An Evaluation of DNA pairing in Bacterial Systematics

by

Trudy Hartford

A Thesis submitted for the degree of Doctor of Philosophy in the Faculty of Science at the University of Leicester.

September 1990.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

This thesis is based on work conducted by the author in the Department of Microbiology at the University of Leicester, U.K., during the period between October 1985 and June 1989.

All the work recorded in this thesis is original unless otherwise acknowledged in the text by references. None of the work has been submitted for another degree in this or any other University.

Trudy Hartford.

ACKNOWLEDGEMENTS.

I would like to thank S.E.R.C. and Philips Scientific for financial support; Professor P.H.A. Sneath for supervision and the Microbiology Department for help and making my time in the Department a happy one. I would also like to thank Seamus for his continued support.

ABSTRACT

The main aim of this work was to evaluate the use of DNA-DNA pairing techniques in bacterial systematics. The genus *Listeria* was chosen for the study because of the small number of biochemical differences between the seven species. Also there has been a limited amount of nucleic acid studies carried out on the group using an endonuclease technique (Rocourt *et al.*, 1982), therefore some comparisons of the two techniques were possible.

Using optical DNA-DNA reassociation on a spectrophotometer with 23 *Listeria* strains from the seven species, a complete matrix of DNA-DNA homology values was produced. The data were analysed for reproducibility and second order kinetics.

Possible distortion of the derived taxonomic structure due to choice of reference strains was investigated by analysing the structure obtained from the complete matrix and comparing it to results obtained from incomplete 'strip' matrices. An analysis was made on a published matrix of complete DNA relationships (Nakamura and Swezey, 1983*a*; Hartford and Sneath, 1988) as well as on the data from *Listeria* species produced in this study. Great distortion in apparent taxonomic structure can result unless reference strains are widely spaced and representative of the clusters present. Problems caused by the choice of reference strains and the use of incomplete matrices was also explored by generating a random normal swarm of OTUs and illustrating the often bizarre effects obtained by using incomplete data sets in bacterial systematics.

DNA-DNA pairing data from a selection of published work were examined for experimental error. The average error from replications lay between 3 and 8.6 %, but the data were very limited.

CONTENTS.

2. METHODS AND MATERIALS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Hybridisation Techniques and their Application to Systematics.

Over the last ten to twenty years nucleic acid pairing studies have played an increasing part in microbial systematics. The major advantage of these techniques is that the entire genome can be examined and that effects of environment are excluded. DNA-DNA homology experiments (using the entire chromosomal DNA) can detect differences between closely related organisms, that is between strains of the same species or closely related species. These homology experiments involve denaturing native i.e. double-stranded DNA to its single-stranded form and monitoring the rate or degree of reassociation with heterologous or homologous DNA. In these experiments, where both nucleic acid strands are DNA, the process is usually called renaturation or reassociation. One variant is to measure the degree of reassociation between single-stranded DNA and RNA, (5sRNA or 16sRNA) here the process is called hybridisation. With the exception of genes that specify ribosomal RNA the bacterial genome does not contain repeated sequences i.e. there is only one copy of each gene (Kohne, 1968; Pace and Pace, 1971).

Hybridisation experiments involving mRNA are not widely used in taxonomy because a large portion of the genome is used for transcribing the mRNA molecules and similar results to DNA:DNA homologies would be obtained. Therefore in theory the nucleotide arrangement in rRNA seems to be more conserved than that of entire DNA, probably due to the role of rRNA in determining the structural and functional aspects of the ribosome (Woese *et al.*, 1975). Therefore hybridisation experiments using rRNA are used to detect similarities between more distantly related organisms than DNA renaturation experiments.

Double-stranded DNA can be separated into single strands by heating at a temperature high enough to break the hydrogen bonds between the

nucleotide base pairs. If the DNA is cooled quickly the strands will remain separated for some time. Incubating the strands at a suitable temperature allows them specifically to reassociate or reanneal with complementary DNA (or RNA) strands from the same or a different organism to form a double-stranded molecule. The rate or degree of reassociation reflects the similarity between the sequences.

Heavy isotope equilibrium centrifugation using caesium chloride, was one of the first methods used to detect hybrid formation of bacterial DNA (Doty *et al.*, 1960; Marmur *et al.*, 1961; Schildkraut *et al.*, 1961; Falkow *et al.*, 1962; DeLey and Friedman, 1964, 1965; Friedman and DeLey, 1965). The technique involved growing the reference strain in "Heavy medium" labelled with ^{15}N. This method is not widely used and is not practical for routine use; not all organisms grow in "Heavy medium" and the resolution of peaks of similar densities is very difficult.

Hybridisation techniques can be divided into two main categories : those using DNA fixed to a porous support and those reacting in free solution.

Nitrocellulose Filter Methods.

The DNA agar technique was the first simple technique used to measure nucleic acid relatedness (Bolton and McCarthy, 1962; McCarthy and Bolton 1963; Hoyer *et al.*, 1964). High molecular weight denatured DNA is added to molten agar, then quickly cooled. The agar is washed to remove any DNA not trapped in the agar. Radiolabelled, sheared, denatured RNA or DNA is incubated with the DNA-agar to allow reassociation. This technique was replaced by the nitrocellulose filter technique (Nygaard and Hall, 1963). In 1965, Gillespie and Spiegelman devised a quantitative assay using sheared single-stranded RNA and high molecular weight filter bound DNA. Several modifications were devised to avoid non-specific adsorption

of single-stranded DNA (Denhardt, 1966; Warnaar and Cohen, 1966; Legault-Demare *et al.*, 1967) so that DNA:DNA hybrids could be detected using the filter method.

There are two ways of using the filters. In the direct method high molecular weight DNA is immobilised on a nitrocellolose filter, then incubated with sheared, denatured labelled DNA. Any unbound DNA is washed away and radioactivity on the membrane counted. The percent homology of heterologous DNA is expressed as the ratio of radioactive counts of the immobilised DNA on the membrane filter in the heterologous system relative to that of the homologous system x 100. These values are normalised to the relative percent binding, using the percent homology of the homologous system. A major disadvantage is the difficulty in obtaining a consistent amount of DNA on the membranes, and leaching of DNA at high temperatures (Okanishi and Gregory, 1970).

The other variant is the competition method first suggested by Hoyer *et al.* (1964). In this method the labelled DNA and immobilised DNA are from the same source. The direct method is followed except that the amount of reduction of bound label is measured by including an excess of unlabelled homologous or heterologous DNA with the labelled DNA. An inhibition reaction occurs as the DNAs compete for complementary sites on the immobilised DNA. The decrease in binding caused by the competitor DNA is compared to the decrease due to using the homologous DNA as the 'competitor'; the reduction in binding is directly related to the complementarity of the two DNA preparations.

Competition experiments are often preferable to the direct filter method as they may be more sensitive in detecting related sequences (McCarthy and Bolton, 1963; Hoyer and King, 1969). The amount of label bound in the direct procedure is highly dependent on incubation conditions; only about 40-50 % of total label will be bound under usual

conditions. Hybridisation values are significantly affected by incubation conditions, especially the temperature (DeLey and Tijtgat, 1970). This is a major problem for organisms with a high % (G+C) ratio which require high temperatures to denature their DNA (Bonner *et al.*, 1967; Legault-Demare *et al.*, 1967; McConaughy *et al.*, 1967; DeLey and Tijtgat, 1970; Rogul *et al.*, 1970) but this has been partially overcome by the use of urea or formamide in reaction solutions (Bonner *et al.*, 1967; McConaughy *et al.*, 1967; Okanishi and Gregory, 1970; Gillespie and Gillespie, 1971; Kourilsky *et al.*, 1971; Schmeckpeper and Smith, 1972).

Hydroxyapatite Method.

Labelled, sheared DNA and unlabelled sheared DNA fragments are denatured and renatured in solution. The concentration of unlabelled fragments is made the rate-determining factor by using a huge excess of unlabelled fragments, at least 2000:1 (Britten and Kohne, 1966). When the reaction is complete, the mixture is passed through a column of hydroxyapatite (a modified calcium phosphate gel) (Bernardi, 1965; Miyazawa and Thomas, 1965; Brenner *et al.*, 1969). Single-stranded DNA is eluted whereas duplexes are adsorbed. Then the duplexed nucleic acids are eluted by raising the ionic strength of the phosphate elution buffer or by raising the temperature until the double strands are dissociated (Brenner *et al.*, 1969). The amount of radioactivity in each fraction is measured and the homology is determined by comparing the ease of elution of the heterologus duplexes with that of homologous duplexes. Homology is expressed as the % RBR, or relative binding ratio (Brenner *et al.*, 1978):

[(% heterologous DNA bound to HA)/(% homologous DNA bound to HA)] x100

Temperatures must be kept $\geq$55°C or non-specific binding of DNA to HA (hydroxyapatite) will occur. This technique has the advantage that no immobilisation of DNA is required and therefore there are no leaching

problems. Also the amount of label bound is very large (85-95 %; Brenner *et al.*, 1969; Staley and Colwell, 1973*a*) ruling out the possibility that the reaction is not representative of the whole DNA molecule. A further advantage is the ability to precisely determine the quantity of labelled and unlabelled reactant involved by spectrophotometric measurements. The number of samples which can be analysed simultaneously is usually a limiting factor; however Brenner *et al.* (1969) used a batch procedure, running up to ten samples at once. Inconsistencies may arise in handling samples as the batch procedure requires that centrifuge rotors and water baths be kept at constant high temperatures (Lachance, 1980). Lachance (1980) devised a method to chromatograph simultaneously numerous samples in microcolumns. This variant was comparable in reproducibility and reduced the volume of elution buffers.

The nucleic acid reassociation studies using the hydroxyapatite technique have also been used in evolutionary studies to establish phyletic classifications and establish lines of evolutionary divergence (Britten and Kohne, 1966).

Endonuclease Technique.

A mixture of labelled and unlabelled DNA is hybridised in solution and then subject to digestion with an endonuclease (usually S1 nuclease) which removes any single-stranded DNA remaining in the hybridisation mixture. S1 nuclease isolated from *Aspergillus oryzae* was first used in relatedness experiments to improve the nitrocellulose membrane DNA hybridisation method (Crosa *et al.*, 1973; Ogasawara-Fujita and Sakaguchi, 1976). After digestion by S1 nuclease the DNA duplexes are separated from the digestion products by either cold trichloracetic acid (TCA) precipitation (Crosa *et al.*, 1973) or by collecting on DEAE cellulose filters (Saltzberg *et al.*, 1977; Popoff and Coynault, 1980). The latter

variant is based on the fact that single nucleotides, not DNA, can be eluted from DEAE-cellulose filters with a phosphate buffer (this is more reproducible than the TCA precipitation method; Grimont *et al.*, 1980). Radioactivity on the filter is determined and values are corrected for radioactivity resistant to endonuclease digestion by self-annealing of labelled DNA. Homology is expressed as the % of radioactivity in the heterologous hybridised DNA relative to that in homologous hybridised DNA. Ogasawara-Fujita and Sakaguchi (1976) reported fairly reproducible results even if labelled reference DNA and unlabelled DNA were exchanged. The technique has the advantage that up to 50 tubes per day may be assayed. It should be noted that there is an S1 resistant core of up to 10% as a consequence of label-label DNA reassociation (Popoff and Coynault, 1980). It is also important to note that many of the agents used in DNA isolation (for example, chelating agents, detergents, alkali) may interfere with the S1 nuclease assay unless carefully removed by dialysis (Crosa *et al.*, 1973).

Optical Techniques.

DNA pairing relationships determined by the optical techniques are based on the assumption that at a given temperature an increase in the number of bases in the paired form is proportional to the decrease in optical absorbance at 260 nm. Applequist (1967) suggested a formula to relate hypochromicity of a helix length n, $H(n)$, to that of a helix of infinite length

$$H(\infty) : \frac{H(n)}{H(\infty)} = 1 - \frac{1}{n} \qquad \text{where } n > 7$$

This implies that hypochromicity is simply proportional to the number of stacking interactions. The reaction

$$\underset{\text{DNA}}{\text{single-stranded}} + \underset{\text{DNA}}{\text{single-stranded}} \xrightarrow{k} \underset{\text{DNA}}{\text{double-stranded}}$$

is $C/C_o = 1/(1 + k_{cot})$ (Britten and Kohne, 1965)

where : C = concentration of single-stranded DNA in moles nucleotides l^{-1}

C_o = total DNA concentration moles nucleotides l^{-1}

k_{cot} = reassociation rate constant dependant on the incubation conditions and DNA complexity.

Britten and Kohne (1966) devised the term $C_o t$ to present measurements of the time course of reassociation. At constant temperature, salt concentration and fragment size, reassociation is determined by DNA concentration and time of incubation. The reaction is almost perfectly second order, although reassociation only occurs to about 90 % due to steric hinderance preventing binding in the short unpaired areas remaining (Britten 1969). DNA homology can be measured by determining the $C_o t\ ^1/2$ values (Britten and Kohne, 1966) of two DNA samples and comparing the values to the $C_o t\ ^1/2$ of a 50:50 mixture of the two DNA samples. The $C_o t\ ^1/2$ is the concentration of DNA, in moles of nucleotide per litre, multiplied by the time in seconds for 50% reassociation. For completely homologous DNAs the $C_o t\ ^1/2$ of the mixture of their DNAs is the same as that of each single DNA. If there are no homologous sequences between two DNAs, the $C_o t\ ^1/2$ of a mixture of the two DNAs will be the sum of the two $C_o t\ ^1/2$ values of the single DNAs. Homology is determined from the following equation (Bradley, 1972) :

$$\% \text{ Homology} = 200(C_o t\ ^1/2A + C_o t\ ^1/2B - C_o t\ ^1/2Mix)/(C_o t\ ^1/2A + C_o t\ ^1/2B)$$

Double stranded or native DNA absorbs at a wavelength of 260 nm with

an extinction coefficient of 1.47 x 10^{-4}M (Caster, 1951; Beaven *et al.*,1955). The concentration is linearly related to the absorbance at 260 nm (Fredericq *et al.*, 1961). As the temperature is increased the DNA begins to separate into single strands, and the absorbance of the solution increases until all the DNA is single-stranded. If the DNA is pure, its absorbance in the single stranded form will be approximately 36 % above that in the double stranded form (Wetmur, 1976); this phenomenon is termed the hyperchromic shift and can be determined from the following equation :

Hyperchromicity = 41.1 - 0.21 % (G+C) (Gillis *et al.*, 1970).

The T_m, melting temperature, is taken to be the temperature at which half the DNA is in the single stranded form. Melting curves appear as smooth transitions when the rate of temperature increase is in the range of 0.5-1.0 °C per minute (Johnson, 1985). The temperature at which DNA dissociates is determined by its % (G+C) (Marmur and Doty, 1962) and the reaction conditions, particularly ionic concentration (Britten and Kohne, 1966). If the temperature is lowered quickly the DNA will remain denatured. Droppping the temperature below T_m allows some bases to hydrogen bond non-specifically (Marmur and Doty, 1961). By holding the mixture at a suitable temperature, usually 25-30°C below the T_m (in the same buffer system) (Marmur and Doty, 1961; Marmur *et al.*, 1963), any non-specific bonds are disrupted causing an initial increase in absorbance followed by a decrease in absorbance at 260 nm as the strands reanneal; this is known as hypochromism. The rate of reassociation, measured as decrease in absorbance units per minute, is linear under given conditions over a given length of time. The rate is faster for homologous DNA than for a mixture of two heterologous DNA preparations.

DeLey *et al.* (1970) used these principles to measure the relatedness

of bacterial DNA. They compared the renaturation rate of a mixture of equal amounts of two heterologous DNA samples *a*, *b*, with the rates of the two homologous reactions, using the following equation when a and b have the same genome sizes .

$$\% \text{ pairing} = \frac{4V_m - (V_a + V_b)}{2\sqrt{(V_a \cdot V_b)}}$$

Where : V_m = renaturation rate of the mixture of *a* + *b*

V_a = renaturation rate of DNA from organism *a*

V_b = renaturation rate of DNA from organism *b*

The most obvious advantage of the optical technique is that no radiolabelling is involved which is expensive and time-consuming. For practical reasons homology studies which involve radiolabelling are usually limited in the number of labelled reference strains, so strains are only compared to the reference strains and not amongst themselves. Organisms with % (G+C) compositions which differ up to 8 % may still be compared with the optical method at an intermediate renaturation temperature (DeLey *et al.*, 1970). If hybridisation is measured by absorbance the effect of open loops and free ends is kept to a minimum. In filter techniques, unpaired bases in loops and unpaired end portions of part reassociated strands would add to the apparent % hybridised strands; absorbance would not be decreased by loops or free ends; therefore they will not affect the % homology determination.

It is very important to carry out the experiments under controled conditions in order to obtain reproducible results. Some of the factors affecting renaturation rates are discussed in section 1.3.

Theoretical Estimation of % Homology.

DeLey (1969) devised a mathematical method for estimating the maximum amount of DNA homology between two species of bacteria. The calculation involved the % (G+C), the molecular weight and the compositional nucleotide distribution of the DNA. The relationship between the % (G+C) difference and maximum possible percent homology is linear for base composition differences up to 11 % but deviates for larger differences. DeLey assumed that the G-C bases were normally distributed with respect to the mean base composition and thus calculated that a difference in 1 % (G+C) base pairs was equivalent to a decrease in common base sequences of 9 %.

1.2 Base Composition

An added advantage of the optical technique is the possibility of determining the base composition of the nucleic acids under study. As already mentioned, the T_m of DNA is related to its base composition. G-C base pairs exhibit a higher thermal stability than A-T base pairs due to the number of hydrogen bonds involved with each pairing. Therefore the greater number of G-C pairs within a DNA duplex the greater the thermal stability (Marmur and Doty, 1962), so the T_m increases with increasing % (G+C).

The correlation of spectrophotometric properties of DNA and DNA components with chromatographic data (Chargaff, 1955; Marmur and Doty, 1962) resulted in the spectrophotometric method of estimating % (G+C) values. There is a linear correlation of T_m and % (G+C) for DNA preparations with a base composition of between 25-75 % (G+C) (Marmur and

Doty, 1962; Owen *et al.*, 1969; DeLey, 1970; Mandel *et al.*, 1970).

T_m is affected by ionic strength of the buffer (Schildkraut and Lifson, 1965; Britten and Kohne, 1966) so this must be kept constant and temperature standardised by including a standard such as *E. coli* strain B or strain K-12.

T_m is also affected by fragment size (Johnson, 1985) and may be 1-2 °C higher for unsheared DNA compared with the same DNA passed through a French pressure cell; this effect is more pronounced in organisms with a low % (G+C) composition (Selin *et al.*, 1983).

1.3 Factors Affecting Renaturation Rates.

The extent of reassociation and the specificity of duplex formation in all pairing techniques are governed by many physical and chemical conditions - these are discussed below.

Temperature

At the T_m the renaturation rate is zero. If the incubation temperature is lowered the renaturation rate will increase until the optimal renaturation temperature (T_{or}) (Marmur and Doty, 1961; Wetmur and Davidson, 1968) is reached; the rate plateau then decreases. Marmur *et al.* (1963) first obtained this bell-shaped rate constant-temperature curve illustrated in Figure 1.1. The T_{or} in 2 x SSC (2 x standard saline citrate) buffer is 22-25°C below the T_m measured in the same buffer (Marmur and Doty, 1961; Marmur *et al.*, 1963). When the T_{or} is used the homologous duplexes formed exhibit a thermal stability similar to that of native DNA (Johnson and Ordal, 1968).

DeLey *et al.* (1970) showed that the homology value is almost independent of temperature over a range of 15°C below the optimal renaturation temperature (later supported by Huss *et al.*, 1983). However Gillis *et al.* (1970) found the flat optimum range extended over only 5-10°C below the T_{or}.

The nature of hybrids formed varies with the incubation temperature and with the DNA used (DeLey *et al.*, 1973). If renaturation is allowed to progress at temperatures higher than the T_{or} unstable duplexes are incapable of forming and only perfectly matched sequences reanneal. These are termed stringent conditions. Stringent conditions are typically T_{or} + 10°C, while non-stringent conditions are typically T_{or} - 10°C. Under the latter conditions a considerable proportion of non-specific bonding

Figure 1.1 The effect of temperature on the renaturation rate and the degree of binding.

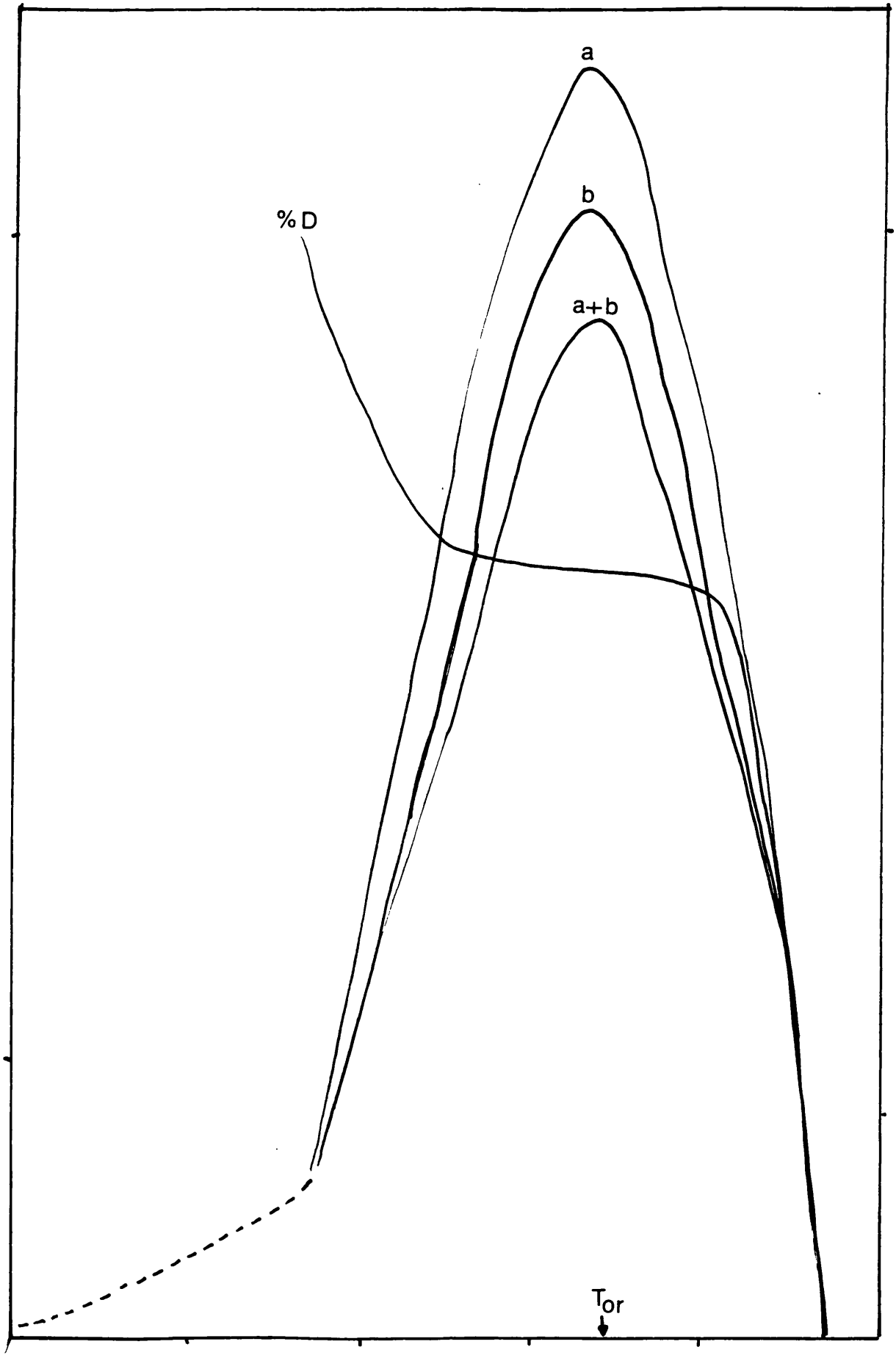
a : Results from DNA isolated from organism a.

b : Results from DNA isolated from organism b.

a + b : Results obtained from an equal mixture of DNA from organisms a and b.

% D : Degree of binding or % homology.

T_{or} : Optimal renaturation temperature.


a
b
%D
a+b
Tor
Renaturation Rate Δ A/t x10⁻³
Degree of Binding %
Hybridization Temperature ℃

occurs and relationships may appear falsely close (McCarthy, 1967; Johnson and Ordal, 1968; Kohne, 1968). In heterologous reactions the extent of binding can be shifted ten fold by varying the temperature of incubation (Brenner *et al.*, 1967; Johnson and Ordal, 1968).

Martin and Hoyer (1966) suggested the use of a thermal binding index (TBI) where the ratio of the degree of binding at T_{or} to that at a higher, more restrictive, temperature is used to aid further separation of similar DNAs and measure the thermal stability of duplexes formed. The TBI indicates the presence or absence of highly related genetic material in heterologous reassociation reactions. A low TBI indicates that most duplexes are not stable and therefore not highly complementary. Brenner *et al.* (1969) used temperatures of 75°C and 60°C to form TBIs. TBIs of duplexes with a degree of binding over 70 % range from 0.85 to 1.0, those of duplexes with a degree of binding below 60 % range from 0.4 downwards (Brenner *et al.*, 1969; Citarella and Colwell, 1970; Brenner *et al.*, 1972*a*). Although many taxonomic studies use two temperatures (one optimal and one stringent) for example, Collins *et al.* (1987), the TBI is not generally used often in taxonomy.

Buffer (ionic) Concentration.

The T_m is almost a linear function of the logarithm of the sodium ion concentration (Schildkraut and Lifson, 1965; Owen *et al.*, 1969) and the T_{or} is related to the T_m. The lower the salt concentration the longer the reaction remains in second order (Marmur *et al.*, 1963; Subirana and Doty, 1966). The dependence of rate on salt concentration is not as marked in high salt concentrations. Generally 2 x SSC buffer produces rates which are constant for 15-30 minutes; but the strength of buffer required may vary with the organisms used (Gillis *et al.*, 1970). Hence if

the renaturation reaction is too quick and does not allow the initial tangent to be accurately traced, then the salt concentration must be reduced.

DNA Concentration.

To maintain a near linear reaction rate long enough to provide accurate measurements, the DNA concentration must not be too high or steric hinderance will cause deviation from second order kinetics (Subirana, 1966; Subirana and Doty, 1966). Gillis *et al.* (1970) found second order kinetics were followed for up to 80 μgml^{-1} of sheared denatured DNA in 2 x SSC buffer at the T_{or}. The optimal concentration is therefore just below 80 μgml^{-1}. Huss *et al.* (1983) suggested 30-40 μgml^{-1} as the optimal concentration and as concentration increased above this, so did deviation from the theoretical degree of binding (DeLey, 1969). At low DNA concentrations unspecific base pairing may occur (Huss *et al.*, 1983).

Fragment Size.

The renaturation rate depends on the DNA fragment size (Marmur and Doty, 1961; Wetmur and Davidson, 1968). In theory the largest possible fragment size would give the most accurate results because it allows high rates. Also a small difference in fragment size would not noticeably affect the rate. In practice a large fragment size is not reproducible and the rate produced is too fast to allow accurate measurement of the initial rate (Gillis *et al.*, 1970). If matching and mismatching sequences alternate along a DNA strand, after one stretch has paired the neighbouring mismatching sequence may prevent the pairing of the next

adjacent matching sequence (DeLey *et al.*, 1970); this would not occur as often with shorter fragments.

The degree of binding is unaffected by fragment size within the range of 300,000 - 550,000 daltons (Huss *et al.*, 1983); however errors could easily arise if the renaturation rate is too fast and thus difficult to measure.

Seidler and Mandel (1971) investigated three methods of shearing bacterial DNA : sonic oscillation, passage through a 27 gauge needle and passage through a French press at 15,000 psi. The latter was found to be the most convenient and reproducible; however Garvie (1978), using streptococci, found the French press yielded fragments which were too small, resulting in renaturation rates that were too curved. She found that 10 passages through a 26 gauge syringe was much better. Owen and Snell (1976) employed sonic oscillation to obtain fragments in the range of 2×10^5 - 3×10^5 daltons, but the fragments were used with the filter technique not for renaturation rate determinations. A fragment size of 300-400 nucleotides per single strand, or 2×10^5 - 3×10^5 daltons, is usual for most techniques.

If the fragments are too small the renaturation reaction may proceed at a rate which prevents accurate measurement of the initial tangent i.e. the reaction is second order for only a very short time. The minimal specific DNA fragment size in bacteria is approximately 15 nucleotides (McCarthy, 1967).

Purity.

The hyperchromicity of a DNA sample observed during T_m determinations depends on the physical condition of the DNA rather than the base composition (Crombach, 1972). Hyperchromicity will decrease if

many hydrogen bonds are already broken before denaturation occurs, for example, by rough isolation and purification of DNA. Owen and Lapage (1976) found hyperchromism less with partly purified DNA than pure DNA although the T_m determinations were the same with both preparations.

The extent of contamination by protein and salt in a DNA sample can be assesed from its absorbance spectra (Marmur, 1961; Kalb and Bernlohr, 1977). The wavelengths 230 and 260 nm correspond to the approximate minimum and maximum of the nucleic acid absorbance spectrum (Kalb and Bernlohr, 1977). It is generally thought that DNA samples must be purer for use in optical renaturation experiments than with other techniques; however 2-5 % protein added to pure DNA samples has no significant effect on thermal denaturation profiles (Crombach, 1972) nor do small amounts of protein appreciably affect renaturation rates (DeLey *et al.*, 1970).

If RNA is present in a DNA sample, the absorbance increases distinctly between 25-60°C due to RNA inter and intra-strand hydrogen-bridges melting (Hastings and Kirby, 1966).

Seidler and Mandel (1971) found that carbohydrate removal had no detectable effect on homology determinations, but this may vary with the organism used as erratic results have been obtained due to polysaccharide production with *Agrobacterium* and *Xanthomonas* strains (DeLey *et al.*, 1970). Heat-denatured DNA forms complexes (hydrogen bonded) with complex carbohydrates; these complexes scatter light and may decrease the 260/280 nm absorbance ratio (Graves, 1968).

Storage of DNA.

Storage at -21°C has no effect on fragment size, thermal denaturation or hybridisation (Crombach 1973) if the preparation is frozen immediately on dissolving. DeLey *et al.* (1970) found DNA samples

could be stored at 4°C for several months with no detrimental effects if the absorbance at 260 nm was at least 20.

Chromosome Replication State and Genome Size.

It is possible that DNA prepared from different stages of growth may effect the degree of homology obtained. If so replication state effects would be reduced if all DNA was prepared at the same stage of growth, preferably the stationary phase. Seidler and Mandel (1971) demonstrated that DNA prepared from log phase cells renatures faster than stationary phase DNA and also departs from second order kinetics.

1.4 The Relationship between Thermal Stability and Base-Pair Mismatching.

The temperature at which double strands separate into single strands is higher for perfectly base-pair matched double-stranded DNA than that for double-stranded DNA which has some base pair mismatches (Laird *et al.*, 1969). Indeed, this is the converse of swifter reannealing of homologous strands than heterologous strands. Bonner *et al.* (1973) made a series of measurements to assay the effect that imperfect matching had on reassociation rates. The greater the degree of mismatches the lower the thermal stability; hence it should be possible to detect imperfect matching by melting after hybridisation. Crombach (1974) defined the term ΔT_{me} as the difference between the remelt T_m of the heterologous duplex with that of the homologous duplex (T_{me} = T_m DNA sample - T_m of same sample after it has been denatured and then reassociated) i.e.:

$$\Delta T_{me} = \text{homologous } T_{me} - \text{heterologous } T_{me}$$

Several correction factors have been proposed: a 1°C drop in T_m per 1.4 % unpaired bases (Laird *et al.*, 1969), 1°C per 1 % unpaired bases (McConaughy *et al.*, 1967; Brenner and Cowie, 1968; Bonner *et al.*, 1973). These may be used to determine the accuracy of the reaction or sequence divergence.

