Identification and Grouping of Bacteria by Numerical Analysis of their Electrophoretic Protein Patterns

By K. Kersters and J. De Ley

Laboratory of Microbiology and Microbial Genetics, Faculty of Science, State University, B 9000 Gent, Belgium

(Received 9 October 1974; revised 2 December 1974)

SUMMARY

Improved methods for the identification and grouping of bacteria by polyacrylamide gel electrophoresis of soluble proteins are described. Electrophoretic protein patterns were obtained in rigorously standardized conditions. The results were much more reproducible than any described previously. Some of the factors affecting reproducibility were: growth conditions, time and speed of centrifugation of extracts, and conditions of gel electrophoresis. Protein patterns were compared by computing correlation coefficients from normalized densitometric tracings and clustering the strains by the unweighted average pair group method. As model systems, both Agrobacterium and Zymomonas were used because of differences in the sharpness of the peaks. The method was applied to 42 Agrobacterium strains. The agreement with the results of clustering by either phenotypic tests or DNA:DNA hybridization was excellent. Computerized comparisons of electrophoretic protein patterns can be a fast, easy and powerful tool for classification and identification of bacteria.

INTRODUCTION

Polyacrylamide gel electrophoresis of cell proteins is used increasingly for the classification and identification of micro-organisms. Protein band patterns are usually compared visually. This type of examination and the listing of the \( R_p \) values of bands are subjective, frequently confusing, and not suitable for large numbers of extracts. Sokal & Sneath (1963) suggested that protein patterns might be suitable for quantitative numerical analysis. A computerized approach developed by Whitney, Vaughan & Heale (1968) consists of identifying homologous bands and calculating percentage similarity values. Applications of this technique to seed proteins of Acer (Ziegenfus & Clarkson, 1971), Abies (Clarkson & Fairbrothers, 1970) and lens proteins of rodents (Day, 1972) were mostly successful. However, similar methods applied to Arthrobacter strains were not satisfactory (Rouatt, Skyring, Purkayastha & Quadling, 1970).

We developed improved techniques, because only very reproducible electropherograms are amenable to successful computerized treatment. Firstly, the reproducibility of protein patterns was increased by establishing a rigorously controlled standard procedure. Secondly, the patterns of closely related organisms were made completely reproducible by a mathematical procedure that we call the 'compensation method'. Johnson & Thein (1970) used a somewhat similar computerized treatment successfully with crude seed extracts of Gossypium. Increased sharpness and narrowness of protein bands could decrease the reproducibility if not corrected properly. We therefore selected two model systems: the broad-banded Agrobacterium and the narrow-banded Zymomonas. The method was applied to a large number of Agrobacterium strains and could be evaluated because the intrageneric relationships of these bacteria are known from DNA:DNA hybridization and phenotypic

METHODS

The methods are described in minute detail in order to make the results reproducible and amenable to computer treatment.

Organisms. The strains of Achromobacter, Chromobacterium, Erwinia, Mycoplana, Pseudomonas, Xanthomonas and Agrobacterium used are mentioned in previous papers (Kersters & De Ley, 1968; Kersters et al. 1973; De Ley et al. 1973). Zymomonas mobilis z3 was isolated by Dr J. Swings (this laboratory) from palm wine in Kinshasa, Zaire. Agrobacterium tumefaciens strains b48, vi, w1, ATCC15955 and ATCC17805 came from Dr J. Lippincott, Department of Biological Sciences, Northwestern University, Illinois, U.S.A.; A. tumefaciens AG19 from Dr C. G. Panagopoulos, Institut Phytopathologique, Benaki, Athens, Greece; A. tumefaciens strains TT4, TT10 and TT107 from the International Collection of Phytopathogenic Bacteria (ICPB), Department of Bacteriology, University of California, Davis, California, U.S.A.; and A. tumefaciens strains Apple 185, s379, 181 and H5 from Dr J. Tempé, INRA, Versailles, France. Agrobacterium radiobacter strains were ATCC6467 and ATCC19358.

Culture conditions. All bacteria, except Zymomonas mobilis, were grown for 38 to 42 h at 28 °C in three Roux bottles on a solid medium containing (g/l tap water): glucose, 10; yeast extract (Nederlandse Gist en Spiritusfabriek, Brugge, Belgium), 10; (NH₄)₂SO₄, 1; KH₂PO₄, 0.25; agar, 25. Cells were suspended in 0.01 M-phosphate buffer pH 7.0, harvested, and washed once by centrifuging in the same buffer and twice in 3 mM-tris-HCl buffer pH 7.0. All buffers were made in bidistilled water. Five ml of 6.4 mM-tris-HCl buffer pH 8.4, containing 0.001 % deoxyribonuclease (Worthington, Freehold, New Jersey, U.S.A.) were added to 5 g (wet wt) bacteria.

