Epidemiol. Infect. (2006), **134**, 1004–1014. © 2006 Cambridge University Press
doi:10.1017/S0950268806006145 Printed in the United Kingdom

Optimal settings of fingerprint-type analysing computer software for the analysis of enterohaemorrhagic *Escherichia coli* pulsed-field gel electrophoresis patterns

E. YOKOYAMA* AND M. UCHIMURA

Division of Bacteriology, Chiba Prefectural Institute of Public Health

(*Accepted 3 February 2006, first published online 28 March 2006*)

SUMMARY

Settings of fingerprint-type analysing computer software were optimized for analysis of enterohaemorrhagic *Escherichia coli* (EHEC) pulsed-field gel electrophoresis (PFGE) patterns. Under the lowest values of parameters, maximum value of similarities calculated using the Dice coefficient were obtained between PFGE patterns from one EHEC strain on the same gel when reference lanes for calibration of distortions during electrophoresis were set to every fourth lane. PFGE patterns of 15 EHEC strains on different gels were investigated. Similarity values calculated using the Pearson product-moment correlation coefficient (Pearson correlation) were significantly higher than those using the Dice coefficient with optimal values of parameters determined by the program ($P<0{\cdot}01$). When PFGE patterns of 45 EHEC strains were analysed by the computer program, EHEC strains from one mass outbreak and three intra-family outbreaks were each clustered and the similarity values within the clusters were $>90\,\%$ using Pearson correlation.

INTRODUCTION

Diffuse outbreaks caused by enterohemorrhagic *Escherichia coli* (EHEC) are one of the most important public health problems. Analysis of EHEC macrorestriction DNA fragment patterns generated by pulsed-field gel electrophoresis (PFGE) is useful for identifying the source of diffuse outbreaks in Japan [1–3]. In Japan, a surveillance system using PFGE pattern analysis is being investigated for the early stage detection of diffuse EHEC outbreaks and collaborative studies were done to compare the quality of PFGE patterns generated by different laboratories [4]. Visual inspection of the gels to identify differences in band patterns was used to pinpoint strain similarities [5]. However, visual inspection is problematic when many PFGE patterns need to be compared [6] and fingerprint-type computer software packages are available for use in these situations.

The fingerprint-type analysing computer software uses a relational database program to identify similarities of PFGE patterns and construct dendrograms. Such computer software programs enable investigators to compare large numbers of complex PFGE patterns in a short period of time [7–9]. However, the distortions that occur during electrophoresis make normalization of PFGE patterns among different gels necessary to correctly evaluate genetic relationships of bacteria [8–13]. The normalization of the PFGE patterns can easily be done by computer programs that use reference lanes on each gel. However, the effect of normalization may be different, depending on the position of the reference lanes on each gel. The common practice of positioning of reference lanes at only the outermost lane of a gel may give inaccurate

* Author for correspondence: Dr E. Yokoyama, Division of Bacteriology, Chiba Prefectural Institute of Public Health, 666-2 Nitona, Chuo, Chiba City, Chiba, 260-8715 Japan.
(Email: e.ykym@ma.pref.chiba.lg.jp)

normalization of PFGE patterns since these lanes are often influenced by smiling and other distortions [7]. The position of the reference lanes is not usually provided in publications.

Generally, two types of calculations are used by fingerprint-type analysing software for determining similarity of PFGE patterns. The Pearson product-moment correlation coefficient (Pearson correlation) [14] calculates similarity values based on densitometric curves of PFGE patterns. The Dice coefficient is obtained by comparison of bands at specific positions and uses probability theory to determine whether the bands are similar or distinct [15]. Whether the Dice and Pearson methods for calculating similarities are comparable is unknown. Similarities of PFGE patterns are calculated using the Dice coefficient in most studies. However, when bands of EHEC PFGE patterns sometimes overlap, they are designated as a single band [16] and calculations of similarity using the Dice coefficient would be incorrect. How to evaluate overlapping bands when using the Dice coefficient is unknown. Moreover, the manual editing that is often needed to obtain the accurate band position after automatic band detection [7] also influences the results calculated using the Dice coefficient because of individual differences in band evaluation by investigators.

Several parameters should be set at each step of the analysis when using fingerprint-type analysing software. The 'tolerance' is the maximal allowable shift in percentage of the pattern length between two bands that considers these bands as matching and applies to all calculations giving a band-matching coefficient [17]. The 'optimization' parameter, present in certain software packages, allows a shift between any two patterns. The software will then look for the best possible matching within the shift. The optimization parameter applies to both band-matching coefficient and densitometric curve-based coefficient [17]. The influence of the parameter settings on similarity calculations in previously reported studies is unknown.

In this study, we examined the conditions for normalizing the PFGE patterns both within the same, as well as among different, gels and compared the similarities using Dice coefficient and Pearson correlation under various setting values of 'tolerance' and 'optimization'. The optimal conditions and parameter settings for normalization were used to evaluate clustering of EHEC PFGE patterns that included some gels with outbreak-derived strains.

