JOURNAL OF CLINICAL MICROBIOLOGY, Nov. 2005, p. 5483–5490 Vol. 43, No. 11
0095-1137/05/$08.00+0 doi:10.1128/JCM.43.11.5483–5490.2005

Assessment of Band-Based Similarity Coefficients for Automatic Type and Subtype Classification of Microbial Isolates Analyzed by Pulsed-Field Gel Electrophoresis

J. A. Carriço,[1]* F. R. Pinto,[1] C. Simas,[2] S. Nunes,[2] N. G. Sousa,[2] N. Frazão,[2] H. de Lencastre,[2,3] and J. S. Almeida[1,4]

Biomathematics Group[1] and Laboratory of Molecular Genetics,[2] Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Oeiras, Portugal; Laboratory of Microbiology, The Rockefeller University, New York, New York[3]; and Department Biostatistics, Bioinformatics, and Epidemiology, Medical University South Carolina, Charleston, South Carolina[4]

Received 5 May 2005/Returned for modification 24 June 2005/Accepted 9 August 2005

Pulsed-field gel electrophoresis (PFGE) has been the typing method of choice for strain identification in epidemiological studies of several bacterial species of medical importance. The usual procedure for the comparison of strains and assignment of strain type and subtype relies on visual assessment of band difference number, followed by an incremental assignment to the group hosting the most similar type previously seen. Band-based similarity coefficients, such as the Dice or the Jaccard coefficient, are then used for dendrogram construction, which provides a quantitative assessment of strain similarity. PFGE type assignment is based on the definition of a threshold linkage value, below which strains are assigned to the same group. This is typically performed empirically by inspecting the hierarchical cluster analysis dendrogram containing the strains of interest. This approach has the problem that the threshold value selected is dependent on the linkage method used for dendrogram construction. Furthermore, the use of a linkage method skews the original similarity values between strains. In this paper we assess the goodness of classification of several band-based similarity coefficients by comparing it with the band difference number for PFGE type and subtype classification using receiver operating characteristic curves. The procedure described was applied to a collection of PFGE results for 1,798 isolates of *Streptococcus pneumoniae*, which documented 96 types and 396 subtypes. The band-based similarity coefficients were found to perform equally well for type classification, but with different proportions of false-positive and false-negative classifications in their minimal false discovery rate when they were used for subtype classification.

Several national and international surveillance studies have been collecting data on the antimicrobial resistance of several bacterial species, namely, *Staphylococcus aureus* and *Streptococcus pneumoniae* (3, 4, 13, 14, 17, 21, 25). In the majority of these studies, pulsed-field gel electrophoresis (PFGE) (20) has been the typing method of choice for clonal type and subtype identification. These large data collection studies provide an excellent resource for the identification of the emergence and the subsequent spread of new clones, which is of particular importance for the tracking of outbreaks as well as obtaining an understanding of the propagation of particular traits, such as resistance to antibiotics. PFGE is also widely used for exchanging clonal identification data between different laboratories, because it has a high interlaboratory reproducibility (6, 17). Its high discriminatory power (24) and relative cost-effectiveness also justify why PFGE is often considered favorably in comparison with complementary typing methods, such as multilocus sequence typing (12).

An enormous variety of band patterns have been found for each bacterial species, with the type and the subtype classification being achieved by the widely used criteria of counting the number of band differences between two lanes proposed by Tenover et al. (23): if two strains differ by up to six bands, counted in both lanes, they are considered the same type. However, these authors pointed out that this method of classification should be used in outbreak studies only and should be backed up with other relevant typing data, such as antibiotic resistance.

Nevertheless, in the majority of longitudinal studies, the use of this criterion (22) yields good discrimination results, particularly when a small number of strains with distinct patterns are being compared (5, 15, 19). This is usually confirmed by visually inspecting the cluster tree to find the cutoff linkage value that agglomerates the band patterns, in accordance with the criteria of Tenover et al. (23).

However, as the number of strains to be clustered increases, this procedure will eventually fail to work because the same difference will span different groupings. This observation is a reflection of the fact that type definitions are arbitrary, in the sense that they reflect the process of strain identification gradually filling a domain of possible band patterns. The loss of a clear distinction between groups produced by hierarchical clustering algorithms (22) will also cause the membership in existing clusters (types) to be rearranged at the previously used cutoff value when a new strain is added to the collection.

A possible solution to the classification instability would be to use a large collection of classified patterns and determine

* Corresponding author. Mailing address: Biomathematics Group, Universidade Nova de Lisboa, Rua da Quinta Grande 6, 2780-156 Oeiras, Portugal. Phone: 351 21 446 98 55. Fax: 351 21 442 87 66. E-mail: jcarrico@itqb.unl.pt.

what similarity value produces the best classification results. New strains would be classified by calculating the band similarity to all the entries in the existing catalog and using the highest similarity value determined to recognize membership in the same type. Such a solution would also require the determination of which band similarity coefficient best reproduces the reference classification. Accordingly, in this paper we evaluate the commonly used band-based similarity coefficients—the Dice, Jaccard, Jeffrey's X, Ochiai, Cosine, and Pearson's correlation coefficients—for use for the automatic classification of both type and subtype. The comparison is performed with reference to a collection of 1,798 isolates of *Streptococcus pneumoniae* visually classified, using the criteria of Tenover et al. (23), into 96 types and 396 subtypes. The assessment of goodness of classification of the different similarity coefficients is performed by using receiver operating characteristic (ROC) curves (8) to determine the ability of the different similarity measures to discriminate the visually recognized groups. The method described in this paper highlights the critical value of large visually classified strain collections as the foundation for the computerized automation of their classification. However, once the most effective similarity measure is found, the prospect is raised that the classification itself may be worth redefinition to adjust it to the natural granularity of the microbial population.

