Journal of General Microbiology (1991), **137**, 2051-2058. *Printed in Great Britain*

Rapid methods in bacterial DNA fingerprinting

KENNETH J. FORBES,* KENNETH D. BRUCE, J. ZOE JORDENS, ALISON BALL and T. HUGH PENNINGTON

Department of Medical Microbiology, University of Aberdeen, Foresterhill, Aberdeen AB9 2ZD, UK

(Received 31 January 1991; revised 16 May 1991; accepted 28 May 1991)

The characterization and comparison of isolates of bacterial species by comparing restriction enzyme digests of their chromosomal DNA (fingerprints) is highly discriminatory for different strains and allows similarities between them to be readily determined. However, the utility of the technique is dependent on the selection of appropriate restriction enzyme(s) and on the method of determining the similarities between the fingerprints generated. We report here a system which circumvents these two problems. The restriction enzyme is selected from amongst those which have a suitable frequency of restriction for given enzyme–genome combinations. The frequencies of restriction enzyme recognition sites are calculated from the frequencies of di- and trinucleotides in sequenced genes from the species of interest using Markov chain analysis. Fingerprints are compared by dividing them up into sections with DNA size standards, scoring the number of bands in a few of these sections, and comparing these scores (numerical profiles) to establish similarities. In this way a single electrophoretic gel yields easily analysable data which can be compared with data from other gels. The time from the acquisition of bacterial isolates to their final characterization is much reduced in comparison to existing methods.

Introduction

The use of directly visualized restriction enzyme digests of chromosomal DNA (fingerprints) from bacteria offers a rapid and reproducible method for the characterization and classification of isolates applicable to ecological, epidemiological and population genetic studies. Two problems soon present themselves however, firstly in the choice of restriction enzyme(s) and secondly in the recognition and grouping together of related isolates; we report a system which circumvents these two problems.

The criteria for the selection of a restriction enzyme for use in DNA fingerprinting are that it should cut the DNA into fragments which are suitable for analysis both in size and in frequency. When separated by conventional agarose gel electrophoresis, fragment sizes in the 1–10 kb range allow short electrophoresis times. Additionally, the fragments should not be so numerous in this size range that they overlap, thus resulting in a crowded fingerprint which is difficult to interpret. In nearly all previous studies the selection of the best restriction enzyme(s) has had to be determined empirically (Ficht *et al.*, 1989; Loos *et al.*, 1989; Owen *et al.*, 1990; Stahl *et al.*, 1990; Wren & Tabaqchali, 1987), as most enzymes cut either too seldom or too frequently, resulting in patterns which may be difficult to interpret. Owen (1989) suggested selecting restriction enzymes on the basis of the mol% G+C content of the enzyme recognition site and of the chromosomal DNA (Nei & Li, 1979) but concluded that it was a poor predictor of restriction frequency (Owen, 1989; Phillips *et al.*, 1987*a*). McClelland *et al.* (1987) discussed the selection of restriction enzymes for use in pulsed-field gel electrophoresis, where enzymes are chosen for their infrequent recognition sites. For particular combinations of restriction enzyme and bacterial species they estimated the frequency of restriction, which is equal to the reciprocal of the mean fragment length in nucleotides, using Markov chain analysis of the frequency of di- and trinucleotides calculated from published DNA sequences. Using a modification of this approach we have developed a more rigorous selection method for the choice of restriction enzymes for chromosomal DNA fingerprinting which is a better predictor of restriction frequency than previous methods.

Where only a few isolates are to be characterized, classified or grouped together by DNA fingerprinting, fingerprints can be directly compared with each other with little difficulty. However, when classifying many isolates, as in population genetic studies, a more

systematic approach has to be adopted. Central to this problem is the dichotomy between fingerprints consisting of a few bands which can be easily compared, and fingerprints with a larger number of bands which, though allowing greater discrimination, are much more tedious to classify. A diversity of organisms have been subjected to DNA fingerprinting, but the technical problem of analysing banding patterns and of recognizing related isolates has usually meant that these studies have been limited to a few isolates. Other studies have sought to limit the number of bands per fingerprint, either by examining small genomes (plasmids, viruses, mitochondria or chloroplasts), or by selecting a few bands from complex fingerprints by probing for particular DNA sequence(s) [rRNA genes (Grimont & Grimont, 1986; Irino *et al.*, 1988), capsular polysaccharide genes (Musser *et al.*, 1990)]. Scoring and classification of fingerprints using scanning densitometry and computer analysis has been adopted (Owen *et al.*, 1990; Sorensen *et al.*, 1985; Stahl *et al.*, 1990) but does not readily allow inter-gel comparisons. Our approach has been to divide each fingerprint up into a number of sections, score the number of fragments visible in each section, and so generate numerical profiles which can be easily compared. In this way it is possible to compare fingerprints of different isolates using up to 20 bands, but still do this manageably.

Methods

Bacteria. Aeromonas hydrophila, Erwinia chrysanthemi, non-typable (non-capsulate) *Haemophilus influenzae, Neisseria meningitidis, Micrococcus luteus, Staphylococcus aureus* and *S. capitis* isolates, held in collections in the Department of Medical Microbiology, Aberdeen University, were used throughout. All were stored at −70 °C or freeze-dried.