The amount of base pair mismatching is affected by the incubation temperature (Johnson and Ordal, 1968; Brenner and Cowie, 1968). At a more stringent temperature the thermal stability of the duplexes formed may increase to give a significant drop in the degree of duplexing. These hybrids of high stability are formed in the initial phase of renaturation (Crombach, 1974).

Thermal stability profiles of reassociated nucleic acids bound to hydroxyapatite can be generated by washing at a series of increasing temperatures. When the temperature exceeds the denaturation temperature

of a duplex it will elute as single stranded DNA (Britten and Kohne, 1966).

1.5 Correlation of Numerical Taxonomic and Nucleic Acid Pairing Studies.

When the DNA-data has been analysed the results must be correlated to those of conventional phenotypic testing and with data from other chemotaxonomic techniques. The degree of correlation of numerical taxonomic and nucleic acid pairing studies seems to vary with the group of organisms. Staley and Colwell (1973*a*) when comparing numerical and nucleic acid data, found a good linear correlation (0.84) for similarity versus pairing values between 75-100 %; however below 50 % pairing and 60 % similarity the phenotypes suggest a closer relationship than the DNA sequences. On this basis they suggested 75-80 % or more for a species and 50-75 % to 'define' a genus, in comparison with Brenner (1973) and Johnson's (1973) 70 %-100 % to define a species. Johnson (1973) suggested that 60-70 % homology corresponded to a sub-species and 30-60 % a 'moderate' relationship. Huss *et al.* (1983) found values under 30 % determined with the optical technique were unreliable as they were rarely obtained and of no use in classification.

The relationship between phenotypic similarity from the percent of shared phenotypic properties (many based on biochemical tests) and DNA:DNA pairing is not straight but sigmoid shaped with a good deal of scatter (Sneath, 1972). The scatter makes it difficult to predict reliability from one scale to the other. There is very little DNA-DNA pairing data at phenetic similarities below about 50 %, but the relationship can be made almost linear by a probit transformation; Sneath (1972) found the relationship to be : probit (% DNA pairing) = 0.4029 + 0.0653 (% Phenetic Similarity).

1.6 Reliability of Nucleic Acid Pairing Methods in Taxonomy.

Whenever information obtained from one method for measurement of homology is to be compared with results from another method the limitations of the techniques and level of agreement between different methods have to be assessed.

The amount of error associated with nucleic acid pairing studies is rarely taken into account; even when the reproducibility is recorded, the effects of the error on the conclusions drawn from the work are often not discussed. However, DNA-DNA pairing has great potential for determining accurate relationships because of its low sampling error and this potential is greatly diminished by experimental error.

Few papers state the number of replications and standard deviation associated with each result. The study of Potts and Berry (1983) is one of the few which lists these.

Comparison of Pairing Techniques

Comparison between values from different methods of nucleic acid pairing has been discussed by several authors (Brenner *et al.*, 1969; Kingsbury *et al.*, 1969; DeLey *et al.*, 1970; Seidler and Mandel, 1971; Gibbins and Gregory, 1972; Crosa *et al.*, 1973; Coykendall and Munzenmaier, 1978; Grimont *et al.*, 1980; Huss *et al.*, 1983; Bouvet and Grimont, 1986). Grimont *et al.* (1980) compared the two S1 nuclease techniques and the hydroxyapatite technique and found a high varience between methods as well as high variance between different bacteria. Error due to the determination of the proportion of reassociated DNA and counting error affects variance within experiments. The S1

nuclease-filter technique was found to be more reproducible than the S1 nuclease-TCA precipitation or hydroxyapatite methods. Curvilinear relationships were found between data obtained with the S1 nuclease procedures and hydroxyapatite procedure.

Data produced from the filter technique are usually higher than that produced by the endonuclease technique for the same set of strains (Coykendall and Munzenmaier, 1978; Bouvet and Grimont, 1986). Bouvet and Grimont (1986) found a curvilinear relationship between these two methods although the two data sets were produced by different working groups and over fifteen years apart.

Confidence limits for DNA homology measurements were determined by Hildebrand *et al.* (1984) calculated from reciprocal homology values produced in labelling methods; they used an average calculated from three different pairing studies and found that the 95 % confidence limit for a given DNA homology measurement is about ± 7.3 % of the expected mean pairing value.

Pairing values of over 100 % (that is greater than the hybridisation of homologous DNA) are often published; values as high as 115 % have been published especially in radiolabelling methods. Theoretically values over 100 % should not exist and point to some form of inconsistency. Often with radiolabelling methods, the homologous reaction will be measured only once even if heterologous reactions are measured in triplicate.

Differences in reciprocal values, where one strain is radiolabelled and then the other, have been noted by Brenner *et al.* (1972*a*); theoretically the differences should be within experimental error. If the differences are very large it suggests that the genome sizes of the strains involved may be substantially different (Brenner *et al.*, 1972*b*;

Staley and Colwell, 1973*b*). Reciprocal relatedness values are only expected to be equal when their genomes have essentially the same molecular weight, and the difference in the values will increase with the relative genome size differences.

Other possible sources of error may be due to extrachromosomal elements, incomplete pairing of base sequences and variations in the reassociation rate due to differences in base composition (Staley and Colwell, 1973*b*).

Choice of Reference Strains.

To uncover true homology relationships between organisms in a study, all strains must be compared with each other to provide a complete matrix of pairing values. These are rarely found due to the cost and effort of obtaining complete matrices. Usually a few strains are employed as reference strains and other strains are compared to this restricted set. Often the second reference strain is chosen from the strains which have a low similarity to the first reference strain and so on. Alternatively reference strains are chosen on the basis of previous (numerical) taxonomic studies. This strategy gives in effect, a few strips of DNA-DNA values in an incomplete inter-strain matrix. If taxonomic structure is derived from this the question arises: how many reference strains are required in order to obtain the 'true' taxonomic structure and how would the choice of reference strains effect the taxonomic structure? The problem is to recover the underlying taxonomic structure, i.e. to recover a structure that is as similar as possible to the structure one would obtain from a complete matrix between all pairs of strains. A level of homology is often selected over which OTUs are grouped to form a species or taxonomic category; this is commonly 70 % for a species in DNA-DNA homology (it would be much higher in RNA-DNA homology studies). An exact

cut off point for a species should not be used unless it is known that all techniques yield comparable results; 70 % in the optical technique may not be the same level of relatedness as 70 % in the filter technique. Also the extent of error and number of replications should be reflected upon.

The problem of a restricted reference has been discussed by Staley and Colwell (1973*b*), Cristofolini (1980), Grimont and Popoff (1980), Sneath (1980; 1983) and Mutters *et al.* (1985). The problem may be illustrated by trying to reconstruct the geography of towns of Britain solely from distances to London and Newcastle. From such distances it is impossible to say whether Bristol and Norwich are close or distant; the required information is not there. However, one can say that Watford and St. Albans are close, because both are very close to London. Further, one might obtain a different reconstruction if one measured only distances from Norwich and from Bristol; one could not now tell if Watford and St. Albans were close or not.

Often a complete matrix is derived from the incomplete relationships and then analysed, this leads to either an explicit matrix of resemblances between all the strains (Sneath, 1980; 1983) or to principal component analysis in which a derived matrix is implicit (Grimont and Popoff, 1980).

Sneath (1983) showed that the choice of reference strains makes a large difference to the taxonomic structure that is recovered from derived matrices. He also showed this was so when the reconstruction was by minimum spanning trees, though this technique has not been used on DNA data. The effects of choice of reference strains has been followed up (see 2.12, 3.9). As part of the reexamination of the reliability of DNA techniques for taxonomy the complete DNA-DNA pairing matrix from Nakamura

and Swezey (1983*a*) using 17 *Bacillus* strains was used.

Despite all the limitations and possible sources of error with DNA pairing techniques, taxonomic studies have often used these methods alongside or without the results of numerical, biochemical studies to derive taxonomic groupings.

1.7 *Listeria* Taxonomy

In 1926 Murray *et al.* described and named what later became the type species of the genus *Listeria* as *Bacterium monocytogenes* after isolating it from laboratory rabbits with mononuclear leucocytosis. The same organism was isolated by Pirie (1927) from the livers of African jumping mice; he named it *Listerella hepatolytica,* but later suggested *Listeria monocytogenes* strictly for nomenclatural reasons (Pirie 1940).

Members of the genus *Listeria* are Gram-positive, non-sporing, motile, rods which are usually catalase positive. Until 1961 there was only one species : *Listeria monocytogenes.* The latest Bergey's Manual of Systematic Bacteriology (Volume 2, 1986) now lists an additional four species :- *L. ivanovii, L. seeligeri, L. welshimeri* and *L. innocua* (all originally included in *L. monocytogenes*) also three species designated as *incertae sedis* : *L. grayi, L. murrayi* and *L. denitrificans.* The last has recently been allocated to a new genus as *Jonesia denitrificans* (Rocourt *et al.*, 1987*a*).

L. monocytogenes is pathogenic for a wide range of animal species including man. All disease caused by *L. monocytogenes* is known as Listeriosis even though the symptoms include meningoencephalitis, septicaemia, abortion, still birth or neonatal death.

In man *Listeria* is an opportunistic pathogen attacking the young, old, and those who are immunocompromised. *L. monocytogenes* is a low grade intracellular pathogen and has therefore been used in many studies on experimental infections in animals and has contributed much to an understanding of the mechanism of the cell-mediated immune system (Mackaness, 1971).

Stuart and Welshimer (1973, 1974) first used DNA-DNA homologies to

further characterise *Listeria*. They noted two DNA homology groups among the *L. monocytogenes* cluster but could not separate them on any other criteria. They found *L. grayi* and *L. murrayi* formed an homogenous group distinct from *L. monocytogenes* and suggested a new genus '*Murraya*' be formed for *L. grayi* and *L. murrayi* strains; this was not widely recognised and was not included in the Approved Lists of Skerman *et al.* (1980). Rocourt *et al.* (1982) later confirmed the low percent homology between *L. grayi, L. murrayi* and the other five species; however Rocourt, Wehmeyer and Stackebrandt (1987*b*) provided genomic evidence to retain them within the genus *Listeria*. *L. grayi* and *L. murrayi* possess the same cell wall, menaquinone and fatty acid composition as other *Listeria* but they have a slightly higher % (G+C) composition and can be separated by electrophoresis of whole cell proteins, a few biochemical reactions (Seeliger and Jones, 1986), the nature of substitutions of lipotechoic acids (Ruhland and Fiedler, 1987) and antigenic patterns (Seeliger and Jones, 1986). Using numerical taxonomy *L. murrayi* and *L.grayi* are closely grouped with *L. monocytogenes* strains at similarity values of 81-87 % (Stuart and Pease, 1972; Stuart and Welshimer, 1974; Jones, 1975; Wilkinson and Jones, 1977). Strains of each species are clustered together at rather higher similarity values of around 90 %.

Listeria monocytogenes sensu lato is divided into 17 serovars (Seeliger and Hohne, 1979), serovar 5 is strongly haemolytic and Ivanov named this serovar '*Listeria bulgarica*', this is now known as *Listeria ivanovii*. Strains of serovars 6, 6a, 6b and undesignated serovars (Seeliger and Hohne, 1979) which are non-haemolytic and non-pathogenic for adult mice (Audurier *et al.*, 1980) were named *Listeria innocua* (Seeliger, 1981). The study of DNA relationships by Rocourt *et al.* (1982) separated *L. monocytogenes* (*L. monocytogenes sensu lato*) into the five closely related species recognised in Bergey's Manualof Systematic

Bacteriology (*L. monocytogenes, L. innocua, L. welshimeri, L.seeligeri, L. ivanovii*). However, only six reference strains were used and the intra-species range of pairing values between strains of *L. monocytogenes* was considerable, over 35 %. The DNA-pairing values thus suggest a much greater variation between species than the phenotypic characteristics, perhaps because a large proportion of the genome is not expressed phenotypically. The large variation in pairing values within the species *L. monocytogenes* raises the question of whether there may be two (or more) clusters within the species, alternatively the strains form a spectrum in terms of their relationships rather than distinct groups and this second possibility would be supported by the small biochemical differences.

Rocourt *et al.* (1983) showed differentiation of the five species by a small number of biochemical tests (Table 1.1). The five species can be separated by relatively few phenotypic tests including haemolysis, CAMP tests with *Staphylococcus aureus* and *Rhodococcus equi* and acid production from D-xylose, L-rhamnose and α-methyl-D-mannoside. *L. monocytogenes* is a short small motile rod with rounded ends in smooth cultures and more elongated in rough cultures. The degree of haemolysis varies between strains and with the species of blood used. *L. innocua* is non-haemolytic; this is the only consistent phenotypic difference separating this species from *L. monocytogenes* strains but Rocourt *et al.* (1982) found that different phage patterns can also distinguish between the two species. The haemolysin is thought not to be the only cause of the pathogenicity because *L. seeligeri* is weakly haemolytic but non-pathogenic. However, one *L. seeligeri* strain has been isolated from a case of meningitis (Rocourt *et al.*, 1986).

L. ivanovii comprises virulent animal strains. Ivanov (1957) first

described this species, isolated from aborted lambs. *L. ivanovii* has more fastidious growth requirments and ferments fewer carbohydrates than other *Listeria* species. The main difference from *L. monocytogenes* is its marked degree of haemolysis; it produces a large haemolytic zone, and a distinctive pattern of multiple zones of haemolysis on bovine, sheep and human blood agar plates.

L. murrayi and *L. grayi* were shown to be very closely related on the basis of numerical taxonomic and serological studies (Welshimeri and Meredith, 1971; Stuart and Pease, 1972; Wilkinson and Jones, 1975, 1977; Jones, 1975, 1986).

Listeria are common in the plant-soil environment, decaying moist vegetation, animal and bird faeces. Rocourt and Seeliger (1985) found all species of *Listeria* except *L. murrayi* to be carried in the intestinal tract of healthy animals, whilst humans were found to carry only *L. monocytogenes, L. ivanovii* and *L. innocua* strains. The majority of *L. welshimeri* strains were isolated in the USA, *L. seeligeri* from Europe; *L. innocua* and *L. ivanovii* from both continents, whereas *L. monocytogenes* has global distribution.

Over the last 25-30 years there has been a vast increase in the amount of pre-packed foods kept at low temperatures; this provides an ideal environment for listeriae as they are able to multiply at low temperatures (Gray and Killinger, 1966) and anaerobic conditions. *Listeria* species also tolerate common preserving agents such as NaCl and sodium nitrite (Shahamat *et al.*, 1980*a*, 1980*b*). It is however possible that the increase in the number of reported cases of listeriosis in man over the last 25-30 years may be due to the development of improved media and techniques for isolating and identifying *Listeria* and this has increased the awareness of their presence in foodstuffs (Gray *et al.*

1948; Gray, 1957; Kramer and Jones, 1969; Ralovich *et al.* 1971; Khan *et al.*, 1973; Durst and Benersci, 1975; Gronstol and Aspoy, 1977). In a review of the genus *Listeria*, McLauchlin suggests there has been a real increase in the numbers of cases in both humans and animals (McLauchlin, 1987; Anon 1983, 1986; Gitter, 1986). Fenlon (1985) suggested the increase in animals may be associated with changes in agricultural practices especially in silage production.

Table 1.1 Biochemical Differences between the species of *Listeria*.

	Species:						
	L. monocytogenes	*L. innocua*	*L. seeligeri*	*L. welshimeri*	*L. ivanovii*	*L. grayi*	*L. murrayi*
β Haemolysis	+[1]	-	+	-	+[2]	-	-
CAMP test *Rhodococcus equi*	-	-	-	-	+	-	-
CAMP test *Staphylococcus aureus*	+	-	+	-	-	-	-
$NO_3 \rightarrow NO_2$ Reduction	-	-	-	-	-	-	+
Acid Production from:							
L-Rhamnose	+	v	-	v	-	-	v
D-Xylose	-	-	+	+	+	+	+
Mannitol	-	-	-	-	-	+	+
Pathogenicity for mice	+	-	-	-	+	-	-
% (G+C)[3]	37-39	36-38	36	36	37-38	41-42	41-42.5

[1] Not all strains of *L. monocytogenes* show β-haemolysis.

[2] A very wide or multiple zone is shown by *L. ivanovii* strains.

[3] Determined by the T_m method (see Bergey's Manual of Systematic Bacteriology, 1986).

Aims of the Study.

This study set out to assess the suitability of DNA-DNA pairing experiments for systematic studies and to analyse the extent of error in these techniques. If DNA pairing data is to be used as a taxonomic tool, then the differences in the techniques must be analysed and accounted for. Methods of estimating different types of error were looked at as well as possible ways of reducing these errors or at least ways of accounting for them in any resulting taxonomic structure. It was also hoped to learn more of the effects of choice of reference strains.

The genus *Listeria* was chosen to illustrate and measure error as well as to assess the optical method of determining DNA homology. The genus *Listeria* was chosen for several reasons. A DNA-DNA pairing study had already been carried out (Rocourt *et al.*, 1982) using a technique other than the optical technique and this study established some of the existing taxonomy of *Listeria*. Also, several of the species of *Listeria* are very closely related, separated by very few biochemical tests (Rocourt *et al.*, 1983). The *Listeria* DNA-DNA pairing study used only a few reference strains and possible distortion due to this was assessed. The effects of choice of reference strains for DNA work needs to be carefully assessed, and distortion on taxonomic structure was evaluated using complete DNA pairing matrices.

The optical technique was chosen because it is a quick, cheap method once a spectrophotometer is available.

2. METHODS AND MATERIALS

2.1 Bacterial Strains

The strains of *Listeria* used in this study are listed in Table 2.1. The collection numbers and source of original isolation are listed where known.

Maintenance of Cultures.

Cultures were maintained on Blood Agar Base Number 2 (BAB2, Difco) and stored at 4°C after overnight incubation at 35°C. Strains were subcultured every 10-14 days.

The composition of Blood Agar Base No. 2 was as follows:-

	Grams per litre
Proteose peptone	15.0
Liver digest	2.5
Bacto-Yeast extract	5.0
Sodium chloride	5.0
Bacto-Agar	12.0

The medium was sterilised by autoclaving at 121°C for 15 minutes, cooled to 55°C and dispensed into Petri dishes.

2.2 Growth of *Listeria* for DNA Isolation.

Listeria strains were inoculated into two-litre conical flasks containing 500 ml of Tryptone Soya broth (Oxoid). Composition (grams per litre) was as follows:

Pancreatic digest of casein	17.0

Table 2.1 : Strains used in DNA:DNA Homology Study.

Species	Culture Collection Numbers	Source
Listeria monocytogenes	C52, NCTC 7973 serotype 1a (type strain)	Guinea-pig mesenteric lymph node
" " "	C200, NCTC 5348 serotype 2	Cerebrospinal meningitis
" " "	C201, NCTC 10357 serotype 1a	Rabbit
" " "	C202, NCTC 5105 serotype 3	Human
" " "	C203, NCTC 4885 serotype 4b	Infant meningitis
" " "	C228 serotype 4a	-
" " "	C231, NCTC 4883 serotype 4c	Fowl myocardial disease
Listeria innocua	C644, SLCC 3379 serotype 6a (type strain)	-
" "	C645, SLCC 3479 serotype 6b	-
" "	JS21	Raw chicken
" "	JS31	Raw chicken
Listeria welshimeri	C1091 (type strain)	-
" "	C1172	-
Listeria seeligeri	C1090, NCTC 11856 (type strain)	-
" "	C1171	-
Listeria ivanovii	C1087, ATCC 19119 (type strain)	-
" "	C663, Pritchard L72	-
" "	C666, Pritchard L234	Foetal-abomasum ovine abortion
" "	C667, Pritchard L102B	-
" "	C659, Pritchard L242	Placenta, ovine abortion
Listeria grayi	C214 a/b, SLCC 332/64 (type strain)	Chinchilla faeces
" "	SLCC 7211	-
Listeria murrayi	C1174, NCTC 10812 (type strain)	-

NCTC : National Collection of Type Cultures;

ATCC : American Type Culture Collection;

SLCC : Special *Listeria* Culture Collection, Seeliger, University of Wurzburg and Pasteur Institute, Paris

JS : Isolated by J. Stephens, Leicester University

C : Leicester University Culture Collection

2.3 Preparation of DNA.

Isolation of DNA from *Listeria*

Cells were harvested by centrifuging at 15,000 g for 10 minutes. The resulting pellet was washed by resuspending and recentrifuging once in distilled water, to remove culture medium, and once in 1 M sodium chloride, to remove extracellular polysaccharides. Cells were resuspended in 20 ml sucrose-tris buffer (0.25 M sucrose, 0.05 M Tris-Cl pH 8.0). The cells were either frozen (-20°C) or disrupted immediately as follows.

A variation of the method of Marmur (1961) was employed for the extraction and purification of the DNA. 2 ml of freshly prepared lysozyme solution (10 $mgml^{-1}$ Lysozyme chloride (Sigma) in sucrose-tris buffer) was added and the suspension incubated at 37°C for an hour. Following incubation 8 ml of Tris-EDTA buffer (0.05 M Tris-Cl, 0.05 M EDTA pH 7.5) and 200 μl proteinase k (BDH) (10 $mgml^{-1}$ made up in 0.15 M NaCl and allowed to digest at 37°C for 30 minutes) were added followed by 2 ml of pre-warmed 20 % w/v sodium dodecyl sulphate in tris-EDTA buffer. Incubation was continued for one hour or until the solution became clear and viscous, showing lysis was complete.

To remove cell debris and dissociate protein from the nucleic acids 1/3 volume freshly prepared sodium perchlorate (66.5 % w/v in tris-EDTA buffer) was added and the solution mixed gently.

Protein was extracted by the addition of an equal volume of phenol:chloroform:isoamyl alcohol mixture (25:24:1 v/v) (see 2.1] Reagents section). The mixture was gently shaken at room temperature until a stable emulsion formed, then it was dispensed into centrifuge tubes. The emulsion was separated into three layers by centrifuging at 7500 g for 30 minutes.

The upper viscous layer, containing the nucleic acids was carefully

removed to clean tubes taking care not to disturb the white layer of protein at the interface. An equal volume of chloroform : isoamyl alcohol (24:1 v/v) was added to remove traces of phenol and further protein. An emulsion was formed and the mixture centrifuged at 7500 g for 15 minutes. The aqueous upper layer was removed to clean tubes and a second chloroform : isoamyl alcohol extraction carried out. The resulting upper phase was carefully pipetted to a sterile 250 ml beaker at 4°C. The upper phase volume was measured and 1/10 volume of 3 M sodium acetate (in 0.01 M EDTA pH 8.0) was added. The nucleic acids were precipitated by carefully overlaying with three volumes of ice-cold absolute ethanol (-20°C). Using a glass rod bent into a large loop shape, the layers were gently mixed; the nucleic acids spooled onto the glass rod as a white thread-like precipitate. The spooled precipitate was pressed onto the side of the beaker to remove excess alcohol; it was briefly air-dried before dissolving overnight at 4°C in 10 ml of 0.1 x SSC buffer (see 2.11 Reagents section).

Heat-treated RNase solution (20 $mgml^{-1}$ ribonuclease A (Sigma) in 0.15 M NaCl, pH 5.0 heat treated at 80°C for 15 minutes to destroy any DNase activity) was added to give a final concentration of 60 μgml^{-1} and the solution incubated at 37°C for an hour. The incubation was continued for a further two hours after the addition of 0.2 ml 10 x SSC (Reagents section) and 100 μl proteinase k (BDH) (10 $mgml^{-1}$ made up in 0.15 M NaCl and self-digested at 37°C for one hour).

An equal volume of chloroform : isoamyl alcohol (24:1 v/v) was added, the layers were shaken gently for five minutes to obtain an emulsion before centrifuging at 7500 g for 15 minutes.

The upper layer was repeatedly washed with chloroform : isoamyl alcohol until no protein was visible at the interface after centrifugation.

The DNA was selectively precipitated with isopropanol by adding 1/10 volume of 3 M sodium acetate in 0.1 M EDTA pH 8.0 followed by 0.55 volume isopropanol added slowly while spooling with a Pasteur pipette. This leaves RNA in solution and in some cases separates polysaccharides from DNA. The DNA was dissolved in a small volume of 0.1 x SSC then reprecipitated with two volumes of ice-cold absolute ethanol (-20°C) as before. The DNA was washed in 70 % ethanol (-20°C) for 10 minutes to allow excess salts to diffuse away from the DNA. The DNA was briefly air-dried and dissolved in a small amount of 0.1 x SSC.

2.4 Fragment Size.

Shearing

Two methods of shearing were investigated. DNA was isolated from *Listeria* species as above. Samples from the same preparation were sheared by one of the two following methods.

i) Using a French press.

The sample was diluted in 0.1 x SSC to a concentration of 250-300 $\mu g ml^{-1}$ DNA. The press was chilled to 4°C for at least two hours before use. Sterile distilled water (4°C) was passed through the press by hand. The DNA solution was passed through the press twice at ten tonnes pressure at a rate of approximately 3 $ml min^{-1}$.

ii) Using a 27 gauge syringe

The DNA solution was diluted to a concentration of approximately 250-300 $\mu g ml^{-1}$ in 0.1 x SSC buffer. The solution was passed repeatedly through a 27 gauge needle on a 1 ml syringe (at least 20 times).

Measurement of Fragment Size.

To determine the approximate range of molecular weight of the DNA prepared and sheared as described above, each sample was run on an agarose gel, using the following method.

75 ml of electrophoresis buffer was added to 0.75 g of Litex Agarose (Miles Laboratories) and heated in a microwave for 5 minutes until totally dissolved. After slight cooling 7.5 μl ethidium bromide (5 $mgml^{-1}$) was added and the gel was poured onto an electrophoresis plate edged with tape. At one end of the plate a plastic comb was balanced about 1 mm above the plate to form wells in the gel. On solidification the comb was removed.

A tank was filled with sufficient elecrophoresis buffer to cover the gel and the plate was positioned in the tank.

For each well to be used an eppendorf was put on ice and 3 μl of gel loading buffer was dispensed into each one. 5-15 μl of sample was added to the appropriate eppendorf; the sample and buffer were mixed well before loading into the relevant wells on the gel using a Gilson pipette.

As a control each gel was loaded with 5 μl of 1 kb ladder, digested at 65°C for five minutes (Bethesda Research Laboratories) and mixed with 3 μl loading buffer. The powerpack was in constant current mode and the gel was ran at 80-100 volts for three hours.

The gel was viewed using an ultraviolet transilluminator. By comparing the distance travelled along the gel by the sample DNA with that travelled by the control bands of the 1 kb ladder, an approximation of the range of fragment size of the sample could be determined.

2.5 Dialysis.

Suitable lengths of tubing were cut and boiled for 10 minutes in a large volume of 2 % Sodium bicarbonate - 1 mM EDTA. The tubing was rinsed well in distilled water then boiled for ten minutes, or autoclaved, in 1 mM EDTA. The tubing was stored submerged at 4°C.

After shearing the DNA samples were dialysed overnight at 4°C in 500 volumes of 2 x SSC buffer.

2.6 Estimation of Concentration and Purity of DNA Solutions by Spectroscopy.

30 μl of sample was dispensed into a semi-micro quartz cuvette followed by 570 μl of 0.1 x SSC buffer (2.11 Reagents section) and the solutions mixed well. Using a PU8700 model spectrophotometer (Philips Scientific) a baseline was set by scanning from 200-300 nm using 0.1 x SSC. The sample was scanned over the same wavelength range and the absorbance noted at 230 nm, 260 nm, and 280 nm. The concentration was estimated by assuming that an absorbance of 1.0 at 260 nm was approximately equivalent to 50 $\mu g ml^{-1}$ of DNA (Wetmur, 1976).

The equations of Kalb and Bernlohr (1977) were used to estimate protein contamination and more accurately to determine the nucleic acid concentration:

$$49.1 \times O.D._{260nm} - 3.48 \times O.D._{230nm} = \mu g\ ml^{-1} \text{ nucleic acids}$$

$$183 \times O.D._{230nm} - 75.8 \times O.D._{260nm} = \mu g\ ml^{-1} \text{ protein}$$

The ratio of absorbances at 230 nm : 260 nm and 280 nm : 260 nm were also used as a purity check. The ratios of pure DNA should be 0.45 and 0.52 respectively (Marmur, 1961).

2.7 Determination of Percent Homology.

i) Preparation of Samples.

DNA was prepared, purified and the purity of the samples were checked by spectroscopy as above. If the ratios were not $\leqslant 0.45$ and $\leqslant 0.55$ respectively the DNA solution was subjected to further deproteinisation with chloroform : *iso*amylalcohol and ethanol precipitation (see 2.3 DNA Isolation) until the requirements were satisfied.

The purified DNA samples were diluted to a concentration of 200-300 $\mu g ml^{-1}$ using 2 x SSC and sheared by passing through the French Press as described above (2.4). All samples were dialysed overnight as described.

The DNA samples were diluted just prior to use with 2 x SSC to give an absorbance of 1.55 ± 0.05 at 260 nm, i.e. a concentration of 75-80 $\mu g ml^{-1}$ DNA.

ii) T_m and T_{or} Determination

A set of eight matched semi-micro quartz cuvettes with false bottoms equipped with Teflon stoppers were used in all experiments. The cuvettes were cleaned by soaking in mild detergent followed by rinses in distilled water. Cuvettes were drained and dried prior to usage. The maximum volume of the cells was 1000 μl and the minimum 550 μl. A volume of 700 μl was used for homology experiments both to allow room for thermal expansion and so the miniscus was safely above the beam. When volumes nearer to 1000 μl were used the stoppers became wet and tended to pop out at high temperatures, over 80°C, and evaporation occurred.

A PU8700 spectrophotometer fitted with a cell programmer and Peltier thermal block was programmed as in Figure 2.1. There are eight cell positions in the cell holder of the PU8700; although the carriage is

supplied with a cover there is still temperature variation over these eight cell positions (Results 3.1). The temperature probe was placed in position four for the renaturation experiments.

Three clean dry cuvettes were required per homology determination : i) cuvette one contained 700 μl of DNA from strain *a*, ii) cuvette two contained 700 μl of DNA from strain *b*, iii) the third cuvette contained 350 μl of DNA from strain *a* and 350 μl of DNA from strain *b*. The cuvettes were stoppered mixed and checked for air-bubbles. As a control, 700 μl of guanine solution (in 2 x SSC buffer) with an absorbance of 2.0 ± 0.1 was placed in a fourth cuvette. The temperature probe cell contained 2 x SSC.

The samples were allowed to equilibrate at 60°C for 5-10 minutes. The temperature was ramped up at a rate of 1°C per minute while the absorbance of each sample was monitored. When the hydrogen bonds of the DNA began to dissociate due to the increase in temperature the absorbance increased with the increase in the proportion of single stranded DNA. When the DNA had all dissociated into the single strand form the absorbance became constant. DNA from *Listeria* becomes completely single-stranded at a temperature of 90-96°C in 2 x SSC depending on the strain. The temperature was held constant for five minutes to ensure complete dissociation before changing to the renaturation program.

The T_m was taken to be the point at which half of the double-stranded DNA has melted to the single-stranded form. An increase in absorbance from 60-80°C indicated the presence of contaminating RNA or single stranded DNA, these samples were discarded.

The optimal renaturation temperature (T_{or}) was taken to be 25°C below the T_m in the same buffer system (2 x SSC).

iii) Determination of Renaturation Rates.