Zymomonas mobilis z3 was grown at 30 °C in Erlenmeyer flasks containing 5 g yeast extract (Difco) and 20 g glucose/l distilled water. The cells were harvested in the early stationary phase, washed in 0.01 M-phosphate buffer pH 7.0, then in 64 mM-tris-HCl buffer pH 9.15, before resuspending in the latter.

Preparation of cell-free extracts. The suspensions were disrupted in a French pressure cell (Aminco, Silver Spring, Maryland, U.S.A.) at 21 000 lb/in² in the liquid. Intact bacteria and debris were removed by centrifuging (4 °C, 15 000 g, 15 min) in a Sorvall RC-2 centrifuge. Four ml of turbid supernatant were centrifuged for 1 h at 4 °C and 80000 g in a Beckman/Spinco model L centrifuge (type 40 rotor). Protein concentration of the supernatant was determined by the method of Lowry, Rosebrough, Farr & Randall (1951) and adjusted to 25 mg protein/ml with 6.4 mM-tris-HCl buffer pH 8.4. Two ml of this extract were again centrifuged for 4 h at 4 °C and 80000 g. The protein concentration of the clear supernatant was adjusted to 12 mg/ml with 6.4 mM-tris-HCl buffer pH 8.4. These cell-free extracts contained the soluble proteins and were stored at -12 °C.

Standard conditions of polyacrylamide gel electrophoresis. A continuous buffer system was used (Hjertén, Jerstedt & Tiselius, 1965). Gels were prepared as follows. Two ml of a solution containing 28 % (w/v) acrylamide (Serva, Heidelberg, Germany) and 0.735 % (w/v) N,N'-methylene bisacrylamide (Serva) were mixed with: (i) 1 ml of a solution containing 0.23 % (v/v) N,N,N',N'-tetramethyl ethylenediamine (Serva) and 6.1 % (w/v) tris (this solution was adjusted with 2 N-HCl to pH 8.8); (ii) 1 ml of double-distilled water; and (iii) 4 ml of 0.14 % (w/v) ammonium persulphate. Glass tubes [(4.92 ± 0.03) x 77 mm] were filled in a thermostated water bath at 24 °C with 1.2 ml gel solution (prewarmed at 24 °C), and overlaid with 0.1 ml double-distilled water. Polymerization was carried out at 24 °C for at least 120 min.
Numerical analysis of electropherograms

Fig. 1. Densitometer tracing of the protein pattern of *Agrobacterium tumefaciens* NCPPB397 minus the reference proteins thyroglobulin and ovalbumin (full line), and part of the scanning of the corresponding protein pattern with the reference proteins (broken line). I to V represent protein bands selected for compensation (see text).

Each protein extract was run at least in duplicate. Approximately 100 μg of soluble protein in 7 to 10 μl, containing 5% (w/v) sucrose, were layered on top of two gels. One gel cylinder was supplemented with 6 μg ovalbumin (grade V, Sigma) and 4 μg bovine thyroglobulin (Sigma). These reference proteins enabled accurate comparison between different protein patterns to be made (see below).

Electrophoresis was performed at 20 °C in a thermostated cabinet containing a circular apparatus for 12 gels, similar to an 'Acrylophor' (Pleuger, Wijnegem, Belgium). In each run, one gel cylinder was loaded with 1 mg human albumin dyed with bromophenol blue (Serva). A small centrifugal pump (Eheim 38 l) circulated electrode buffer (64 mM-tris-HCl pH 8.7) from the lower (700 ml) to the upper (200 ml) electrode chamber through a glass coil immersed in an ice bath. The temperature of the electrode buffer was 8 to 9 °C. Electrophoresis was performed with the anode in the lower electrode chamber at a constant current of 1.25 mA/gel for 15 min, followed by 3.8 mA/gel until the albumin marker band had moved 43 mm through the gel, which usually took 110 min. Gels were removed from the glass tubes, stained overnight for protein with 0.6% (w/v) Amido black 10 B (Merck) in 7% (v/v) acetic acid, and destained electrophoretically in 7% (v/v) acetic acid at room temperature and 4 mA/gel.