MATERIALS AND METHODS

Pulsed-field gel electrophoresis. In this study we analyzed a collection of *Streptococcus pneumoniae* strains collected from children attending day care centers in Lisbon, Portugal, between 2000 and 2004 (9, 21). Chromosomal DNA preparation, restriction with SmaI endonuclease, and PFGE were done as described previously (19).

Visual similarity group (VSG) assignments. PFGE patterns were assigned to types and subtypes by visual inspection of the macrorestriction profiles by using currently accepted criteria (23). Two strains are considered of the same subtype if they have an exact match of band patterns and are considered of the same type if they have up to six band differences on both lanes. In the rare case that a strain could have less than six differences from two types, the type assignment was done by comparison of the strain to all the strains of the two types, and the strain was then assigned to the type with a fewer overall number of band differences. In these cases the type assignment was also supported by other epidemiological information, such as antibiotic resistance patterns and, more recently, multilocus sequence typing information.

The type and subtype names were assigned one or more capital letters. The first pattern identified for a subtype in a type was assigned only a capital letter (e.g., A), and the remaining subtypes were named with capital letters and numbers (e.g., A2 and A3).

Gel analysis. A database of the PFGE patterns was created with Bionumerics software (version 3.0, Applied Maths, Ghent, Belgium). The gel photos were scanned and imported into a Bionumerics database as inverted 8-bit gray-scale TIF images. For each image, spectral analysis included in the software was used to determine the disk size that should be used in "rolling disk" background subtraction (background scale) and the cutoff threshold for least-squares filtering (Wiener cutoff scale). Furthermore, a median filter was used in the image to further smooth the densitometric curves.

After this image preprocessing, intergel and intragel normalizations of the PFGE runs were done with the *S. pneumoniae* R6 strain as a molecular marker. All the gels had three markers: one in the second lane, one lane in the middle, and in the lane before the last. Fifteen bands from 16,320 bp to 340,914 bp were used. The existence of these bands was verified, and their sizes were calculated by virtual digestion of the gel by using a perl script to recognize the restriction sequence of SmaI (CCC^GGG) in the GenBank file of the complete sequence (10). A cubic spline curve was used for the normalization and calibration of each gel. Strain R6 was obtained from the Rockefeller University culture collection.

On all gel images, band assignment was manually curated after automatic band detection. This step is of paramount importance, since there are band intensity variations from gel to gel, which cause errors in the automatic band assignment. Bands ranging from 14 kbp to 400 kbp were considered in this study.

TABLE 1. Band-based similarity coefficients between any two gel band patterns, i and j

Coefficient	Formula[a]
Dice	$S_{ij} = \frac{2n_{ij}}{2n_{ij} + n_i + n_j}$
Jaccard	$S_{ij} = \frac{n_{ij}}{n_{ij} + n_i + n_j}$
Jeffrey's X	$S_{ij} = \frac{n_{ij}}{N_i} + \frac{n_{ij}}{N_j}$
Ochiai	$S_{ij} = \frac{n_{ij}}{\sqrt{(n_{ij} + n_i)\cdot(n_{ij} + n_j)}}$

[a] Similarity (S_{ij}) is calculated as described, where n, is the number of bands occurring only in pattern i,n_{ij} is the number of bands shared between the two patterns, and N_i is the total number of bands in pattern i.

The software was then used to calculate the alternative band pattern similarity coefficients. For the 1,798 isolates used in this study, a comparison was created and the corresponding similarity matrices were exported by using the four different band-based similarity coefficients (the Dice, Jaccard, Jeffrey's X, and Ochiai coefficients) and two curve-based correlation coefficients (the Pearson and Cosine coefficients). For the comparative evaluation of the different band-based coefficients, the optimization parameter was evaluated with a range of band position tolerances of from 0% to 8%.

Band-based similarity coefficients. The four most popular band-based similarity coefficients were considered in this study for quantification of the similarities between PFGE band patterns: the Dice (7, 22), Jaccard (22), Jeffrey's X (18a), and Ochiai (18) coefficients (Table 1). All these coefficients exclude negative band matches, which is a necessary compromise, since all possible band positions are unknown.

Also, the Pearson and Cosine correlation coefficients were considered for reference. Generally, these two methods yield lower similarity values than band-based methods, since they take into account all the densitometric curve values, which causes them to be more sensitive to small variations. This makes them the methods of choice for the comparison of strain similarity by typing methods in which the intensities of bands are to be considered, such as AFLP (16, 26), but they can produce erroneous conclusions when only the presence or the absence of a band is important and the band intensity varies among the strains being compared.

ROC curves. ROC curves were used to assess the classification by use of the different similarity coefficients. This method, created in signal detection theory, is frequently used in classification problems and is widely applied in medical diagnosis and psychometric analysis (8). This method is commonly employed for the binary classification of continuous data, usually categorized as positive and negative cases. In our study, the correct classification was considered the VSG assignment; for each coefficient, the VSG assignment thus classified each case at each threshold as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). The classification accuracy of each coefficient was then measured by plotting for the different threshold values the ratio of the number of true-positive classifications over the total number of positive classifications, also named the sensitivity or the true-positive rate, versus the false-positive rate, or 1 − specificity (Table 2), which is the ROC curve.

The area under a ROC curve (AUC) is the parameter employed to quantify the goodness of classification of the classifier being tested, since it is a threshold-independent performance measure. For a perfect classifier the AUC is 1, and for a random classifier the AUC is 0.5. Additional results and a comprehensive discussion of the AUC measure are provided elsewhere (1, 2).