If the T_m and T_{or} values of the samples had been accurately

Figure 2.1 Spectrophotometer Settings for T_m Determinations.

PU8700 Series UV/Vis Spectrophotometer

I.D. 0 0 TM DETERMINATION
Mode ABSORBANCE — High 2.250 ABS
Wavelength 260.0 — Low 1.500 ABS
Bandwidth 2.0 nm — No Of Cycles 1
Smoothing MEDIUM — Cycle Time CONTINUOUS
Integration 0:02
Low Limit OFF — High Limit OFF

Plotter Mode DATA — Line Format AUTO
Scan Abscissa Scale 10.0 — Fixed Abscissa Scale 60.0
Sample Identification ON — Grid ON
Annotated Scales ON — Analytical Conditions ON

Cell Status	OFF	ON	ON	OFF	ON	ON	OFF	OFF
Sample/Reference	S	S	S	S	S	S	S	S

Mode NORMAL No of Cycles 1 Cycle Time CONT

Temperature Profile T1-T2
Start Temperature T1(°C) 60.0 — Stop Temperature T2(°C) 102.0
Ramp Rate 1.0 °C/min — Hold Time 2:00 mins

Figure 2.2 Spectrophotometer Settings for the Reassociation of single stranded DNA.

PU8700 Series UV/Vis Spectrophotometer

I.D. 0 0 RENATURATION
Mode ABSORBANCE — Range AUTO ABS
Wavelength 260.0 — Slope NEGATIVE
Bandwidth 2.0 nm — Display SEQUENTIAL
Smoothing MEDIUM
Delay 0:00 — Integration 0:02
Factor 1.000

Plotter Mode GRAPHICS AND DATA — Line Format AUTO
Scan Abscissa Scale 10.0 — Fixed Abscissa Scale 60.0
Sample Identification ON — Grid ON
Annotated Scales ON — Analytical Conditions ON

Cell Status	OFF	ON	ON	OFF	ON	ON	OFF	OFF
Sample/Reference	S	S	S	S	S	S	S	S

Interim results PLOTTED Mode NORMAL No of Cycles 30 Cycle Time 0:45

Temperature Profile FIXED
Start Temperature T1(°C) 67.0
Temperature Output OFF

determined previously then samples of the DNA were denatured in the cuvettes by rapidly raising the cell block temperature to 98.5°C then holding for 5 minutes to ensure complete denaturation.

The temperature was dropped rapidly by setting the temperature programmer to 0°C. When the cell temperature registered T_{or} + 10°C the programmer was set to the desired renaturation temperature; the whole cooling process takes 2.5 - 3 minutes. The renaturation program is listed in Figure 2.2. When the T_{or} was reached the change in absorbance was followed for 20-30 minutes.

The rate of renaturation was determined from the straight line portion of the graph. The interim values (absorbance values read every 45-60 seconds) were fed into the MINITAB package on the Leicester University Vax cluster mainframe and a linear regression analysis carried out. Using this package the change in guanine absorance, if any, could easily be subtracted from the change in sample absorbance to detect any alteration to the rate of renaturation.

After renaturation, if no visible evaporation had occurred the samples were remelted as in the T_m determination, starting from the T_{or}. An increase in T_m indicated evaporation had occurred; a decrease in T_m in an homologous sample, indicated contamination with particulate matter and the data from this sample was discarded. In hybrid samples a change in T_m could also be indicative of mismatching.

iv) Calculation of Percent Homology

The percent homology was determined by the following equation (DeLey et al. 1970):

$$\text{Percent Homology} = \left[\frac{4V_M - (V_A + V_B)}{2\sqrt{(V_A \times V_B)}} \right] \times 100$$

Where V_M = rate of renaturation of the mixture in Δ absorbance $minute^{-1}$

V_A, V_B = rate of renaturation of DNA from strains *a* and *b* respectively (Δ absorbance $minute^{-1}$)

2.8 Base Composition Determination.

Measurement of Temperature Variation in the PU8700

There are eight cell positions in the cell holder of the PU8700 spectrophotometer (Philips Scientific). Although the carriage is supplied with a cover, there is still temperature variation over these eight cell positions. The temperature probe was placed in position 4 or 5 for the renaturation experiments. To detect the degree of variation over the holder, an *Escherichia coli* DNA preparation (Sigma) of known T_m was dissolved in 1 x SSC and dialysed overnight. The DNA concentration was adjusted to 50 $\mu g ml^{-1}$, or an absorbance of 1.0 ± 0.05 at 260 nm. 700 μl of DNA sample was aliquoted into seven matched quartz cuvettes. The T_m of each cell was determined as above and the variation of temperature across the cell holder estimated.

Determination of Base Composition of *Listeria* species

Pure unsheared DNA was dissolved in 0.1 x SSC. The samples were dialysed against 500 volumes of 0.1 x SSC pH 7.0 for 24 hours with one change of buffer. *E.coli* strain B DNA (Sigma) of known % (G+C) was included as a control in each determination. The concentration of each sample was adjusted to 50μg ml^{-1}, i.e. an absorbance of 1.0 ± 0.05 at 260 nm.

700 μl aliquots of the samples were placed in clean dry quartz

cuvettes equipped with Teflon stoppers. The PU8700 spectrophotometer was programmed as in Figure 2.1. The probe cell was filled with 0.1 x SSC buffer. Up to seven samples, including the *E. coli* standard, could analysed per run. The samples were placed in the cell holder and allowed to equilibrate at 45°C for 5 - 10 minutes. The T_m was determined as in the homology determination experiments except that the temperature was ramped up at a rate of 0.5°C per minute. The absorbance was monitored until all DNA was in the single stranded form and the T_m calculated as before. The determinations were carried out in triplicate.

The % (G+C) was estimated from the following equation:-

$$\% (G+C) = \% (G+C)_{std} + 2.08 (T_m - T_{mstd})$$

Where $_{std}$ = values of the *E. coli* standard.

2.9 Polysaccharide Removal

To determine whether polysaccharide affected the degree of binding between *Listeria* DNAs samples of DNA were prepared as above; half of the preparation was subjected to one of the following treatments.

i) Precipitation using CTAB

DNA was ethanol precipitated, as in the DNA isolation procedure, then dissolved in 0.4 M NaCl to a concentration of 0.5 $mgml^{-1}$. For every 5 ml of DNA solution 2 ml of 5 % Cetyl-trimethyl-ammonium bromide (CTAB) in 0.4 M NaCl was added. The mixture was left at room temperature for 15 minutes to allow the CTAB-DNA to precipitate. A few drops of CTAB solution was added to ensure complete precipitation. The DNA was spooled

onto a Pasteur pipette or spun down in a microfuge then washed twice 0.4 M NaCl.

The DNA was dissolved in 1 M NaCl and washed with an equal volume of chloroform : isoamyl alcohol (24:1 v/v). After centrifuging in a microfuge for five minutes the aqueous upper layer was removed to a clean tube. The DNA was precipitated with two volumes of ethanol as in the DNA isolation procedure and resuspended in a small volume of 0.1 x SSC before dialysing overnight in 0.1 x SSC.

ii) Using 2-Methoxyethanol (Bellamy and Ralph, 1968).

DNA was precipitated with two volumes of ethanol, as in the isolation procedure, then dissolved in a small amount of 0.1 x TNE buffer (see Reagents) and an equal volume of 2.5 M potassium phosphate buffer (see Reagents) added. After mixing, the same volume of 2-methoxyethanol was added and the phases mixed. The layers were separated in a microfuge on high speed for two minutes. The viscous upper layer containing the DNA was carefully removed. The oily lower layer was reextracted by adding one volume of sterile distilled water, one volume of 2.5 M potassium phosphate buffer and one volume of 2-methoxyethanol, mixing after each addition. The phases were separated as before. The pooled upper phases were dialysed against 0.1 x TNE buffer at 4°C overnight.

Detection of Carbohydrate (Umbreit and Burris, 1964)

Anthrone reacts with all carbohydrates to give a characteristic blue colour; however the colour yield is not the same for different carbohydrates. DNA samples which had undergone one of the polysaccharide removal treatments were compared with samples from the same DNA preparation which were 'untreated'. Glucose standards were used for comparison in each experiment.

20 ml of concentrated sulphuric acid was cautiously added to 1 ml of

distilled water. After cooling 40 mg of anthrone was added and allowed to dissolve. Capped test tubes were aliquoted with 0.5ml volumes of DNA samples of known concentration. Glucose standards ranging from 20 μg to 100 $\mu g ml^{-1}$ were aliquoted into capped test tubes in duplicate in 0.5 ml amounts. 1 ml of the anthrone reagent was added to each tube then mixed thoroughly by swirling. The tubes were put in a boiling water bath for 3 minutes then cooled. The optical density at 620 nm was recorded for each sample. The amount of carbohydrate per μg of DNA was determined.

2.10 Storage of DNA Samples

DNA samples were stored for short periods at 4°C over a drop of chloroform. Over longer periods DNA in 0.1 x SSC was stored at 4°C at a concentration of at least 300 $\mu g\ ml^{-1}$. Before use in hybridisation experiments, the DNA was precipitated with two volumes of ethanol (see DNA isolation procedure) and resuspended in 0.1 x SSC.

Sheared DNA was stored at 4°C and used within 7 days.

All glassware and solutions were nuclease free as far as possible.

2.11 Reagents.

Stock solutions

i) 0.5M EDTA

186.12g of EDTA, disodium salt, was dissolved in 750 ml of distilled waterby raising the pH to 8.0 with sodium hydroxide while stirring. When the salt had dissolved the volume was made up to 1 litre with distilled water.

ii) 0.5M Tris-Cl

60.55g Trizma base was dissolved in 750 ml of distilled water and the pH adjusted to 7.5 with concentrated HCl then the volume made up to 1 litre.

Standard Saline Citrate (SSC)

	0.1 x SSC	2 x SSC	10 x SSC
Sodium chloride	0.015M	0.3M	1.5M
tri-Sodium citrate	0.0015M	0.03M	0.15M

The solution was made very accurately in a volumetric flask. Solid ingredients were dissolved in 800 ml distilled water and the pH adjusted to pH 7.0 ± 0.05 with 2 or 3 drops of 1 N HCl. The volume was made up to one litre and sterilised by autoclaving.

Phenol - chloroform mixture.

1 kg Phenol (AR, Fisons) was dissolved in 100 ml of distilled water over 3-4 hours. 200 ml NaCl-EDTA buffer (0.1 M NaCl, 0.01 M EDTA pH 8.0) was added and the mixture allowed to equilibrate.

Chloroform and *iso*amyl alcohol were mixed 24:1 v/v and 0.1 % 8-hydroxyquinoline (BDH) added to help elimination of harmful peroxides and prevent them building up during storage. This was mixed with an equal

volume of the lower layer of phenol-water and allowed to equilibrate before use.

Tris-acetate buffer (TAE)

48.4g Tris base

20 ml 0.5 M EDTA pH 8.0

The Tris base was dissolved in 800 ml of distilled water and the EDTA added. The pH was corrected to pH 7.7 by the addition of glacial acetic acid. The volume was then made up to 1 litre with distilled water.

Agarose gel

0.75g Litex agarose

75 ml 1 x TAE buffer

Ethidium bromide (10 $mgml^{-1}$, stored in a light-proof bottle at 4°C)

The agarose was melted in the buffer in a microwave oven for 5 minutes. The agarose was allowed to cool to 50°C and ethidium bromide was added to a final concentration of 5 μgml^{-1}.

Gel Loading Buffer

0.25 % Bromophenol blue

0.25 % Xylene cyanol

30 % Glycerol

in distilled water

10 $mgml^{-1}$ Ribonuclease A (Sigma) in 0.15 M NaCl, pH 5.0 (heat treated at 80°C for 15 minutes to destroy any DNase activity).

The ingredients were mixed and 1/15 volume of ribonuclease solution was added. The buffer was stored at 4°C.

Table 2.2 List of Suppliers

*Iso*amyl alcohol	BDH
Cetyl trimethyl ammonium bromide	BDH
Chloroform	BDH
1 kb DNA Ladder	Bethesda Research Laboratories
(EDTA) sodium salt	Fisons
Ethidium bromide	Sigma
Glucose	Fisons
Hydrochloric acid	Fisons
8-hydroxyquinoline	BDH
Lysozyme	Sigma
2-Methoxyethanol	Fisons
Phenol	Fisons
Potassium dihydrogn orthophosphate	BDH
diPotassium hydrogen orthophosphate	BDH
Proteinase k (fungal)	BDH
Ribonuclease, bovine pancreatic	Sigma
Sodium acetate	BDH
triSodium citrate	BDH
Sodium chloride	BDH
Sodium dodecyl sulphate (SDS)	BDH
Sodium hydrogen carbonate	Fisons
Sodium perchlorate (AnalaR)	BDH
Sucrose	Fisons
Trizma base (reagent grade)	Sigma

2.12 Distortion of Taxonomic Structure due to Choice of Reference Strains.

The symbolism in Sneath (1983) was used throughout the study. It was assumed that a complete matrix of DNA-DNA values was available for t strains, and that any replicates have been averaged to give a single pairing value between a pair of strains j and k. It was further assumed that for those methods in which values of j versus k may be different from values of k versus j such reciprocal pairs have been averaged. This allows construction of a symmetrical t x t matrix (which is more convenient here than the usual lower triangle matrix). The values were then represented as distance between strains, $d_{j,k}$, with appropriate transformation if required. Values in the principal diagonal are set to zero. Then c reference strains are chosen and only the tc values representing the c columns are retained. The remaining values are treated as unknown.

The derived matrices were obtained either by principal component analysis of the DNA-DNA pairing values or by a single iteration of formula (1) in Sneath (1983). It is shown below that the two methods are algebraically identical when principal components are obtained in one particular way, Principal Coordinate Analysis of (Gower, 1966), and this way was employed. The formula for derived distances is

$$d_{jk}^{*} = \left[\frac{1}{c} \sum_{r=1}^{r=c} (d_{rj} - d_{rk})^2 \right]^{1/2}$$

where r is a refrence strain, but in this study the summed squares were not divided by c as shown above, so as to retain the algebraic relations with principal components. The only effect, however, is to introduce a constant scaling factor of $1/\sqrt{c}$ that affects all relationships alike.

The taxonomic structure was represented in two ways. The first is the three-dimensional ordination from the first three principal axes of

principal component analysis. The second is a dendrogram from cluster analysis.

2.13 Cluster Analysis

Clustering of strains into groups occurs as a stepwise process. Step one is to find the pair of strains with the highest similarity. This pair are then joined to form a cluster and the similarities between this group and all the other strains, or OTUs is determined. The highest similarity is again sought; this may be between a single pair of strains or a single strain and a group, or two groups. The process continues in this cycle until all strains have joined to form one group, i.e. after $t-1$ cycles. The clustering method varies in the way that the similarity between a group or two groups is defined. The two methods used here were:

a) Unweighted Average Linkage (UPGMA - unweighted pair group method with arithmetic averages) (Sokal and Michener, 1958; Sneath and Sokal 1973, p.230) which defines the similarity between the two groups as the arithmetic average of the similarities across the two groups, each similarity having an equal weight.

b) Single Linkage (Sneath, 1957) which defines the similarity between two groups as the similarity of the two most similar strains or OTUs, one from each group.

The clustering results are presented in the form of a dendrogram. The similarity values shown by the diagram approximately represent the values of the original similarity matrix, or distance matrix from which it was derived. To ascertain the level of agreement between the dendrogram and the original matrix the cophenetic correlation coefficient may be determined. A cophenetic correlation value of 1 indicates perfect correlation; in practice coefficients of 0.7 or more are considered to be

satisfactory (Sneath, 1978).

Principal Component Analysis (Sneath and Sokal, 1973)

Principal component analysis (PCA) involves placement of t OTUs in a space of dimensionality varying form 1 to t-1. PCA involves computing eigenvalues and eigenvectors i.e. to solve the equation $(R - \lambda I)v = 0$ to obtain r non zero, positive, scalar quantities $\lambda_1, \lambda_2 \ldots \lambda_r$. Where $r \leqslant t$-1. There will be an equal number of associated eigenvectors. An eigenvalue is equal to the variance along its corresponding axis. The second axis accounts for the second largest amount of variance from the sample etc. Often as few as three principal axes will be responsible for most of the variance.

The ordination gives the most convenient visual representation of salient features. The dendrogram gives more reliable information, because it is based on the distances in the full space of c dimensions (not simply in the first three dimensions): it is, however, less easy to interpret by eye.

Taxonomic structure cannot be satisfactorily represented if the number of dimensions is reduced too much. A suitable measure of the effective dimensionality of the derived configurations is therefore needed. If points lie in a straight line the dimensionality is 1. This is true even if the points are embedded in a space of many dimensions. If they lie almost in a straight line, but show small displacements from it in numerous dimensions, the points cannot be represented exactly in one dimension. The effective dimensionality, n' however, is only a little greater than 1, and it may be, for example, 1.13.

The measure of n' is $1/\Sigma p_i^2$, where p_i is the proportion $\lambda_i / \Sigma\lambda_i$, where λ_i values are the non-negative eigenvalues from principal component

or principal coordinate analysis (Sneath, 1983). A simpler formula is for n' is $(\Sigma\lambda_i)^2 \ / \ \Sigma\lambda_i^2$. It is necessary to exclude negative eigenvalues because these represent "imaginary" or "non-euclidean" dimensions. Then n' cannot be more than the lesser number of characters n and t-1; it is maximal for a hyperspherical configuration.

When a model of lower dimensionality is prepared this removes some of the variation. The effective dimensionality is therefore calculated as m', where summation is only over the m non-negative eigenvalues of the m axes in the model.

Grimont and Popoff (1980) and Rocourt *et al.* (1982) have employed principal component analysis of DNA pairing values to obtain taxonomic structure from data on reference strains, whereas Sneath (1983) employed principal coordinate analysis (Gower, 1966) of euclidean distances, d^*_{jk}, between strains. The equivalence of principal coordinates with one form of principal components is illustrated in Table 2.3. Strains 1 and 3 are reference strains, with hypothetical DNA percent dissimilarity values as shown in Table 2.3a. It should be noted that in Table 3a reciprocal distances are not identical, and also that the triangle inequality does not hold for all cases. Thus the sum of distance between 1 and 2 and 1 and 3 is either 23 or 28, depending on whether 11 % or 17 % is chosen to represent the distance from 1 to 3. The distance from 2 to 3 is far greater than either 23 or 28 at 41, so the points 1, 2 and 3 cannot be represented as a triangle in euclidean space. However, such features are not uncommon in DNA data, and the analyses show that they can be accomodated by principal axis methods.

Distances between strains are shown in Table 2.3b. For example $d^*_{1,2} = \sqrt{[(0-12)^2 + (17 - 41)^2]} = 26.8328$. On analysing Table 2.3b by principal coordinates one obtains a new distance matrix scaled in the manner given by Gower (1966), and this matrix has four eigenvalues; λ_1 =

1147.66, λ_2 = 131.09, and the other two are zero. On scaling the eigenvectors of this new matrix so that the sum of squares of each column equals the corresponding eigenvalue, one obtains the coordinates in Table 2.3c. These coordinates represent a rigid rotation about the centroid of points representing the strains. Note, however, that the positive and negative ends of the axes are arbitary, because this information is lost when calculating interstrain distances. Thus the configuration may appear reflected about the centroid when compared with that from principal components (Table 2.3e).

If one performs principal component analysis on Table 2.3a using sums of squares and crossproducts, the same eigenvalues are obtained. Scaling the eigenvectors so that the sum of squares of each column is unity gives the principal component matrix Table 2.3d. This represents a rotation matrix such that if one centres the values of Table 2.3a by subtracting column means, and then matrix multiplies by Table 2.3d one obtains the coordinates in Table 2.3e. For example, strain 1 on axis 1 has the coordinate (0-10.25) x 0.1925 + (17-24) x0.9813 = 8.8419, and on axis 2 (0-10.25) x 0.9813 + (17-24) x -0.1925 = -0.8711. It can be seen that these coordinates are the same (within machine accuracy) as those in Table 2.3c, except for change of sign as mentioned above. It was this form of principal component analysis that was used in this thesis.

However, if other forms of principal components are used the resulting configurations can be very different (Hope, 1968). One common practice is to scale each principal axis so that its sum of square is unity. If this is done, the coordinates become those in Table 2.3f: the resulting plots or models show equal variance on each principal axis, and, for example, a mainly linear configuration can be turned into a mainly circular or spherical one.

Another variant of principal components employs correlations in

Table 2.3 Comparison of Principal Coordinate and Principal Component Analyses.

		Strains	Strains 1	2	3	4
Original Data	(a)	1	0		17	
		2	12		41	
		3	11		0	
		4	18		38	
		Mean	10.25		24	

		Strains	Strains 1	2	3	4
Distance Between Strains	(b)	1	0			
		2	26.8328	0		
		3	20.2485	41.0122	0	
		4	27.6586	6.7082	38.6394	0

		Strains	Axes 1	2	3	4
Principal Coordinates	(c)	1	8.8419	8.7111	0	0
		2	-17.0190	1.5347	0	0
		3	23.4069	5.3552	0	0
		4	-15.2298	-4.9105	0	0
		Sum	0	0	0	0
		Sum of Squares	1147.6583	131.0912	0	0

		Old Variates	New Variates 1	2
Principal components form sums of squares & products	(d)	1	.1925	.9813
		2	.9813	-.1925
		λ	1147.66	131.09

		Strains	Axes 1	2
Coordinates from components in (d) scaled to eigenvalues	(e)	1	-8.8419	-8.7111
		2	17.0190	-1.5547
		3	-23.4069	5.3552
		4	15.2298	4.9105
		Sum	0	0
		Sum of Squares	1147.6583	131.0912

Table 2.3 continued.

		Axes	
	Strains	1	2
Coordinates from (f)	1	-.2610	-.7608
(d) scaled to	2	.5024	-.1358
unity on each axis	3	-.6909	.4677
	4	.4496	.4289
	Sum	0	0
Sum of Squares		1.0	1.0

		New Variates	
	Old Variates	1	2
Principal (g)	1	.7071	-.7071
components from	2	.7071	.7071
correlations	λ	1.4436	.5564

		Axes	
	Strains	1	2
Coordinates (h)	1	-.4230	.1086
from (g)	2	.4598	.5097
scaled to unity	3	-.5702	-.9272
on each axis	4	.5334	.2089
	Sum	0	0
Sum of Squares		1.0	1.0

replace of sums of squares and crossproducts. Correlations do not yet yield a rigid rotation, because the relations are distorted before rotation takes place, so that the final coordinates bear no simple relation to the original configuration. Table 2.3g shows the principal components from correlations after scaling so that sums of squares are unity. The resulting coordinates are shown in Table 2.3h, and are obviously very different from Tables 2.3e and 2.3f.

It should be emphasized that only Table 2.3c and 2.3e represent the data of Table 2.3a in the manner that is normally desired for taxonomy.

The DNA pairing data between 17 strains of *Bacillus circulans* (Nakamura and Swezey 1983*a*) were used to illustrate the effects of choice of reference strains on taxonomic structure. The data were transformed to % dissimilarities (Appendix 1) which are the equivalent of distances between strains.

A multivariate random swarm was constructed (using a program written by P. H. A. Sneath on Leicester University Vaxcluster) using 17 points scaled to have a similar dimensionality to the data from Nakamura and Swezey, 1983*a*. This was also examined for effects of choice of reference strains.

2.14 Estimating Error from published DNA Homology Data.

Published Standard Deviations.

The average error was determined from standard deviations after correcting for degrees of freedom.

The average error, as a standard deviation s_E, was obtained as follows. From individual standard deviations, s_i, and numbers of replicates, n, on which s_i was based; $s_E = \sqrt{[\ \Sigma\ s_i^2\ (n_i - 1)\ /\ \Sigma(n_i - 1)]}$.

In a few instances it was evident from internal evidence that the published s_i, issues had not been corrected for degrees of freedom, so they were then recalculated.

Reciprocal Pairs.

For methods involving radiolabelling techniques (membrane filter or endonuclease techniques) often a square matrix is published where (for strains *a* and *b*) the corresponding values i.e. *a* versus *b* and *b* versus *a* are not duplicates, but where first strain *a* was the labelled nucleic acid, and then strain *b*. Theoretically the relation of *a*:*b* should equal that of *b*:*a* but this is not always so. Error was calculated as the standard deviation between the reciprocal pair. In spectrophotometric techniques this error does not arise, as the experiment for measuring pairing between *a*:*b* and *b*:*a* is identical.

The standard deviation, s, for such a pair of reciprocal values, $X_{a:b}$ and $X_{b:a}$, has 1 degree of freedom and s is

$$\sqrt{[[(X_{a:b} - \text{mean})^2 + (X_{b:a} - \text{mean})^2\ ]\ /\ 1]}$$

and this reduces to

$$\sqrt{[(X_{a:b} - X_{b:a})^2\ /\ 2]}.$$

The average s_E for m such pairs is $\sqrt{(\Sigma s^2/m)}$.

Use of Triangles

This method involves looking at all possible combinations of each of three strains in turn. For convenience the data are converted to dissimilarities, e.g. 90 % DNA pairing corresponds to 10 % dissimilarity or
a "distance" of 10. Thus any three strains can be represented as apices of a triangle with the lengths of the sides corresponding to the distances between strains.

With a square matrix (not using the optical method) there will be eight possible triangles per three strains a, b and c, assuming it is a complete matrix.

	a	b	c
a	0	27	30
b	25	0	42
c	35	40	0

1. $X_{a:b}$, $X_{a:c}$, $X_{b:c}$
2. $X_{a:b}$, $X_{a:c}$, $X_{c:b}$
3. $X_{a:b}$, $X_{c:a}$, $X_{b:c}$
4. $X_{a:b}$, $X_{c:a}$, $X_{c:b}$
5. $X_{b:a}$, $X_{a:c}$, $X_{b:c}$
6. $X_{b:a}$, $X_{a:c}$, $X_{c:b}$
7. $X_{b:a}$, $X_{c:a}$, $X_{b:c}$
8. $X_{b:a}$, $X_{c:a}$, $X_{c:b}$

Such data permit two kinds of analysis. The first is to estimate test error from triangles with one zero side. The second is to determine whether the values satisfy the triangle inequality and thus have properties consistent with a Euclidean metric, and therefore well suited to spatial geometric representations of taxonomic structure.

Most published tables of DNA pairing data are very incomplete,

consisting only of a limited number of the possible strain comparisons. Therefore a computer program (Appendix 2) was written to list the complete triangles (i.e. the cases where data for three sides were available) and then to determine the number where the triangle inequality was violated and to compute errors.

Triangles with zero sides

Strains which appear identical within the sensitivity of the experiment will form, with a third strain, a triangle with one side zero. Theoretically the other two sides should be equal, but they are frequently unequal and the discrepancy may be used as a measure of error. Error was determined as $\sqrt{[(X_{a:c} - X_{b:c})^2 /2]}$ for a triangle with $X_{a:b} = 0$, and averaged as for reciprocal pairs.

Analysis of variance (ANOVA) was used to examine the amount of variation between and within comparable sets of data.

2.15 Triangle Inequalities

The triangle hypothesis was studied by counting the proportion of triples which do not satisfy the triangle inequality. Thus, if $X_{a:b}$ is 25 and $X_{a:c}$ is 25 then $X_{b:c}$ cannot be greater than 50, i.e. the largest side of the triangle must be equal to or less than the sum of the other two sides if the metric used is Euclidean (as is desirable for taxonomy). Any triangle with a zero side will violate the triangle inequality if any error is present (though perhaps only to a small extent) because then the other two sides will not be exactly equal.

All distances were square-rooted and the resulting number of violating triangles determined. To test for a significant reduction in the number of

violating triangles after this treatment chi-square with Yate's correction was used, and the probability determined for one degree of freedom (except where any value in the 2x2 table was less than 5, when Fisher's Exact method was used; Conover 1971).

For complete triangles the number of non-violating and violating triangles, before and after square-rooting are tabulated as:-

	Before taking square-root	After taking square-root	
Violating	a	b	(a + b)
Not violating	c	d	(c + d)
	(a + c)	(b + d)	

$$X^2 = n([ad-bc]-n/2) \;/\; (a+b)(c+d)(a+c)(b+d)$$

Triangle inequalities and error estimations (2.14) were calculated using a computer program (TRUDNA.pas) listed in Appendix 2.

3. RESULTS

3.1 Temperature Control in the Spectrophotometer

Effects of Sample Position.

The T_m was found to vary according to the position of the sample in the cell holder. It was also noted that the extent of the variation changed with a different temperature probe. Temperature variation was calibrated using *E. coli* strain B DNA (Sigma), which has a known T_m, and the results are listed in Appendix 3. The average variation, with respect to cell position five, (the probe is usually placed in position five) using the probe used to do the majority of percent pairing experiments is shown below:

Cell Position:	2	3	4	5	6	7
	+0.8	-0.1	0	0	-0.1	+0.7

Temperature Settings

The actual temperature of the PU8700's thermal block was shown on the screen of the spectrophotometer and this did not always correlate with the temperature at which the machine had been programmed.

The actual temperature was found to have a linear correlation with the set temperature (Figure 3.1). The discrepancy between the set and actual final temperature was taken into account when programming the spectrophotometer for renaturation experiments.