MATERIALS AND METHODS

PFGE

Briefly, strains were grown with agitation in Luria–Bertani (LB) broth at 37 °C overnight. A volume of 70 μl of the suspensions was centrifuged at 9000 ***g*** for 5 min and resuspended in 100 μl of Salt–EDTA buffer (75 mM NaCl, 0·1 M EDTA) with 4 μl of lysozyme (25 mg/ml, Merck, Darmstadt, Germany). A volume of 100 μl of 1% low melt agarose (Bio-Rad, Hercules, CA, USA) was added and the mixture was dispensed into a disposable plug mould (Bio-Rad). After solidification, the plugs were transferred to microtubes containing lysis solution [1 M NaCl, 0·1 M EDTA (pH 8·0), 0·5% Briji-58, 0·2% deoxycholate, 0·5% sarkosyl, 1 mg/ml lysozyme] and incubated at 37 °C for 2 h. After lysis, the plugs were transferred to microtubes containing proteinase K solution [0·25 M EDTA (pH 8·0), 1% sarkosyl, 1 mg/ml proteinase K (Merck)] and incubated at 50 °C for 18 h. After incubation, the plugs were washed for 30 min once in Tris–EDTA (TE) buffer [10 mM Tris, 1 mM EDTA (pH 8·0)] containing 1 mM phenylmethlsulphonyl fluoride (PMSF) (Sigma, St Louis, MO, USA) and three times in TE buffer without PMSF. The plugs were stored at 4 °C in TE buffer until use. Before digestion, the plugs were equilibrated in 0·1× TE buffer for 30 min and digested by *Xba*I (Roche, Basel, Switzerland) at 37 °C for 18 h. Restriction fragments were separated in 1% pulsed field certified agarose (Bio-Rad) using the CHEF DRIII system (Bio-Rad). A 20-well sample comb (Bio-Rad) was used to make 20 lanes on a gel. A λ DNA ladder marker (BMA, Rockland, ME, USA) was used as molecular weight standard. The conditions of switching and run time were from 4 to 8 s for 9 h and from 8 to 50 s for 13 h at 200 V. Images of banding patterns were saved as TIFF files using Gel Print200i/VGA (Bio Image, MI, Ann Arbor, USA).

Computerized analysis of EHEC PFGE patterns

Fingerprinting II version 3 (Bio-Rad) software was used to analyse PFGE patterns. Spectral analysis of the densitometric curve was carried out to obtain setting values of background subtraction and least-square filtering. After those values were set, a band search was done automatically under default setting parameters. Active reference was set to the 48·5–630·5 kbp fragments of the molecular-weight markers to normalize PFGE patterns. PFGE patterns on different gels were normalized for compression or

https://doi.org/10.1017/S0950268806006145 Published online by Cambridge University Press

Lane interval for references	Setting patterns of reference lanes										
8	R				R				R		X
6	R			R			R			R	
4	R		R		R		R		R		X
2	R	R	R	R	R	R	R	R	R	R	

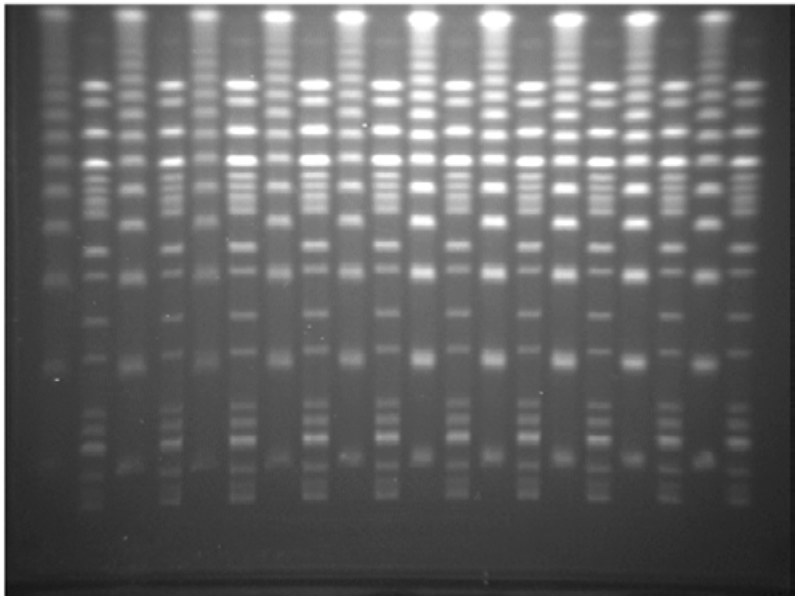

Fig. 1. Location of molecular-weight marker and setting of reference lanes. R, Lane was set as reference lane; X, lane was excluded on calculation of similarity.

stretch by matching reference lanes on individual gels to the active reference lane.

Similarities were calculated by both Dice coefficient and Pearson correlations. The optimal values of tolerance and optimization were calculated by the software. In Dice coefficient analysis, the bands that were present from 48·5 to 630·5 kbp were subjected to similar calculation. Following the automatic detection of bands, the bands that were outside the 48·5–630·5 kbp range were excluded from analysis by manual deletion. In Pearson correlation analysis, the densitometric curves between 48·5 and 630·5 kbp were subjected to similarity calculations. The resulting dendrogram was made by the unweighted pair-group method using arithmetic averages (UPGMA).

