Numerical Analysis of Normalized Whole-cell Protein Profiles after Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis

By BRIAN D. PLIKAYTIS,1* GEORGE M. CARLONE2 AND BONNIE B. PLIKAYTIS3

Statistical Services Activity,1 Molecular Biology Laboratory, Meningitis and Special Pathogens Branch,2 and Immunology Laboratory, Respiratory Diseases Branch,3 Division of Bacterial Diseases, Center for Infectious Diseases, Centers for Disease Control, Atlanta, Ga. 30333, USA

(Received 27 January 1986; revised 21 April 1986)

A technique is described for mathematically normalizing whole-cell protein profiles after sodium dodecyl sulphate-polyacrylamide gel electrophoresis to obtain standardized absolute migration distances using two internal $M$, standards. A soft laser scanning densitometer was used to measure protein band migration distances in wet, silver-stained gels. The normalized values were superior to the unnormalized migration distances and common $R_f$ values in reducing the inter- and intragel variability of the protein band positions. A procedure is described for clustering normalized bacterial protein profiles using a sample data set obtained from the type strains of four Legionella species.

INTRODUCTION

Gel electrophoresis is increasingly being used to great advantage as an aid in differentiating and classifying micro-organisms. This method has been used both singly (Jackman, 1985; Kersters, 1985) and in combination with biochemical and phenotypic characteristics (Izard et al., 1981; Vera Cruz et al., 1984) to classify and identify different species of bacteria. This technique has also been shown to be in good agreement with DNA–DNA hybridization experiments (Kersters, 1985; Jackman, 1985).

Kersters & De Ley (1980) reported that electrophoregrams (scans of protein banding patterns or profiles) of bacterial strains produced under well-defined and standardized conditions can be considered unique and that genetically closely related bacteria display similar or almost identical electrophoregrams. However, experimental variability in factors associated with gel electrophoresis precludes the comparison of absolute migration distances among gels (Seiter & Jay, 1980). Many investigators standardize or normalize the migration distances of the protein bands by computing $R_f$ values, i.e. the ratio of the migration distances of the protein bands to that of a tracking dye at the bottom of the gel (Hames, 1981). However, $R_f$ values are non-reproducible in common practice, and they do not completely eliminate the variability of the protein band positions within gels (Seiter & Jay, 1980; Chrambach et al., 1976). Kersters & De Ley (1975, 1980) incorporated two internal standards in each sample preparation to normalize the resulting protein band migration distances, thus facilitating gel-to-gel comparisons of electrophoregrams. Additional techniques involve the inclusion of reference or type strains on each gel to normalize the remaining profiles on that gel (Seiter & Jay, 1980), but it is unclear to what extent this approach reduces the intragel variability.

Once the electrophoregrams have been normalized, several protein profiles may be compared for taxonomic classification and discrimination. Although photographs or tracings of a small number of profiles may be examined visually, this form of comparison is subjective. Large numbers of samples must be examined quantitatively, as it is impossible to distinguish mentally, and remember differences among, large numbers of samples (Kersters, 1985).
Correlation or similarity coefficients are sometimes calculated between pairs of normalized protein profiles (Izarid et al., 1981; Vera Cruz et al., 1984) and taxonomic structure may be determined by clustering the data using widely available computer programs (Kersters & De Ley, 1975). The success of the classification is largely determined by how well the normalization techniques reduce the variability of the band positions and increase the reproducibility of the electrophoretic runs (Kersters & De Ley, 1975). This concern is as important when making comparisons between pairs of profiles on one gel as it is among many gels.

We report a technique for mathematically normalizing whole-cell protein profiles after sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) to obtain standardized absolute migration distances, using two internal $M_t$ standards, and a procedure for clustering normalized bacterial protein profiles. Type strains of four *Legionella* species were used as a sample data set.