Restriction enzyme selection. Using as much published DNA sequence data as was available (EMBL and GenBank databases) for the chosen species, or a closely related one, the frequencies of mono-, di- and trinucleotides were determined by computer (Composition Program, Sequence Analysis Software Package for VAX computers, Genetics Computer Group, University of Wisconsin). The mol% G+C content of the total sequence was calculated from the mononucleotide frequencies and compared to published frequencies for that species (Krieg & Holt, 1984; Sneath *et al.*, 1986) to determine how representative of the genome were the sequences used. Estimations of the mean fragment lengths for particular restriction enzyme and genome combinations were determined using the di- and trinucleotide frequencies. Given that mean fragment length in nucleotides is the reciprocal of the frequency of a particular recognition sequence for a given combination of restriction enzyme and genome, then using the frequency (p) of di- and trinucleotides and second-order Markov chains (Phillips *et al.*, 1987*a*), for restriction enzymes with a tetranucleotide recognition sequence ($N_1N_2N_3N_4$), mean fragment length in nucleotides =

$$\frac{p(N_2N_3)}{p(N_1N_2N_3)\cdot p(N_2N_3N_4)} \quad (1)$$

Similarly for the hexanucleotide $N_1N_2N_3N_4N_5N_6$, mean fragment length in nucleotides =

$$\frac{p(N_2N_3)\cdot p(N_3N_4)\cdot p(N_4N_5)}{p(N_1N_2N_3)\cdot p(N_2N_3N_4)\cdot p(N_3N_4N_5)\cdot p(N_4N_5N_6)} \quad (2)$$

DNA fingerprinting. Chromosomal DNA was isolated from *E. chrysanthemi, H. influenzae* and *N. meningitidis* using a modification of the method of Pitcher *et al.* (1989). Bacteria were harvested from a single agar plate which had been incubated overnight, using a sterile dry swab, and suspended in TE8 (10 mM-Tris/HCl, 1 mM-EDTA, pH 8·0) to give a final volume of about 100 µl. Following lysis and DNA extraction by the original method, DNA precipitated by ethanol (2 vols, −20 °C) was dissolved in 100 µl TE8 and reprecipitated once with cold ethanol and finally dissolved in TE8. DNAs from *A. hydrophila, M. luteus* and *Staphylococcus* spp. were kindly provided by, respectively, A. Ball, A. Leanord and P. Carter, Aberdeen University. DNA (5 µg) was digested overnight with the appropriate restriction enzyme (5–10 units) according to the manufacturers' instructions. Electrophoresis was in 0·8% agarose at 3 V cm^{-1} for about 7 h. Gels were stained with ethidium bromide (0·5 µg ml^{-1}) in distilled water for 0·5 h, destained in distilled water for 1 h, and photographed (Polaroid 665 positive-negative). The DNA size standards were 1 kb-ladder DNA (Gibco-BRL).

Fingerprint sectioning, scoring and classification. 1 kb-ladder DNA (5 µl of 1:50 dilution of stock) was added to each digested DNA sample after the addition of loading buffer [10×: 200 mM-EDTA (pH 8·0), 2·5 mM-bromophenol blue, 25% (w/v) Ficoll (type 400) in distilled water; EDTA prevented the digestion of the ladder DNA]. Following electrophoresis, the numbers of DNA bands in three (or four) of the 1 kb-ladder rungs were counted and these numerical profiles arranged in ascending order by computer (SuperCalc4, Computer Associates International, Slough, UK). Similarities between pairs of isolates were estimated by summing the difference in the number of bands in each of these three (or four) homologous pairs of rungs. For pairs of isolates where this sum was zero, one or two, their actual similarities were determined by comparing their digest patterns by eye, using the 1 kb-ladder as a guide. The recognition of groups of isolates with lower similarities, and the opportunity to allow for minor errors in the reading of the gel, were possible when pairs of isolates with numerical profiles with greater differences were compared again at this stage. Groups were electrophoresed again, in adjacent lanes, without the 1 kb-ladder, to determine their actual similarities more quantitatively (e.g. using Dice coefficients: Dice, 1945).

Results

Restriction enzyme selection

The expected frequencies of restriction of several enzymes were determined for chromosomal DNA of *A. hydrophila, E. chrysanthemi, H. influenzae, N. gonorrhoeae, M. luteus* and *S. aureus*. These species represent both Gram-negative and Gram-positive organisms, as well as genomes with widely differing G+C contents: 33 to 75 mol%.

For *H. influenzae* a total of 3205 bp of sequence from the type b capsule expression gene (*bexA*) (740 bp), outer-membrane protein P1 (*ompP1*) (1598 bp) and outer-membrane protein P6 (*ompP6*) (867 bp) genes was used to calculate the frequencies of the mono-, di- and trinucleotides (Table 1). The G+C content of these sequences was 39 mol%, the same value as that reported

Table 1. *Frequencies of mono-, di- and trinucleotides in three H. influenzae genes*

Frequencies calculated from 3205 nucleotides of sequence from the genes *bexA*, *ompP6* and *ompP1*, are correct to two significant figures.