Normalization and analysis of parameter settings for similarity calculation of PFGE patterns on the same gel

Molecular-weight markers and one arbitrarily chosen EHEC serovar O157:H7 strain (VT1) were located adjacent to each other on a gel. Four different setting patterns of reference lanes were tested. First, lanes 1, 3, 5, 9, 11, 13, 15, 17 and 19 were used to set molecular-weight markers as reference, representing references set at every second lane. Second, lanes 1, 5, 9, 13 and 17 of molecular-weight markers were set as reference lanes, representing reference lanes set at every fourth lane. Third, lanes 1, 7, 13 and 19 of molecular-weight markers were set as reference lanes, representing references set at every sixth lane. Fourth, lanes 1, 9 and 19 of molecular-weight markers were set as reference lanes, representing references set at every eighth lane (Fig. 1).

The similarities between two lanes of EHEC PFGE patterns on a gel were analysed by reading the similarity values from the dendrogram. Because lane 20 was not captured by two reference lanes and was not adjoined to one reference lane when settings were every fourth or every eighth lane (Fig. 1), it was excluded from similarity calculations when those settings were used.

Normalization and analysis of parameter settings for similarity calculation on different gels

Fifteen EHEC serovar O157:H7 strains, isolated from both epidemiologically unrelated patients

https://doi.org/10.1017/S0950268806006145 Published online by Cambridge University Press

and asymptomatic carriers, were investigated. Out of the 15 strains, three strains produced VT1, one strain produced VT2, and 11 strains produced VT1 and VT2. All 15 strains were analysed twice using PFGE. The most appropriate setting patterns of reference lanes were used for this experiment. The similarities of the PFGE patterns from the same strain between lanes on two different gels were calculated.

Clustering analysis of EHEC PFGE patterns

Forty-one EHEC strains which were isolated from patients and asymptomatic carriers during 2002 in the Chiba prefecture of Japan were studied. Out of 41 strains, 25 strains were serovar O157:H7, six strains were serovar O26:H11, and 13 strains were serovar O103:H2. Out of the 25 serovar O157:H7 strains, one strain produced VT1, 14 strains produced VT2, and 10 strains produced VT1 and VT2. One VT1-producing strain was derived from a sporadic case. In the 14 VT2-producing strains, two strains were derived from 'intra-family A' outbreak, two strains were derived from 'intra-family B' outbreak, and 10 strains each were derived from sporadic cases. In the 10 VT1- and VT2-producing strains, two were derived from 'intra-family C' outbreak and eight were each derived from each sporadic cases. All of the six serovar O26:H11 strains produced VT1. Out of the six strains, three were derived from 'intra-family D' outbreak and three were derived from each sporadic case. All of the 13 serovar O103:H2 strains produced VT1 and were derived from a single mass outbreak. The tested strains were analysed by PFGE with five different gels. The most appropriate setting pattern of reference lanes was used for this experiment. The similarities were calculated and a dendrogram was constructed.

Statistical analysis

The similarities calculated by the Dice coefficient under computer analysing parameters and with the arbitrarily increased parameters were compared using Wilcoxon signed-rank test analysis. The similarities calculated by either Dice coefficient or Pearson correlation were compared using Wilcoxon signed-rank test analysis. Statistical significance was at the $P<0{\cdot}01$ level.

RESULTS

Normalization and analysis of parameter settings for similarity calculation on the same gel

The values of 0·21% 'tolerance' and 0% 'optimization' were obtained using automatic software analysis. When the similarities were calculated using the Dice coefficient under the computer analysed parameters, the median value of the similarities was 72% even though references were set at every second lane. The similarities were significantly increased with all of the reference lane setting patterns when tolerances were raised to 0·5, 0·75 and 1·0%. The similarities between all combinations of lanes on a gel were 99·99% under 0·75% tolerance and 0·5% optimization when references were set at every fourth lane, under 1·0% tolerance and 0% optimization when references were set at every second, or under 1·0% tolerance and 0·5% optimization when references were set at every sixth and eighth lane.

Using Pearson correlation and the computer analysed optimization, the similarities were over 90% with all reference lane setting patterns. The similarities using Pearson correlation were significantly higher than that using Dice coefficient. The similarities using Pearson correlation were slightly improved when optimization was raised to 0·25%, but not significantly (Table 1).

Normalization and analysis of parameter settings for similarity calculation on different gels

In this analysis, the outermost lanes of a gel were not used and every fourth lane was a reference. The values of 0·32% tolerance and 0% optimization on two gels were obtained automatically by the computer software. When the similarities were calculated using Dice coefficient under the computer analysing parameters, the median value of the similarities was 63%. A significant increase of the similarities was observed when tolerance was raised to 0·5, 0·75 and 1·0%. Significantly higher similarity values were obtained using Pearson correlation with the computer analysing optimization when compared to those obtained using Dice coefficient under tolerances of 0·32 and 0·5% (Table 2).