RESULTS

The PFGE patterns of 1,798 distinct strains of *S. pneumoniae* were visually classified into 96 types and 396 subtypes, with the assistance of Bionumerics software, at the Laboratory of Molecular Genetics of Instituto de Tecnologia Química e

TABLE 2. ROC curve parameters

Parameter	Formula[a]
Sensitivity, or true-positive rate	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
1 − specificity, or false-positive rate	$\frac{FP}{TN + FP}$

[a] TP, number of samples with a true-positive classification; TN, number of samples with a true-negative classification; FP, number of samples with a false-positive classification; FN, number of samples with a false-negative classification.

Biológica. This manually curated repository was analyzed for objective assessment of alternative measures of similarity for automation of PFGE-based classification of new isolates. Figure 1 displays the visual classification of strains into types and subtypes along with the normalized patterns. The strains are sorted alphabetically according to the naming protocol detailed in the Materials and Methods section. The 10 most represented types are delimited by vertical lines.

The VSG classification of the PFGE band patterns into types and subtypes was used as a reference to assess the goodness of classification of the similarity coefficients typically considered in comparisons of PFGE band patterns: the Dice, Jaccard, Jeffrey's X, Ochiai, Pearson, and Cosine coefficients (Table 1). This assessment was first performed by using ROC curves, which measure the classification by plotting, for different similarity coefficient threshold values, the ratio of TP matches over the total number of positive samples versus the ratio of FP matches over the total number of negative samples. The AUC was then used as the threshold independent measure of goodness of the classification (see Materials and Methods). Second, for each band-based similarity coefficient, different band position tolerance settings were compared to determine the optimal parameter settings, i.e., band position tolerance values. In this study, this is the most important parameter for accurate band matching between two different lanes.

For example, in Fig. 2, ROC curves are plotted for the comparison of the visual type assignments of the band and Dice coefficient values for different band position tolerance settings. This illustrates how best the tolerance value for the Dice coefficient can be determined. The table inset in Fig. 2 provides the corresponding AUC values. The Pearson correlation coefficient AUC value is also included to illustrate the relative inefficient classification of correlation similarity coefficients (AUC of 0.901 versus an AUC up to 0.984 for the Dice coefficient). Even worse performance was found for the Cosine correlation coefficient, with an AUC value of 0.882 (not plotted).

The goodness of classification, as assessed by the AUC, for the alternative similarity coefficients considered in this study is represented in Fig. 3. All the band-based similarity coefficients (the Dice, Jaccard, Jeffrey's X, and Ochiai coefficients) behave remarkably similarly in the classification of types and subtypes. As was to be expected, when properly selected band position tolerance values are used, band-based similarity coefficients have higher AUC values than correlation coefficients.

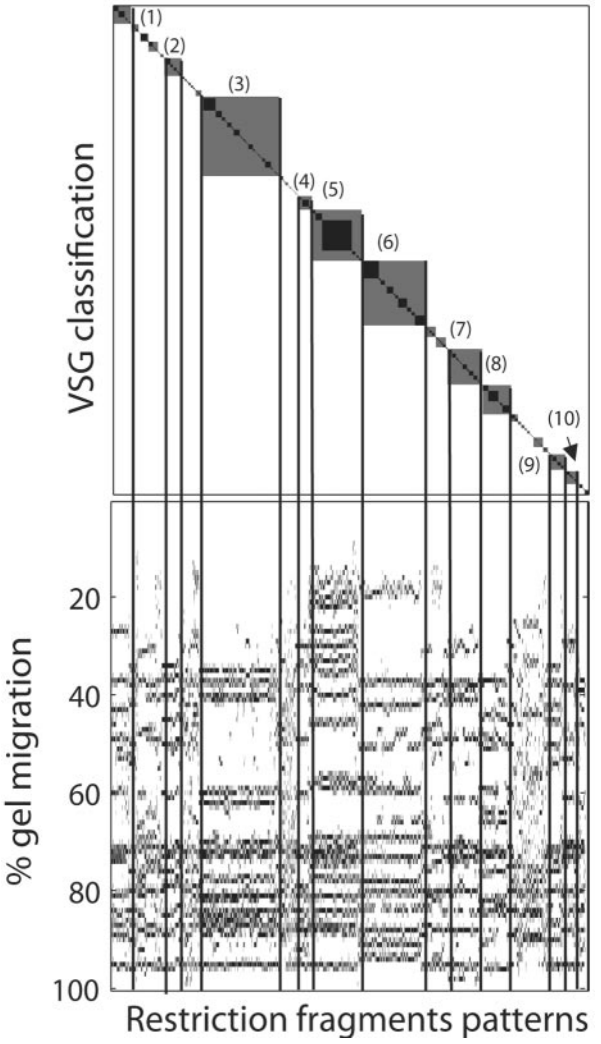

FIG. 1. Representation of VSG classification and band patterns for the 1,798 strains of *S. pneumoniae*. In the upper part (visual similarity group classification matrix), the black areas represent PFGE subtypes and the gray areas represent PFGE types. The most represented groups (PFGE types) are (point 1) A (67 isolates), (point 2) AO (65 isolates), (point 3) B (292 isolates), (point 4) DDD (51 isolates), (point 5) E (187 isolates), (point 6) FF (238 isolates), (point 7) M (131 isolates), (point 8) MM (107 isolates), (point 9) R (57 isolates), and (point 10) SI (47 isolates). The lower part of the figure includes the corresponding PFGE band patterns. The lines were drawn to help the reader isolate the PFGE patterns visually.

As shown in Fig. 2 (and also in Fig. 3B), for type classification the optimal band position tolerance was found to be 1.7% for all band-based similarity coefficients, with an AUC of 0.984. For subtype classification (Fig. 3A), the optimal settings were found for higher band position tolerance values, 2.5%, which also correspond to a higher AUC of 0.995, which is the same for all band-based similarity coefficients Again, correlation coefficients yielded a lower AUC of 0.906 for the Pearson correlation coefficient and 0.898 for the Cosine coefficient.