Mononucleotide frequencies			
A: 0·30	C: 0·17	G: 0·21	T: 0·31
10^3 × Dinucleotide frequencies			
GG: 44	GA: 53	GT: 63	GC: 51
AG: 59	AA: 110	AT: 83	AC: 49
TG: 72	TA: 80	TT: 110	TC: 50
CG: 36	CA: 62	CT: 52	CC: 25
10^3 × Trinucleotide frequencies			
GGG: 6·9	GGA: 8·7	GGT: 19	GGC: 8·7
GAG: 8·7	GAA: 19	GAT: 17	GAC: 7·8
GTG: 16	GTA: 16	GTT: 24	GTC: 7·8
GCG: 11	GCA: 18	GCT: 16	GCC: 6·6
AGG: 12	AGA: 14	AGT: 18	AGC: 15
AAG: 19	AAA: 44	AAT: 33	AAC: 17
ATG: 19	ATA: 18	ATT: 31	ATC: 15
ACG: 9·4	ACA: 17	ACT: 14	ACC: 8·1
TGG: 15	TGA: 22	TGT: 14	TGC: 21
TAG: 15	TAA: 28	TAT: 21	TAC: 12
TTG: 26	TTA: 32	TTT: 36	TTC: 16
TCG: 10	TCA: 16	TCT: 17	TCC: 6·6
CGG: 9·7	CGA: 7·8	CGT: 12	CGC: 6·9
CAG: 16	CAA: 22	CAT: 12	CAC: 11
CTG: 11	CTA: 11	CTT: 20	CTC: 11
CCG: 5·9	CCA: 11	CCT: 4·7	CCC: 3·7

for the whole genome (Krieg & Holt, 1984). From the frequencies of the di- and trinucleotides and formulae 1 and 2 the mean fragment lengths for various enzymes were calculated. These are shown in Table 2.

For *Neisseria* a total of 9770 bp of sequence from the *N. gonorrhoeae* genes for azurin (1610 bp), IgA protease (4899 bp), penicillin-binding protein 2 (2049 bp) and outer-membrane protein PIIc (1212 bp) was used. The G+C content calculated from the mononucleotide frequency of 49 mol% compares to published values of 49–53 mol% for *N. gonorrhoeae* and 50–52 mol% for *N. meningitidis* (Krieg & Holt, 1984). The calculated mean fragment lengths are shown in Table 2.

M. luteus sequence data from a gene homologous to *Escherichia coli uvrA* (2286 bp) were used to predict mean fragment lengths for particular restriction enzymes with *M. luteus* DNA. Similarly sequence data from *S. aureus* genes for staphylocoagulase (3042 bp), serine protease (1634 bp) and alpha-toxin (1485 bp) were used to predict mean fragment lengths for DNA from staphylococcal species and selected restriction enzymes. The G+C contents of these sequences were 69 and 33 mol% for *M. luteus* and *S. aureus*, respectively; Sneath *et al.* (1986) report values of 70–75 mol% and 32–36 mol% respectively.

Table 2. *Calculated, theoretical values of the mean fragment lengths between restriction enzyme recognition sequences in different species*

		Mean fragment length (kb)†					
	Species*...	*S. a.*	*H. i.*	*N. g.*‡	*E. c.*	*A. h.*	*M. l.*
	Mol% G+C...	32–36	39	50–52	55–57	58–62	70–75
Restriction enzyme	Recognition sequence						
*Alu*I	AGCT	0·28	0·21	0·44	0·28	0·21	0·29
*Taq*I	TCGA	(0·48)	0·45	0·39	(0·32)	(0·30)	0·14
*Hae*III	GGCC	0·99	0·91	0·15	(0·15)	(0·10)	0·10
*Dra*I	TTTAAA	0·70	0·67	1·7	3·7	21	>100
*Hin*dIII	AAGCTT	3·1	1·7	2·5	4·6	(4·2)	16
*Eco*RI	GAATTC	3·0	3·1	5·3	6·0	(7·5)	>100
*Bgl*II	AGATCT	5·5	3·6	4·6	(5·2)	(6·6)	10
*Eco*RV	GATATC	3·6	5·9	3·0	(3·7)	(2·0)	>100
*Pst*I	CTGCAG	(11)	3·4	4·0	2·3	2·2	2·0
*Kpn*I	GGTACC	9·3	8·3	5·5	3·0	8·5	1·5
*Bam*HI	GGATCC	(16)	15	4·8	(7·8)	3·0	1·2
*Sac*I	GAGCTC	(14)	7·0	20	(8·3)	4·1	(1·4)
*Sal*I	GTCGAC	32	20	7·7	(3·9)	(3·7)	(1·5)
*Apa*I	GGGCCC	(59)	38	4·5	9·0	1·4	(1·8)

* *S. a.*, *Staphylococcus aureus*; *H. i.*, *Haemophilus influenzae*; *N. g.*, *Neisseria gonorrhoeae*; *E. c.*, *Erwinia chrysanthemi*, *A. h.*, *Aeromonas hydrophila*; *M. l.*, *Micrococcus luteus*.

