microbiol Methods 2016;128(Supplement C):108–17.
[5] Pavel AB, Vasile CI. PyElph - a software tool for gel images analysis and phylogenetics. BMC Bioinf 2012;13:9.
[6] Khakabimamaghani S, Najafi A, Ranjbar R, Raam M. GelClust: A software tool for gel electrophoresis images analysis and dendrogram generation. Comput Methods Programs Biomed 2013;111(2):512–8.
[7] Heras J, Domínguez C, Mata E, Pascual V, Lozano C, Torres C, et al. GelJ – a tool for analyzing DNA fingerprint gel images. BMC Bioinf 2015;16(1):270.
[8] Fuhrmann DR, Krzywinski MI, Chiu R, Saeedi P, Schein JE, Bosdet IE, et al. Software for automated analysis of DNA fingerprinting gels. Genome Res 2003;13(5):940–53.
[9] Heras J, Domínguez C, Mata E, Pascual V, Lozano C, Torres C, et al. A survey of tools for analysing DNA fingerprints. Brief Bioinform 2016;17(6):903–11.

[10] Intarapanich A, Kaewkamnerd S, Shaw PJ, Ukosakit K, Tragoonrung S, Tongsima S. Automatic DNA diagnosis for 1D gel electrophoresis images using bio-image processing technique. BMC Genom 2015;16(Suppl 12):S15.
[11] Skutkova H, Vitek M, Krizkova S, Kizek R, Provaznik I. Preprocessing and classification of electrophoresis gel images using dynamic time warping. Int J Electrochem Scopy 2013;8(2):1609–22.
[12] Edgar RC, Batzoglou S. Multiple sequence alignment. Curr Opin Struct Biol 2006;16(3):368–73.
[13] Chatzou M, Magis C, Chang J-M, Kemena C, Bussotti G, Erb I, et al. Multiple sequence alignment modeling: methods and applications. Brief Bioinform 2016;17(6):1009–23.
[14] Skutkova H, Vitek M, Sedlar K, Provaznik I. Progressive alignment of genomic signals by multiple dynamic time warping. J Theor Biol 2015;385:20–30.
[15] Rosenberg MS. Multiple sequence alignment accuracy and evolutionary distance estimation. BMC Bioinf 2005;6:278.
[16] Phillips A, Janies D, Wheeler W. Multiple sequence alignment in phylogenetic analysis. Mol Phylogenet Evol 2000;16(3):317–30.
[17] Feng D-F, Doolittle RF. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. J Mol Evol 1987;25(4):351–60.
[18] Versalovic J, Koeuth T, Lupski JR. Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. Nucleic Acids Res 1991;19(24):6823–31.
[19] Schmidt M, Kutzner A, Heese K. A novel specialized single-linkage clustering algorithm for taxonomically ordered data. J Theor Biol 2017;427:1–7.
[20] Nurhayati Priyambada ID, Radjasa OK, Widada J. Repetitive element palindromic PCR (Rep-PCR) as a genetic tool to study diversity in amylolytic bacteria. Adv Sci Lett 2017;23(7):6458–61.
[21] Ishii S, Sadowsky MJ. Applications of the rep-PCR DNA fingerprinting technique to study microbial diversity, ecology and evolution. Environ Microbiol 2009;11(4):733–40.