Although the different band-based similarity coefficients are surprisingly equivalent regarding the goodness of classification, the proportions of true-positive and false-positive subtype clas-

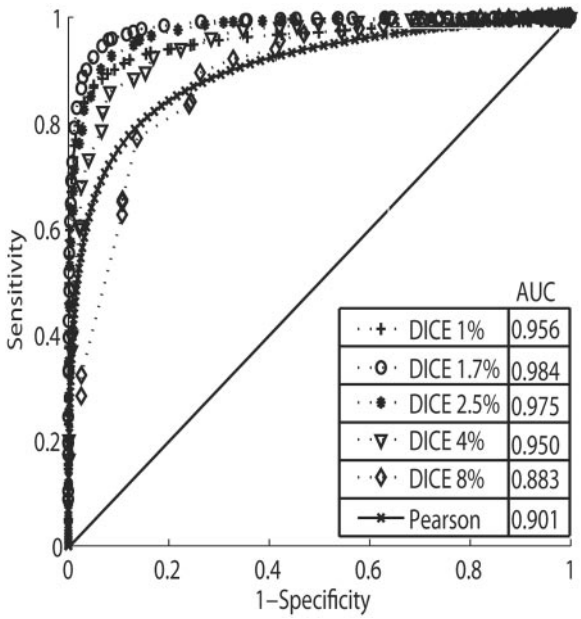

FIG. 2. ROC curves for several band position tolerances of the Dice coefficient in type classification. The maximum AUC value, 0.984, was found for a band position tolerance of 1.7%. The random classification (straight diagonal; AUC, 0.5) and the underperforming Pearson's correlation coefficient (AUC, 0.901) are plotted for reference.

sifications differ. Figure 3C and D represents the contribution of false-positive or false-negative classifications on the total classification error.

For each band position tolerance, the point where the similarity coefficient threshold had a minimum absolute classification error (a minimum of false-positive plus false-negative classifications) was plotted. For example, by using the Dice coefficient for type classification, the similarity threshold value with minimal classification error was found to be 81% for a 1.7% band position tolerance (Fig. 4).

For subtype classification at a 2.5% band position tolerance, the Dice and Jaccard coefficient (Fig. 3C) classifications resulted in fewer false-negative classifications and, conversely, more false-positive classifications than the Ochiai and Jeffrey coefficients. The same is true for the absolute numbers of misclassifications (data not shown).

Regarding the type classification, band-based similarity coefficients also performed equally well (Fig. 3B), but the heterogeneity of band patterns included in each type is reflected by the persistence of false-negative classifications for wider band position tolerance values. At a band position tolerance of 1.7%, the four band-based similarity coefficients are nearly indistinguishable in terms of the contribution of false-positive and false-negative classifications to the type classification error.

These calculated optimal position tolerance settings apply only to the data analyzed in this study, although it is a very good starting point for data obtained by the same PFGE protocol, since the running conditions should be similar and should generate similarly resolved band patterns.

As suggested by the results plotted in Fig. 3, the fact that the four band-based similarity coefficients performed equally well for the same band position tolerance implies that there are equivalent, but not necessarily similar, threshold values between each of the band pattern similarity measures. This equivalence is confirmed in Fig. 4, where, for optimal band tolerance (1.7% for type; 2.5% for subtype), the ROC curves and corresponding threshold values are displayed. Figure 4, as discussed in the next section, can be used to determine the appropriate threshold values for the desired proportion of false-positive and false-negative classifications in the total classification error. Figure 4 can be analyzed to produce optimal threshold values for arbitrary cost-benefit ratios. For example, if FP and FN classifications are equally undesirable, the four band-based similarity coefficient should be used with the band identity tolerance values indicated in Table 3.

DISCUSSION

The classification of *Streptococcus pneumoniae* isolates by PFGE has followed the typical method of visual recognition of similar patterns by the absolute number of band differences within existing isolates. New isolates are classified by incrementally assigning them to a type or a subtype of the previously classified isolates already described in databases. This solution pragmatically produces guidelines for group recognition, as prescribed by the widely used criteria of Tenover et al. (23) detailed in the introduction. However, when enough isolates have been processed in this fashion, the collection of results can be analyzed to identify an equivalent computational procedure. In order to achieve that goal of automation of manual classification, it is necessary to assess alternative metrics to quantify band pattern dissimilarity and also to determine its most discriminant settings: the band identity tolerance and the similarity threshold value for positive classification in the same group as another band pattern. This work was made possible by the extensive collection of *S. pneumoniae* isolates that had been manually annotated. Accordingly, the collection of 1,798 *S. pneumoniae* isolates was analyzed for determination of the settings that maximize the goodness of classification by use of the alternative band-based similarity coefficients—the Dice, Jaccard, Jeffrey's X, and Ochiai coefficients—and also, for reference, densitometric curve-based correlation coefficients—the Pearson and Cosine coefficients.

As expected, discrete band-based similarity coefficients clearly outperformed the correlation coefficients, leading to a much higher goodness of classification, as assessed by the area under the ROC curve. Surprisingly, all of the band-based similarity coefficients tested were found to be equally discriminant for both type and subtype (Fig. 3). That is, all of the four similarity coefficient band-based formulations (Table 1) will produce the same percentage of erroneous classifications for a given band identity tolerance value (Fig. 3A and B). However, this does not necessarily imply that the erroneous classifications will include the same proportion of false-negative and false-positive classifications.