† Values in parenthesis indicate enzyme–genome combinations in which electrophoretic gels have not been performed.

‡ Tabulated values are calculated from *N. gonorrhoeae* DNA sequences. Electrophoretic gels with *N. meningitidis* DNA were performed for all the enzyme–genome combinations shown, as discussed in Results.

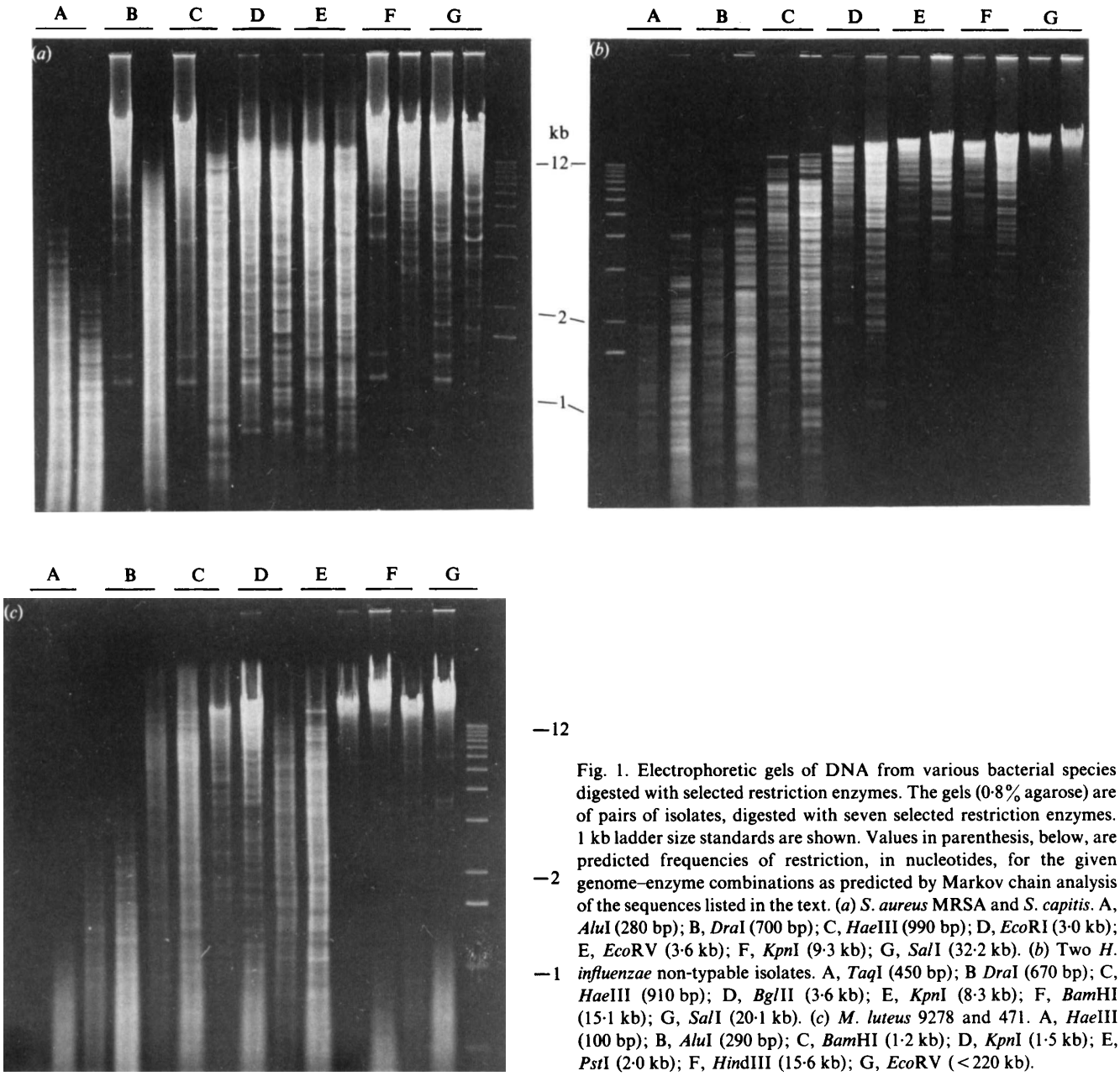

Fig. 1. Electrophoretic gels of DNA from various bacterial species digested with selected restriction enzymes. The gels (0·8% agarose) are of pairs of isolates, digested with seven selected restriction enzymes. 1 kb ladder size standards are shown. Values in parenthesis, below, are predicted frequencies of restriction, in nucleotides, for the given genome–enzyme combinations as predicted by Markov chain analysis of the sequences listed in the text. (*a*) *S. aureus* MRSA and *S. capitis*. A, *Alu*I (280 bp); B, *Dra*I (700 bp); C, *Hae*III (990 bp); D, *Eco*RI (3·0 kb); E, *Eco*RV (3·6 kb); F, *Kpn*I (9·3 kb); G, *Sal*I (32·2 kb). (*b*) Two *H. influenzae* non-typable isolates. A, *Taq*I (450 bp); B *Dra*I (670 bp); C, *Hae*III (910 bp); D, *Bgl*II (3·6 kb); E, *Kpn*I (8·3 kb); F, *Bam*HI (15·1 kb); G, *Sal*I (20·1 kb). (*c*) *M. luteus* 9278 and 471. A, *Hae*III (100 bp); B, *Alu*I (290 bp); C, *Bam*HI (1·2 kb); D, *Kpn*I (1·5 kb); E, *Pst*I (2·0 kb); F, *Hin*dIII (15·6 kb); G, *Eco*RV (<220 kb).