Decreases of the similarities were observed in 11 out of 15 strains, according to the increase of parameter values. When the tolerance was raised from 0·5 to 0·75%, the similarities of EHEC strains 1, 2, 4, 5, 6, 7 and 11 were decreased. This adverse effect was

https://doi.org/10.1017/S0950268806006145 Published online by Cambridge University Press

Table 1. *Impact of setting patterns of reference lanes, calculating method, and parameter settings on similarity of banding patterns*

Reference lane	Calculating method	Parameter setting: Tolerance	Optimization: 0	0·25	0·5	0·75	1·0
Every 2nd lane	Dice coeff.	0·21	72 (35–91)*	78 (77–91)	78 (77–91)	78 (77–91)	78 (77–91)
		0·5	97 (91–100)	96 (96–100)	99 (96–100)	99 (96–100)	99 (96–100)
		0·75	98 (98–100)	100 (99–100)	100 (99–100)	100 (99–100)	100 (99–100)
		1·0	100 (99–100)	Identical†	Identical	Identical	Identical
	Pearson corr.	—	96 (90–98)	97 (96–98)	97 (96–98)	97 (96–98)	97 (96–98)
Every 4th lane	Dice coeff.	0·21	50 (24–92)	78 (76–92)	79 (77–92)	79 (77–92)	79 (77–92)
		0·5	94 (84–100)	95 (95–100)	100 (96–100)	100 (96–100)	100 (96–100)
		0·75	97 (97–100)	100 (100–100)‡	Identical	Identical	Identical
		1·0	100 (99–100)	Identical	Identical	Identical	Identical
	Pearson corr.	—	94 (92–94)	96 (96–99)	96 (96–99)	96 (96–99)	96 (96–99)
Every 6th lane	Dice coeff.	0·21	54 (54–86)	84 (76–97)	84 (76–97)	84 (76–97)	84 (76–97)
		0·5	97 (82–100)	100 (94–100)	100 (94–100)	100 (94–100)	100 (94–100)
		0·75	100 (87–100)	100 (97–100)	100 (99–100)	100 (99–100)	100 (99–100)
		1·0	97 (97–100)	99 (99–100)	Identical	Identical	Identical
	Pearson corr.	—	96 (94–98)	96 (96–99)	96 (96–99)	96 (96–99)	96 (96–99)
Every 8th lane	Dice coeff.	0·21	68 (67–88)	78 (74–90)	78 (74–90)	78 (74–90)	78 (74–90)
		0·5	96 (81–100)	98 (91–100)	98 (93–100)	98 (93–100)	98 (93–100)
		0·75	100 (92–100)	99 (98–100)	98 (98–100)	98 (98–100)	98 (98–100)
		1·0	99 (98–100)	Identical	Identical	Identical	Identical
	Pearson corr.	—	95 (95–97)	98 (95–99)	98 (95–99)	98 (95–99)	98 (95–99)

* Median (10 percentile–90 percentile) values of similarity of EHEC PFGE patterns.
† All values of similarities were 99·99%.
‡ The similarities of three combinations of lanes on a gel were 99%.

Table 2. *Change in similarities of 15 EHEC strains associated with different methods of calculation and parameter settings*

Calculating method	Parameter setting: Tolerance	Optimization: 0	0·25	0·5	0·75	1·0
Dice coeff.	0·32	63 (23–87)*	75 (37–93)	72 (60–93)	76 (63–93)	76 (63–93)
	0·5	76 (37–95)	96 (54–100)	96 (71–100)	96 (85–100)	96 (88–100)
	0·75	95 (58–100)	95 (75–100)	96 (85–100)	96 (90–100)	96 (90–100)
	1·0	96 (74–100)	96 (87–100)	97 (91–100)	97 (92–100)	97 (93–100)
Pearson corr.	—	96 (83–97)	96 (92–98)	97 (92–98)	97 (92–98)	97 (92–98)

* Median (10 percentile–90 percentile) values of similarity of EHEC PFGE patterns.

also observed when tolerance was raised from 0·75 to 1·0% in EHEC 5, 6, 8 and 9 and in EHEC 3, 6, 8, 9 and 10 when 'optimization' was raised (Table 3).

It was impossible to obtain the 100% maximum value of the similarities calculated using the Dice coefficient in nine out of 15 strains, even though 1·0% tolerance and 1·0% optimization were set for the calculation. Out of the nine strains tested, there were differences in two bands in one strain and one band in eight strains with automatic band detection. These phenomena were due to different sensitivity of band detection at peak-shoulder in four strains and

https://doi.org/10.1017/S0950268806006145 Published online by Cambridge University Press

Table 3. *Adverse effect similarities of EHEC strains according to changes in parameters*