Normalized photographs. Gels were placed on a uniformly illuminated frosted glass plate and photographed with Agfaortho 25 negative film. Negatives were enlarged on a transparent Rapidoprint TP5F positive film (Agfa-Gevaert), always with a distance of 47 mm between the top of the gel and the reference ovalbumin peak.

Densitometry. Up to eight gels were inserted between two glass plates of a Plexiglass sample holder and scanned with a Joyce Loebl MK IIIB microdensitometer. The optimal settings of the instrument were as follows: objective, ×10; effective slit width, 40 μm; red filter, Ilford 204; ratio of record/sample travel, 5; wedge range, $E = 0$ to 1.2.
Normalized densitometer scans. The mobility of protein preparations differed slightly from one electrophoresis run to another, even in our most reproducible conditions. The addition of the internal reference proteins ovalbumin and thyroglobulin allowed normalization and subsequent direct comparison of different scans. The densitometer tracings of the same protein preparation, with and without reference proteins, were superposed and the position of both reference peaks was marked on the scan without reference proteins (Fig. 1). A horizontal baseline was drawn 5 mm below the lowest point of the latter scan and the distance between ovalbumin and thyroglobulin divided into 90 equal parts. The average mobility of thyroglobulin was $\frac{1}{90}$ of the mobility of ovalbumin. Position numbers 6 and 96 were therefore assigned to thyroglobulin and ovalbumin, respectively (Fig. 1). This allowed the top of the gel (position 0) to be found exactly. We call this procedure ‘normalization’ (N) of the scans.

Compensation of normalized densitometer scans. Even after normalization of the scans, small variations of a few tenths of a millimetre in the position of homologous bands cannot be eliminated (Fig. 2). These technical imperfections are corrected by ‘compensation’ of normalized densitometer scans. This is a mathematical imitation of the visual compensation that is performed unconsciously when comparing gels with similar, but not identical, band patterns. In a group of protein profiles from the same or similar bacterial strains, all easily recognizable homologous protein bands are visually selected (e.g. bands I, II, III, IV and V in Figs. 1 and 2b) and the average position number of each band is calculated from the scans. In each scan, each selected band is moved to its calculated average position number. The distance between each set of consecutive bands has therefore to decrease or increase. This is effected by omission or addition, respectively, of the lowest extinction values in the valley floor between these bands. When comparing stained gels visually, more weight is indeed given to protein bands than to zones with low protein concentration. The results are ‘compensated normalized scans’ (NC).

Computation. The extinction, expressed in mm height, of each position on a scan was punched on IBM cards. Each normalized protein pattern of Agrobacterium, compensated if required, was thus converted to a sequence of 107 numbers. Only 90 positions were needed for protein patterns of Zymomonas mobilis Z3. The Pearson product–moment correlation coefficient, $r$, between any pair of densitometric tracings of protein patterns was computed. The resulting $r$ matrix of correlation coefficients was converted into a $z$ matrix by Fisher transformation (Rohlf & Sokal, 1969), and the $z$ values were clustered by unweighted average linkage (Sokal & Michener, 1958). The $z$ values of all clustering levels were again transformed into $r$ values.

Calculations were performed on a IBM 360–30 computer at the Centraal Digitaal Rekencentrum, State University, Gent.

RESULTS AND DISCUSSION
Reproducibility of the gel electrophoretic method

The reproducibility of the gel electrophoretic method was studied with two types of bacteria: Z. mobilis Z3, displaying very sharp protein bands (Fig. 2a); and Agrobacterium strains with more diffuse protein bands (Figs. 1, 2b). Soluble protein was prepared from three batches of Z. mobilis and A. tumefaciens, grown independently in the same conditions, and each protein extract was analysed on polyacrylamide gels under standard conditions, in at least three separate electrophoretic runs. Clustering levels were at least $r = 0.82$ for Z. mobilis Z3 (Fig. 2a, N), and $r = 0.92$ to $r = 0.97$ for 12 Agrobacterium strains. One example
Numerical analysis of electropherograms

Fig. 2. Effect of three independently grown cultures, A, B and C, of (a) *Z. mobilis* Z3 and (b) *A. tumefaciens* NCPPB397, on the reproducibility of gel electrophoretic protein patterns. Each extract was analysed under standard conditions in three separate electrophoretic runs A1 to A3, B1 to B3 and C1 to C3. Numerical analysis of normalized protein profiles was performed with and without compensation. N, normalized; NC normalized and compensated; T, thyroglobulin; OV, ovalbumin; I to V, protein bands selected for compensation (see text). The gel patterns in this and all subsequent Figures were normalized photographically.
Fig. 3. Effect of the composition of the culture medium on the protein electrophoretic pattern of *A. tumefaciens* CIP 86. (A) Cultures grown for 38 h at 28 °C on solid media as described in Methods. (B) Cultures grown on a minimal basal medium (Graham & Parker, 1964) supplemented with (% w/v): NH₄H₂PO₄, 0.05; yeast extract, 0.2; and glucose, 1, on a reciprocal shaker at 28 °C for 20 h. (C) As (B) but with 1% mannitol instead of glucose.