A. hydrophila genes for aerolysin (2346 bp) and extracellular amylase (1536 bp) allowed estimation of restriction site frequencies in this species, for which the calculated G+C content of 58 mol% compares to published values of 58–62 mol% (Krieg & Holt, 1984). The *E. chrysanthemi* shikimate kinase gene (2450 bp) was used to estimate restriction site frequencies in this species, for which the calculated G+C content of 54 mol% compares to published values of 55–57 mol% (Krieg & Holt, 1984).

To determine whether there was a correlation between these predicted fragment lengths and the range of fragment sizes observed in fingerprints, a selection of enzyme–genome combinations were tested (Table 2). Up to 14 different enzymes were used with up to 15 isolates from each of the six genera under examination. Representative gels for two isolates each from three of the genera (*Staphylococcus, Haemophilus* and *Micrococcus*) with seven appropriately selected restriction enzymes loaded in order of predicted restriction frequency are shown in Fig. 1.

There was generally quite close agreement between predicted and observed frequencies of restriction. As conventional gel electrophoresis with 0·8% agarose can

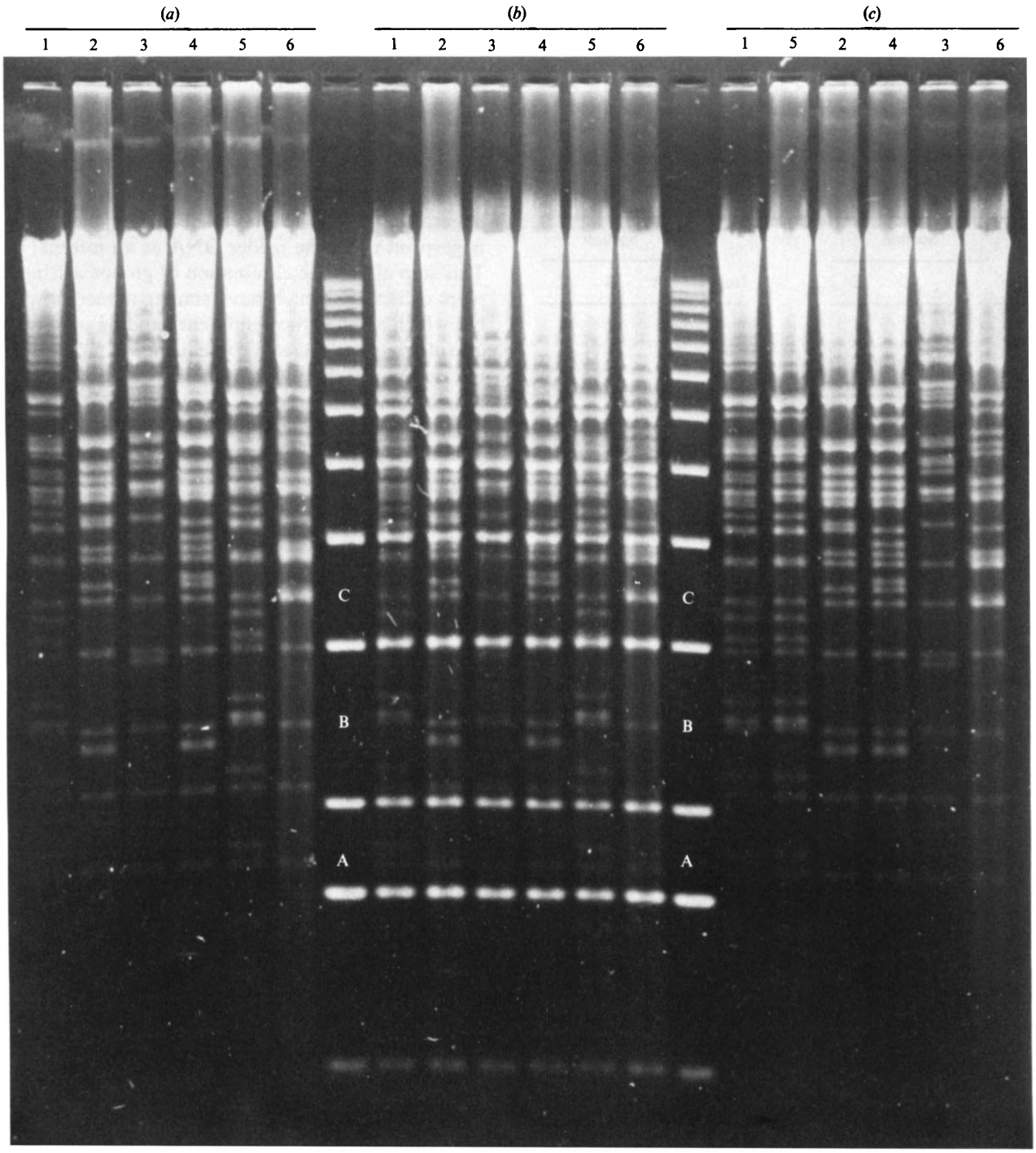

Fig. 2. Fingerprinting gel of *H. influenzae* DNA digested with *Bam*HI. (*a*) DNA fingerprints of six isolates, unordered. (*b*) As (*a*) but with the addition of 1 kb-ladder. The sections that were scored for the construction of the numerical profile are indicated (A, B, C). (*c*) Isolates regrouped after numerical profile classification.