		Parameter setting					
			Optimization				
Strain (toxigenicity)	Method	Tolerance	0	0·25	0·5	0·75	1·0
EHEC 1 (VT1)	Dice coeff.	0·32	23	37	73	73	73
		0·5	37	80	87	93	93
		0·75	58	75	100	100	100
		1·0	74	100	100	100	100
	Pearson corr.	—	84	94	97	97	97
EHEC 2 (VT1 & 2)	Dice coeff.	0·32	63	87	87	87	87
		0·5	75	100	100	100	100
		0·75	93	94	96	96	96
		1·0	94	98	100	100	100
	Pearson corr.	—	94	94	94	94	94
EHEC 3 (VT1 & 2)	Dice coeff.	0·32	63	64	72	63	63
		0·5	75	94	94	94	94
		0·75	100	100	100	100	100
		1·0	100	100	100	100	100
	Pearson corr.	—	96	96	96	96	96
EHEC 4 (VT1)	Dice coeff.	0·32	76	76	76	76	76
		0·5	97	97	97	97	97
		0·75	95	95	95	95	95
		1·0	95	95	95	95	95
	Pearson corr.	—	98	98	98	98	98
EHEC 5 (VT1 & 2)	Dice coeff.	0·32	64	75	67	76	76
		0·5	78	97	97	97	97
		0·75	97	97	96	96	96
		1·0	97	96	97	97	97
	Pearson corr.	—	97	97	97	97	97
EHEC 6 (VT1 & 2)	Dice coeff.	0·32	64	75	67	76	76
		0·5	78	95	95	95	95
		0·75	93	96	92	92	92
		1·0	96	93	93	93	93
	Pearson corr.	—	97	97	97	97	97
EHEC 7 (VT2)	Dice coeff.	0·32	72	72	72	72	72
		0·5	72	96	96	96	96
		0·75	95	95	96	96	96
		1·0	96	96	96	96	96
	Pearson corr.	—	97	97	97	97	97
EHEC 8 (VT1 & 2)	Dice coeff.	0·32	35	37	43	51	51
		0·5	37	54	63	85	86
		0·75	58	91	97	95	95
		1·0	74	95	91	91	91
	Pearson corr.	—	62	93	93	93	93
EHEC 9 (VT1 & 2)	Dice coeff.	0·32	23	37	67	79	76
		0·5	37	54	71	85	93
		0·75	58	75	81	87	87
		1·0	70	80	84	82	82
	Pearson corr.	—	83	89	91	91	91
EHEC 10 (VT1 & 2)	Dice coeff.	0·32	65	64	64	64	64
		0·5	76	100	100	100	100
		0·75	100	100	100	100	100
		1·0	100	100	100	100	100
	Pearson corr.	—	96	96	96	96	96
EHEC 11 (VT1)	Dice coeff.	0·32	87	87	87	87	87
		0·5	93	93	93	93	93
		0·75	92	95	95	95	95
		1·0	95	95	95	95	95
	Pearson corr.	—	96	96	96	96	96

https://doi.org/10.1017/S0950268806006145 Published online by Cambridge University Press

difference of high-density bands in the remaining five strains.

Clustering analysis of EHEC PFGE patterns

In this analysis, the outermost lanes of a gel were not used and every fourth lane was set as a reference. The values of 1% tolerance and optimization were set when similarities were calculated using the Dice coefficient. The similarities among strains in intra-family A, B, C, and D cases were >90%, while the similarity of mass outbreak-derived strains were <90% (Fig. 2). In contrast, similarities of mass outbreak-derived strains, as well as all intra-family case-derived strains, were >90% using Pearson correlation and 0·5% optimization (Fig. 3).

DISCUSSION

Our results indicated that calculating similarities of PFGE patterns using the Dice coefficient gave conflicting data regarding epidemiological relationships among EHEC strains. When the Dice coefficient was used for calculations, the PFGE patterns of the same strain on a gel were not normalized sufficiently under the computer analysis parameters, even though every second lane was set for reference. Moreover, similarities of same strains on two different gels were low when tolerance and optimization obtained by automatic software analysis were used for calculation. Increases in stringency of the analysing parameters were necessary to improve the normalization and obtain accurate information about the similarities of PFGE patterns. The same tendency was observed when the similarities of strains derived from a mass outbreak were calculated. However, it is difficult to determine the level of tolerance and optimization. In the experiment using two gels, the similarities of the same strains increased when the values of tolerance and optimization were raised, but in nine strains, the similarities were decreased when the values of tolerance were increased from 0·5 to 0·75% or from 0·75 to 1%. This phenomenon should be called the 'tolerance paradox' and is probably due to forced matching of unrelated bands of unrelated strains by the increases in tolerance. We suggest that the setting of a high tolerance value may lead to a false clustering of unrelated strains. Davis *et al.* [16] also reported that the Dice coefficient gave a poor estimate of genetic relatedness between two isolates when PFGE of EHEC was carried out using only one restriction enzyme. Calculating similarity using the Dice coefficient also presents the problem of bands of EHEC PFGE patterns that overlap each other giving the impression of a single band [16]. The overlapped bands would lead to incorrect similarity and an imprecise dendrogram if calculations were done. The problems with the Dice coefficient are likely to be present in almost all previous investigations, leaving in question the real genetic relatedness of the EHEC strains that were tested. All band-matching coefficient calculations may be subject to the same source of inaccuracy.