As noted above, the results presented in Fig. 3 for the dependence of goodness of classification, as assessed by the corresponding AUC value, suggest not only that the four band-based methods will perform equally well but also that they will perform equally well for the same band tolerance values (Fig. 3A and B). This observation was observed to be valid for both

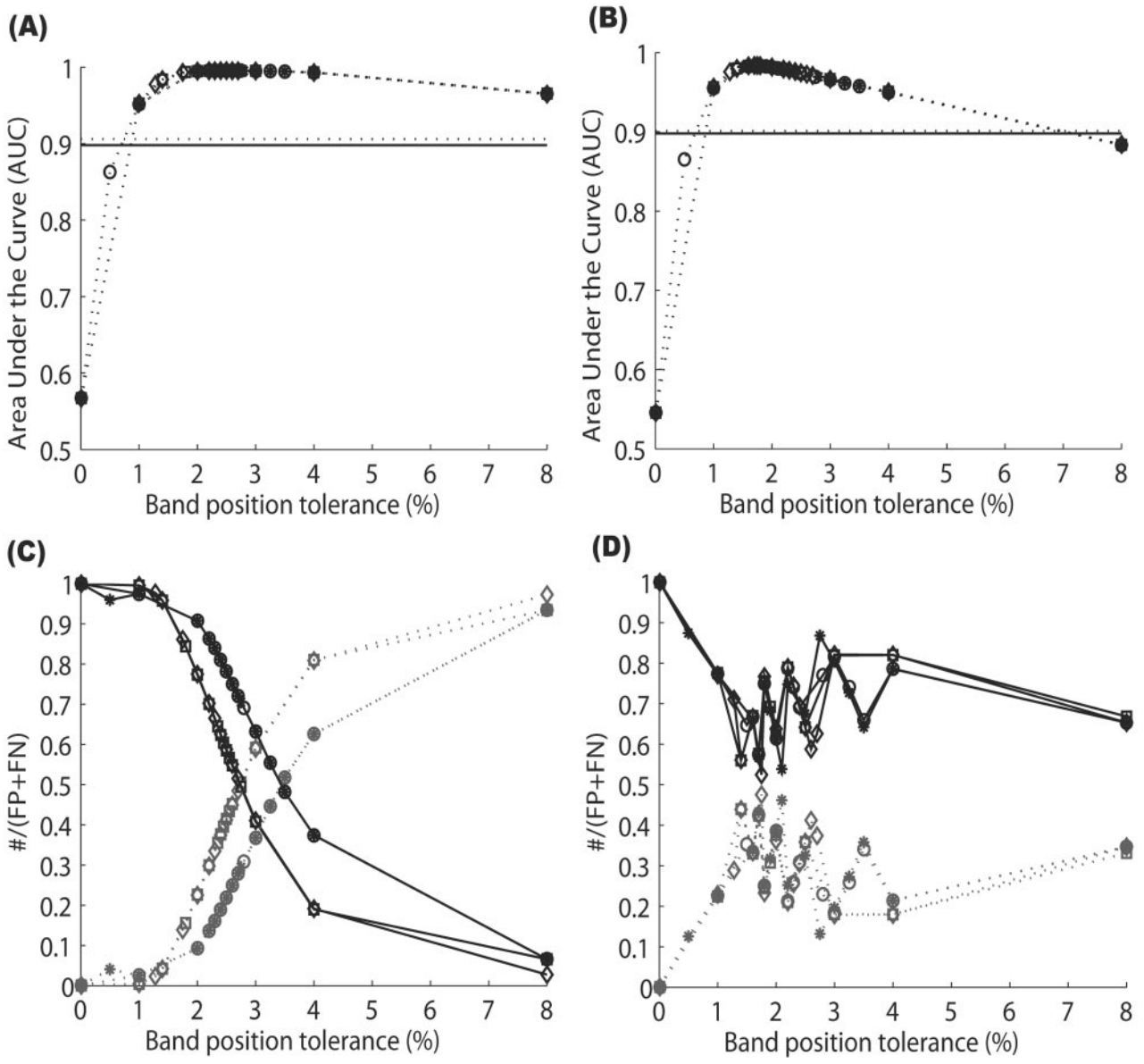

FIG. 3. Area under the curve of ROC curves of the coefficients tested for different band position tolerances for subtype (A) and type (B) classification. Contribution of false-positive and false-negative classifications for the total classification error in subtype (C) and type (D). The Dice coefficient is identified by squares, the Jaccard coefficient is identified by diamonds, the Ochiai coefficient is identified by asterisks, the Jeffrey's X coefficient is identified by circles, the Pearson coefficient is identified by a dotted line without markers, and the Cosine coefficient is identified by a solid line without markers. For panels C and D, FP classifications are represented by gray dotted lines, and FN classifications are represented by black solid lines.

type and subtype classifications. However, inspection of the corresponding proportions of FP and FN classifications (Fig. 3C and D) shows that, for subtype classifications (Fig. 3C), the Dice and the Jaccard coefficients will yield comparatively more FP classifications and fewer FN classifications than Jeffrey's X or the Ochiai coefficient. This distinction is the most pronounced when the goodness of classification (AUC) is maximal. It is also interesting that for subtype classification with exaggerated band identity tolerance values, the erroneous classifications will be heavily dominated by false-positive classifications. In contrast, neither of these observations is valid for type classification (Fig. 3D), where the proportion of false-positive and false-negative classifications is not noticeably different between the band-based methods assessed, and high band tolerance values do not cause false-positive classifications to predominate. It is also noteworthy that the proportions themselves (Fig. 3D) are somewhat erratic, which is a reflection of the fact that any of the two band patterns classified as the same type can have up to six band differences, allowing for a great heterogeneity of patterns.