only resolve DNA fragments up to about 8 kb (Sambrook *et al.*, 1989) the co-migration of larger fragments at the top of the gel gives an artificial upper size limit (Fig. 1). This apart, there was a generally observed trend of decreasing average size of fragments on gels with enzymes which were predicted to restrict that DNA more frequently (Fig. 1). In some enzyme–genome combinations the test DNA was not restricted, presumably due to appropriate methylation of bases in that recognition sequence in the DNA of these strains (*S. aureus* with *Dra*I, *Hae*III and *Kpn*I, Fig. 1*a*; or *H. influenzae* with *Hin*dIII, not illustrated). In other in-

Table 3. *Numerical profiles from fingerprint-sectioned H. influenzae DNA*

The data are from Fig. 2. Isolates 6 and 3 each show no similarity to other isolates. Isolates 2 and 4 differed by a single band in the sections scored, and in several other bands also. Isolates 1 and 5 have identical numerical profiles, though they do show a few band differences in the larger fragments.

Unordered numerical profiles				Ordered numerical profiles			
	Section				Section		
Isolate	A	B	C	Isolate	A	B	C
1	2	5	5	6	1	2	4
2	1	4	4	3	1	3	2
3	1	3	2	2	1	4	4
4	1	4	5	4	1	4	5
5	2	5	5	1	2	5	5
6	1	2	4	5	2	5	5

stances the DNA was restricted, but at a lower frequency than predicted (*M. luteus* with *Kpn*I, Fig. 1*c*). This could be due to methylation of a sub-population of these recognition sequence sites in the genome; for example through occasional overlap of this recognition sequence with another sequence, the latter being methylated appropriately. Conversely, and critically, we did not observe greater frequencies of restriction than would be predicted. Thus restriction frequency depends primarily on the sequence composition of the DNA and of the restriction enzyme recognition sequence, and secondarily, in a few cases, on the presence of protective methylation of an enzyme recognition sequence in a particular isolate.

Fingerprint sectioning, scoring and classification

Restriction enzymes suitable for fingerprint sectioning with 1 kb-ladder should cut the chromosomal DNA into fragments between 1 and 4 kb, as this part of the ladder has well-separated rungs. From Table 2 and Fig. 1, enzymes which cut DNA into predicted mean fragments of 3–15 kb in length are therefore appropriate, as these generally yield fragments in the 1–4 kb range. The final choice of the most suitable enzyme(s) for fingerprint sectioning must be done empirically.

The technique for scoring sectioned fingerprints is illustrated in Fig. 2, using a collection of 25 non-typable *H. influenzae* isolates from Aberdeen Royal Infirmary. Fig. 2(*a*) shows the DNA fingerprints of six of these isolates. After the addition of 1 kb-ladder (Fig. 2*b*) the numbers of bands in three sections (Fig. 2*b*, A, B, C) were scored. The results on reordering (Table 3) allow the rapid recognition of presumptively similar or identical groups of isolates. Scoring of the sectioned fingerprints benefits greatly from a consistency of technique, but was assisted to a degree by including potential pairs in the classification which have a greater number of mismatches. This will reduce problems that may arise in the scoring, as for example whether a band is from a single or two co-migrating fragments. Further confirmation of the similarities between these isolates using the whole fingerprint can also be made from the sectioned fingerprint using the ladder DNA as an internal guide. This step allows the elimination of groups arising from pairs of isolates which have similar numerical profiles but which actually have different banding patterns, and also of groups arising as a consequence of minor errors in the scoring technique. Homology within groups of isolates can be determined over whole fingerprints after the isolates are electrophoresed together in these new groupings (Fig. 2*c*). Other than the two pairs of related non-typable *H. influenzae* isolates shown in Fig. 2(*c*), all of the other isolates were readily distinguishable from each other.