**METHODS**

**Bacterial cultures and antigen preparation.** Legionella pneumophila strain Philadelphia 1, *L.* dumoffii strain NY-23, *L.* bozemanii strain WIGA and *L.* micdadei strain TAYLOCK were obtained from the Immunology Laboratory, Respiratory Diseases Branch, Division of Bacterial Diseases, Centers for Disease Control (CDC), Atlanta, Ga., USA. Three independent cultures of each strain were grown and processed separately. Antigens were prepared by harvesting cells from buffered charcoal/yeast extract agar (BCYE) plates (Carr Scarborough Microbiologicals, Decatur, Ga., USA) that had been incubated at 37 °C for 48 h. Cells harvested from each plate were suspended in 2 ml sterile distilled water, centrifuged at 2000 g for 20 min at 4 °C and the cell pellets resuspended and stored at 4 °C in 2 ml sterile distilled water for approximately 2 weeks. The protein concentration was determined for each cell suspension by a modified Lowry procedure (Markwell et al., 1978) and cell suspensions were adjusted to 1:2 mg protein ml⁻¹ in sterile distilled water.

**Gel electrophoresis and silver stain.** SDS-PAGE was done by the discontinuous method of Neville & Glossman (1974) as modified by Tsang et al. (1983), using a 12% resolving gel. Equal volumes of sample buffer [5% (w/v) SDS, 10% (v/v) 2-mercaptoethanol and 20% (v/v) glycerol in 0-01 M-Tris HCl, pH 8-0] and cell suspension (50 μl) were mixed, heated at 100 °C for 3 min, and a 5 μl sample was applied to one of 15 wells.

Replicate samples of the three whole-cell protein preparations were made for each of the four species of *Legionella*. The replicate preparations were electrophoresed in four separate SDS-polyacrylamide gels, with the twelve lanes randomly assigned in each gel. Solutions of 2 mg ml⁻¹ each of β-galactosidase (116-25 kDa) (Sigma) and lysozyme (14.4 kDa) (Pentex) in 50% (v/v) glycerol and 0.5 M-NaCl were used for internal normalization. β-Galactosidase, lysozyme and sample buffer were mixed (1:1:3), heated at 100 °C for 3 min, and a 2 μl sample applied to each well.

β-Galactosidase and low $M_t$ standards (Bio-Rad) phosphorylase B (92.5 kDa), bovine serum albumin (66.2 kDa), ovalbumin (45 kDa), carbonic anhydrase (31 kDa), soybean trypsin inhibitor (21.5 kDa) and lysozyme (14.4 kDa) were used to quantify the reproducibility of the electrophoretic procedure. Standards were mixed with sample buffer (1:20) and treated in the same manner as the sample preparations. Equal volumes (3 μl) of the treated internal and Bio-Rad standards were applied to each well of four 10-lane gels, resulting in 2.4 and 1.8 μg protein, respectively. Proteins were separated in the stacking gel at 10 mA. The current was increased to 20 mA once the tracking dye reached the separating gel, and was maintained at 20 mA until the dye migrated to the bottom of the gel (approx. 2.5 h). The gels were stained with silver (Morrissey, 1981) and were analysed with a soft laser scanning densitometer (Biomed Instruments, Fullerton, Calif., USA) (Carlone et al., 1985). The binary data were collected using an Apple IIIC personal computer operated with an Auto-stepover program (Biomed Instruments) for data capture and a Videophoresis II program (Biomed Instruments) for data integration.

**Normalization.** To decrease the inter- and intragel variability of protein band positions and increase the resultant reproducibility of the electrophoretic runs, we normalized each protein profile to a fixed distance based on the mean position of the two internal $M_t$ standards (β-galactosidase and lysozyme). The revised or normalized distances were calculated using a derivative of a ranging technique proposed by Gower (1971):

\[
MD_{\text{norm}} = MD_{\text{vm}} + \left( (MD_{\text{vm}} - MD_{\text{gm}}) \times (MD - MD_{\text{gm}}) / (MD_{\text{vm}} - MD_{\text{gm}}) \right)
\]

where $MD$ = the actual migration distance to be normalized; $MD_{\text{vm}}$ and $MD_{\text{vm}}$ = the average mean migration distances, over all profiles on all gels, of the β-galactosidase and lysozyme internal $M_t$ standards, respectively; $MD_{\text{gm}}$ and $MD_{\text{gm}}$ = the individual migration distances for the β-galactosidase and lysozyme internal $M_t$ standards, respectively, in the lane being normalized; and $MD_{\text{norm}}$ = the final normalized migration distance. This transformation scaled the actual migration distances of each protein profile so that the internal standards were positioned on their calculated mean values and the remaining bands were adjusted to range about them. Thus, a lane which ran slightly higher or lower than the mean was expanded or contracted, by a proportional amount, to a fixed distance.
Once the revised or normalized migration distances were calculated, they were used to compute similarities, based on matching co-migrating band positions, between pairs of protein profiles using the Dice coefficient (Dice, 1945):