The discussion above highlights the observation that if bands that discriminate between types are in close proximity to each other and are possibly bands of lower molecular size (from approximately 19 kbp to 100 kbp), misclassification will even-

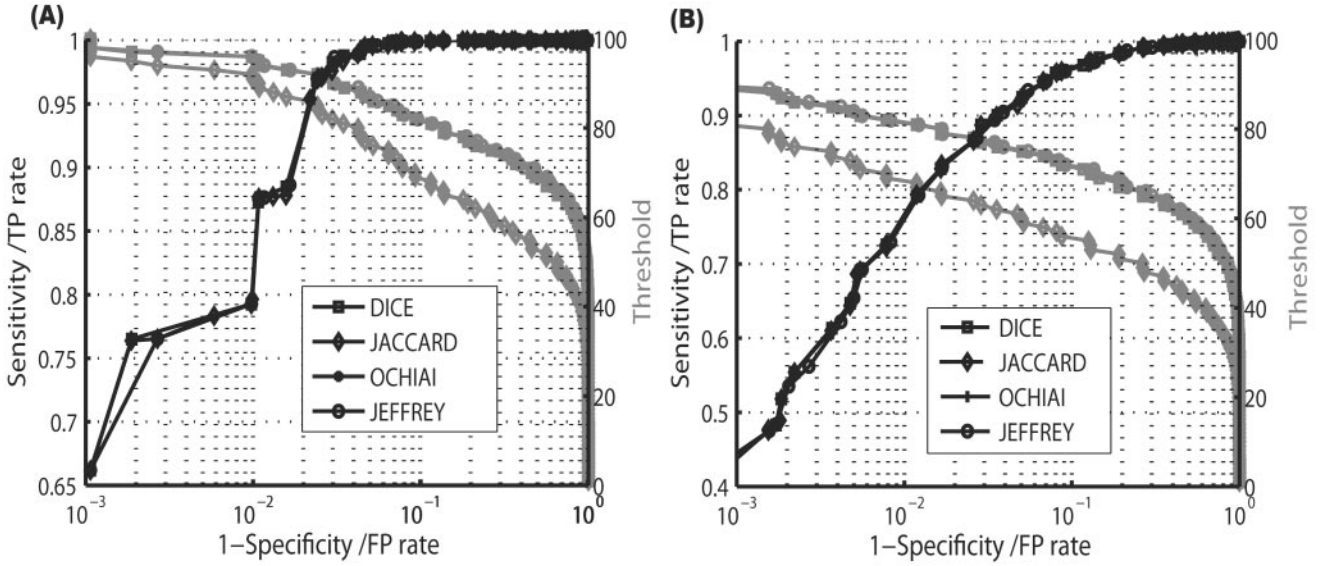

FIG. 4. ROC curves and threshold representation for subtype (A) and type (B). This figure allows the choice of a threshold value as a function of the false-positive rate/true-positive rate, for the optimal band position tolerance settings that provide a maximum discrimination between types. Note that the false-positive rate (which corresponds to 1 − specificity) is represented on a logarithmic scale.

tually occur as more subtypes are identified. This heterogeneity of patterns for isolates of the same type and the blurring of arbitrary type distinctions by new isolates justify why the number of false-negative classification contributions did not decrease for higher band identity tolerance values. This is in sharp contrast to what happens in subtype classification (Fig. 3C), where band patterns should be exactly equal (band-based similarity coefficient value of 100%) between strains of the same subtype. In practice, experimental conditions can cause small distortions in the gel that are not compensated for by software or visual classification, and that is why the determination of the optimal band identity tolerance is critical for automation of band pattern classification of subtypes. Accordingly, the similarity levels for strains of the same subtype oscillate in the 95 to 100% interval, even after selection of an optimal band position tolerance setting. These results suggest that while automation of the classification of PFGE band patterns of visually recognized subtypes and types is achieved with considerable accuracy by the proposed method (maximum AUC values of 0.9954 and 0.9837, respectively, for the Dice coefficient), visual assignment mostly delimits arbitrary groupings of subtype patterns. On the contrary, the automated classification of PFGE band patterns of subtypes confines defined groups where the band positions oscillate only very slightly around a reference value.

TABLE 3. Threshold similarity values for the point where there are a minimum of misclassifications (minimum of false positives and false negatives) of subtype and type

Coefficient	Subtype			Type		
	FP rate	TP rate	Threshold	FP rate	TP rate	Threshold
Dice	0.002	0.76	97	0.012	0.79	81
Jaccard	0.002	0.76	95	0.012	0.79	67
Jeffrey's X	0.001	0.66	98	0.012	0.79	81
Ochiai	0.001	0.66	98	0.012	0.79	81

The immediately useful result of this paper is delivered in Fig. 4. It plots, for the optimal band position tolerance value, the logarithm of the false-positive rate versus the true-positive rate and the respective threshold values. The logarithm of the false-positive rate provides easier reading of the values for the lower false-positive rates (from 0.001 to 0.1). Figure 4 allows the choice of a threshold value as a function of the false-positive rate/true-positive rate for the optimal band position tolerance settings that provide a maximum discrimination (measured by the AUC). This choice weighs the relative cost of having a false-positive or a false-negative classification. For example, if the objective of a study is to recognize membership in a specific PFGE type, the threshold should be chosen to minimize the number of false-negative assignments. If the Dice similarity coefficient was the metric chosen and the acceptable false-positive classification was only 1%, then the threshold obtained by inspecting Fig. 4 would be about 80%. If, instead, the goal was the maximization of the true discovery rate, then the appropriate threshold for the same method would be 97% (this result is also listed in Table 3). This exercise also illustrates the conclusion that although the similarity coefficients perform equally well, they are not interchangeable, as different proportions of false-negative and false-negative classifications may result. Conversely, Fig. 4 can also be used to determine what threshold values will render two similarity coefficients equally discriminant for the optimal band position tolerance value.

Over the past few years large databases of genotyped clinical strains have been assembled. These repositories contain a unique record documenting both the diversity and the dynamics of the emergence of new strains. Furthermore, it has been consistently shown that PFGE has a higher discriminatory power than newer sequence-based methods, such as multilocus

sequence typing, which justifies the prospect that the cost-effective use of PFGE will be seamlessly integrated with other genotyping methods in even larger repositories. In that regard, the study reported here leads to the following conclusions.