3.2 Shearing and Fragment Size.

Figure 3.2 shows the renaturation of a sample of C52 which was syringe sheared. The rate is very fast and that of the straight

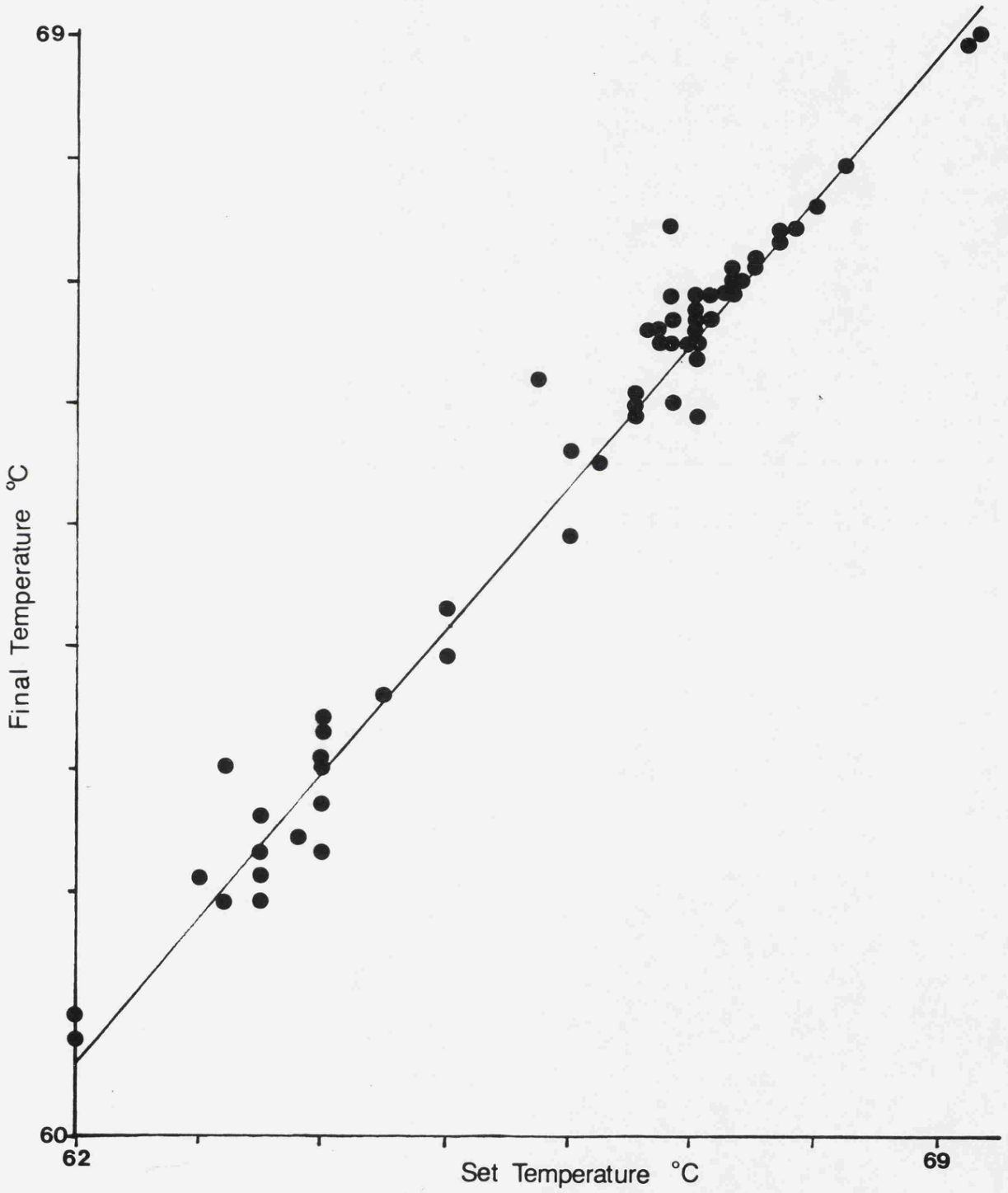
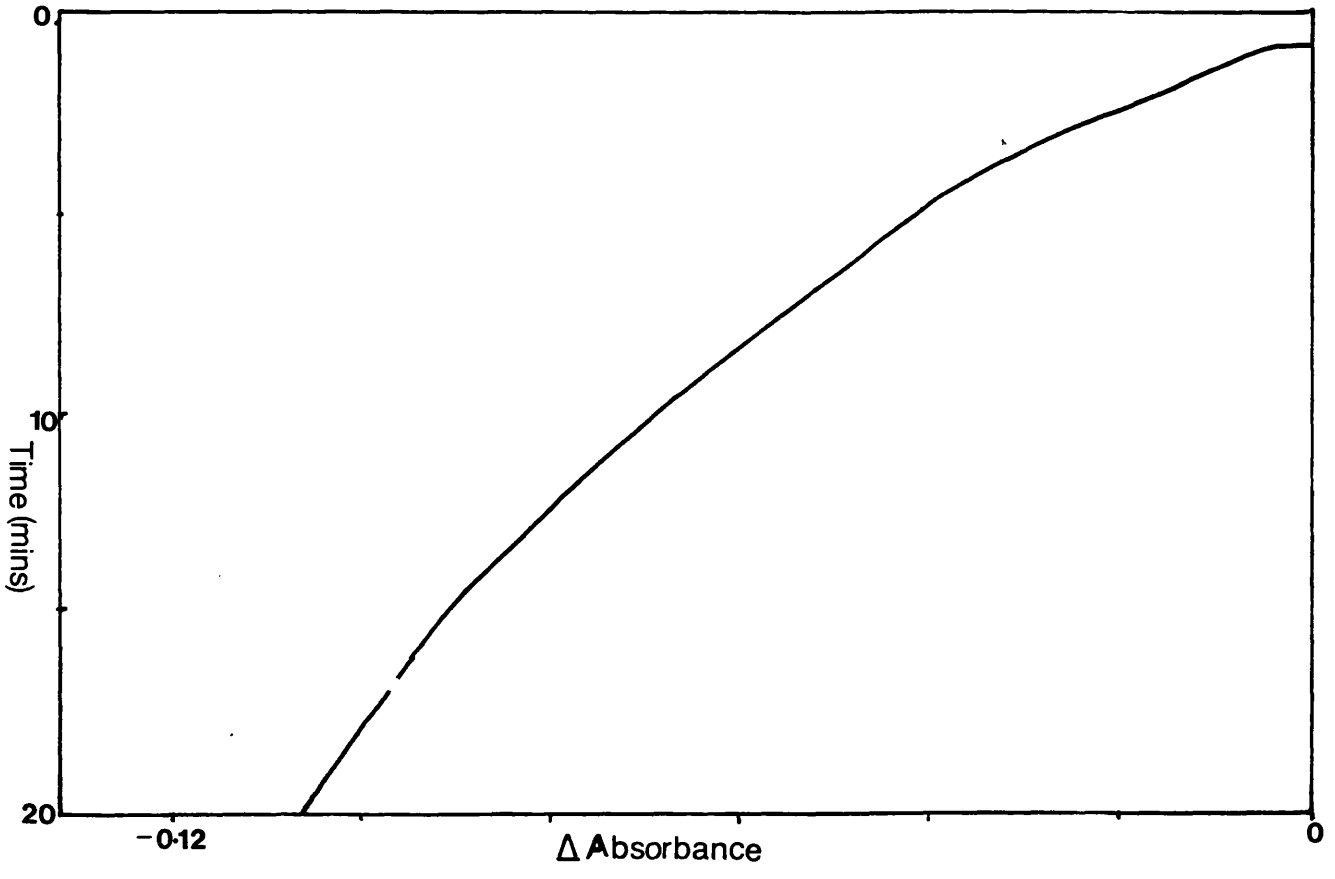


Figure 3.1 A plot of the set temperature on the spectrophotometer versus the resulting (final) temperature. A regression analysis showed the relationship :

Final temperature = 1.16(set temperature) - 11.3

The correlation, r, is 95.4 %.

Figure 3.2 The reassociation curve from a sample of syringe-sheared DNA isolated from strain C52. The DNA was melted as in section 2.7(ii) and renatured at the T_{or}, as in section 2.7(iii).



line portion indeterminable. The plot deviates from the second order too quickly, suggesting that the fragment size is too large, causing a rapid zippering effect between homologous strands.

Agarose gel electrophoresis confirmed the fragment size. Syringed samples banded at the same level as unsheared DNA (from the same isolation) and had a molecular weight of >12,000 base pairs.

DNA passed through the French press was found to have a fragment size of 344-514 base pairs. On renaturation DNA sheared in this way produced rates which were straight, i.e. of second order kinetics, for over 15 minutes allowing the initial tangent to be easily traced. This method of shearing was used for all other experiments in this thesis.

3.3 Effects of Salt Concentration.

The effects of ionic concentration are well documented (see 1.3). In renaturation experiments it is important that the renaturation buffer (SSC) is made up precisely.

Renaturation experiments were carried out pre- and post-overnight dialysis in 0.1 or 2 x SSC buffer, using DNA samples dissolved in accurately made up buffer. The T_m and renaturation rate results are shown in Table 3.1. There was no significant difference in the T_ms of the DNA which had been dialysed overnight and that which was just dissolved in the accurately made up buffer. A sample of which part was dialysed overnight and part dialysed for 3-4 hours produced different renaturation rates although the T_ms were similar.

Table 3.1 Effects of Salt Concentration.

	Sample: C644 pre-dialysis	C644 dialysed overnight	dialysed 3-4 hours
T_m[1]	87.9 °C	88.4 °C	
	86.7 °C	87.4 °C	
	86.7 °C	85.5 °C	
	84.7 °C	84.9 °C	
Average (n=4)	86.5 °C	86.6 °C	
s.d.	1.33	1.63	
		87.6 °C	86.5 °C
		88.6 °C	87.2 °C
Average (n=2)		88.1 °C	86.9 °C
s.d.		0.71	0.49
Renaturation[2] Rates	.00218	.00236	
	.00220	.00205	
Average (n=2)	.00219	.00220	
		.00220	.00163
		.00213	.00152
Average (n=2)		.00217	.00158

[1] Not corrected for cell position.

[2] Absorbance per minute

3.4 Polysaccharide Removal.

Appendix 4 contains a table of T_m data from samples pre and post carbohydrate removal. Carbohydrate removal had no significant effect on the T_m or on the rates of renaturation of DNA samples of *Listeria*. In fact carbohydrate removal procedures greatly reduced the yield of DNA per litre of culture and extensive dialysis was required to obtain pure samples, particularly with the CTAB-removal procedure. Scans of DNA samples before and after carbohydrate removal are shown in Figure 3.3. Using the colorimetric technique (2.9) and a set of glucose standards the μg of carbohydrate per μg of DNA was determined on a sample of C644 DNA, half of which had been treated with CTAB. The untreated sample had 0.08 μg of carbohydrate per μg DNA (n = 2 ±0.03), compared with 0.19 μg carbohydrate per μg DNA (n = 2 ±0) for the CTAB treated sample.

3.5 Storage.

Two samples of DNA isolated from *Listeria innocua*, C644, were used to investigate the effects of storage for long periods at -20°C. One preparation was prepared, sheared by two passages through the French press and stored at -20°C for over one year. Results are summarised in Table 3.2.

There was found to be no marked difference between the C644 sample stored sheared and diluted at -20°C and a freshly prepared C644 DNA sample. The percent homology of C644-stored with C644-fresh was 105.6 %. The percent hyperchromicity of the stored sample was 4.2 % less than that of the fresh sample. This was not significant; the average percent hyperchromicity of C644 DNA samples was 31.4 % ± 3.02 (Appendix 6)

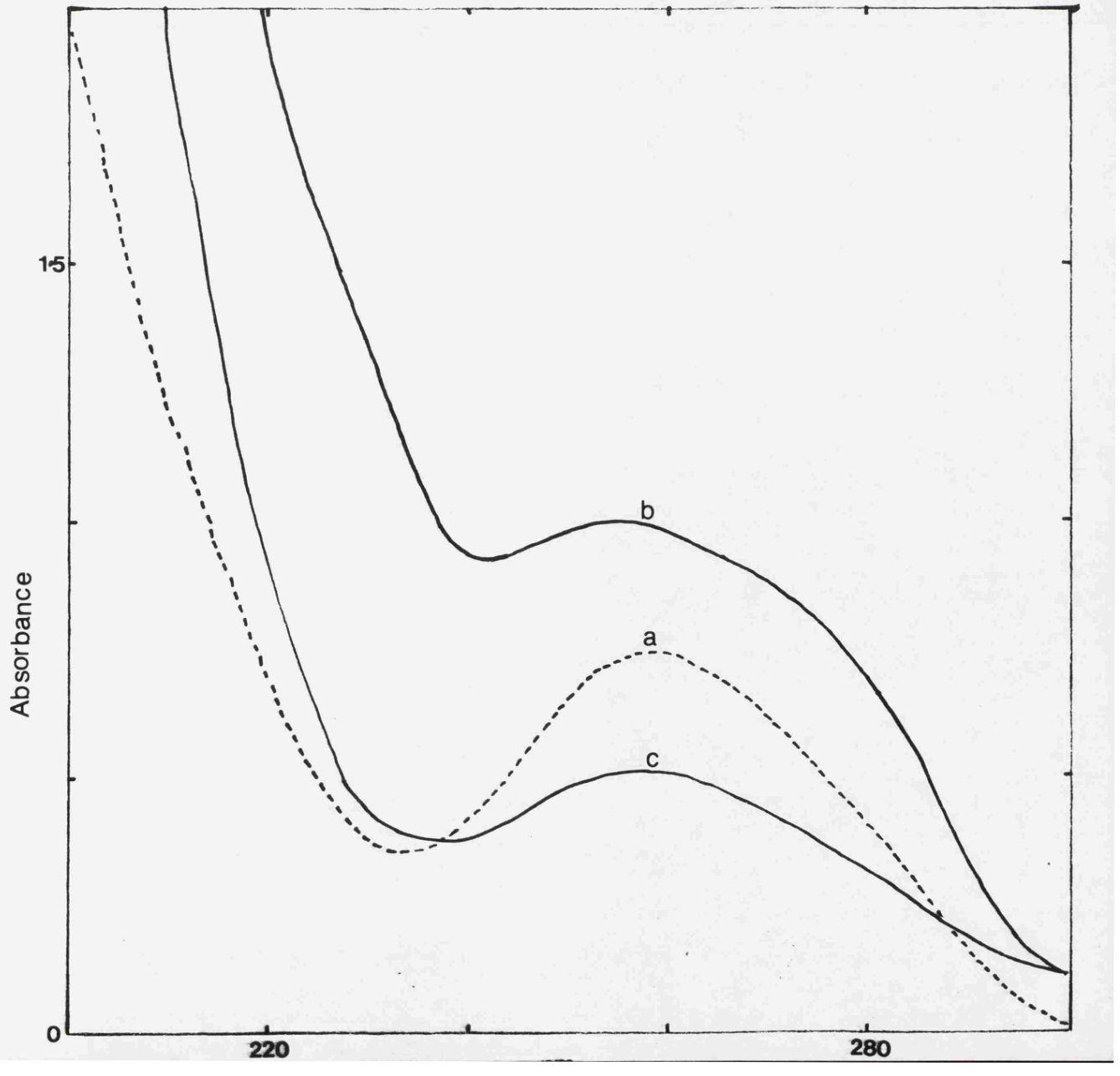
Figure 3.3 Scans of DNA samples before and after carbohydrate removal using CTAB.

a : DNA, from strain C52, unsheared, pre-dialysis, dissolved in 0.1 x SSC prior to treatment with CTAB. The 280/260 ratio is 0.55.

b : DNA sample a after treatment with CTAB but before dialysis. The 280/260 ratio is 0.70.

c : DNA sample b dialysed overnight. The 280/260 ratio is 0.63.

Over half the DNA was lost during the cleaning process.



which encompasses both the value of the fresh and stored sample. The quality of the renaturation plot was not affected by storage (MINITAB analysis gave a 99.6 % fit for the stored sample and 99.3 % fit for the fresh sample). Reducing the concentration of the stored sample from 74.2 μgml^{-1} to 30 μgml^{-1} did not affect the T_m value and the rate was reduced as expected (Huss *et al.* 1983; Gillis *et al.* 1970).

A sample of *Klebsiella pneumoniae* stored for over two years at -20°C in dilute buffer had a reproducible T_m of 95.2 ± 0.77 (n = 9, no correction for cell position) after storage.

Table 3.2 Storage Experiments

Sample	% Hyperchromicity	% fit (MINITAB)	% Homology	Expected % Homology
C644, stored	28.3	99.8		
C644, fresh	32.5	99.3		
Mixed	32.6	99.3	105.6	100

T_m Determinations:

	Concentration ($\mu g\ ml^{-1}$)	T_m	Rate of reassociation
C644, stored	74.2°C	89.2°C	0.00236
	30.0°C	89.3°C	0.00076
K. pneumoniae after 2 years storage.	76.8°C	96.1°C	
	74.7°C	95.4°C	
	74.3°C	96.1°C	
	56.8°C	94.0°C	
	58.1°C	95.2°C	
	56.0°C	95.2°C	
	56.7°C	94.2°C	
	56.7°C	95.9°C	
	55.3°C	94.8°C	

3.6 Stringent Conditions.

Several renaturation experiments were carried out at stringent and non-stringent temperatures, for closely related strains and strains from different species. Results are summarised in Table 3.3. A huge variation in pairing results was obtained for strains C214*a* and C214*b* under stringent conditions. Strains JS21 and JS31 showed only a 5.4 % variation between stringent and optimal conditions.

3.7 Base Composition Determinations.

The complete data for base composition determination experiments is tabulated in Appendix 5 and summarized in Table 3.4. All % (G+C) determinations were higher than those obtained by Ferusu, 1980. The slower ramp rate gave a lower standard deviation and a lower % (G+C) value which was closer to the value obtained by Ferusu, 1980.

Table 3.3 Stringent Conditions.

Strains:	Renaturation Temperature (°C)	± T_{or} (°C)	% Homology Observed	% Homology Expected[1]
C214*a*, C214*b*	78.6	+10	78.6	95.5
" "	"	"	49.1	"
" "	78.8	"	91.2	"
JS21, JS31	76.8	"	89.6	95.0
C52, C1091	71.9	+10	39.5	52.6
" "	69	+7	36.3	"
" "	69	+7	38.9	"
" "	66.5	+4	42.1	"
" "	55	-10	59.1	"

[1] Based on the average homology determined from DNA pairing results in Appendix 6.

Table 3.4 Summary of % (G+C) Determinations

Strain	% (G+C) present work	Standard deviation	Previous Determinations % (G+C)	Reference
C644	37.5	1.34	36.2	Ferusu, 1980
C644*	36.6	0.4	"	" "
C1091	36.0	0.7		
C1174	39.9	0.92		
C214*b*	46.1	1.24	42.4	Ferusu, 1980

* A ramp rate of 0.5°C min^{-1} was used as opposed to 1°C min^{-1} in other experiments.

3.8 Pairing Data from *Listeria* species.

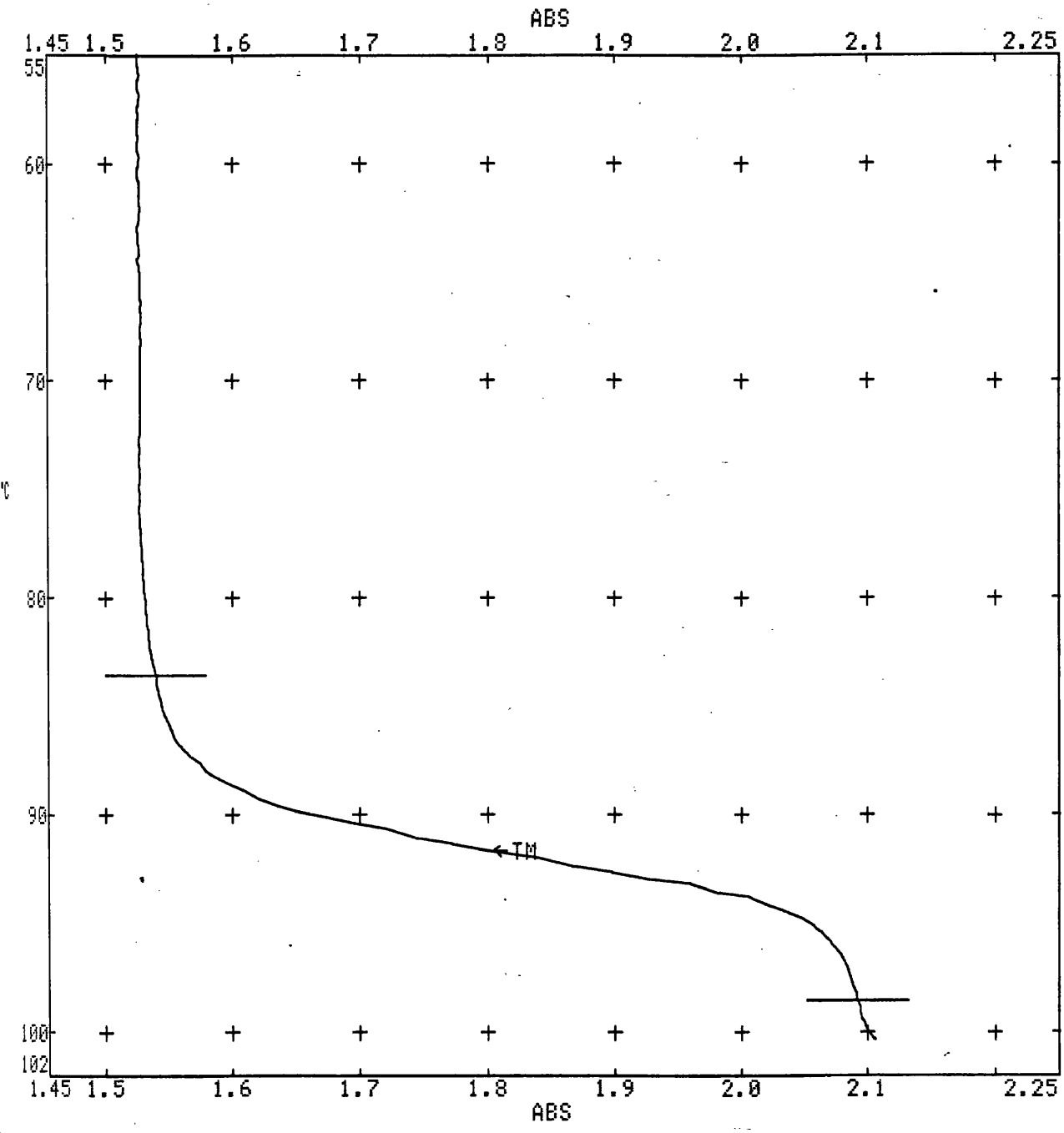
DNA isolated from 22 strains of *Listeria*, sheared by two passages through the French press and dialysed overnight was used to obtain the pairing data tabulated in Appendix 6.

DNA was stored pre-sheared, at -20°C when necessary, but not refrozen after defrosting. Shearing was carried out the day before use and all sheared DNA samples were used within 10 days. The T_{or} of each sample was determined from the T_m, an example of a T_m curve from DNA of a strain of *Listeria* is shown in (Figure 3.4).

Absorbance readings output by the PU8700 spectrophotometer during renaturation experiments were typed into columns in the MINITAB package available on the VAX cluster mainframe at Leicester University. Columns of absorbance readings were regressed against the time of renaturation; the gradient of the regression line corresponded to the rate of renaturation. The percent fit of the linear regression line is shown in Appendix 6; this illustrates the purity of the samples and confirms the period over which the reaction is second order. Figures 3.5 and 3.6 show renaturation experiments from two very closely related strains and two distantly related strains.

The average percent pairing between each pair of strains was determined. (The standard deviations from replicates are summarised in Table 3.18, see section 3.13). A complete 16 x 16 matrix of average homologies was derived from these results (Table 3.5). By subtracting from 100, the matrix of relatedness was converted to a matrix of taxonomic distances (Table 3.6). The matrices in Tables 3.5 and 3.6 were used in the principal components computer program, TRUPCA.bas (Appendix 7).

Figure 3.4 Example of a T_m curve from *L. innocua*, strain C645, DNA as produced by the PU8700 spectrophotometer.



SAMPLE C645 REFERENCE _____

CELL PATH _____ OPERATOR _____

Cell 3	A(I)	°C(I)	A(F)	°C(F)	ΔA	TM
T1 - T2	1.540	83.6	2.092	98.5	0.553	91.7

Figure 3.5 The renaturation curves for two closely related strains:
reassociation of DNA from strain C52
reassociation of DNA from strain C201
reassociation of a 50:50 mixture of DNA from strains C52 and C201.

$$\% \text{ Pairing} = \frac{(4 \times 0.00299) - (0.00375 + 0.00265)}{2\sqrt{(0.00375 \times 0.00265)}} = 88.2\ \%$$

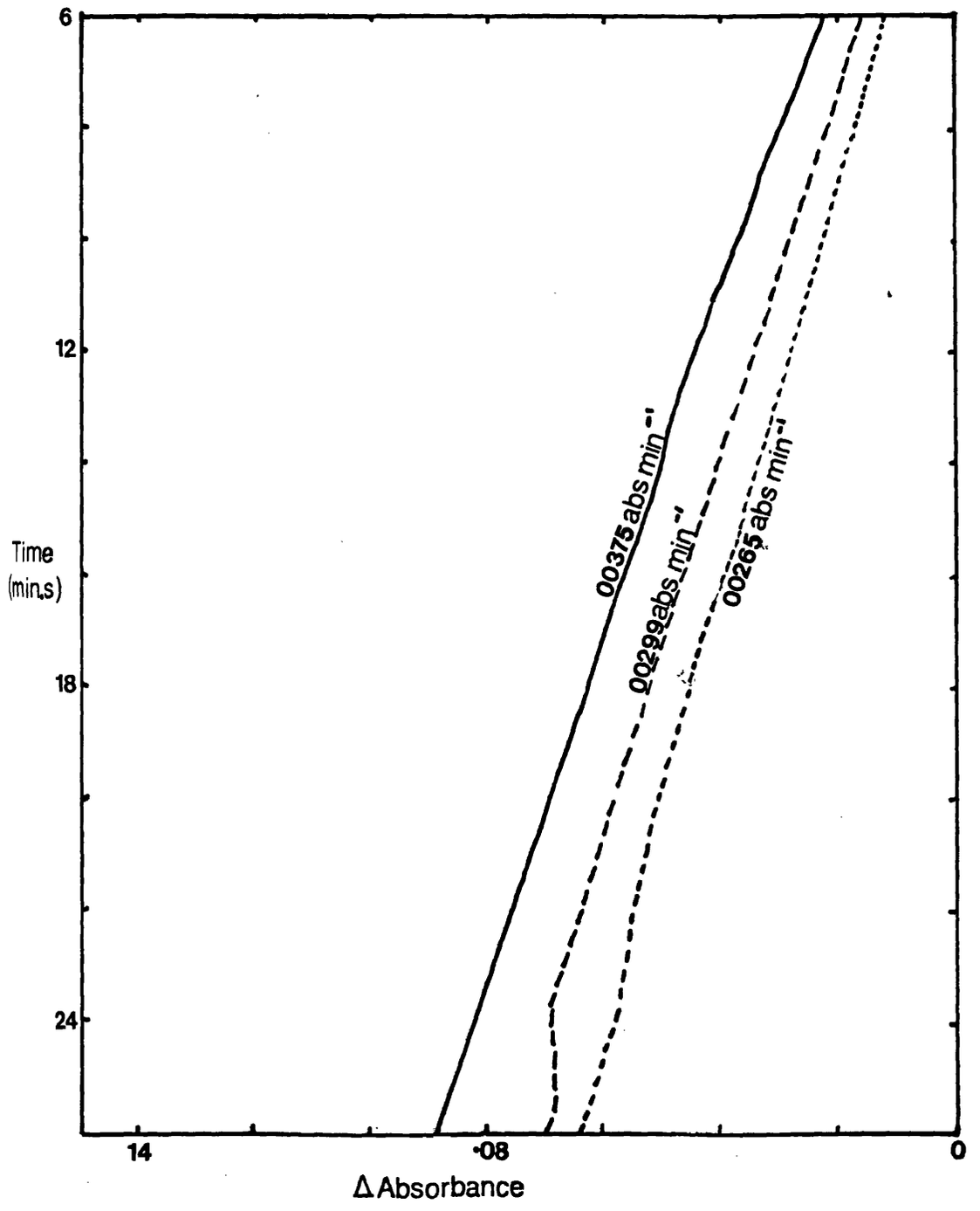


Table 3,5 Matrix of % Homology Values,

	Strains																		
	C52	C200	C202	C203	C231	C644	C1090	C1171	C1091	C1172	C1087	C663	C667	C659	C214*a*	C1174	C201	C228	31
C52	100																		
C200	67,0	100																	
C202	70,6	81,5	100																
C203	67,1	75,2	85,1	100															
C231	69,3	73,8	79,3	71,4	100														
C644	71,9	51,2	49,7	62,3	44,6	100													
C1090	52,2	55,0	39,6	43,7	59,8	50,3	100												
C1171	57,0	50,6	50,3	40,1	53,4	49,3	93,5	100											
C1091	52,6	57,4	49,4	55,9	42,3	49,7	46,2	43,6	100										
C1172	44,1	53,4	45,6	52,3	50,8	45,5	53,5	42,5	91,6	100									
C1087	44,7	34,6	44,4	29,9	38,2	26,9	37,3	37,2	34,9	42,4	100								
C663	30,4	41,8	33,6	34,5	41,2	27,8	43,0	35,9	24,3	27,3	69,4	100							
C667	36,8	31,0	34,0	36,4	27,1	35,2	38,9	38,6	31,5	36,1	70,7	83,8	100						
C659	33,0	30,4	34,3	30,4	30,7	30,8	35,4	24,0	24,9	32,0	86,2	58,4	83,4	100					
C214*a*	11,8	15,7	29,3	18,9	30,1	4,3	13,6	11,7	14,7	6,7	8,9	30,4	28,4	27,5	100				
C1174	21,6	14,9	4,5	26,3	25,2	15,9	16,9	12,9	24,3	11,4	14,3	9,0	8,0	17,2	84,3	100			
C201	89,4					79,0											100		
C228	56,8	71,6	54,8				47,9		51,7	57,8	32,9				9,8			100	
31	46,6						50,7		56,2		38,9				24,9	7,2		49,8	100
C645	80,6	55,4	19,1	49,8	54,8	94,9	61,6	44,9							0,0	-3,6			
21									71,1						12,9				51,2
C666	35,7				38,4								84,9	94,6					
C214*b*				11,5											95,5				
SLCC 7211	23,4					25,2			10,5		17,2			35,9			25,2		

Table 3.6 Matrix of DNA Distances

	Strains																	
	C52	C200	C202	C203	C231	C644	C1090	C1171	C1091	C1172	C1087	C663	C667	C659	C214*a*	C1174	C201	C228	31
C52	0																		
C200	33,0	0																	
C202	29,4	18,5	0																
C203	32,9	24,8	14,9	0															
C231	30,7	26,2	20,7	28,6	0														
C644	28,1	48,8	50,3	37,7	55,4	0													
C1090	47,8	45,0	60,4	56,3	40,2	49,7	0												
C1171	43,0	49,4	49,7	59,9	46,6	50,7	6,5	0											
C1091	47,4	42,6	50,6	44,1	57,7	50,3	53,8	56,4	0										
C1172	55,9	46,6	54,4	47,7	49,2	54,5	46,5	57,5	8,4	0									
C1087	55,3	65,4	55,6	70,1	61,8	73,1	62,7	62,8	65,1	57,6	0								
C663	69,6	58,2	66,4	65,5	58,8	72,2	57,0	66,1	75,7	72,7	30,6	0							
C667	63,2	69,0	66,0	63,6	72,9	64,8	61,1	61,4	68,5	63,9	29,3	16,2	0						
C659	67,0	69,6	65,7	69,6	69,3	69,2	64,6	76,0	75,1	68,0	13,8	41,6	16,6	0					
C214*a*	88,2	84,3	70,7	81,1	69,9	95,7	86,4	88,3	85,3	93,3	91,1	69,6	71,6	72,5	0				
C1174	78,4	85,1	95,5	73,7	74,8	84,1	83,1	87,1	75,7	88,6	85,7	91,0	92,0	82,5	15,7	0			
C201	10,6					21,0											0		
C228	43,2	28,4	45,2				52,1		48,9	42,2	67,1				90,2			0	
31	53,4						40,3		43,8		61,1				75,1	92,8		50;2	0
C645	19,4	44,6		50,2	45,2	5,1	38,4	55,1							100,0	103,6			
21									28,9						87,1			48,8	5,0
C666	64,3				61,6								15,1	5,4					
C214*b*				88,5											4,5				
SLCC 7211	76,6					74,8			89,5		82,8			64,1			74,8		

3.9 Clustering and Presentation of DNA Pairing Data.

The relationships between the 16 strains are illustrated in Figures 3.7-3.9. Figure 3.7a is the UPGMA dendrogram, and is the best representation, because it takes into account the distances in the full space of $t - 1 = 15$ euclidean dimensions.

Figure 3.7b, the three dimensional model from principal component analysis, only represents the first three of the 15 dimensions, and therefore neglects some of the information, but it does allow an easier appreciation of the salient features than Figure 3.7a. In this case it gives broadly the same information.

The UPGMA dendrogram shows all strains have clustered together at a distance of 47.3. Strains C214*a* and C1174 (*L. grayi* and *L. murrayi* respectively) are easily separated from the other strains by eye. The four *L. ivanovii* strains are grouped together, however the remaining strains appear very closely related. It would be very difficult to separate the *L. innocua, L. monocytogenes, L. welshimeri, L. seeligeri* strains from each other without previously knowing the numerical taxonomic data already documented. The two *L. welshimeri* strains and the two *L. seeligeri* strains do join together before attaching to the *L. monocytogenes-L. innocua* 'group'. The most interesting branch is the association of the type strains of *L. monocytogenes,* C52, and *L. innocua,* C644, before attachment to the other *L. monocytogenes* strains.