In contrast, our results indicated that the Pearson correlation was suitable for comparing EHEC PFGE patterns. When the Pearson correlation was used for calculation, the PFGE patterns of the same strain on a gel were sufficiently normalized in all setting patterns of reference lanes under the computer analysis parameters. The similarities of PFGE patterns in two different gels was >90% in all 15 strains under the values of tolerance and optimization generated by automatic computer software analysis. These similarities were higher than those calculated using the Dice coefficient. The similarities using the Pearson correlation were slightly improved by an increase in optimization, without observing the 'tolerance paradox'. Pearson correlation has two advantages compared to the Dice coefficient since similarities are calculated based on the densitometric curves of PFGE patterns and is independent of band definition [14]. The first advantage is that Pearson correlation is not affected by overlapping bands that may lead to incorrect similarities when the Dice coefficient is used. Overlapped bands are reflected in the densitometric curve, which is the basis for calculating similarities using the Pearson correlation. The second advantage is that the peak-shoulder mismatches often found with the Dice coefficient [12] are not a factor in the Pearson correlation. In this study, the peak-shoulder mismatches with the Dice coefficient were found in four out of nine strains in which we could not obtain 100% similarities between two different gels. The advantages of using the Pearson correlation may be observed with other curve-based coefficient calculation methods. A report suggesting that the Pearson correlation is influenced when background intensities are different [12] is contrary to our findings. It is clear that a high-quality PFGE pattern with low background and high contrast is important for adequate analysis. However, the majority of software programs can remove background noise,

https://doi.org/10.1017/S0950268806006145 Published online by Cambridge University Press

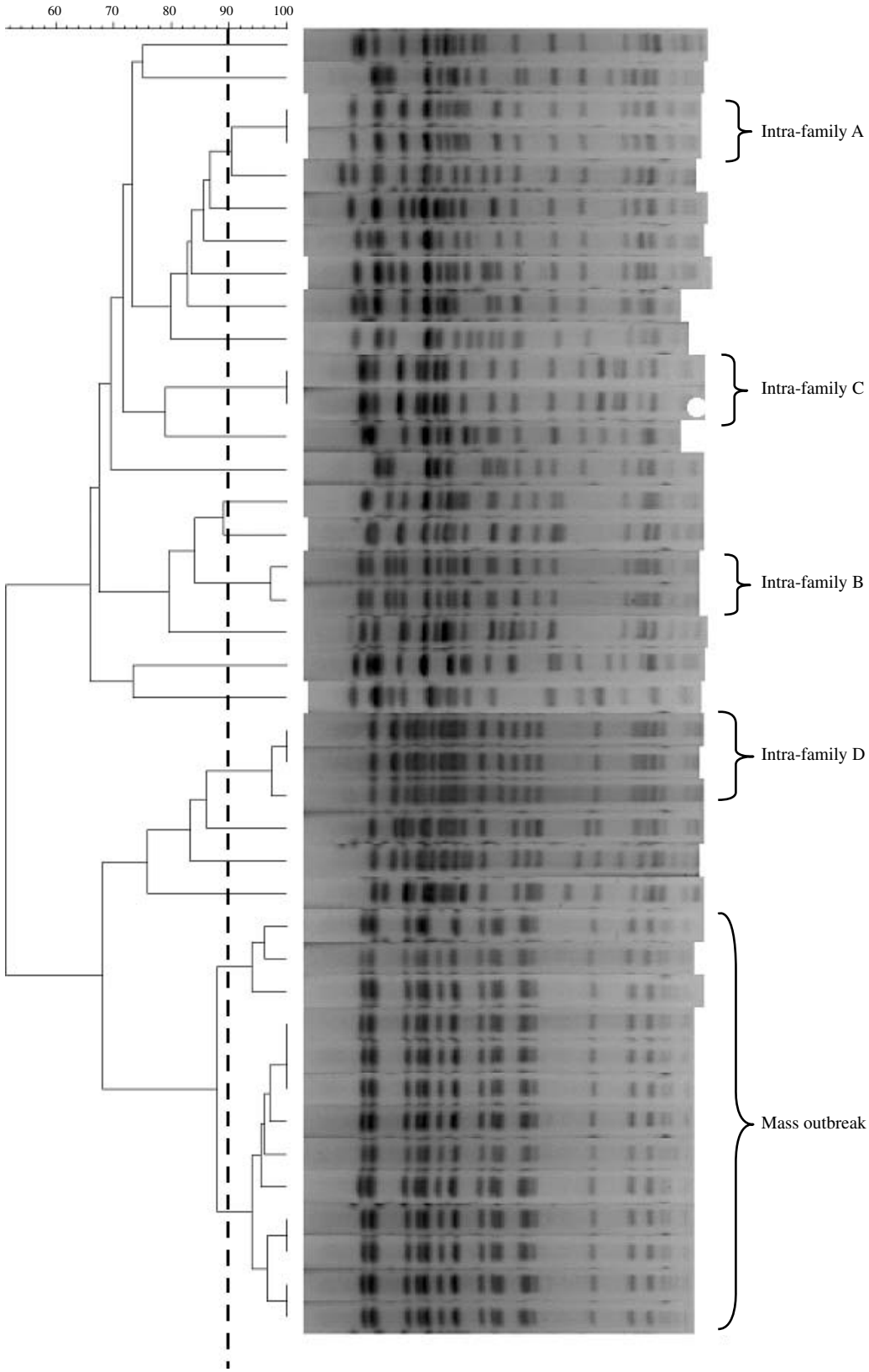

Fig. 2. Dendrogram of EHEC strains using the Dice coefficient under 1·0% tolerance and 1·0% optimization.

https://doi.org/10.1017/S0950268806006145 Published online by Cambridge University Press