Discussion

With DNA sequences becoming available for a large and diverse number of bacterial species, it is now possible to select restriction enzymes for bacterial chromosome fingerprint analysis of a particular species on a theoretical basis and thus avoid the tedious, and expensive, alternative of experimentally testing randomly selected enzymes. In the simplest case whereby DNA sequence is available for the species of interest a confident selection of the restriction enzymes for fingerprinting can be made. However, such a concordance may not always be possible, either through a lack of sufficient, suitable sequence data, or indeed none at all; in such cases the use of what limited sequence is available would probably lead to spurious predicted frequencies of restriction. Where more extensive sequences from a related species are available these would seem more appropriate. Care should be exercised that the substitute species has a similar mol% G+C content to the species under study and here comparison of the calculated mol% G+C content from the sequence can be useful. Given the strong correlation between the mol% G+C content of bacteriophage and their host species (McClelland, 1988), appropriate bacteriophage sequence might be another source of suitable sequence data. Where possible, sequences of low-expression genes should be used as these constitute the majority of the genome. High-expression genes often have biased codon usage and this may affect the frequencies of oligonucleotides calculated from these sequences (Phillips *et al.*, 1987*b*). Adequate

lengths of sequence to obtain meaningful predicted restriction frequencies must be representative of the genome as a whole, and the mol% G+C content gives some indication of this. The occurrence of unrepresented trinucleotides in the analysis would probably not be critical, as only restriction enzymes with comparatively high frequencies of restriction, and hence frequencies of restriction site, are used.

A very good correlation between the predicted frequencies of restriction based on gene sequences and those observed experimentally with the chromosomal DNA digests was found with all of the organisms used in the present study. Good correlations were found between observed fragment lengths of restricted DNA and the predicted frequencies of restriction from sequence data where these were from related species. This is of great practical help, as the selection of enzymes for fingerprint analysis of species with no sequence data can be based on sequences available from related species, thus greatly extending the utility of the method.

Some general trends were noted. As might be anticipated, the G+C content of genomes greatly affects the range of restriction enzymes that would be useful in fingerprint analyses. The equality with which all four nucleotides are present in species with a G+C content of 50 mol% (e.g. *Neisseria*) results in near-equal frequencies of restriction with most restriction enzymes which is tempered only by variation in the frequencies of the di- and trinucleotides in the DNA of that species. In contrast, restriction enzymes for species with high or low G+C contents will restrict the DNA over a much greater range of fragment sizes (e.g. *Staphylococcus* and *Micrococcus*). In practice then, the selection of enzymes will be more difficult for species with about 50 mol% G+C. This G+C bias is also seen with respect to the recognition sequence of the enzyme in combination with the G+C content of the species. Thus A+T-rich recognition sequences such as that of *Dra*I (TTTAAA) are much more frequent in G+C-poor genomes, and conversely G+C-rich recognition sequences, such as those of *Hae*III (GGCC) or *Apa*I (GGGCCC) are much more frequent in G+C-rich genomes. This correlation is also found with respect to strings of A+T or G+C in the recognition sequence. For example, the *Eco*RI recognition sequence (GAATTC), with a string of four A+T nucleotides, is increasingly rarer in G+C-richer genomes (e.g. *Micrococcus*) than those of enzymes which have the same frequencies of nucleotides but in a discontinuous A+T sequence – *Hin*dIII (AAGCTT) or *Bgl*II (AGATCT).

Moderately high frequencies of restriction were observed with the four-base recognition sequence restriction enzymes *Alu*I and *Taq*I, irrespective of the genome (Table 2). Where a rapid comparative fingerprint is required, as for example in one-off comparisons or where suitable DNA sequence data are unavailable, these may be ideal first-choice enzymes.

The sectioning of the fingerprints, and the classification of isolates on the basis of the number of bands present in three or four of the sections, allows a much speedier comparison of large numbers of isolates. This is facilitated not only by the ability to group together isolates on the basis of similar, or identical, numerical profiles after a single electrophoresis run, but to then further quantify any differences between non-adjacent pairs of isolates by using the internal ladder as a marker, again using this first gel. Similarity coefficients need only then be calculated for those isolates established as having very similar patterns. The very high frequency of chromosomal polymorphisms found in non-typable *H. influenzae*, one of the most genetically diverse organisms known, made this an ideal organism to test this method of strain grouping, and is being found to be most useful.

Where a greater degree of discrimination is needed or where isolates are clonally related the discrimination of fingerprint-sectioning can be increased by a second DNA digest with a different restriction enzyme, joining the two numerical profiles together and identifying groups from this larger numerical profile. For those species with a high degree of genetic diversity, such as non-typable *H. influenzae* or *N. meningitidis*, this should not be necessary.

Discrimination can also be increased by scoring larger fragments. In this way a greater proportion of the genome can be scored; up to 10% of the genome can be used as compared to less than half this where fragments of a few kilobases in size are used. This is most easily achieved by using higher-frequency restricting enzymes and using the upper portion of the fingerprint (e.g. Fig. 1*b*, *H. influenzae*, tracks B, C). Such enzyme–genome combinations can be selected using Table 2 and Fig. 1, or their equivalent for the chosen species. Ladders with bands in this higher molecular mass region need to be used, but since the inter-rung distances are purely arbitrary and do not need to be of equivalent size, suitable ladders can be constructed from digests of λ-phage DNA or of a plasmid, using a suitable restriction enzyme.

The techniques discussed here are all technically simple and allow a very rapid characterization of a collection of isolates to a degree not readily achieved by other methods. The time-consuming selection of suitable restriction enzymes is greatly reduced, as is the financial cost of testing a large selection of enzymes. Comparison of the isolates is similarly rapid, with quantitatively meaningful information available after running a single electrophoretic gel. These techniques should prove invaluable not just in situations where the identification

of a single strain is sought, as in epidemiological studies, but also where measures of similarities are required between large numbers of isolates, as in population genetic studies.