$$S_D = \frac{2a}{2a + u} = \frac{2a}{n_1 + n_2}$$

where $a$ = the number of bands in common between a pair of profiles; $u$ = the number of unmatched bands between the pair; and $n_1$ and $n_2$ = the total number of bands, respectively, in the first and second profiles. The similarity values were clustered using the unweighted pair group method with arithmetic means (UPGMA) and CLUSTAN 1C (Wishart, 1978).

For comparative purposes, $R_F$ values were computed using the lysozyme internal $M_f$ standard:

$$R_F = \frac{MD}{MD_{in}}$$

where $MD$ = the actual migration distance to be normalized; and $MD_{in}$ = the individual migration distance for the lysozyme internal $M_f$ standard in the profile being normalized.

**RESULTS AND DISCUSSION**

There were two common sources of variability inherent in our electrophoretic studies: gel-to-gel variability was introduced while attempting to terminate the electrophoretic run at a fixed distance, and within gel variability was observed when the middle lanes occasionally ran further than the outside lanes, creating a very slight but noticeable curved effect in the stained banding patterns. In an effort to decrease this variability, the protein migration distances of each protein profile were normalized using two internal $M_f$ standards present in each sample lane. This technique mathematically describes the normalization process of Kersters and De Ley (see Kersters, 1985) with photographic negative enlargements and, as such, is not subject to the variability introduced from the manual manipulation of the negatives. The method is also automated, obviating the time and effort involved with the photographic processing.

**Protein profile normalization**

To quantify how well the normalization process decreased the variance of band migration distances within and among gels, we electrophoresed a series of four gels, each with ten replicates of known $M_f$ standards, and analysed them with a soft laser scanning densitometer (Carlone et al., 1985). Each protein profile was normalized to the overall means of the $\beta$-galactosidase and lysozyme migration distances (0.813 and 15.565 cm, respectively). Use of these mean values over a set of arbitrarily selected numbers resulted in the least amount of adjustment of the protein band positions while retaining the overall scale of measurement. The means and standard deviations of the actual migration distances, the normalized distances and the $R_F$ values for each gel and the total of all four gels for each of the $M_f$ standards are detailed in Table 1. The coefficient of variation $[CV = (standard\ deviation/mean) \times 100]$ is also presented as a gauge of the relative variability of the measurements. This statistic is often used to compare the variability of different samples which have different arithmetic means (Rosner, 1982). A low CV is an indication that the sample variability is small in relation to the sample mean.

The standard deviations for each of the four gels for each $M_f$ standard is an indication of the within gel variability of the protein band positions (Table 1). The standard deviations for the totals of all four gels for each of the standards is an indication of the overall variability of the band positions among the four gels. The standard deviations of the actual migration distances were similar for all of the standards, indicating that the electrophoretic process was consistent from the top to the bottom of the gel. The standard deviations for the migration distances decreased in all but one case when comparing the normalized to the actual distance values. However, since the means for each standard are different and the $R_F$ value is of a different scale than the other two measurements, the CV is a more revealing statistic for these data. The CVs are quite large for the heavier standards (phosphorylase B and bovine serum albumin) with small migration distance means, and they decrease with decreasing $M_f$ (increasing migration distance means). This indicates that while the absolute standard deviations are similar from
standard to standard, without regard to $M_r$, the relative variability of the band positions is much greater for the heavier standards, since their mean migration distances are smaller, relative to their standard deviations, than the lighter standards. The CVs for the normalized migration distances are demonstrably less than those of the actual migration distances, indicating that the normalization process has reduced the variability of the measured band positions. This is especially evident with the phosphorylase B and bovine serum albumin standards. The CVs for the $R_F$ values confirm the observations of other authors who question the stability of $R_F$ values within and among gels (Seiter & Jay, 1980). The CVs for the $R_F$ values tend to be of the same magnitude as the CVs for the actual migration distances and are, in all cases but one, larger than those of the normalized migration distances. Again, this is clearly seen with the heavier $M_r$ standards. These results support the use of the present normalization technique over the use of $R_F$ values as a means of decreasing the variability of protein band positions and increasing reproducibility of results both within and among gels.