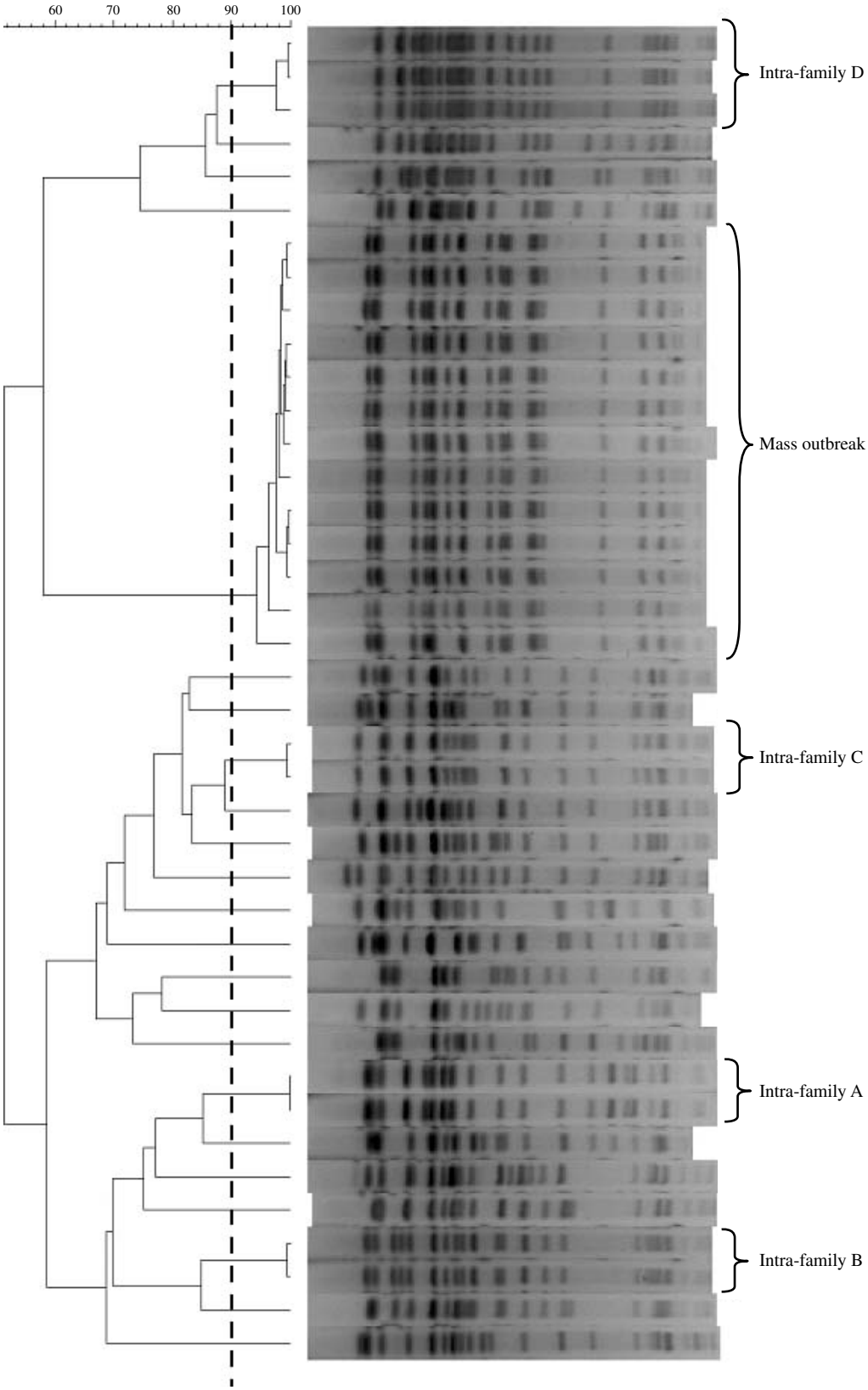

Fig. 3. Dendrogram of EHEC strains using the Pearson correlation under 0·5% optimization.

https://doi.org/10.1017/S0950268806006145 Published online by Cambridge University Press

and minimize the influence on Pearson correlation calculations.

Our results also indicated that a 90% cut-off value is appropriate when similarities of EHEC PFGE patterns are calculated by Pearson correlation with 0·5% optimization. In this study, the similarities among all epidemiologically related strains were >90% and <90% for all epidemiologically unrelated strains. If clustering computer software does not have an optimization parameter, the cut-off value should be set lower than 90% since similarities calculated by Pearson correlation were slightly improved when the optimization was increased. However, a cluster possessing over 90% similarities does not always indicate occurrence of an outbreak since PFGE does not sufficiently resolve bands that are near in size [16]. The epidemiological information should be accompanied to genetic analysis, even if computer analysis of EHEC PFGE patterns is carried out under optimal conditions.