Single Linkage and Complete Linkage cluster analysis gave almost the same results as the UPGMA dendrogram. In the single link dendrogram (Figure 3.8a) C52 does not join to C644 first - however the branch lengths between the *L. monocytogenes* and *L. innocua* strains are very

Key

o *L. monocytogenes*

● *L. innocua*

0 *L. ivanovii*

s *L. seeligeri*

w *L. welshimeri*

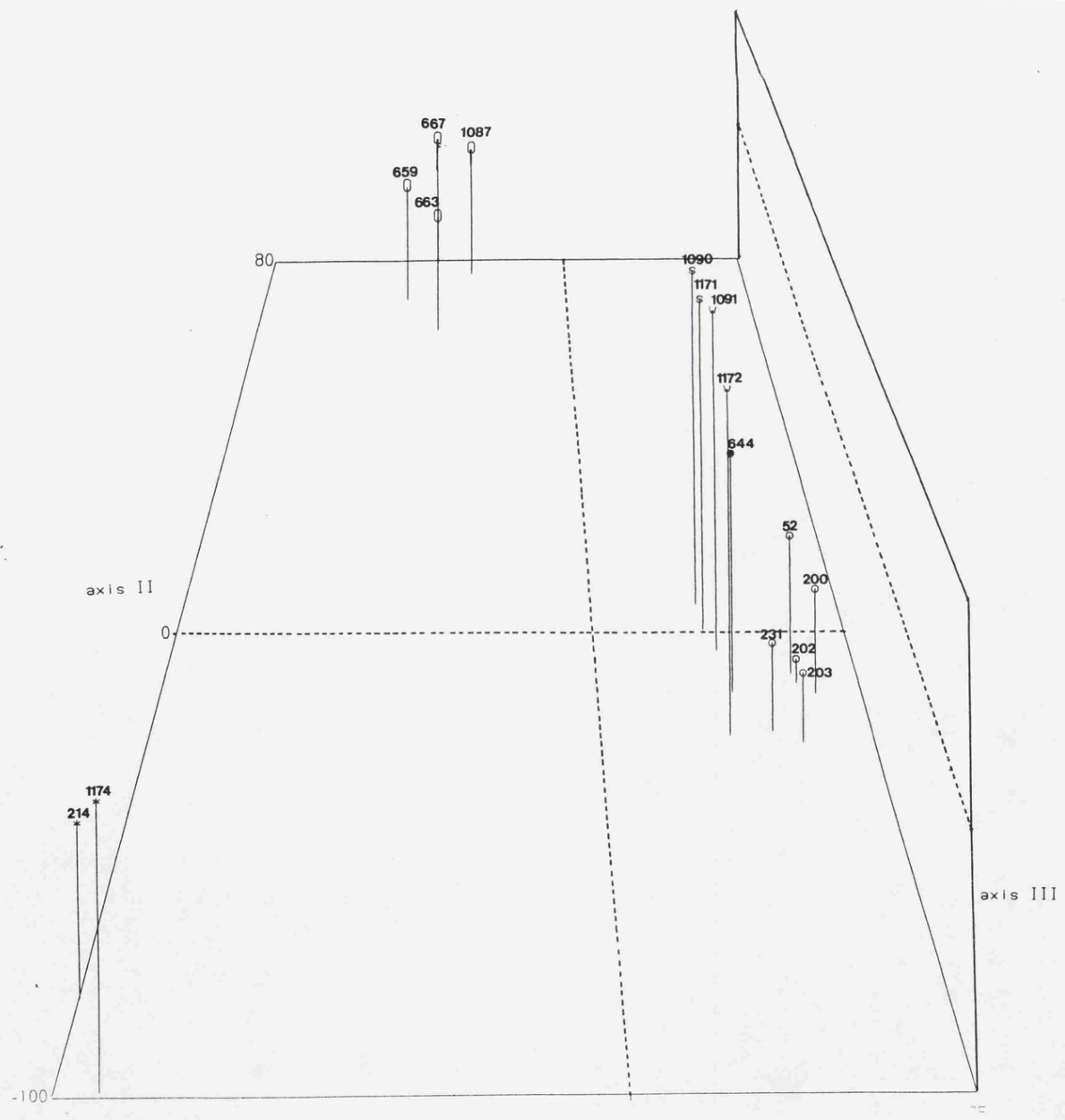
✱ *L. grayi* and *L. murrayi*

Figure 3.7

The UPGMA dendrogram and three-dimensional ordination from Principal Component Analysis of the 16 x 16 complete matrix (Table 3.5) of % DNA pairing values from strains of *Listeria*.

Figure 3.8 (p.87)

Single Link (3.8a) and Complete Link (3.8b) cluster analysis of the 16 x 16 complete matrix of *Listeria* % pairing values.



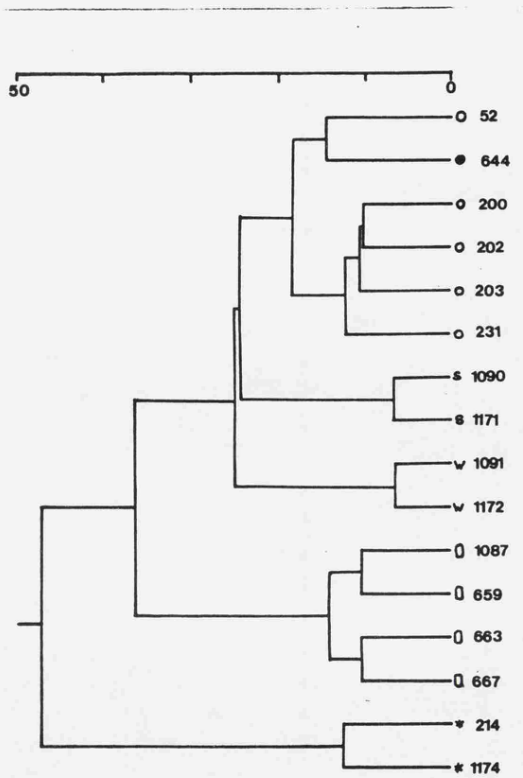
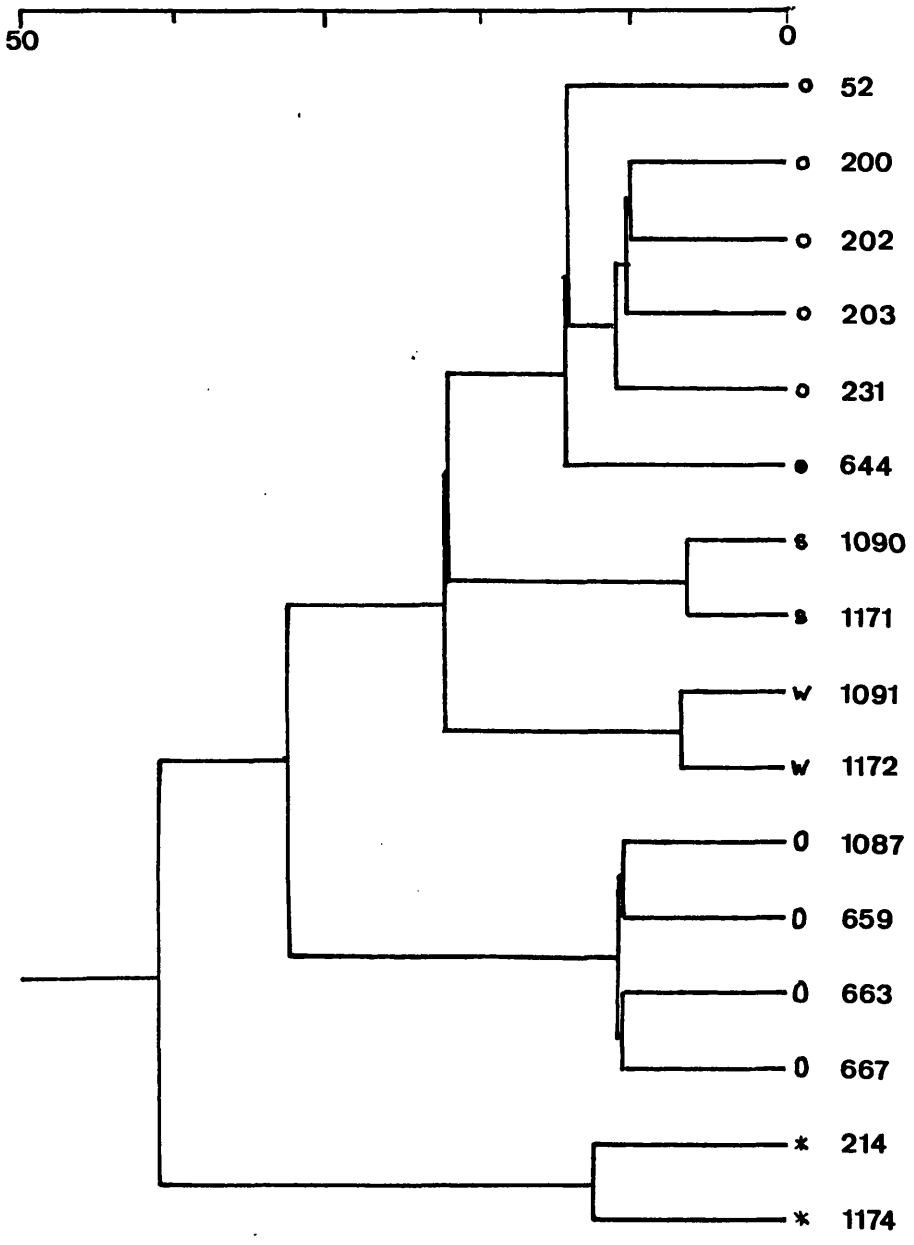


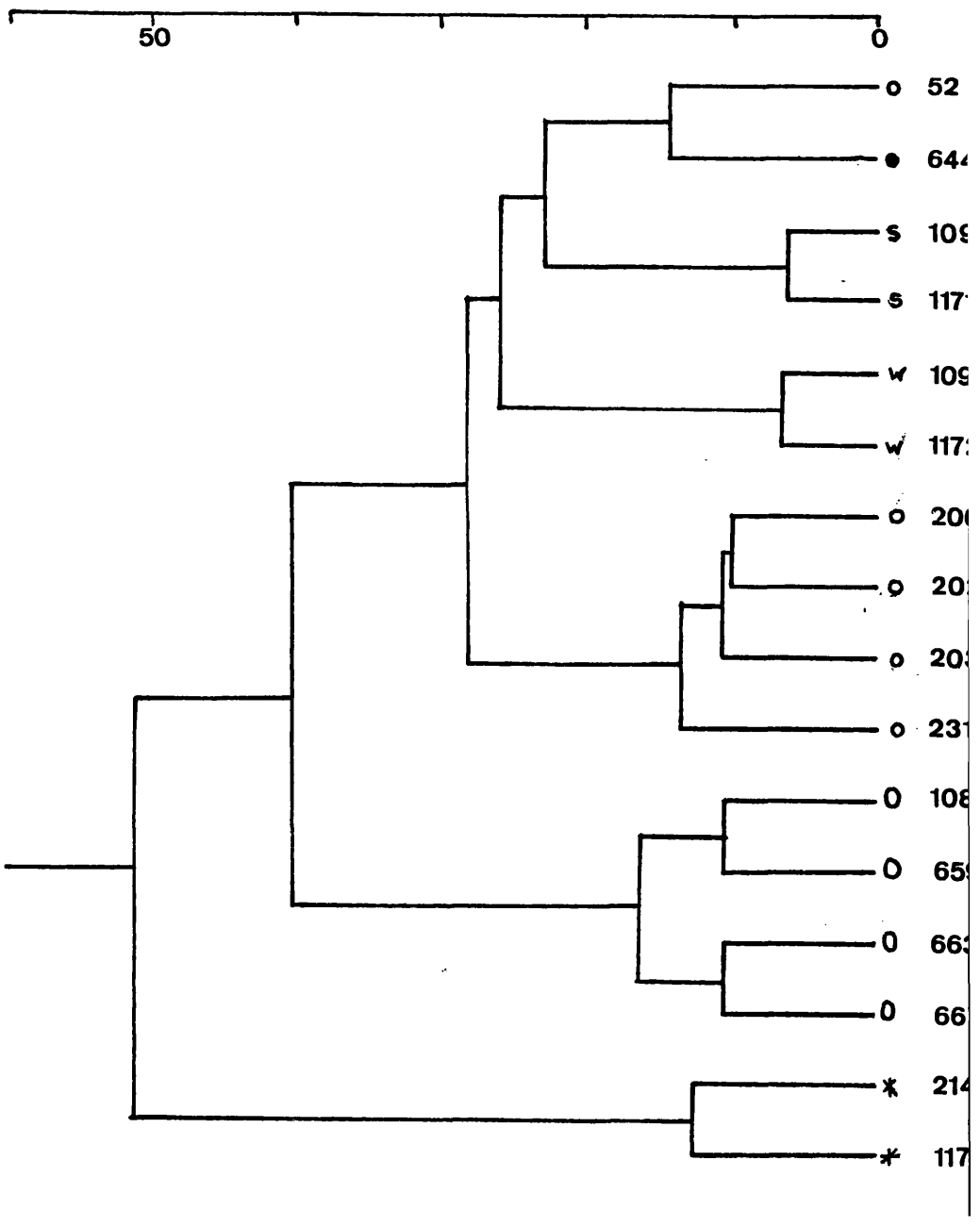
Figure 37 a

Figure 37 b

3.8a


50
0
o 52
o 200
o 202
o 203
o 231
● 644
s 1090
s 1171
w 1091
w 1172
0 1087
0 659
0 663
0 667
* 214
* 1174

3.8b


50
0
o 52
● 644
s 109
s 117
w 109
w 117
o 20
o 20
o 20
o 23
0 108
0 65
0 66
0 66
* 214
* 117

short. The Complete Link dendrogram (Figure 3.8b) showed further separation of C52 and C644 from the other *L. monocytogenes* strains, joining with *L.seeligeri* and *L. welshimeri* strains first.

Figure 3.7b, the three dimensional model from principal component analysis shows clear separation of the *L. murrayi- L. grayi* cluster and the *L. ivanovii* cluster from the other strains which are grouped closely together and not clearly separable into clusters. The *L. monocytogenes* strains form a loose cluster. The strains in the *L. seeligeri - L. welshimeri - L. innocua* complex span away from the *L. monocytogenes* group and are not clearly separable into clusters.

The percentage of variation accounted for by the first three dimensions is 83.2 % (Table 3.7). This is rather high for taxonomic structures; values of about 50 % are more usual (Bridge and Sneath, 1983; Sneath, 1983; Sneath and Stevens, 1985), though these refer to complex taxonomies from phenotypic analyses, not from DNA data. The effective dimensionality, n', is low(3.33, Table 3.7) and this phenomenon has been noted before (Sneath, 1983).

Reduction to three dimensions reduces the effective dimensionality to $m' = 2.38$ (Table 3.7).

The results from principal coordinate analysis of d^* coefficents using all 16 strains as reference strains are shown in Figure 3.9. The taxonomic structure is essentially correct and the distortion is small. The *L. welshimeri* and *L. seeligeri* strains are further apart and inseparable from the *L. monocytogenes* and *L. innocua* strains. The effective dimensionality is only slightly larger for the principal coordinate analysis and consequently there is approximately 2 % less variation in the first three axes (81.1 %, Table 3.7).

Table 3.7 Eigenvalues of the first three axes of Figures 3.7-3.20, together with the Effective Dimensionality and that of the three dimensional ordination.

Figure No.	c	λ_1	λ_2	λ_3	Percent variation in first 3 axes	n'	m'
3.9	NA	10869	8830	3849	81.1	4.75	2.63
3.7	16	62234	39728	13793	83.2	3.33	2.38
3.10	14	29898	22973	16218	73.5	4.79	2.83
3.11	7	35184	13829	5237	87.9	2.57	2.02
3.12	11	41118	37038	10675	90.1	3.03	2.48
3.13	14	56679	39329	12913	88.5	3.04	2.41
3.14	14	55527	39536	13023	88.2	3.08	2.43
3.15	12	60864	15794	10947	85.5	2.54	1.88
3.16	2	11837	4428	0	100	1.66	1.66
3.17	3	20985	4429	2807	100	1.70	1.70
3.18	2	16355	395	0	100	1.05	1.05
3.19	2	10090	5473	0	100	1.84	1.84

Key

o *L. monocytogenes*

● *L. innocua*

0 *L. ivanovii*

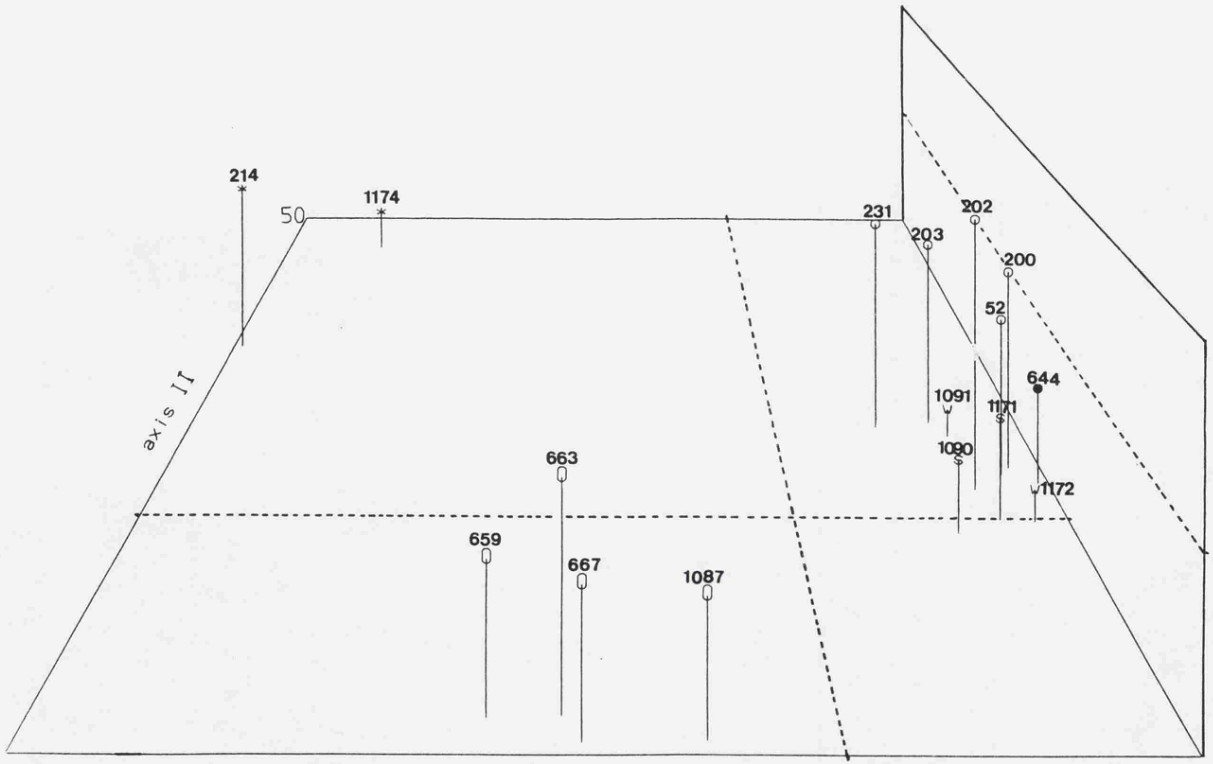
s *L. seeligeri*

w *L. welshimeri*

x *L. grayi* and *L. murrayi*

Figure 3.9

UPGMA dendrogram and three-dimensional ordination from Principal Coordinate Analysis of the 16 x 16 complete matrix (Table 3.6) of DNA distances.


214
1174
50
231
202
203
200
52
644
1091
1171
1090
663
1172
659
667
1087
axis II
axis I
axis III

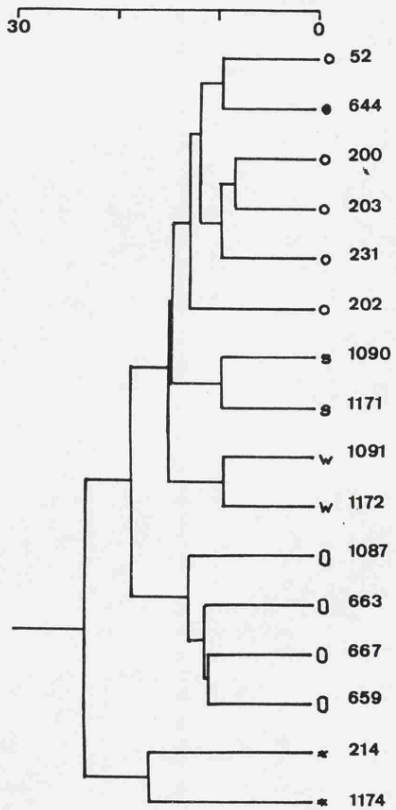

30
0
52
644
200
203
231
202
1090
1171
1091
1172
1087
663
667
659
214
1174

Figure 39

3.10 Distortion of DNA Relationships due to Choice of Reference Strains.

In order to investigate these close relationships further, the *L. grayi* and *L. murrayi* strains were removed from the matrix, leaving a 14 x 14 matrix the resulting dendrogram and ordination are shown in Figure 3.10. Without the 'influence' of the two distantly related strains (C214*a* and C1174) the configuration becomes sprawling and more confusing. The effective dimensionality is larger than that of Figure 3.7, but the variation in the first three axes is still fairly large (73.5 %). There are no distinct clusters, the principal component analysis gives the image of a spectrum of relatedness, or one large loose cluster. The UPGMA phenogram shows the closeness of the branching; however strains from the same species remain together.

Figure 3.11 shows the principal component ordination and UPGMA dendrogram when only the type strains are used as reference strains. The ordination is therefore, derived from a 7 x 16 matrix. The configuration is much more compact than that of Figure 3.7, the 'true' configuration. Both the *L. grayi-L. murrayi* and *L. ivanovii* clusters remained distinct, although the latter is noticeably closer to the main cluster. The remaining strains form a loose cluster with reference strains C52 and C644 on the periphery of the group. The *L. seeligeri* and *L. welshimeri* strains are separable in the third dimension (height) but not sufficiently distinct to be isolated as clusters or separate groups. Again the image of a spectrum of relatedness is suggested by the loose cluster, *L. monocytogenes* and *L. innocua* strains 'hovering' between the *L. seeligeri* and *L. welshimeri* strains in the third dimension, though inseparable by the first and second dimension. This image is also reflected in the UPGMA dendrogram, portrayed alongside the configuration

Key

o *L. monocytogenes*

● *L. innocua*

0 *L. ivanovii*

s *L. seeligeri*

w *L. welshimeri*

✱ *L. grayi* and *L. murrayi*

Figure 3.10

UPGMA dendrogram and three-dimensional Principal Component ordination for the 14 x 14 matrix (i.e. not including *L. grayi* and *L. murrayi* strians).

Figure 3.11 (p. 95)

UPGMA and Principal Component ordination when only the 7 type strains are used as reference strains. The reference strains are : C52, C644, C1087, C1090, C1091, C214*a*, C117*a*.

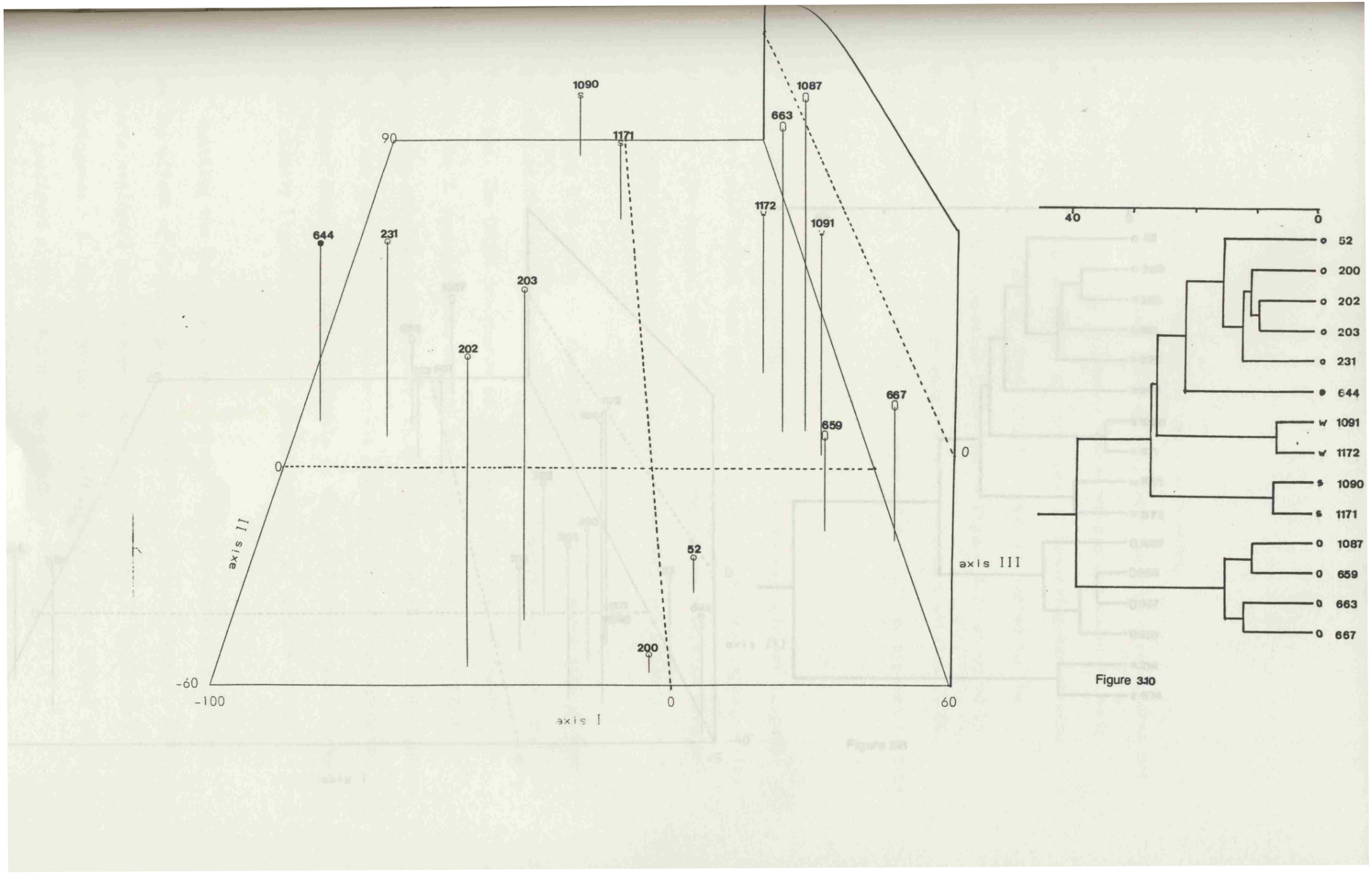

1090
1087
663
1171
90
1172
1091
644
231
203
202
667
659
0
axis III
52
axis II
200
-60
-100
0
60
axis I
40
0
52
200
202
203
231
644
1091
1172
1090
1171
1087
659
663
667

Figure 3.10

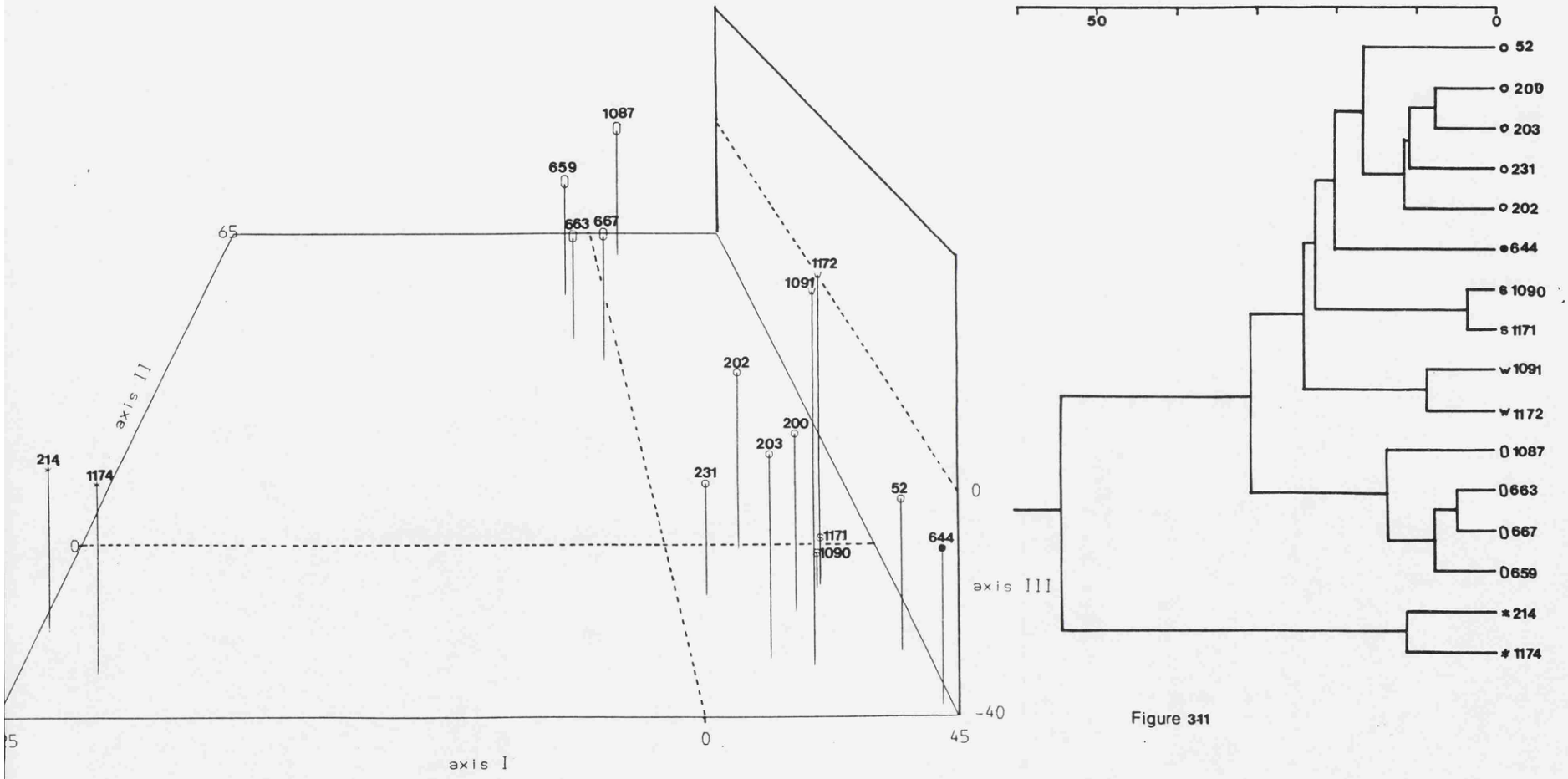

axis II
65
0
214
1174
659
1087
663
667
1172
1091
202
200
203
231
52
644
1171
1090
0
-40
axis III
axis I
45
50
0
52
209
203
231
202
644
1090
1171
1091
1172
1087
663
667
659
214
1174

Figure 311

(Figure 3.11), in the closeness of the branch nodes.

Listeria monocytogenes, C52, derived from the type strain, does not seem to be a very typical *L. monocytogenes* strain and shows as much affinity for *L. innocua* (C644) as for other *L. monocytogenes* strains.

Reference strains C52, C644, C1087, C214*a* are pushed to the periphery of the configuration (relative to Figure 3.7) as noted by Sneath (1983). This is not as apparent with reference strains C1090, C1091, which have been pushed to the edges of axis III as opposed to axes I and II.