First, we have found that the perception that band-based similarity coefficients are superior to correlation methods is correct, provided that they are correctly parameterized. This observation puts a prize not only on the correct parameterization method but also on the use of robust image analysis software for gel lane alignment and band recognition.

Second, we have used a repository of 1,798 PFGE types isolates of *S. pneumoniae* to assess the relative merits of the different band-based similarity coefficients: the Dice, Jaccard, Jeffrey's X, and Ochiai coefficients. Surprisingly, they were all found to be equally able to classify the isolates from the reference database, with equivalent performances occurring for distinct thresholds but the same band position tolerances. The goodness of classification was assessed by use of the AUC of the ROC curve.

Third, the equivalence in AUC with the same proportion of erroneous classifications was found to correspond to different proportions of false-positive and false-negative classifications, which will play a role in the selection of a similarity coefficient for use in a fully automated bioinformatic implementation. Consequently, the assessment and parameterization of PFGE similarity coefficients are delivered as ROC curve plots with the corresponding threshold values (Fig. 4), where the cost-benefit assigned to the different types of erroneous classifications can be weighted quantitatively and the most appropriate method and threshold values can be selected.

Fourth, the automated procedure was found to perform satisfactorily, with an optimal AUC of 0.984. This result supports the conclusion that the implementation of automated classification is highly advantageous, particularly since multiparametric statistics can be associated to select those patterns that warrant subsequent visual inspection.

The optimal parameterization of band-based similarity coefficients opens the prospect of revisiting the identification of types as a dynamic entity defined by unsupervised classification algorithms such as nearest means (*K* means) or self-organized maps. Therefore, the identification of similarity metrics that reproduce and automate the classification of typing results enables the redefinition of heterogeneous types in *S. pneumoniae* with time-dependent identities that converge to the confinements of the natural populations as more isolates are characterized. The tracking of how the definitions evolve could be solved automatically by the implementation of repositories that can be queried by use of the shortest similarity coefficient value. The methods used in this paper can be used in any database to determine which similarity metric is more adequate to describe the data and also which parameters optimize the classification procedure.

ACKNOWLEDGMENTS

We thank Alexander Tomasz, The Rockefeller University, for the gift of strain *S. pneumoniae* R6. We also acknowledge Susana Vinga for help on ROC curves and Luc Vauterin for bibliographic help on the similarity coefficients.

Partial support for this work was provided by contracts EURIS (QLK2-CT-2000-01020) and PREVIS (LSHM-CT-2003-503413 from the European Community) awarded to H. de Lencastre and J. S. Almeida. J. A. Carriço and F. R. Pinto were supported by grants SFRH/BD/3123/2000 and SFRH/BD/6488/2001, respectively, both from the Fundação para a Ciência e Tecnologia of Portugal. S. Nunes and N. G. Sousa were supported by grants 011/BIC/01 and 043/BIC/00, respectively, from contract QLK2-CT-2000-01020; S. Nunes, N. G. Sousa, and N. Frazão have been supported since March 2004 by grants 010/BIC/2004, 009/BIC/2004, and 011/BIC/2004, respectively, from contract LSHM-CT-2003-503413. C. Simas was supported by a grant from IBET, project WLP (grant 31 CEM/NET); and N. Frazão was also supported by IBET grant 28/12/02.