Table 1. Comparisons of normalization and $R_F$ value computations for data of protein $M_r$ standards

<table>
<thead>
<tr>
<th>Gel</th>
<th>Actual migration distances (cm)</th>
<th>Normalized migration distances (cm)</th>
<th>$R_F$ values (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std dev.</td>
<td>CV</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.487</td>
<td>0.121</td>
<td>8.134</td>
</tr>
<tr>
<td>2</td>
<td>1.360</td>
<td>0.284</td>
<td>20.907</td>
</tr>
<tr>
<td>3</td>
<td>1.200</td>
<td>0.125</td>
<td>10.410</td>
</tr>
<tr>
<td>4</td>
<td>1.345</td>
<td>0.118</td>
<td>8.735</td>
</tr>
<tr>
<td>Total</td>
<td>1.348</td>
<td>0.199</td>
<td>14.739</td>
</tr>
</tbody>
</table>

Phosphorylase B

- Bovine serum albumin
- Ovalbumin
- Carbonic anhydrase
- Soybean trypsin inhibitor
**Bacterial SDS-PAGE numerical analysis**

*Fig. 1.* Dendrogram displaying whole-cell protein profile relationships among *L. dumoffii, L. micdadei, L. bozemanii* and *L. pneumophila* after SDS-PAGE. Analysis was done, after normalizing the data and computing the Dice similarity coefficient ($S_D$), using the unweighted pair group method with arithmetic means (UPGMA). The numbers 1–4 represent the gel which contained the samples and the letters A–C represent the three independent culturings of each strain.

**Clustering protein profiles**

The type strains of four species of *Legionella* were examined to evaluate how effectively the normalized data could be used to cluster them (Fig. 1). The Dice similarity coefficient ($S_D$) (Dice, 1945) was calculated using the 30 most prominent bands of each silver-stained protein profile, based on the percentage area measurements calculated by the densitometer. Due to the high sensitivity of this instrument, different numbers of bands were detected from profile to profile for a given strain, the differences being most noticeable with the lightly stained bands. This was dictated by the intensity of silver stain and the quantity of protein sample loaded onto the gel. These factors were controlled as precisely as possible but, nevertheless, were still variable in nature. As a result, the mean number of detected bands over the four gels, with their ranges, for *L. dumoffii, L. micdadei, L. bozemanii* and *L. pneumophila* were, respectively, 37, 33–45; 35, 31–40; 38, 33–41; and 33, 30–35. This resulted in an overall mean of 36 with a range of 30–45 bands. Those samples which appeared overall more densely stained contained more of the minor bands than the less densely stained samples. This phenomenon of detecting differing numbers of minor bands from scan to scan has also been reported by Seiter & Jay (1980). It has been reported that too high a sensitivity of measurement, which may lead to the detection of many minor bands, can also reduce reproducibility to an unacceptably low level (Jackman,
1983). To counteract this source of variability and to prevent the many lighter, less distinct bands from hindering the analysis, a fixed number of bands was chosen from each lane. Since the number of bands ranged from 30–45 among all the scans, we chose 30 bands for our analysis, as this was the maximum number contained in all the scans. A number less than 30 could have been chosen but it is preferable to use as many bands as possible without sacrificing reproducibility. This resulted in setting the denominator of the formula for the Dice similarity coefficient equal to 60. When comparing two profiles, a match was recorded if the normalized migration distance of a band in the first profile was within a window of plus or minus 0.284 normalized migration distance units of a band in the second profile; this window setting yielded maximum similarities between protein profiles in this particular data set.