The effects of removing the *L. grayi* - *L. murrayi* cluster (Figure 3.10) were more marked than expected so the experiment was repeated with the other five species. Each species was removed (as reference strains) from the matrix in turn and the principal component ordination plotted.

Figure 3.12 shows the result when no *L. monocytogenes* strains were included as reference strains, i.e. an 11 x 16 matrix was input into TRUPC.bas. The UPGMA dendrogram is very similar to that of figure 3.7 except the *L. monocytogenes* strains are closer together, with C644, *L. innocua*, on the edge of the group. In the ordination the *L. monocytogenes* strains are very close and inseparable from C644; the *L. seeligeri* and *L. welshimeri* strains are distinguishable in the third dimension, although still closely linked with the *L. monocytogenes* - *L. innocua* group.

Omitting the two *L. seeligeri* strains as references (Figure 3.13) has the effect of distancing *L. seeligeri* from the main cluster. The *Listeria seeligeri* strains appear as outliers to the *L. welshimeri* - *L. monocytogenes* - *L. innocua* group. This produces a similar effect to the two *L. seeligeri* strains as the only reference strains i.e. it

isolates them.

Omitting the two *L. welshimeri* strains as reference otus (Figure 3.14) has a similar effect. The *L. welshimeri* strains are isolated from the main cluster and pushed towards the centroid, as the *L. seeligeri* strains are in Figure 3.13.

When the four *L. ivanovii* strains are not used as reference strains the *L. ivanovii* species is still easily identifiable (Figure 3.15). The seven species are all distinct unlike the figure derived from the complete matrix (Figure 3.7)

Key

o *L. monocytogenes*

● *L. innocua*

0 *L. ivanovii*

s *L. seeligeri*

w *L. welshimeri*

✱ *L. grayi* and *L. murrayi*

Figure 3.12

UPGMA dendrogram and Principal Component ordination without employing any *L. monocytogenes* strains as reference strains.

Figure 3.13 (p.100)

UPGMA dendrogram and Principal Component ordination without employing any *L. seeligeri* strains as reference strains.

Figure 3.14 (p.101)

UPGMA dendrogram and Principal Component ordination without employing any *L. welshimeri* strains as reference strains.

Figure 3.15 (p.102)

UPGMA dendrogram and Principal Component ordination without employing any *L. ivanovii* strains as reference strains.

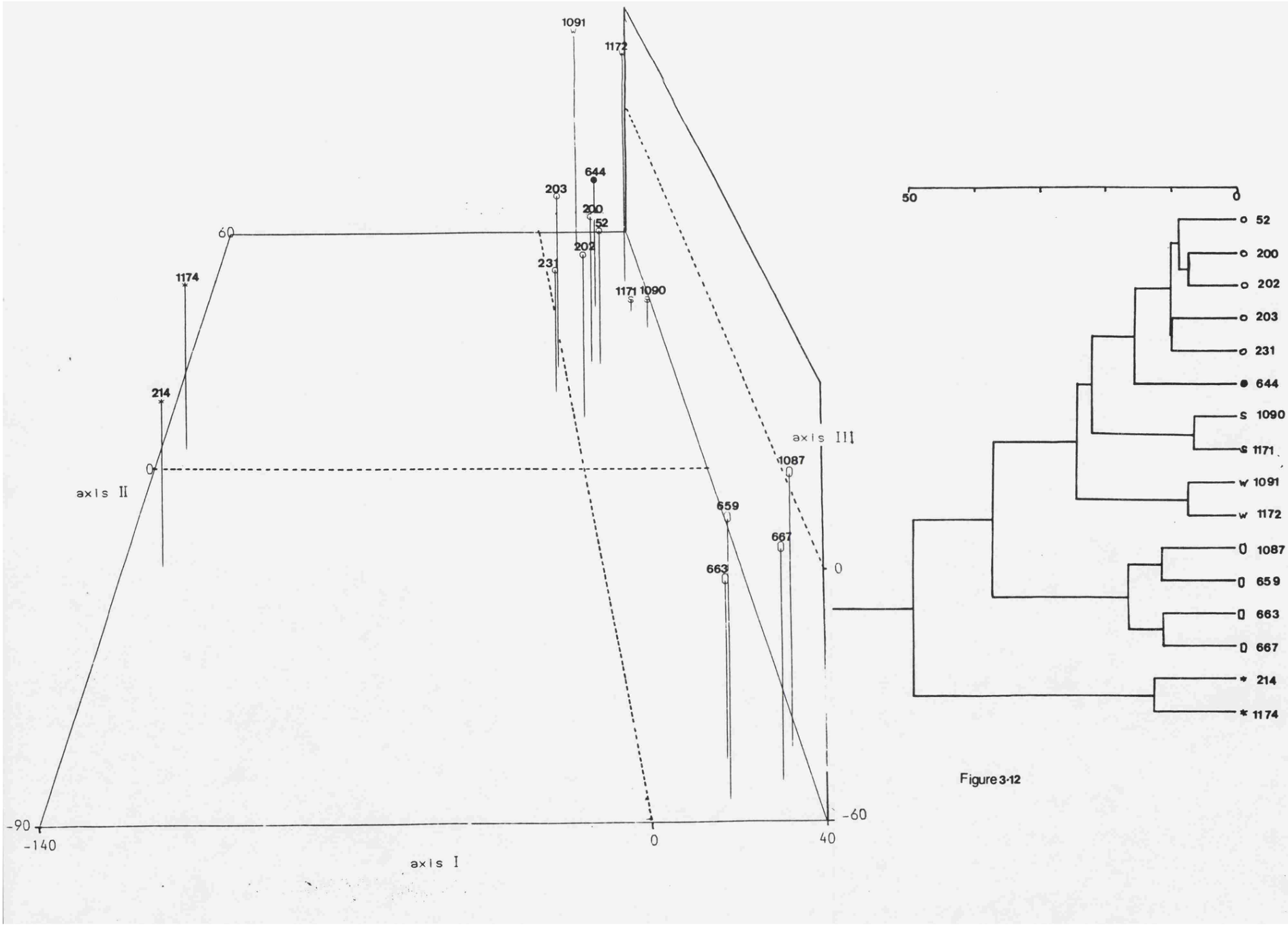

1091
1172
644
203
200
52
202
231
1174
1171
1090
214
60
0
axis II
axis III
1087
659
667
663
0
-60
-90
-140
0
40
axis I
50
0
o 52
o 200
o 202
o 203
o 231
● 644
s 1090
s 1171
w 1091
w 1172
0 1087
0 659
0 663
0 667
* 214
* 1174

Figure 3·12

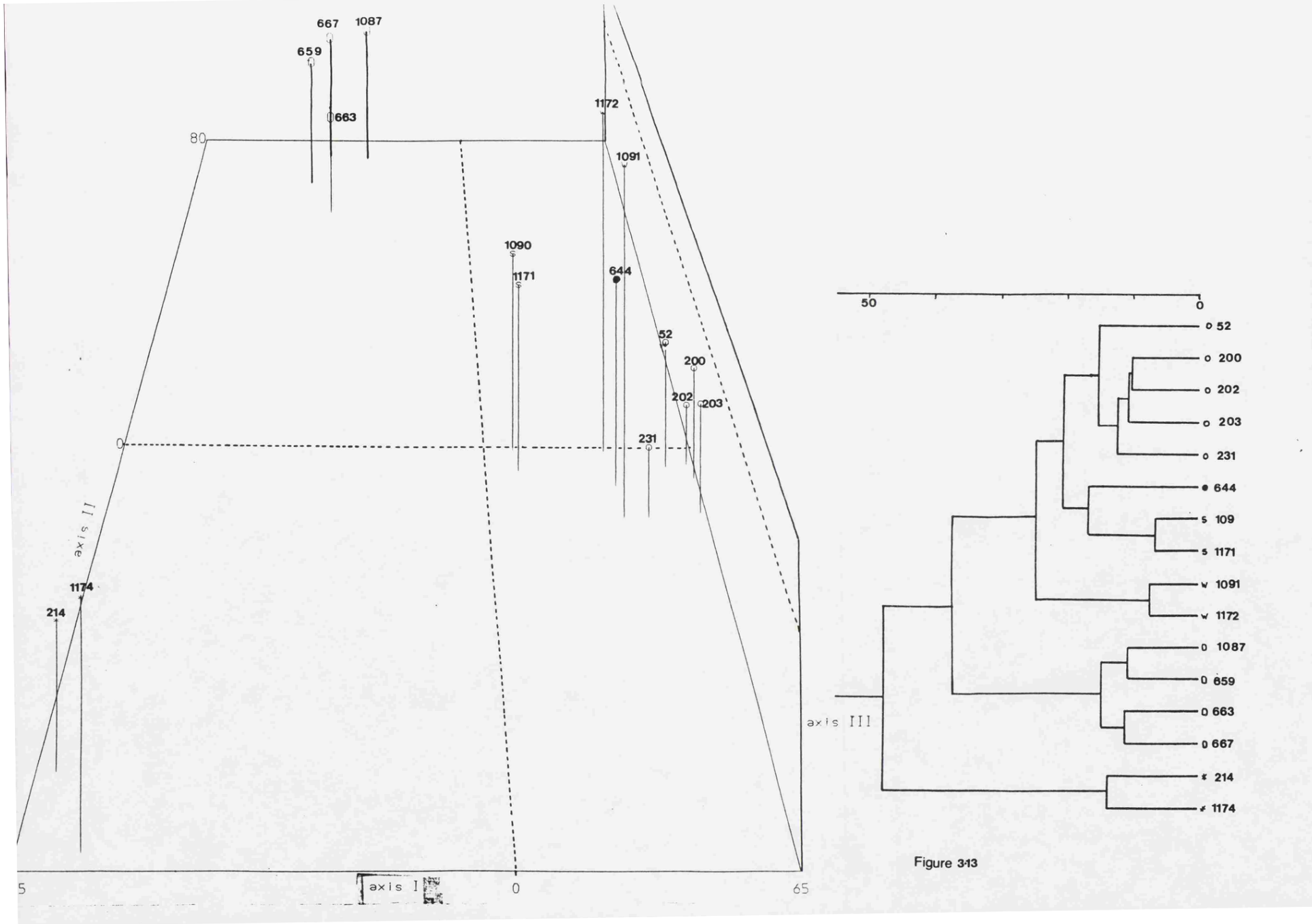

667
1087
659
663
1172
80
1091
1090
1171
644
52
200
202
203
231
0
axis II
1174
214
axis I
5
0
65
50
0
52
200
202
203
231
644
109
1171
1091
1172
1087
659
663
667
214
1174
axis III

Figure 313

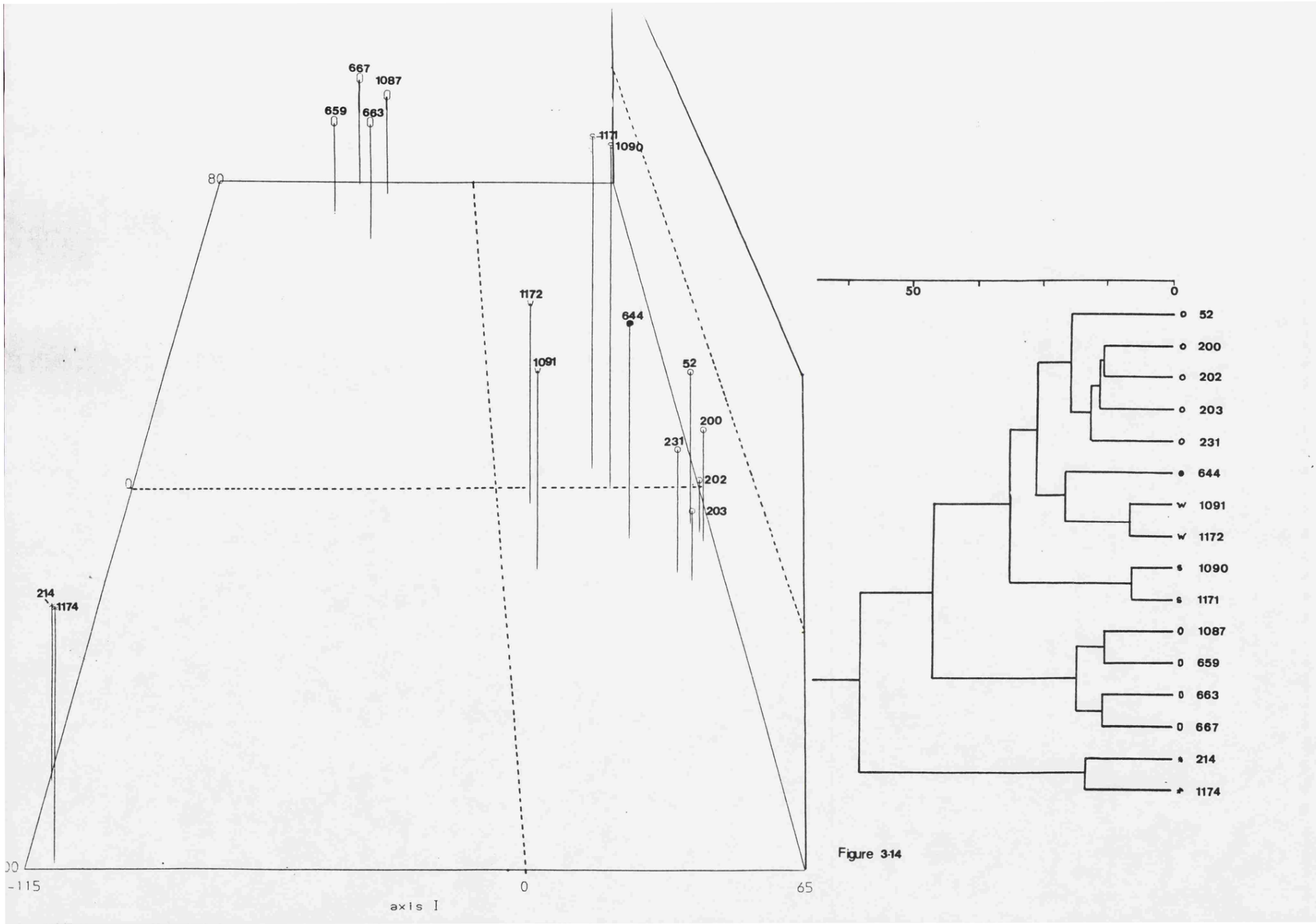
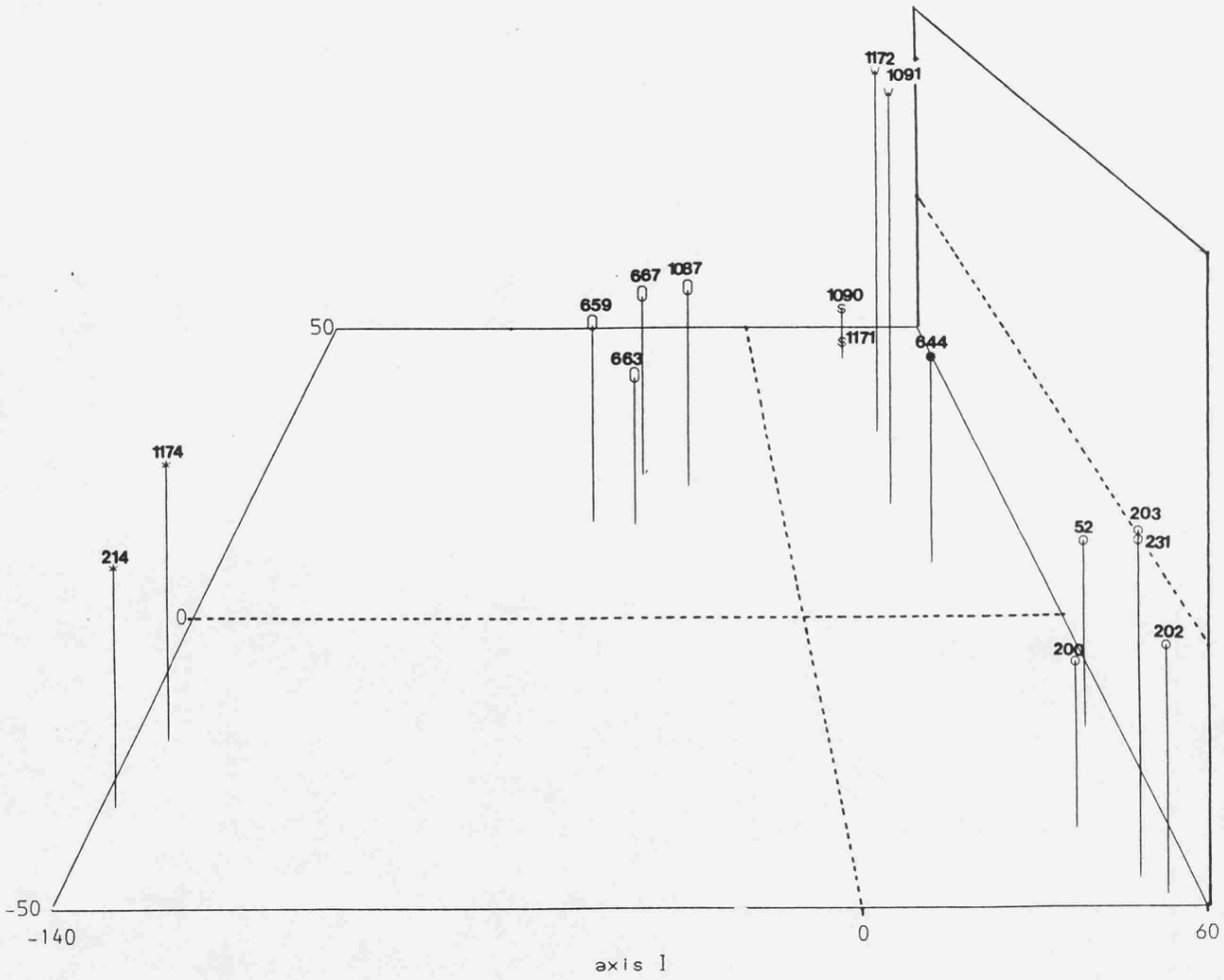

667
1087
659
663
1171
1090
80
1172
644
1091
52
200
231
202
203
214
1174
0
-115
0
65
axis I
50
0
52
200
202
203
231
644
1091
1172
1090
1171
1087
659
663
667
214
1174

Figure 3-14


1172
1091
667
1087
659
1090
1171
644
663
50
1174
214
203
52
231
0
202
200
-50
-140
0
60
axis I

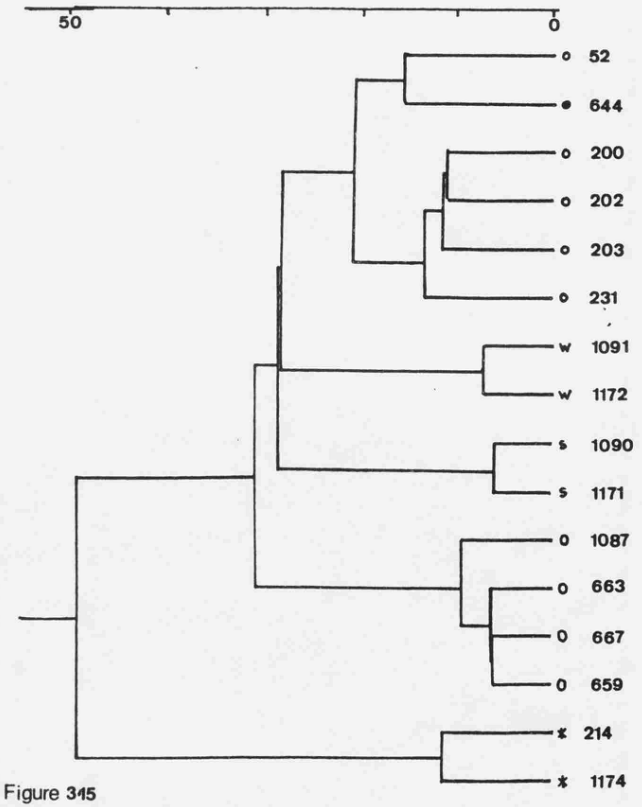

50
0
o 52
● 644
o 200
o 202
o 203
o 231
w 1091
w 1172
s 1090
s 1171
0 1087
0 663
0 667
0 659
✱ 214
✱ 1174

Figure 345

Frequently the choice of reference strains in DNA pairing studies is based on a previously determined taxonomic structure, or one reference strain is chosen and the second chosen because of its relationship with reference strain one, and so on. With *Listeria* the former case would produce a figure similar to Figure 3.11. The latter case was simulated by choosing C52 as reference strain one and C1090 - a moderately related strain (52.2 % pairing) as the second reference strain. From Table 3.5 this provided a 2 x 19 matrix. This incomplete matrix was used to produce the UPGMA dendrogram and three dimensional ordination in Figure 3.16. As there are only two reference strains used the ordination may only be drawn in a maximum of two dimensions, therefore all strains are of constant height in the third dimension. The two reference strains are outlying. *L. ivanovii* strains are clustered together, although closer to the *L. welshimeri, L.monocytogenes, L. innocua* strains than in Figure 3.7. The *L. monocytogenes* strains are well spread because of their large variation in % pairing with the reference strain, C52; they are also inseparable from the *L. welshimeri* and *L. innocua* strains.

A third reference strain, C214*a* was chosen; first using only the strains involved in the 16 x 16 complete matrix (Table 3.5) the principal component ordination and UPGMA were produced (Figure 3.17), then the extra three strains were inserted and any additional effects noted.

The *L. grayi* and *L. murrayi* strains remained distinct although not as close to each other as in Figure 3.7. The other strains in the study were compacted. The *L. ivanovii* strains, except C1087, formed a tight cluster on the edge of the main group. The two *L.seeligeri* strains, one of which was a reference strain, were distinct and on the perimeter of the main cluster. The other reference strain, *L. monocytogenes,* C52, was also pushed to the edge. *L. innocua,* C644 outlay from the main cluster of

Key

o *L. monocytogenes*

● *L. innocua*

0 *L. ivanovii*

s *L. seeligeri*

w *L. welshimeri*

* *L. grayi* and *L. murrayi*

Figure 3.16.

UPGMA dendrogram and Principal Component ordination employing C52 and C1090 as reference strains. All strains are of equal height in the third dimension as there are only two reference strains.

Figure 3.17 (p.106)

UPGMA dendrogram and Principal Component ordination employing C52, C1090, C214*a* as reference strains.

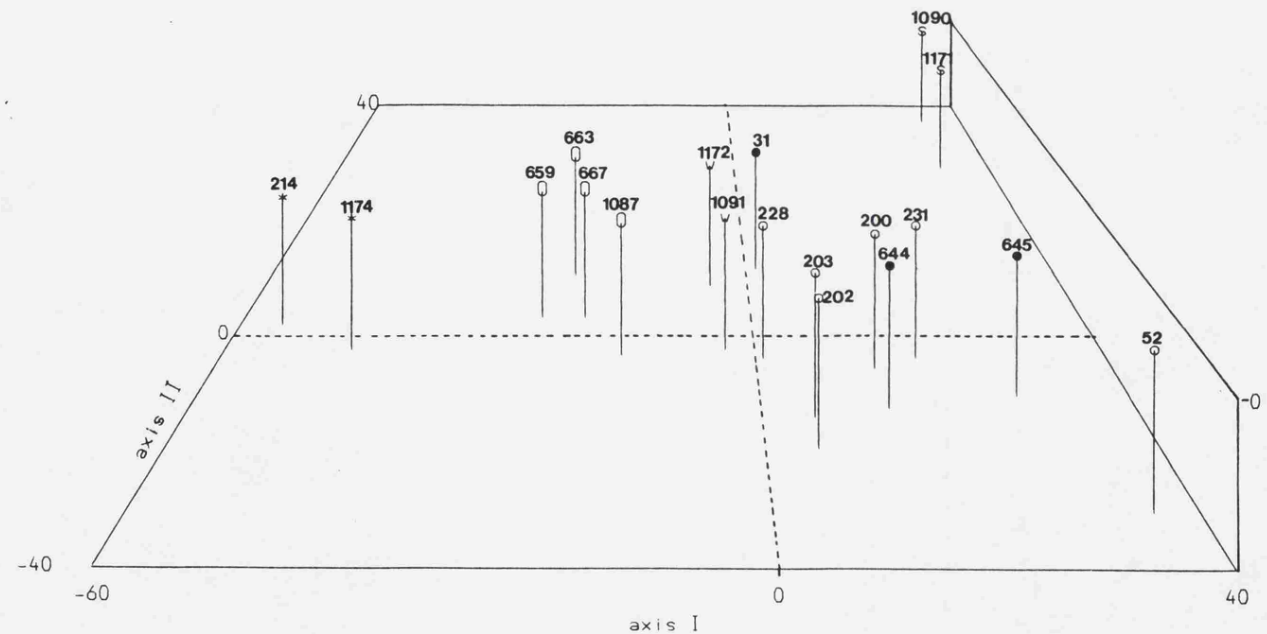
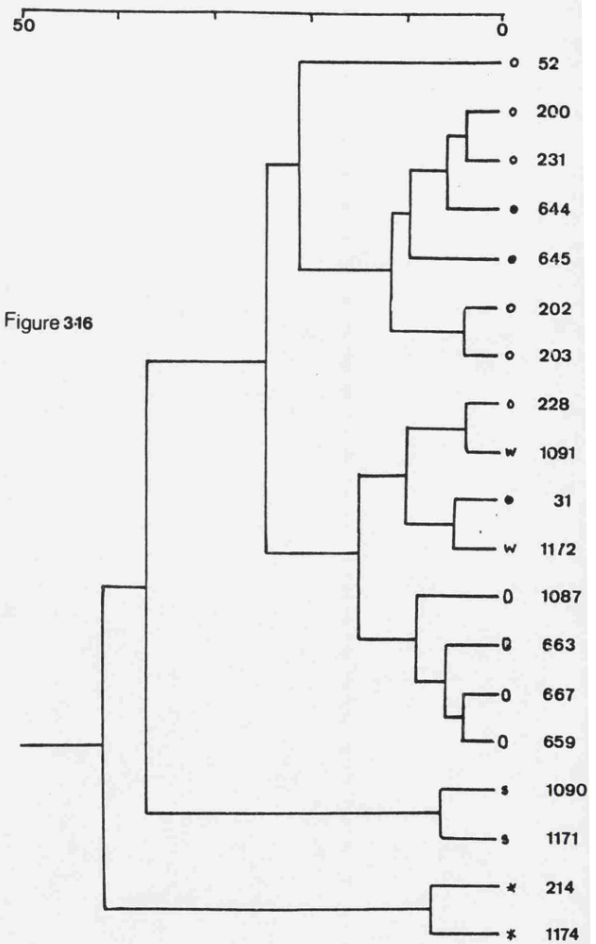
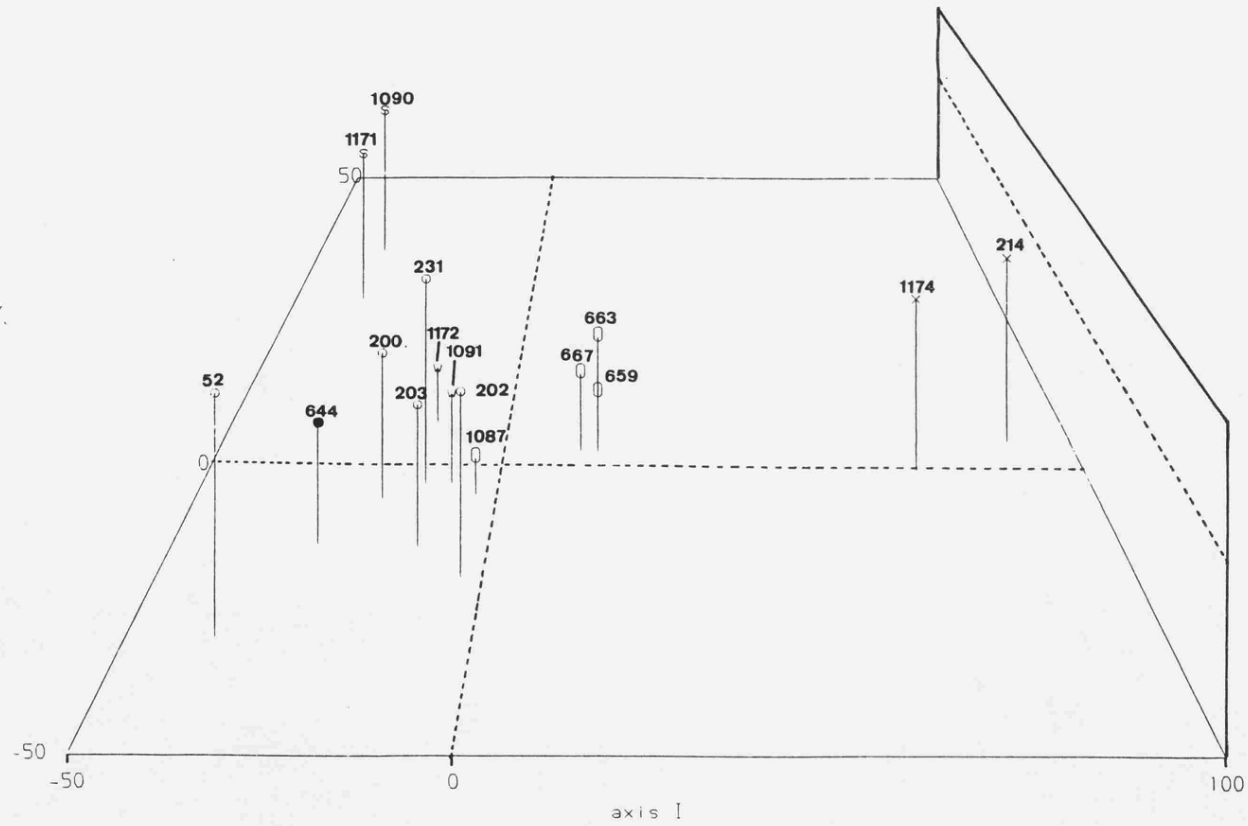

1090
1171
40
663
659
667
1087
1172
31
1091
228
214
1174
200
231
644
203
202
645
52
0
-40
-60
0
40
axis II
axis I

Figure 3.16


50
0
52
200
231
644
645
202
203
228
1091
31
1172
1087
663
667
659
1090
1171
214
1174


1090
1171
50
231
214
1174
663
200
1172
1091
667
659
52
203
202
644
1087
0
-50
-50
0
100
axis I

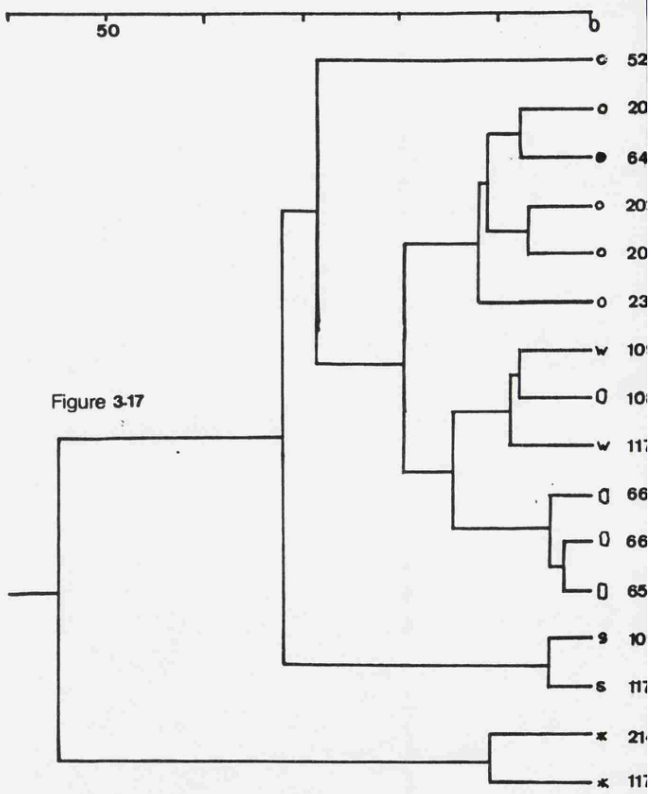

50
b
c 52
o 20
● 64
o 20
o 20
o 23
w 10
0 10
w 11
0 66
0 66
0 65
s 10
s 11
x 21
x 11

Figure 3-17

strains towards C52. All other strains were loosely grouped together including C1087 and the *L. welshimeri* strains.