We acknowledge the use of the Seqnet node at the SERC Daresbury Laboratory, Warrington, UK, for the provision of DNA sequence analysis computer packages.

References

DICE, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302.

FITCH, T. A., TEMPLETON, J. W. & ADAMS, L. G. (1989). Homology of *Brucella abortus* strains 19 and 2308. *American Journal of Veterinary Research* **50**, 655–661.

GRIMONT, F. & GRIMONT, P. A. D. (1986). Ribosomal nucleic acid gene restriction patterns as potential taxonomic tools. *Annales de l'Institut Pasteur, Microbiologie,* **137B**, 165–175.

IRINO, K., GRIMONT, F., CASIN, I., GRIMONT, P. A. D. & BRAZILIAN PURPURIC FEVER STUDY GROUP (1988). rRNA gene restriction patterns of *Haemophilus influenzae* biogroup aegyptius strains associated with Brazilian purpuric fever. *Journal of Clinical Microbiology* **26**, 1535–1538.

KRIEG, N. R. & HOLT, J. G. (1984). *Bergey's Manual of Systematic Bacteriology*, vol. 1. Baltimore & London: Williams & Wilkins.

LOOS, B. G., BERNSTEIN, J. M., DRYJA, D. M., MURPHY, T. F. & DICKINSON, D. P. (1989). Determination of the epidemiology and transmission of nontypable *Haemophilus influenzae* in children with otitis media by comparison of total genomic DNA restriction fingerprints. *Infection and Immunity* **57**, 2751–2757.

MCCLELLAND, M., JONES, R., PATEL, Y. & NELSON, M. (1987). Restriction endonucleases for pulsed field mapping of bacterial genomes. *Nucleic Acids Research* **15**, 5985–6005.

MCCLELLAND, M. (1988). Recognition sequences of Type II restriction systems are constrained by the G+C content of host genomes. *Nucleic Acids Research* **16**, 2283–2294.

MUSSER, J. M., KROLL, J. S., GRANOFF, D. M., MOXON, E. R., BRODEUR, B. R., CAMPOS, J., DABERNAT, H., FREDERIKSEN, W., HAMEL, J., HAMMOND, G., HOIBY, E. A., JONSDOTTIR, K. E., KABEER, M., KALLINGS, I., KHAN, W. H., KILIAN, M., KNOWLES, K., KOORNHOF, H. J., LAW, B., LI, K. I., MONTGOMERY, J., PATTISON, P. E., PIFFARETTI, J. -C., TAKALA, A. K., THONG, M. L., WALL, R. A., WARD, J. I. & SELANDER, R. K. (1990). Global genetic structure and molecular epidemiology of encapsulated *Haemophilus influenzae*. *Reviews of Infectious Diseases* **12**, 75–111.

NEI, M. & LI, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 5269–5273.

OWEN, R. J. (1989). Chromosomal DNA fingerprinting – a new method of species and strain identification applicable to microbial pathogens. *Journal of Medical Microbiology* **30**, 89–99.

OWEN, R. J., FRASER, J., COSTAS, M., MORGAN, D. & MORGAN, D. R. (1990). Signature patterns of DNA restriction fragments of *Helicobacter pylori* before and after treatment. *Journal of Clinical Pathology* **43**, 646–649.

PHILLIPS, G. J., ARNOLD, J. & IVARIE, R. (1987*a*). Mono- through hexanucleotide composition of the *Escherichia coli* genome: a Markov chain analysis. *Nucleic Acids Research* **15**, 2611–2626.

PHILLIPS, G. J., ARNOLD, J. & IVARIE, R. (1987*b*). The effect of codon usage on the oligonucleotide composition of the *Escherichia coli* genome and identification of over- and underrepresented sequences by Markov chain analysis. *Nucleic Acids Research* **15**, 2627–2638.

PITCHER, D. G., SAUNDERS, N. A. & OWEN, R. J. (1989). Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. *Letters in Applied Microbiology* **8**, 151–156.

SAMBROOK, J., FRITSCH, E. F. & MANIATIS, T. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

SNEATH, P. H. A., MAIR, N. S., SHARPE, M. E. & HOLT, J. G. (1986). *Bergey's Manual of Systematic Bacteriology*, vol. 2. Baltimore & London: Williams & Wilkins.

SORENSEN, B., FALK, E. S., WISLOFF-NILSEN, E., BJORVATN, B. & KRISTIANSEN, B. E. (1985). Multivariate analysis of *Neisseria* DNA restriction endonuclease patterns. *Journal of General Microbiology* **131**, 3099–3100.

STAHL, M., MOLIN, G., PERSSON, A., AHRNE, S. & STAHL, S. (1990). Restriction endonuclease patterns and multivariate analysis as a classification tool for *Lactobacillus* spp. *International Journal of Systematic Bacteriology* **40**, 189–193.

WREN, B. W. & TABAQCHALI, S. (1987). Restriction endonuclease DNA analysis of *Clostridium difficile*. *Journal of Clinical Microbiology* **25**, 2402–2404.