The similarity values were clustered (Fig. 1). The four species of Legionella clustered into four separate, well demarcated groups. The normalized migration distance window was varied to determine how sensitive the clustering would be to this factor. The four species of Legionella retained the same expected groupings as the window ranged from 0.074 to 0.284 normalized migration distance units. This wide range indicates that the correct discrimination of the species is robust, in that the procedure is not over-dependent on the magnitude of the window. The operational window size selected for successful clustering is dictated by the range of the co-migrating protein normalized migration distances chosen for analysis. Since the intragel variability is commonly greater than the intergel variability of the protein band positions, a narrow window tends to maximize the similarity coefficients among samples on a given gel. Therefore, samples from four different preparations on the same gel would cluster more closely together than duplicate samples from the same preparation on different gels within a species of Legionella. As the window widens, the influence of the intragel variability on the similarity coefficient lessens and duplicate samples from the same preparation on different gels will sometimes cluster more closely together than samples from different preparations on the same gel. Fig. 1 illustrates the clustering pattern with a window of 0.284 normalized migration distance units and demonstrates a combination of both within gel and among gel clustering. The magnitude of the similarity among protein profiles within a species of Legionella also increased to a maximum using the widest window setting. Fig. 1, therefore, illustrates the optimal case for these data, in terms of maximum similarities among profiles within a species.

Jackman and coworkers (Jackman et al., 1983; Jackman, 1983) detailed a normalization scheme where the electrophoregram scans are shrunk or stretched to a fixed length between the start of the separation gel and the tracking dye front. They computed similarities by recording matching co-migrating band positions and by calculating $S_D$, as in the present study. They incorporated a misalignment factor that is used to search for the best possible fit between two different profiles by shifting one entire scan unilaterally, in small increments, in one direction and then the other, with $S_D$ being recalculated at each increment. The final similarity coefficient chosen is the maximum derived from this series of calculations. One disadvantage of this technique is that the misalignment factor can only operate in one direction at a time and operates in the same direction for all bands. If the band positions shift slightly from one profile to the next, this shift may occur at random about the mean position of each band and may not be in one direction throughout the electrophoregram. In this instance, the calculations of Jackman and coworkers (Jackman et al., 1983; Jackman, 1983) may not reveal the true magnitude of similarity between two profiles. The present method is more flexible in that a window is used for each band comparison individually by searching for a match while bracketing each band in both directions using the particular window setting. In this manner the influence of small shifts in position, in either direction, may be eliminated by examining each band independently.

Kersters & De Ley (1975, 1980) and Kersters (1985) use the correlation coefficient to determine similarities between pairs of samples. These statistics are calculated using the density values of their normalized electrophoregrams at fixed intervals. However, Swings et al. (1976) reported that slight and sometimes hardly noticeable and unavoidable shifts of very sharp and narrow protein bands from Zymomonas strains strongly affected the correlation coefficient and resulting reliability of the final clustering. Also, Jackman (1983) hypothesized that the correlation coefficient, which measures the similarity of the overall shapes of the electrophoreograms, gives undue weighting to major bands in the scans.
Given these considerations, it is desirable to develop an alternative method for clustering which examines each band independently and is not dependent on the correlation coefficient. The present method of normalization and computation of similarities based on matching co-migrating band positions through the use of a window setting is much less susceptible to the very small shifts in protein band positions and is not influenced by the magnitude or shape of the peaks in the scans when clustering bacterial strains. The investigator is able to experiment with a range of window settings to determine the optimal window size for the final clustering.

The amount of variability inherent with the $R_F$ values, as detailed in Table 1, is also a cause for concern when considering the accuracy of other measurements derived from electrophoretic studies, such as $M_r$ determinations. The present normalization technique leads to much more stable estimates of band location, especially for the heavier $M_r$ values, and will result in less variable estimates of $M_r$ values when using such methods as described in Plikaytis et al. (1986). Comparisons of $M_r$ values derived from protein profiles normalized with the present techniques may, therefore, be compared with much greater confidence than those derived from profiles normalized using $R_F$ values. Also, in addition to analysis of the silver-stained whole-cell proteins, these procedures have been successfully applied to outer membrane protein preparations stained with Coomassie blue (Carlone et al., 1985).

Any protein profile may be successfully normalized regardless of the technique used for staining. A silver-stained protein profile cannot necessarily be directly compared to a Coomassie blue stained profile since the degree of silver staining is characteristic for each protein and is not directly related to the concentration of that protein. However, Coomassie blue staining must be used instead of silver staining if cluster analysis is to be based on proteins produced in highest concentration. We chose to use a silver-staining technique because we have extensive laboratory experience comparing silver-stained Legionella protein profiles. As long as a distinctive protein banding pattern can be demonstrated as being reproducible, as illustrated by the successful clustering of the four species of Legionella in this study, the staining technique becomes secondary in consideration.

**References**