The three additional strains affected the overall ordination producing a mirror image of the 3 x 16 plot. The additional strains were: C645 - a *L. innocua* strain - on the UPGMA this inserted close to C644 as expected; JS31 - also a *L. innocua* strain - this showed a greater affinity with the *L. welshimeri* strains and C1087 than with the other *L. innocua* strains; C228 - a *L. monocytogenes* strain - this branched with C1091 (*L. welshimeri*) and did not appear close to the other *L. monocytogenes* strains - probably because C52 was the only *L. monocytogenes* reference strain and this only shares 56.8 % pairing with C228. From the three-dimensional ordination the reference strains remained peripheral, C1087 rejoined other *L. ivanovii* strains to form a loose cluster. The other strains formed a range of relatedness with C645 outlying.

A different choice of three reference strains C52, C1087, C214*a* caused less distortion to the original structure. All three reference strains became peripheral but otherwise the ordination was not greatly effected.

When two closely related strains are used as reference strains the differences between the two strains are magnified. Figure 3.18 illustrates this, where only the two *L. seeligeri* strains are used as reference strains. Only the *L. grayi* and *L. murrayi* strains are in a comparable position to that in Figure 3.7. The *L. welshimeri, L. innocua* and *L. monocytogenes* strains are loosely clustered together. The ordination is almost one-dimensional (1.05, Table 3.7) due to the high % pairing between the two reference strains; the scatter about the axis represents the small differences in the relationships of C1090 and C1171 to the other strains.

The close relationship of C52 and C644 was explored using these two strains as the only reference strains, Figure 3.19. Here the dimensionality is 1.84, considerably more than that of Figure 3.18. The distance between the two reference strains has been exaggerated by the ordination and the *L. monocytogenes* group is easier to distinguish with respect to *L. innocua*.

Key

o *L. monocytogenes*

● *L. innocua*

0 *L. ivanovii*

s *L. seeligeri*

w *L. welshimeri*

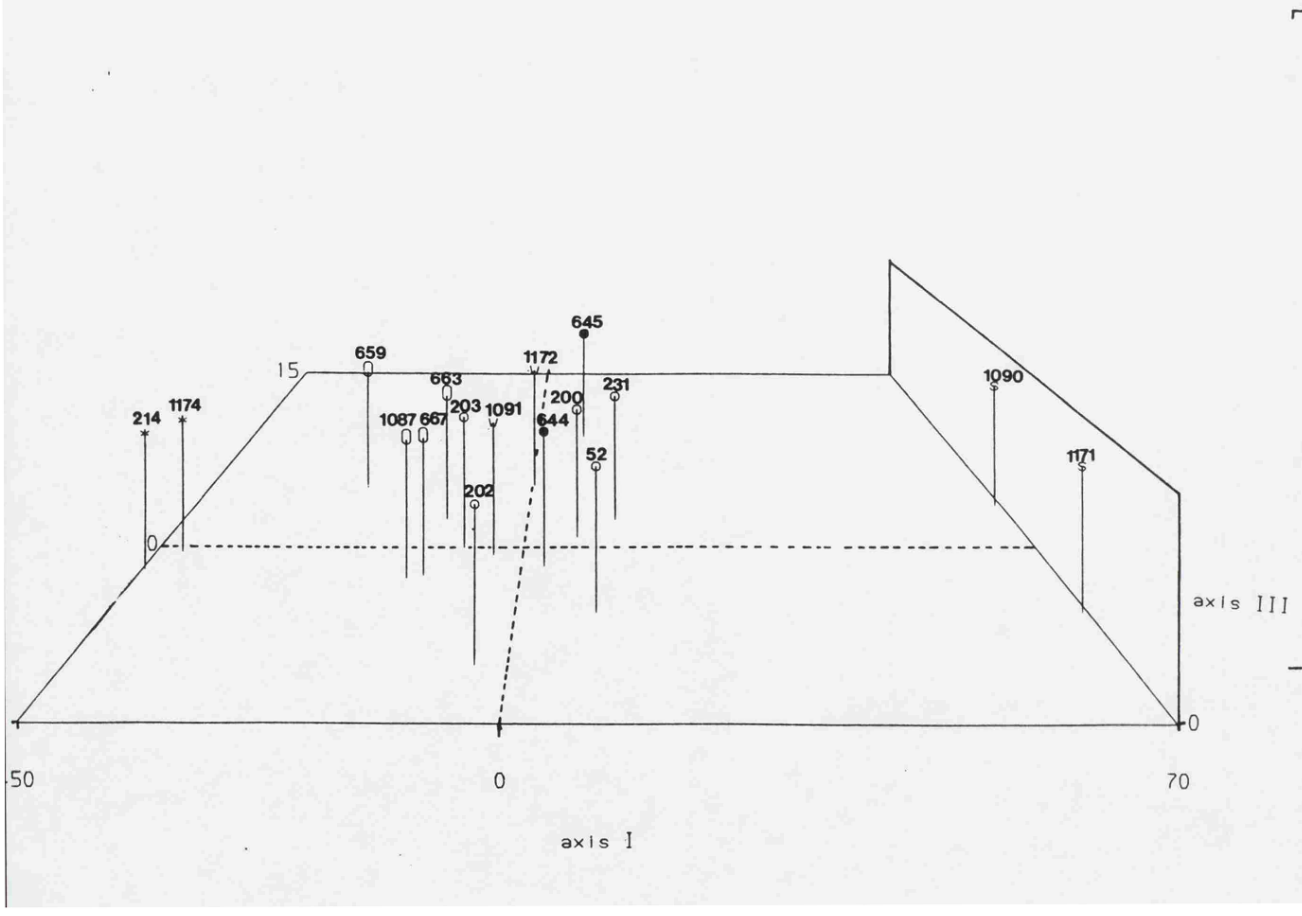
✱ *L. grayi* and *L. murrayi*

Figure 3.18.

UPGMA dendrogram and Principal Component ordination employing C1090 and C1171 as reference strains. All strains are of equal height in the third dimension as there are only two reference strains.

Figure 3.19 (p.111)

Principal Component ordination employing C52, and C644 as reference strains. All strains are of equal height in the third dimension.


15
0
214
1174
659
1087
667
663
203
1091
1172
645
644
200
231
52
202
1090
1171
axis III
0
50
0
70
axis I

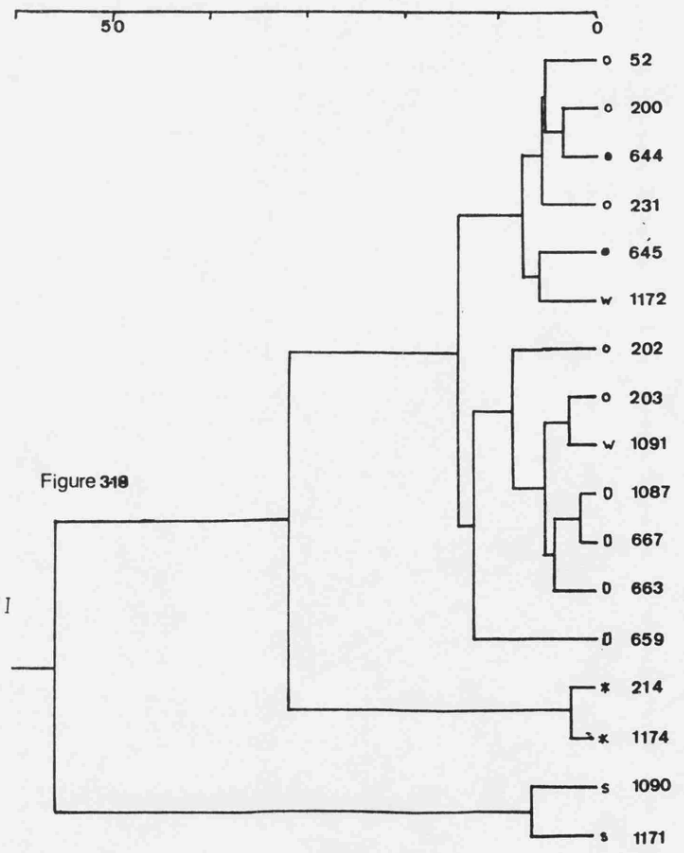

50
0
52
200
644
231
645
1172
202
203
1091
1087
667
663
659
214
1174
1090
1171

Figure 318

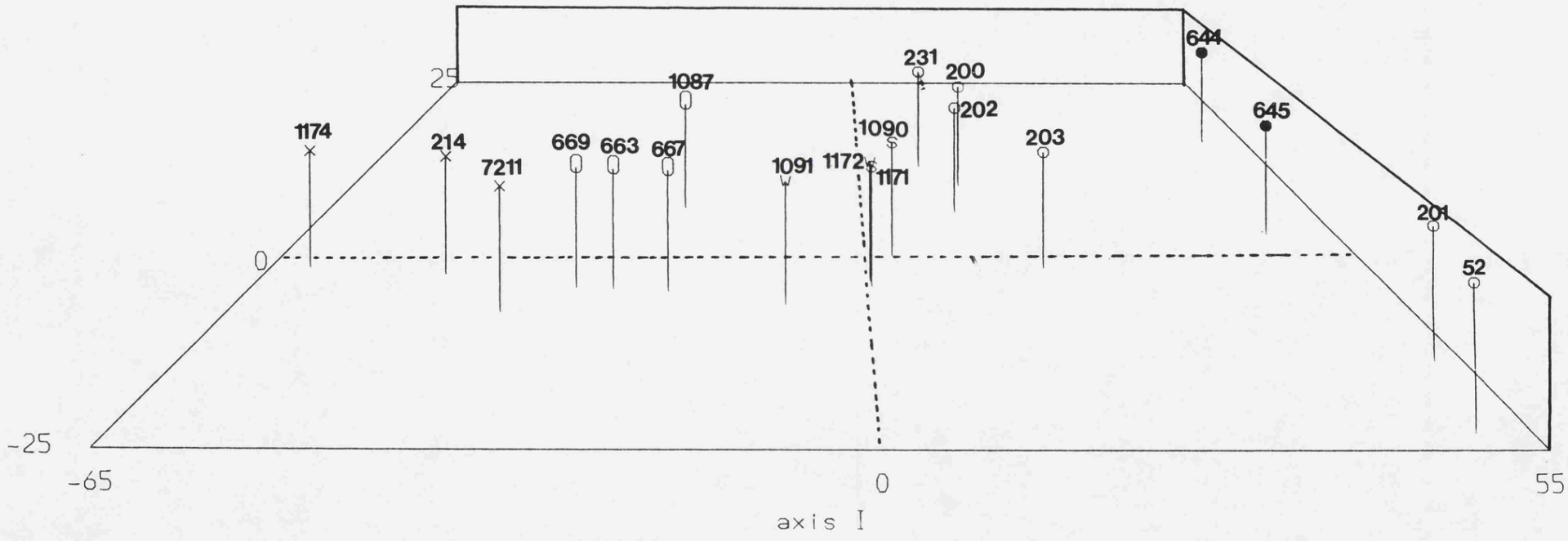

644
231
200
25
1087
202
645
1174
214
669
663
667
1090
203
7211
1091
1172
1171
201
0
52
-25
-65
0
55
axis I

Figure 3.19

DNA Pairing Data from Rocourt *et al.* 1982.

Figure 3.20 is a reconstruction of the data of Rocourt *et al.* 's data (1982) based on six reference strains (Appendix 8). Strains are divided into four groups or clusters: *L. murrayi* and *L. grayi* strains are distinct from other species although not from each other. *L. ivanovii* strains also form a distinct but loose cluster. The third group comprises six *L. seeligeri* strains, four tightly clustered and two outlying towards the fourth group. *L. innocua, L. monocytogenes* and *L. welshimeri* made up the largest cluster, however the *L. innocua* strains can be isolated from the *L. monocytogenes* and *L. welshimeri* strains in the third dimension. The *L. monocytogenes* reference strain represented as an otu is peripheral to the cluster as are the *L. welshimeri* and *L. innocua* reference strains.

In order to make comparisons between these data and my data similar figures were constructed. The *L. grayi* and *L. murrayi* strains were removed from the ordination (although neither species was previously represented by a reference strain); this caused some rotation about the axes. *L. seeligeri* strains remain separable from the other strains although some strains are now outlying (Figure 3.21).

The three dimensional ordination produced by removing the *L. monocytogenes* reference strain (Figure 3.22) (i.e. the ordination was based on a 5 x 47 matrix) separates the five species into distinguishable groups. As the data are in the form of an incomplete matrix, in some situations it is not possible to have all of the reference strains illustrated on the ordination (for example, if the pairing value between two reference strains is not available). Reference strains, where

represented, tend to be peripheral to their clusters; the *L. welshimeri* reference strain is pushed towards the *L. monocytogenes* group.

Removal of the two *L. ivanovii* strains to provide a 4 x 47 strip matrix (Figure 3.23) results in even clearer separation of the five species. *L. monocytogenes* strains form a long sprawling cluster but it is easily separated from the *L. welshimeri* and *L. innocua* strains. *L. seeligeri* strains also form an almost linear cluster.

A similar picture was obtained when the *L. seeligeri* reference strain is removed (Figure 3.24); the *L. seeligeri* cluster has moved towards the centroid instead of the *L. ivanovii* group.

When the *L. welshimeri* reference strain was removed (Figure 3.25), the *L. welshimeri* strains are drawn towards the *L. monocytogenes* group and are barely distinguishable as a group. *L. innocua* strains can be separated from *L. monocytogenes* in the third dimension. *L. seeligeri* strains are a distinct group.

Removal of the *L. innocua* reference strain brought *L. innocua* and *L. monocytogenes* strains very close together; the two species are separated by a very small distance on the third axis. *L. seeligeri* strains form a tight cluster and a looser cluster of two strains as a 'satellite group'. The distance between the two *L. seeligeri* 'groups' is larger than the distance between some *L. innocua* and some *L. monocytogenes* strains.

Table 3.8 shows the dimensionality of Figures 3.21-3.27. The dimensionality is low, and in most cases that of the 3-dimensional ordination (m') is not much lower than that of the 4-6 dimensions available with the maximum use of the information given (n').

Key

o *L. monocytogenes*

● *L. innocua*

D *L. ivanovii*

s *L. seeligeri*

w *L. welshimeri*

✱ *L. grayi* and *L. murrayi*

Figure 3.20.

A reconstruction of the data of Rocourt *et al.* (1982) using strains of *Listeria.* A strip matrix of 6 x 52 strains was formed from the published data and the Principal Component ordination produced.

Figure 3.21 (p.116)

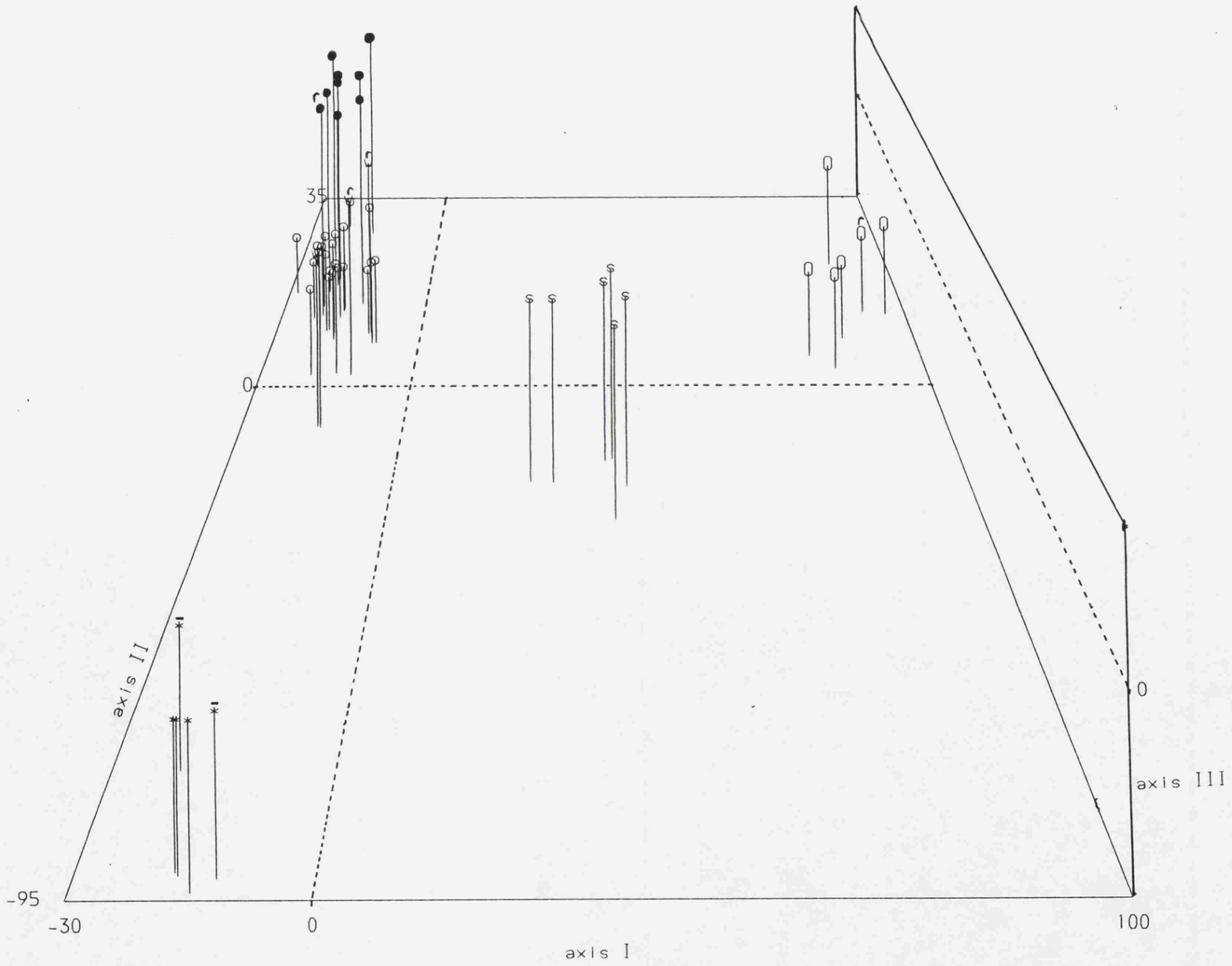
The Principal Component ordination of Rocourt *et al.* (1982) without featuring any *L. grayi* or *L. murrayi* strains.

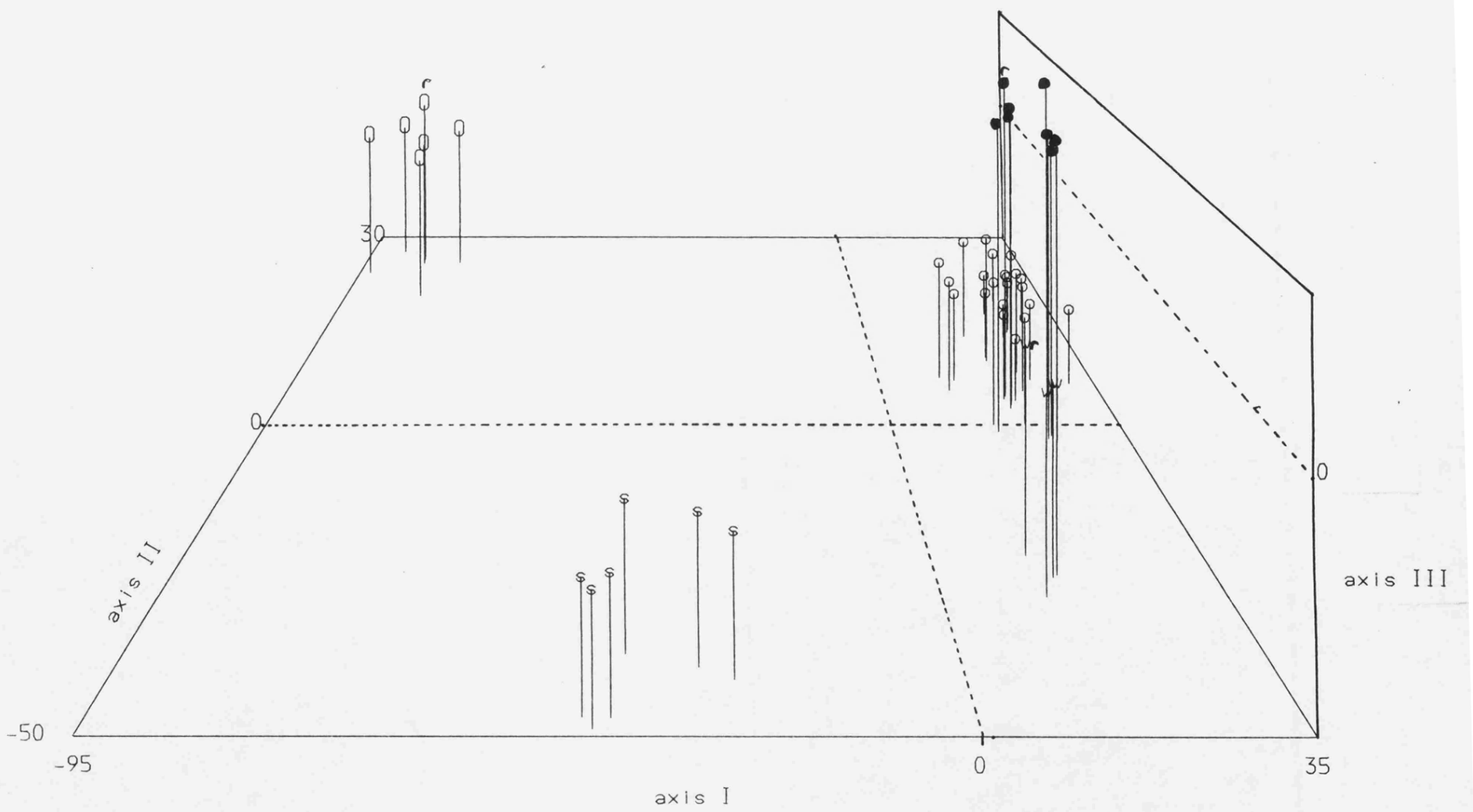
Figure 3.22 (p.117)

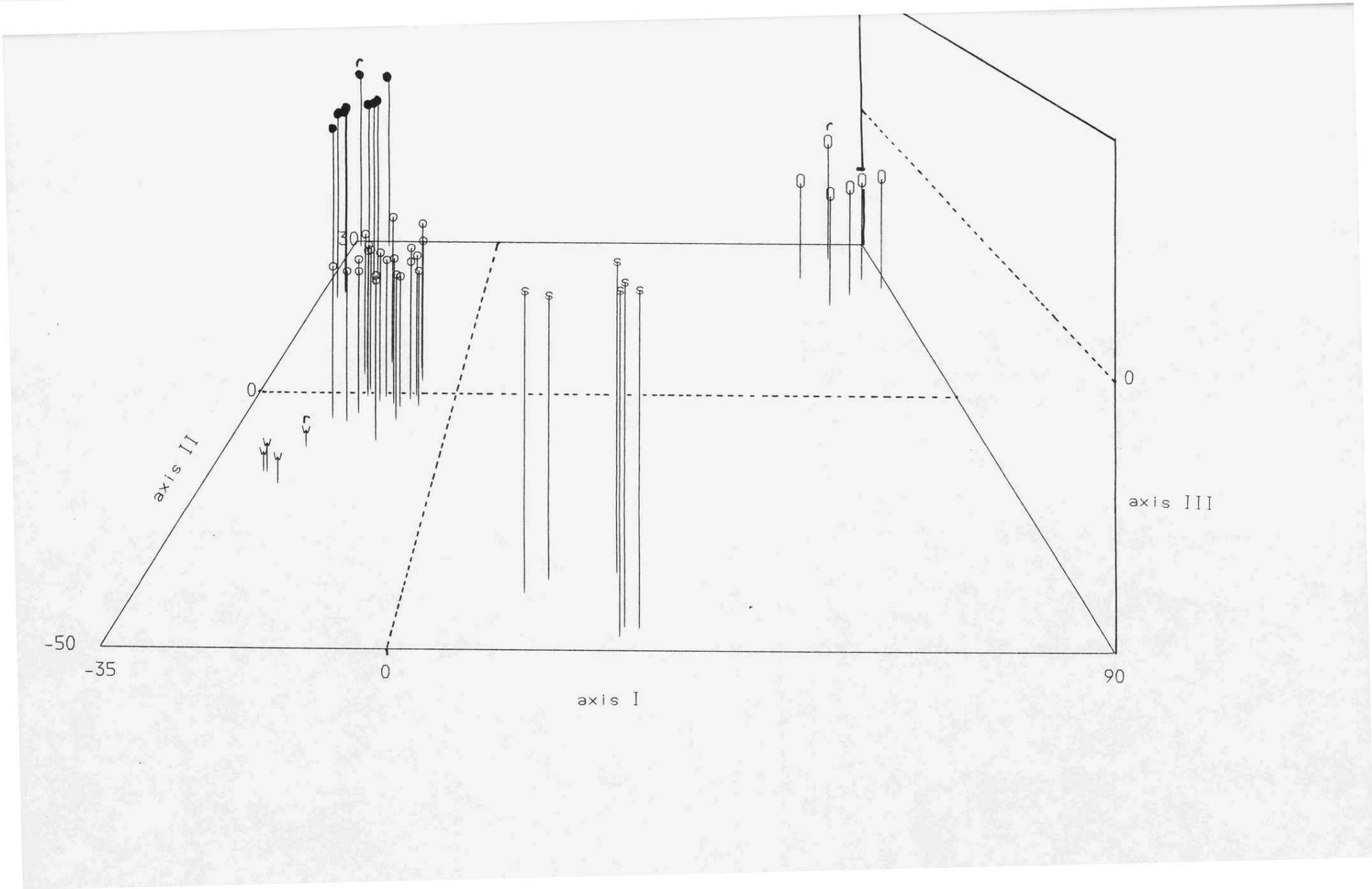
The Principal Component Ordination constructed from the data of Rocourt *et al.* (1982) without employing *L. monocytogenes* as a reference strain (5 x 47 matrix).

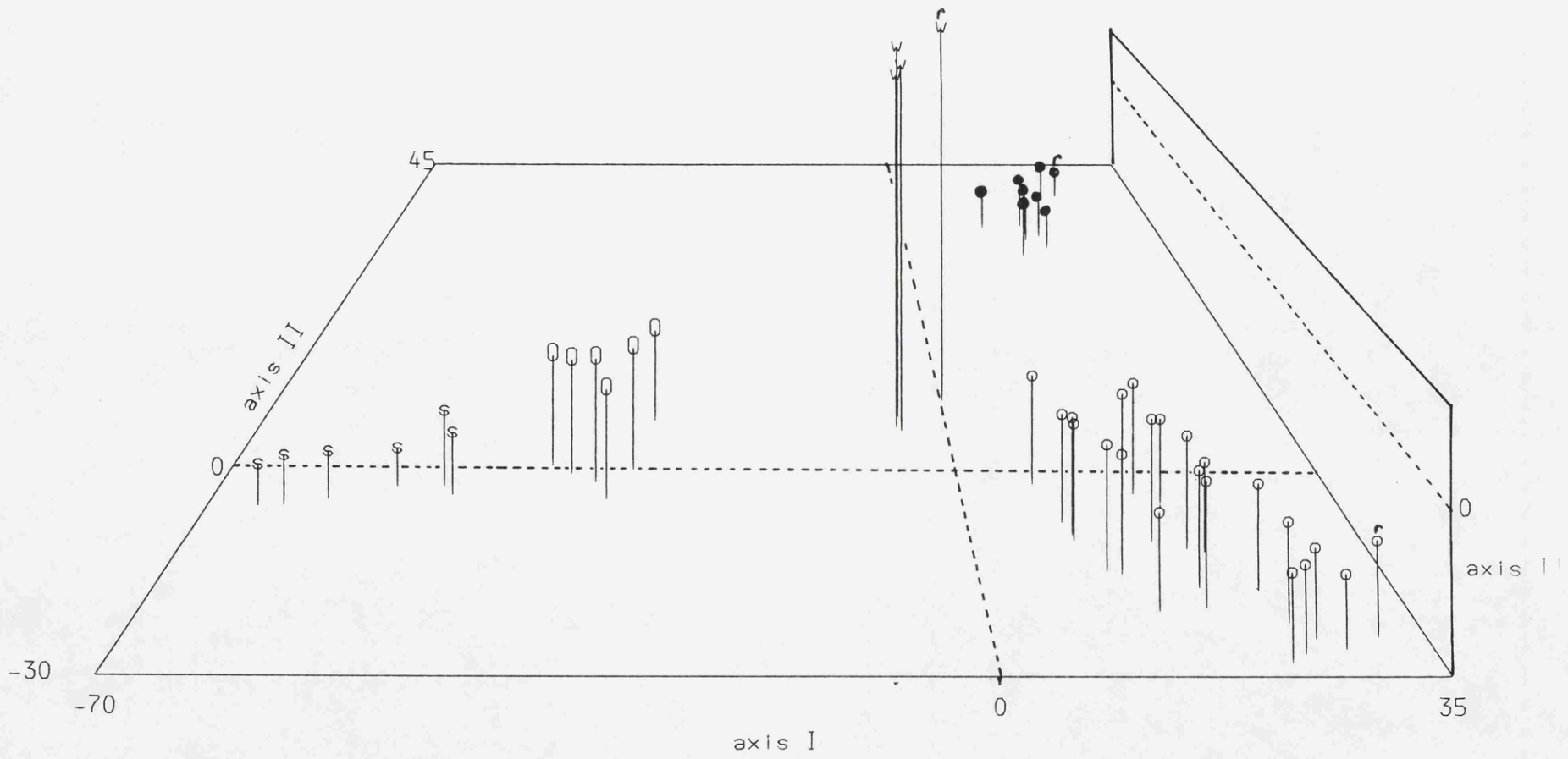
Figure 3.23 (p.118)

The Principal Component ordination constructed from the data of Rocourt *et al.* (1982) without employing *L. ivanovii* as a reference strain (4 x 47 matrix).


35
0
axis II
-95
-30
0
axis I
0
axis III
100


30
0
axis II
-50
-95
0
35
axis I
0
axis III


30
0
axis II
-50
-35
0
axis I
90
0
axis III


45
axis II
0
-30
-70
0
35
axis I
0
axis III

Key

o *L. monocytogenes*

● *L. innocua*

0 *L. ivanovii*

s *L. seeligeri*

w *L. welshimeri*

Figure 3.24.