Fig. 4. Effect of centrifugation time of the extract on the gel electrophoretic protein pattern of *Z. mobilis* Z3. The numbers represent hours of centrifugation at 80,000 g and 4 °C.

of the latter is given in Fig. 2(b), N. Protein patterns of independently grown cultures of each strain are intertwined in the dendrograms of Fig. 2. Compensation improved the reproducibility (see below). The numerical treatment indicted that experimental variation is much greater for protein patterns with sharp bands (Fig. 2a) than for patterns with more diffuse bands (Fig. 2b). Reproducibility of electropherograms from one culture was no higher than those from different batches.

The effect of the following factors on the reproducibility of protein patterns was investigated.

*Composition of the growth medium.* Alterations in the composition of the growth medium changed the protein pattern of *A. tumefaciens* CIP 86 (Fig. 3).

*Centrifugation of the extracts.* Different centrifugation times of extracts at 80,000 g led to differences in the upper part of the protein patterns. The mobility of the slow-moving protein bands of *Zymomonas mobilis* Z3 decreased with increasing centrifugation time (Fig. 4). Centrifugation time and temperature, protein concentration and volume of sample have to be kept constant.

*Type of micro-organisms used.* With some groups of micro-organisms our method may require slight modifications, e.g. in buffer systems and/or concentrations of gel components. Dialysis of the soluble protein fraction against 6.4 mM-tris-HCl buffer pH 8.4 can improve the sharpness of the starting zones. We advise that the level of reproducibility be checked carefully whenever electropherograms of new groups of bacteria need to be clustered numerically.
Factors with little effect on the reproducibility. There were no significant differences between protein patterns from Agrobacterium cells, whether they were grown on liquid or solid media. Protein patterns of middle and late exponential phase, and of early stationary phase cells of Agrobacterium showed only small differences in minor protein bands in the upper third of the gel cylinder; clustering levels were always higher than 0.94, not compensated.

The following factors had little effect on the reproducibility of protein patterns from Agrobacterium and \textit{Z. mobilis} z3: method of breakage (French pressure cell, Ultrason, MSK-shaker); electrophoresis time between 90 and 120 min; storage of extracts at $-12^\circ C$ for up to 1 month; the amount of protein applied per gel cylinder, provided it was between 50 and 150 $\mu$g. Clustering levels for scans of the same gel, repeatedly placed in the densitometer, were $r = 0.95$ for \textit{Z. mobilis} z3 and $r = 0.995$ for \textit{A. tumefaciens} NCPPB397. Thus the error introduced by the densitometer was negligible.

Ionic impurities, unreacted monomer molecules, persulphate, etc., can occasion artefacts (Brewer, 1967; Maurer, 1971). Pre-electrophoresis of the gel cylinders, or addition of reducing agents (e.g. mercaptoethanol or dithiothreitol), was reported to suppress such artefacts (Brewer, 1967; Besemer & Clauss, 1968). In our hands, protein patterns of several Agrobacterium strains were not altered, and reproducibility was not improved, by pre-electrophoresis.

Computerized clustering of electrophoretic protein patterns

Several approaches for computerized numerical analysis of gel electrophoretic protein patterns were tried. Distance and similarity coefficients were unsatisfactory. Correlation coefficients were most suitable because they are insensitive to the experimental differences in amount of protein applied. Comparing profiles only by protein bands, either qualitatively as plus or minus (Rouatt \textit{et al.} 1970) or quantitatively by peak height, is unsatisfactory. All information on peak and valley shape and on relative position is preserved by slicing the profile in a large number of columns and listing the heights, as outlined under Methods. The unweighted average linkage pair group method gave the most satisfactory clustering.

The compensation method is only useful for a group of similar protein patterns. It corrects the inevitable small differences in the position of the major bands, which decrease $r$-values considerably in the case of protein profiles with sharp, narrow and dense bands (Fig. 2a). When nine protein patterns of the sharp-peakz \textit{Z. mobilis} z3 were compensated and clustered numerically, the reproducibility increased considerably from $r = 0.82$ to $r = 0.92$ (Fig. 2a, NC). In contrast compensation made little difference to the broad protein bands of \textit{A. tumefaciens} NCPPB397 (from $r = 0.965$, to 0.975; Fig. 2b, NC).