In this investigation, we tried to exclude the manual operation by automatic settings of the computer software as much as possible. We only used manual editing to delete some of the detected bands that existed out of the size range of the active reference markers when similarities were calculated using the Dice coefficient. The densitometric curves that were outside the range of active reference markers were also excluded when similarities were calculated using the Pearson correlation. Comparison of PFGE patterns using similarities between bacterial strains is widely used to investigate bacterial genetic relationships. However, most papers give no information on how to normalize PFGE patterns and the condition of analysing parameters. The setting patterns of reference lanes are also important to normalize PFGE patterns on either the same gel or different gels [8–13]. In this study, for analysis of parameter settings on similarity calculation and cut-off value, the two outermost lanes were not used, and every fourth lane was used for reference. Such conditions provided the lowest tolerance for obtaining the maximum similarities calculated using the Dice coefficient for comparison with similarities calculated using the Pearson correlation. Recently, it was reported that a *Salmonella enterica* serovar Braenderup strain (H9812) was used instead of a λ DNA ladder marker because it yields a wide range of DNA restriction fragments and facilitates accurate comparison of PFGE patterns [18]. This new marker may provide better normalization than a λ DNA ladder marker. However, the conditions for normalizing both inter- and intra-gel PFGE patterns should be investigated in all laboratories for similarity analysis. The degree of distortion during electrophoresis may be different among laboratories because of variability in levels of use and different ages of PFGE equipment.

In conclusion, Pearson correlation should be used to calculate similarity to compare genetic relationship of EHEC using computer software analysis of PFGE patterns. Calibration of the distortion of PFGE is required for proper analysis.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Asai Y, *et al.*** Isolation of shiga toxin-producing *Escherichia coli* O157:H7 from processed salmon roe associated with the outbreak in Japan, 1998 and a molecular typing of the isolates by pulsed-field gel electrophoresis [in Japanese with English summary]. *Journal of the Japanese Association of Infectious Diseases* 1999; **73**: 20–24.
2. **Ozeki Y, *et al.*** A diffuse outbreak of enterohemorrhagic *Escherichia coli* O157:H7 related to Japanese-style pickles in Saitama, Japan [in Japanese with English summary]. *Journal of the Japanese Association of Infectious Diseases* 2003; **77**: 493–498.
3. **Yokoyama E, Uchimura M, Koiwai K.** A diffuse outbreak by enterohemorrhagic *Escherichia coli* serovar O157:H7 associated with contaminated lightly roasted beef [in Japanese]. *Japanese Journal of Food Microbiology* 2004; **21**: 156–159.
4. **Watanabe H, *et al.*** PulseNet Japan: surveillance system for the early detection of diffuse outbreak based on the molecular epidemiological method [in Japanese with English summary]. *Journal of the Japanese Association of Infectious Diseases* 2002; **76**: 842–848.
5. **Chung M, *et al.*** Molecular typing of methicillin-resistant *Staphylococcus aureus* by pulsed-field gel electrophoresis: comparison of results obtained in a multilaboratory effort using identical protocols and MRSA strains. *Microbial Drug Resistance* 2000; **6**: 189–198.
6. **Tenovor FC, *et al.*** Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. *Journal of Clinical Microbiology* 1995; **33**: 2233–2239.
7. **Smidt PG, *et al.*** Computerized analysis of restriction fragment length polymorphism patterns: comparative evaluation of two commercial software packages. *Journal of Clinical Microbiology* 1998; **36**: 1318–1323.
8. **Garaizar J, *et al.*** Suitability of PCR fingerprinting, infrequent-restriction-site PCR, and pulsed-field gel

electrophoresis, combined with computerized gel analysis, in library typing of *Salmonella enterica* serovar Enteritidis. *Applied and Environmental Microbiology* 2000; **66**: 5273–5281.

9. **Rementeria A, *et al.*** Comparative evaluation of three commercial software packages for analysis of DNA polymorphism patterns. *Clinical Microbiology and Infection* 2001; **7**: 331–336.
10. **Duck WM, *et al.*** Optimization of computer software settings improves accuracy of pulsed-field gel electrophoresis macrorestriction fragment pattern analysis. *Journal of Clinical Microbiology* 2003; **41**: 3035–3042.
11. **Cardinali G, *et al.*** Multicenter comparison of three different analytical systems for evaluation of DNA banding patterns from *Cryptococcus neoformans*. *Journal of Clinical Microbiology* 2002; **40**: 2095–2100.
12. **de Boer P, *et al.*** Computer-assisted analysis and epidemiological value of genotyping methods for *Campylobacter jejuni* and *Campylobacter coli*. *Journal of Clinical Microbiology* 2000; **38**: 1940–1946.
13. **Houang ETS, *et al.*** Study of the related ness of isolates of *Shigella flexneri* and *Shigella sonnei* obtained in 1986 and 1987 and in 1994 and 1995 from Hong Kong. *Journal of Clinical Microbiology* 1998; **36**: 2404–2407.
14. **Pearson K.** On the coefficient of racial likeliness. *Biometrika* 1926; **18**: 105–117.
15. **Dice LR.** Measures of the amount of ecological association between species. *Journal of Ecology* 1945; **26**: 297–302.
16. **Davis MA, *et al.*** Evaluation of pulsed-field gel electrophoresis as a tool for determining the degree of genetic relatedness between strains of *Escherichia coli* O157:H7. *Journal of Clinical Microbiology* 2003; **41**: 1843–1849.
17. **Applied Maths.** GelCompar II Version 2.0 Manual.
18. **Hunter SB, *et al.*** Establishment of a universal size standard strain for use with the PulseNet standardized pulsed-field gel electrophoresis protocols: converting the national database to the new size standard. *Journal of Clinical Microbiology* 2005; **43**: 1045–1050.