REFERENCES

1. **Baldi, P., S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen.** 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics **16:**412–424.
2. **Bradley, A. P.** 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition **30:**1145–1159.
3. **Bronzwaer, S. L., U. Buchholz, J. L. Kool, J. Monen, and P. Schrijnemakers.** 2001. EARSS activities and results: update. Euro. Surveill. **6:**2–5.
4. **Bronzwaer, S. L., O. Cars, U. Buchholz, S. Molstad, W. Goettsch, I. K. Veldhuijzen, J. L. Kool, M. J. Sprenger, and J. E. Degener.** 2002. A European study on the relationship between antimicrobial use and antimicrobial resistance. Emerg. Infect. Dis. **8:**278–282.
5. **Brueggemann, A. B., D. T. Griffiths, E. Meats, T. Peto, D. W. Crook, and B. G. Spratt.** 2003. Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. J. Infect. Dis. **187:**1424–1432.
6. **Chung, M., H. de Lecanstre, P. Matthews, A. Tomasz, I. Adamsson, M. Aires de Sousa, T. Camou, C. Cocuzza, A. Corso, I. Couto, A. Dominguez, M. Gniadkowski, R. Goering, A. Gomes, K. Kikuchi, A. Marchese, R. Mato, O. Melter, D. Oliveira, R. Palacio, R. Sá-Leão, I. Santos Sanches, J.-H. Song, P. T. Tassios, and P. Villari.** 2000. Molecular typing of methicillin-resistant *Staphylococcus aureus* by pulsed-field gel eletrophoresis: comparison of results obtained in a multilaboratory effort using identical protocols and MRSA strains. Microb. Drug Resist. **6:**189–198.
7. **Dice, L. R.** 1945. Measures of the amount of ecological association between species. Ecology **26:**297–302.
8. **Egan, J. P.** 1975. Signal detection theory and ROC-analysis. Academic Press, Inc., New York, N.Y.
9. **Frazão, N., A. Brito-Avô, C. Simas, J. Saldanha, R. Mato, S. Nunes, N. G. Sousa, J. A. Carriço, J. S. Almeida, I. Santos-Sanches, and H. de Lencastre.** 2004. Effect of the seven-valent conjugate pneumococcal vaccine on carriage and drug resistance of *Streptococcus pneumoniae* in healthy children attending day-care centers in Lisbon. Pediatr. Infect. Dis. J. **24:**243–252.
10. **Hoskins, J., W. E. Alborn, Jr., J. Arnold, L. C. Blaszczak, S. Burgett, B. S. DeHoff, S. T. Estrem, L. Fritz, D. J. Fu, W. Fuller, C. Geringer, R. Gilmour, J. S. Glass, H. Khoja, A. R. Kraft, R. E. Lagace, D. J. LeBlanc, L. N. Lee, E. J. Lefkowitz, J. Lu, P. Matsushima, S. M. McAhren, M. McHenney, K. McLeaster, C. W. Mundy, T. I. Nicas, F. H. Norris, M. O'Gara, R. B. Peery, G. T. Robertson, P. Rockey, P. M. Sun, M. E. Winkler, Y. Yang, M. Young-Bellido, G. Zhao, C. A. Zook, R. H. Baltz, S. R. Jaskunas, P. R. Rosteck, Jr., P. L. Skatrud, and J. I. Glass.** 2001. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J. Bacteriol. **183:**5709–5717.
11. Reference deleted.
12. **Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA **95:**3140–3145.
13. **McDougal, L. K., C. D. Steward, G. E. Killgore, J. M. Chaitram, S. K. McAllister, and F. C. Tenover.** 2003. Pulsed-field gel electrophoresis typing of oxacillin-resistant *Staphylococcus aureus* isolates from the United States: establishing a national database. J. Clin. Microbiol. **41:**5113–5120.
14. **McGee, L., L. McDougal, J. Zhou, B. G. Spratt, F. C. Tenover, R. George, R. Hakenbeck, W. Hryniewicz, J. C. Lefevre, A. Tomasz, and K. P. Klugman.** 2001. Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network. J. Clin. Microbiol. **39:**2565–2571.
15. **Miragaia, M., I. Couto, S. F. Pereira, K. G. Kristinsson, H. Westh, J. O. Jarlov, J. Carrico, J. Almeida, I. Santos-Sanches, and H. de Lencastre.** 2002. Molecular characterization of methicillin-resistant *Staphylococcus epidermidis* clones: evidence of geographic dissemination. J. Clin. Microbiol. **40:**430–438.
16. **Mueller, U. G., and L. L. Wolfenbarger.** 1999. AFLP genotyping and fingerprinting. Trends Ecol. Evol. **14:**389–394.
17. **Murchan, S., M. E. Kaufmann, A. Deplano, R. de Ryck, M. Struelens, C. E. Zinn, V. Fussing, S. Salmenlinna, J. Vuopio-Varkila, N. El Solh, C. Cuny, W. Witte, P. T. Tassios, N. Legakis, W. van Leeuwen, A. van Belkum, A. Vindel, I. Laconcha, J. Garaizar, S. Haeggman, B. Olsson-Liljequist, U. Ransjo, G. Coombes, and B. Cookson.** 2003. Harmonization of pulsed-field gel electro-

phoresis protocols for epidemiological typing of strains of methicillin-resistant *Staphylococcus aureus*: a single approach developed by consensus in 10 European laboratories and its application for tracing the spread of related strains. J. Clin. Microbiol. **41:**1574–1585.

18. **Ochiai, A.** 1957. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. Bull. Jpn. Soc. Fish Sci. **22:**526–530.

18a. **Pena, S. D. J., R. Chakraborty, J. T. Epplen, and A. J. Jeffreys.** 1993. DNA fingerprinting: the state of the science, p. 1–19. Birkhauser Verlag, Basel, Switzerland.

19. **Sá-Leão, R., A. Tomasz, I. Santos-Sanches, S. Nunes, C. R. Alves, A. B. Avo, J. Saldanha, K. G. Kristinsson, and H. de Lencastre.** 2000. Genetic diversity and clonal patterns among antibiotic-susceptible and - resistant *Streptococcus pneumoniae* colonizing children: day care centers as autonomous epidemiological units. J. Clin. Microbiol. **38:**4137–4144.

20. **Schwartz, D. C., and C. R. Cantor.** 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell **37:**67–75.

21. **Silva, S., R. Gouveia-Oliveira, A. Maretzek, J. Carrico, T. Gudnason, K. G. Kristinsson, K. Ekdahl, A. Brito-Avo, A. Tomasz, I. S. Sanches, H. de Lencastre, and J. Almeida.** 2003. EURISWEB—Web-based epidemiological surveillance of antibiotic-resistant pneumococci in day care centers. BMC Med. Inform. Decision Making **3:**9.

22. **Sneath, P. H., and R. R. Sokal.** 1973. Numerical taxonomy. W. H. Freeman & Co., San Francisco, Calif.

23. **Tenover, F. C., R. D. Arbeit, R. V. Goering, P. A. Mickelsen, B. E. Murray, D. H. Persing, and B. Swaminathan.** 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J. Clin. Microbiol. **33:**2233–2239.

24. **van Belkum, A., M. Struelens, A. de Visser, H. Verbrugh, and M. Tibayrenc.** 2001. Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. Clin. Microbiol. Rev. **14:**547–560.

25. **van Belkum, A., W. van Leeuwen, M. E. Kaufmann, B. Cookson, F. Forey, J. Etienne, R. Goering, F. Tenover, C. Steward, F. O'Brien, W. Grubb, P. Tassios, N. Legakis, A. Morvan, N. El Solh, R. de Ryck, M. Struelens, S. Salmenlinna, J. Vuopio-Varkila, M. Kooistra, A. Talens, W. Witte, and H. V. L. Verbrugh.** 1998. Assessment of resolution and intercenter reproducibility of results of genotyping *Staphylococcus aureus* by pulsed-field gel electrophoresis of SmaI macrorestriction fragments: a multicenter study. J. Clin. Microbiol. **36:**1653–1659.

26. **Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, et al.** 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23:**4407–4414.