Evaluation, limitations and application

The method is being used successfully in this laboratory for the grouping and identification of large numbers of strains of Agrobacterium, Achromobacter, Alcaligenes and Zymomonas. One example is given here, demonstrating at the same time the usefulness and the limitations of the method (Fig. 5). Protein patterns of 42 Agrobacterium strains cluster into three groups, called B6, TT111 and 'cluster 2'; they separate clearly from representatives of six other bacterial genera (\textit{Achromobacter}, \textit{Chromobacterium}, \textit{Erwinia}, \textit{Mycoplasma}, \textit{Pseudomonas} and \textit{Zymomonas}; broken lines in Fig. 5). Agrobacterium strains of 'cluster 2' show essentially identical protein patterns ($r \geq 0.93$, Fig. 5), whereas more variation is observed within the B6 and TT111 groups. All this corresponds perfectly with the inter- and intra-groupings obtained through DNA:DNA hybridizations (De Ley, 1972; De Ley \textit{et al.} 1973), and with numerical analysis of phenotypic features of the agrobacteria (Kersters \textit{et al.} 1973). Com-
Fig. 5. Clustering of normalized gel electrophoretic protein patterns from 42 Agrobacterium strains (full lines) and 11 strains of nine other bacterial genera (broken lines).
Numerical analysis of electropherograms

Computerized processing helps considerably in grouping the protein patterns but does not always reveal the finer distinction possible by visual or photographic comparison. For example, protein patterns of Agrobacterium strains ATCC143 and ICPI TT107 cluster at \( r = 0.98 \), whereas some distinct differences are visible in the region 2 to 3 cm (Fig. 5). Compensation did not noticeably improve the clustering of the agrobacteria in Fig. 5, but was very useful for grouping Zymomonas strains (J. Swings and J. De Ley, unpublished).

When large numbers (over 50) of protein patterns from various bacteria were compared some strains occasionally clustered in groups where they did not belong, either phenotypically or genotypically. Figure 5 was selected because it is our worst example. Protein patterns of *Erwinia herbicola* PA, *Rhizobium leguminosarum* USDA316C1OA and *Xanthomonas pelargonii* ICPI P121 ‘intrude’ in, or between, the B6 and TT111 groups of the Agrobacterium strains. Although the protein pattern of *R. leguminosarum* USDA316C1OA shows two characteristic bands at 3.0 and 3.4 cm (Fig. 5), it clusters at \( r = 0.92 \) within the agrobacteria. *Erwinia herbicola* PA clusters at \( r = 0.96 \) with the agrobacteria Kerr 14 and AG19 in the B6 group. While sharing similarities with both agrobacteria in the region 2.5 to 4.5 cm, *E. herbicola* PA differs from these and all other strains of the B6 group with respect to protein bands at 1.9 and 4.9 cm. *Rhizobium leguminosarum* USDA316C1OA and *X. pelargonii* P121 are located between the B6 and the TT111 groups of *Agrobacterium*. All these strains probably intrude because of the overall pattern similarity. Numerical analyses of phenotypic tests not infrequently show similar intrusions. A final check is therefore made by grouping the normalized photographs according to the computer results and inspecting visually for abnormal inclusions.

The main advantages of numerical analysis of electrophoretic protein patterns of large numbers of bacteria are: (i) rapid grouping; (ii) allocation of an unknown micro-organism to a group and its possible identification; (iii) storage of large numbers of patterns in data banks for reference; (iv) a quick decision on whether two colony types in a culture are due to variation or contamination; (v) information on epidemiological spreading of animal and plant pathogens; (vi) DNA:DNA homology determination of large numbers of strains can be reduced to hybridization of DNAs from the typical representatives of each group, previously established by gel electrophoresis. We stress that useful results can only be obtained when the conditions for preparation of extracts and their electrophoresis are rigorously controlled.

We thank Dr R. De Caluwe and Dr B. Becue, Centraal Digitaal Rekencentrum, State University, Gent, for help in programming and computer facilities, Mr D. Dewettinck for skilful technical assistance, and Professor P. H. A. Sneath for reading the manuscript. J.D.L. is indebted to the Fonds voor Kollektief en Fundamenteel Onderzoek for research and personnel grants.

REFERENCES