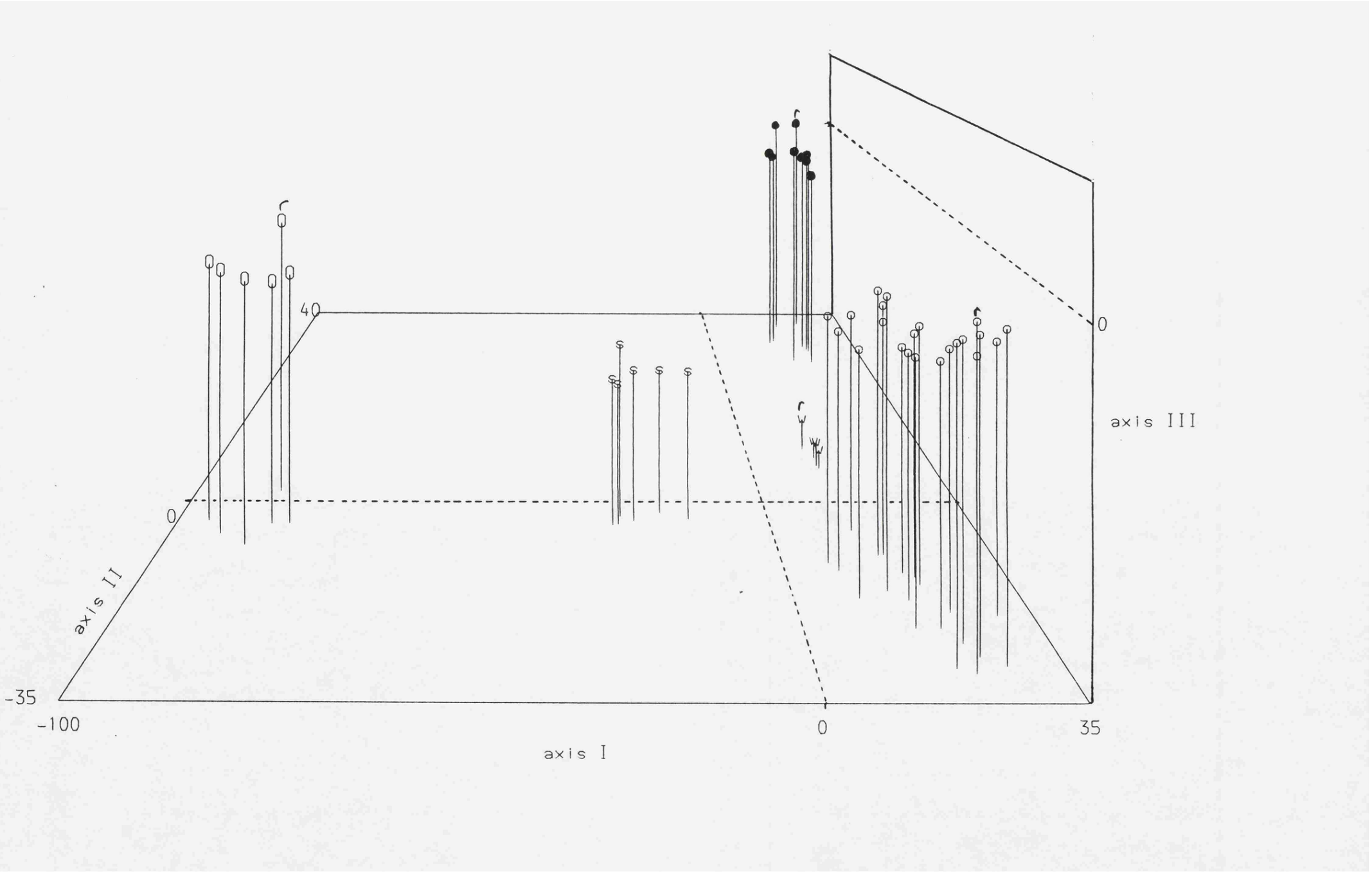
A reconstruction of the data of Rocourt *et al.* (1982) using strains of *Listeria* without employing *L. seeligeri* as a reference strain; the Principal Component ordination produced is based on a 5 x 47 strip matrix.

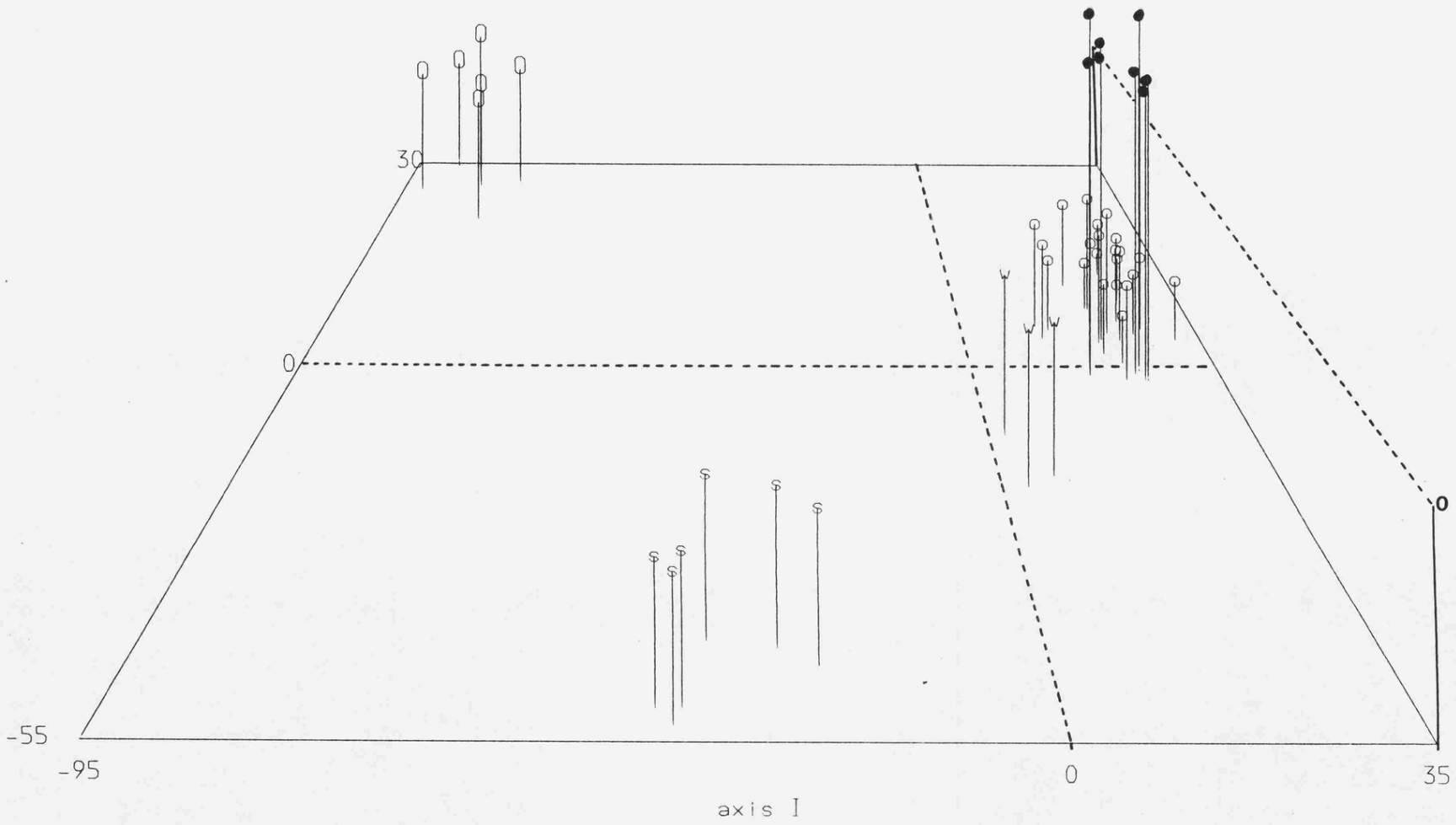
Figure 3.25 (p.121)

The Principal Component ordination of Rocourt *et al.* (1982) without featuring any *L. welshimeri* strains as reference strains.

Figure 3.26 (p.122)

The Principal Component Ordination constructed from the data of Rocourt *et al.* (1982) without employing *L. innocua* as a reference strain.


40
0
axis II
-35
-100
0
35
axis I
0
axis III


30
0
-55
-95
0
35
axis I

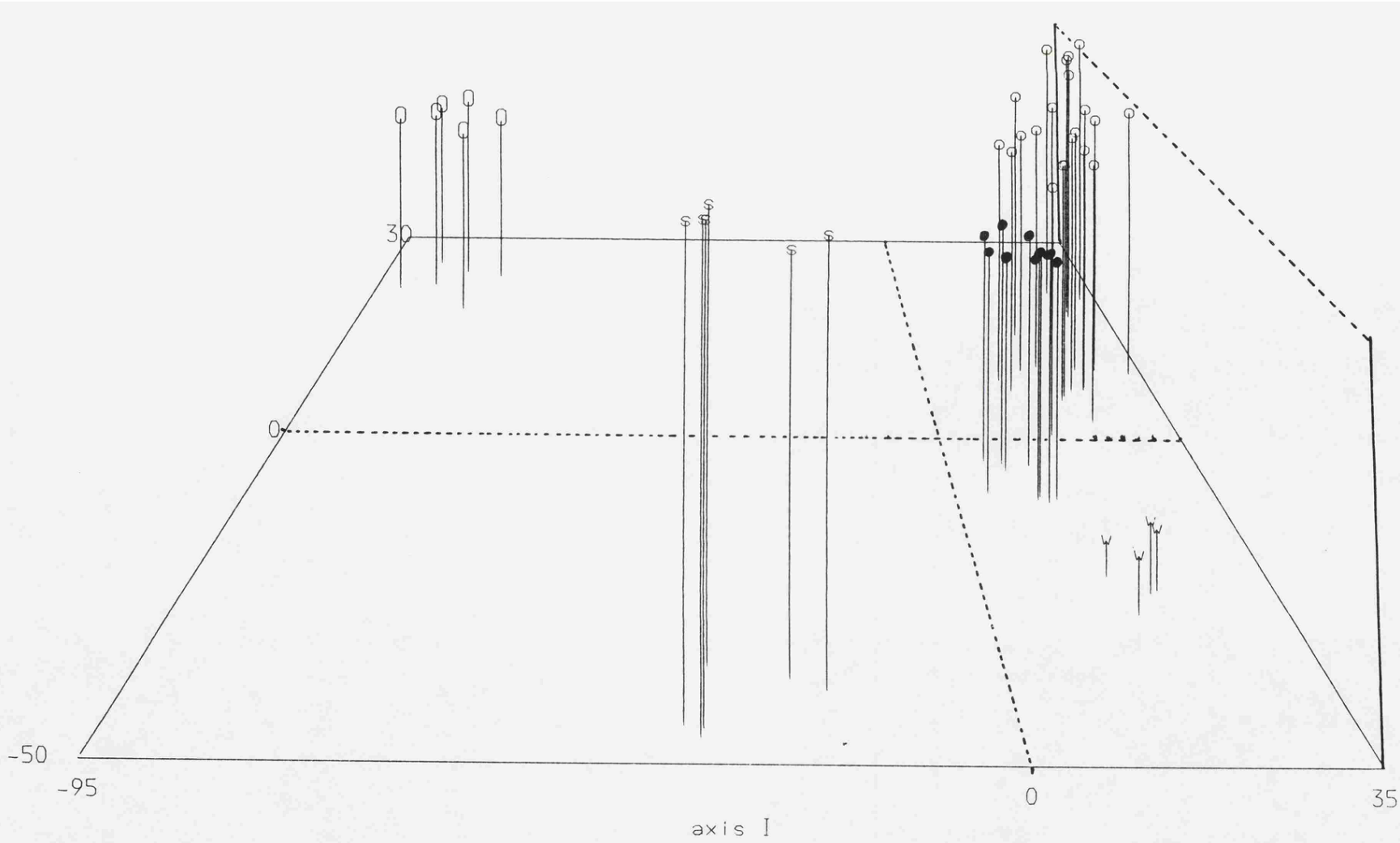

30
0
-50
-95
0
35
axis I

Table 3.8 Eigenvalues and % Variation for *Listeria* from Rocourt *et al.* 1982.

Figure number	*c*	λ_1	λ_2	λ_3	% variation in first 3 axes	*n'*	*m'*
3.21	6	66736	50379	18569	82.3	3.51	2.51
3.22	6	66253	20170	18213	87.8	2.69	2.14
3.23	5	55315	19396	13088	93.1	2.42	2.12
3.24	4	37392	18261	12550	93.4	2.79	2.46
3.25	5	48633	32958	25311	68.8	3.00	2.79
3.26	5	61813	18311	15195	93.0	2.47	2.24
3.27	5	61073	18310	15155	93.0	2.38	2.08

Bacillus circulans Data. (Nakamura and Swezey 1983*a*).

The 'true' relationships between the 17 strains, obtained from the complete matrix of d_{jk} values (Appendix 1) is shown in Figure 3.27. The three-dimensional model, Figure 3.28, represents the first three of the 16 possible dimensions.

Strains 1, 2, 6, 7, 8, 9, 10, 11 and 13 form a major cluster. The first eight are a tight cluster, whereas strain 13 is a satellite of the cluster, lying some distance away. Strains 3 and 4 form a minor, looser, cluster and strain 5 is a satellite of this. Strains 12, 14, 15, 16 and 17 are outlying singletons. Of these, 17 is the most outlying. The percentage of variation accounted for by the first three dimensions is 77.9 (Table 3.9). The effective dimensionality is only 4.5, a good deal less than the nominal dimensionality of 16; reduction to three dimensions reduces the effective dimensionality to $m' = 2.4$ (Table 3.9).

It is against the configurations of Figure 3.28 that the others were judged. It was noted that strains 1 and 9 both derive from ATCC 4516, and the differences in value for these two in Appendix 1 are probably due to the experimental error of estimating DNA pairing. Strains 6 and 13 both derive from Ford 26 but I am less confident that experimental error completely accounts for the differences between values for these strains [see 3.10].

The results from Principal Coordinate analysis of d^* coefficients using all 17 strains as reference strains is shown in Figure 3.30. The taxonomic structure is essentially correct, and the distortion is small. Within the major cluster strain 8 is now closer and strain 13 relatively a little less close. The effective dimensionality has been reduced, and consequently there is more variation in the first three axes (91.3 %,

Table 3.9 Eigenvalues and % Variation over the First Three axes of the 3-Dimensional Ordinations for *Bacillus circulans* data.

Figure Number	c	λ_1	λ_2	λ_3	% variation in first 3 axes	n'	m'
3.28	NA	16688	7362	5392	77.9	4.51	2.40
3.29	17	239490	19234	10847	91.3	1.54	1.26
3.30	3	18203	6972	1997	100	1.92	1.92
3.31	2	27856	6713	0	100	1.46	1.46
3.32	2	28101	6010	0	100	1.41	1.41
3.33	2	51171	380	0	100	1.01	1.01
3.34	3	32576	6440	3666	100	1.63	1.63
3.35	8	36354	15267	9985	75.3	3.80	2.29

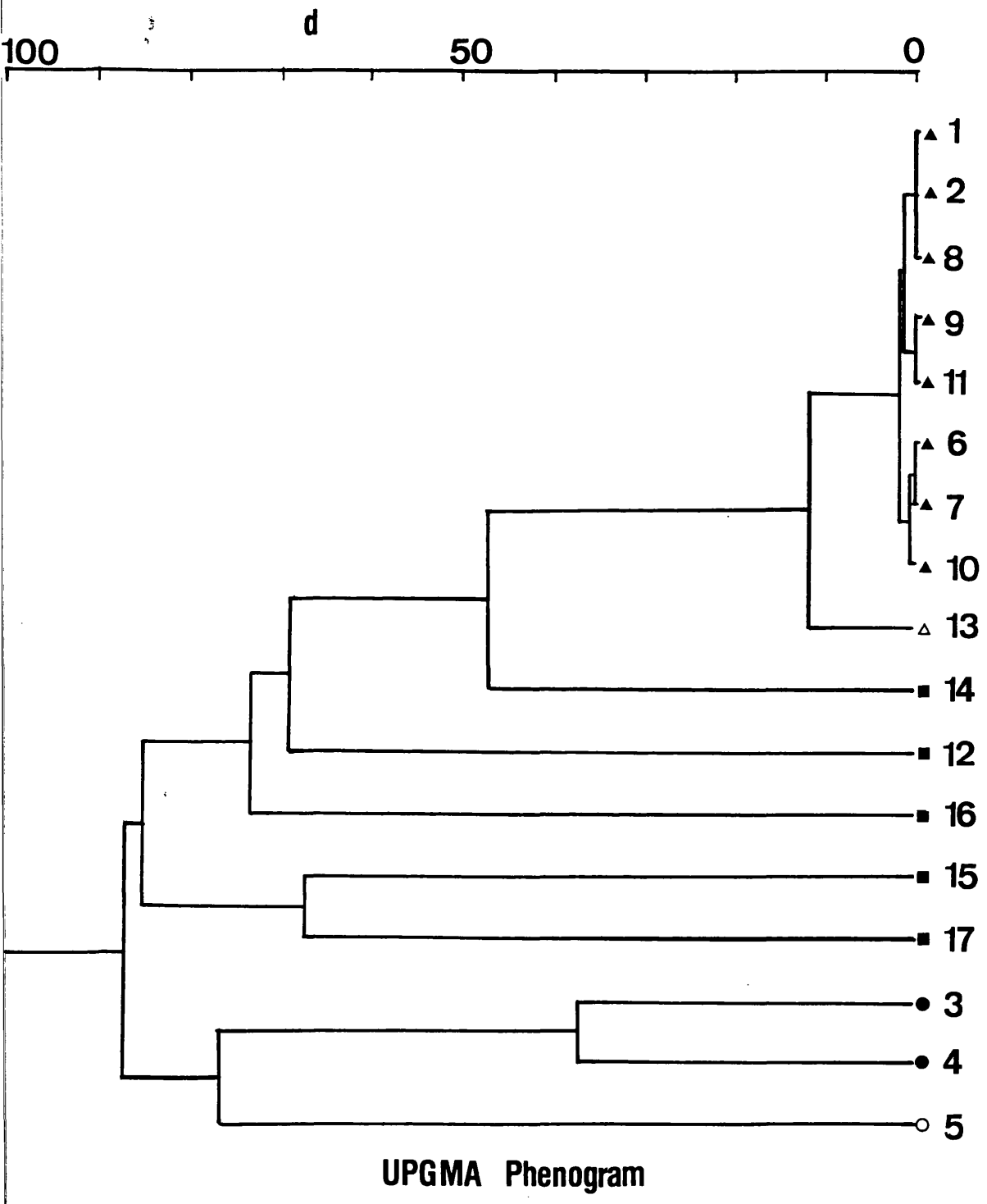


Figure 3.27 The UPGMA dendrogram of dissimilarities, treated as distances, *d* when all 17 strains are used as reference strains. Symbols : solid triangles, members of main cluster; open triangle, satellite of main cluster; solid circles, members of minor cluster; open circle, satellite of minor cluster; solid squares, outlying singletons.

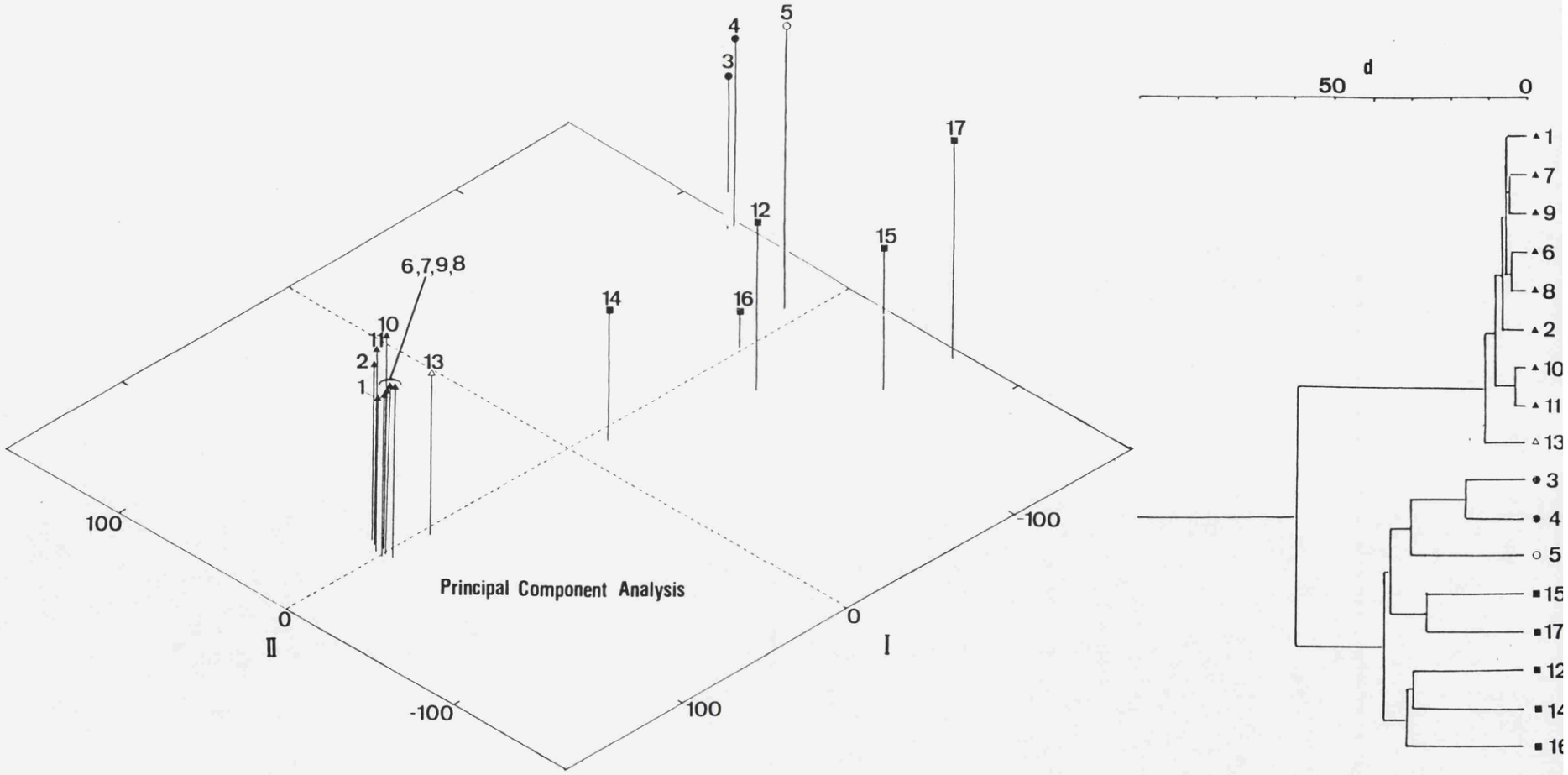


Figure 3.28 Principal Component analysis of percent DNA-DNA dissimilarities of Appendix 1 when all 17 strains are reference strains. Third axis vertical, baseplate -90. Symbols as in Figure 3.27.

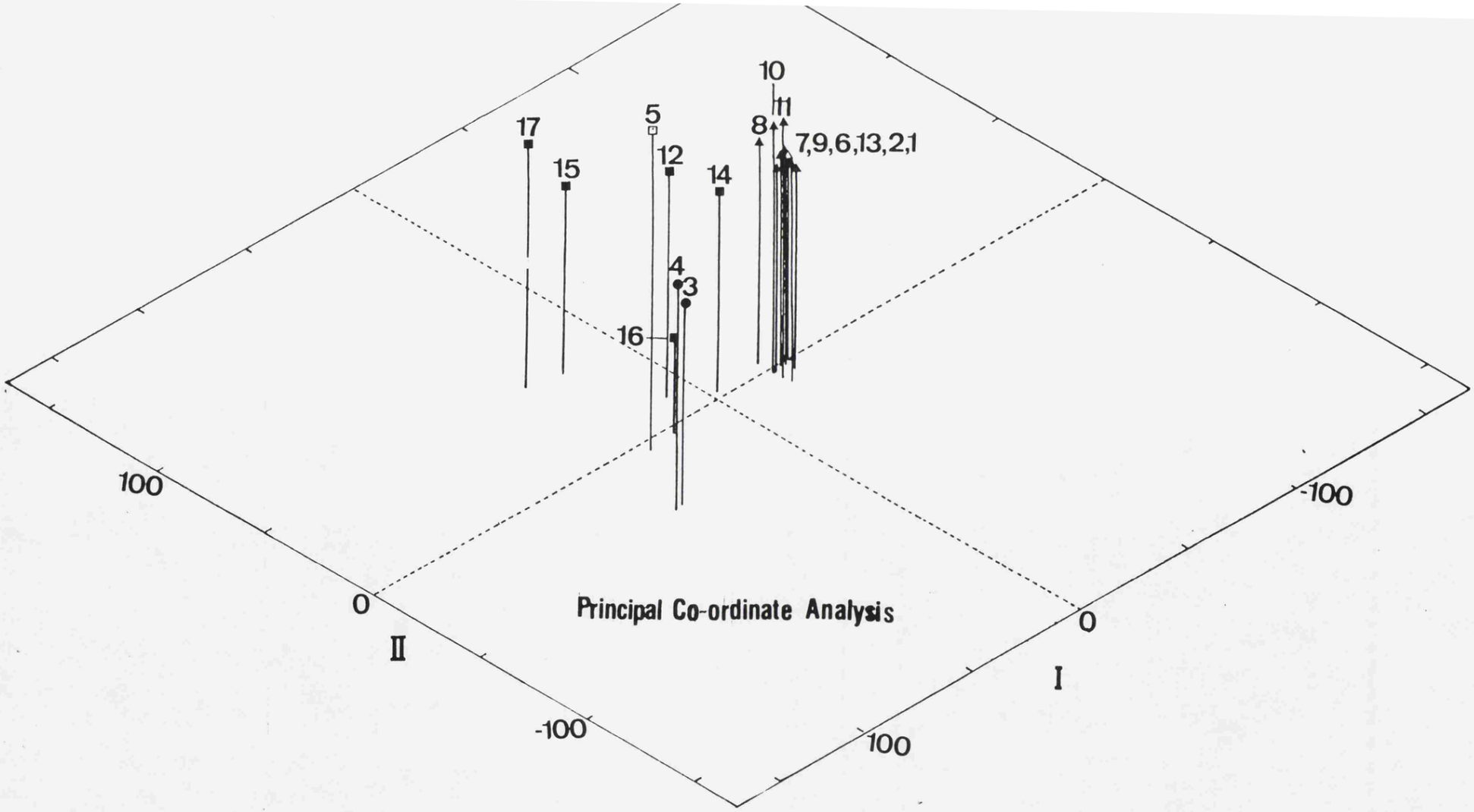


Figure 3.29 Three dimensional representation of relationships from Principal Coordinate analysis of percent DNA-DNA dissimilarities of Appendix 1. First two principal axes are horizontal, the third axis is vertical, baseplate -90. Symbols as in Figure 3.27.

3.9). The arbitrary reflection on axes I and II can be seen by comparing Figure 2.3c with Figure 2.3e (see Methods and Materials, 2.13).

Figure 3.30 shows the results from three reference strains, the members of the minor clusters, 3 and 4, and its satellite, 5 (marked with asterisks). There is bizarre distortion. The minor cluster has greatly expanded, and all the remaining strains, including singletons, have been compressed into an apparently tight but false group near the centroid. This behaviour is particularly significant, because a choice such as this could easily occur if the first strains examined happened to be from a loose cluster.

When strains 3 and 4 were employed, without strain 5, the results were similar: the two strains of the minor cluster became widely separated and all other strains (including strain 5) were in one compact group.

Figure 3.31 shows the results from two reference strains, one from the major cluster, 1, and a singleton, 15. The structure is remarkably good: both clusters are easily recognised and the other strains are placed appropriately. Because $c = 2$ all the variation is in the first two axes, and the points are all at a constant height above the baseplate.

There is notable compression of the loose minor cluster together with its satellite strain 5. The reference strain 15 is now very peripheral. The singletons 12 and 16 are close, giving the false impression that they form the nucleus of a cluster.

Another similar choice, strain 10 from the major cluster and the singleton 17, resulted again in compression of the other singletons and the minor cluster into one group; in this instance one might easily be misled into thinking that those singletons belonged to the minor cluster.

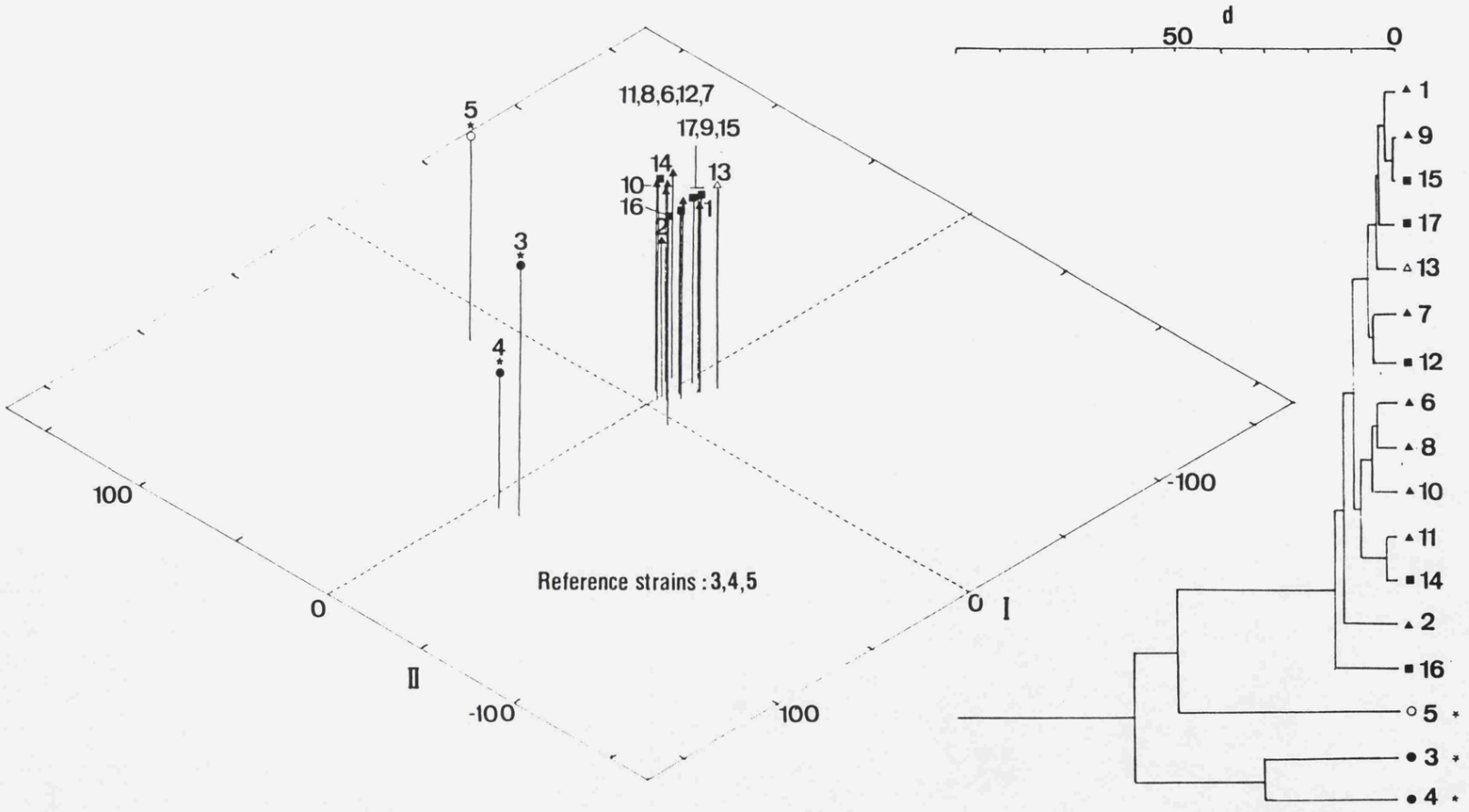


Figure 3.30 Principal Component analysis of Appendix 1 employing strains 3, 4 and 5 as reference strains (marked with asterisks), and corresponding UPGMA dendrogram from derived distances. Other symbols and conventions as Figures 3.27 and 3.28.

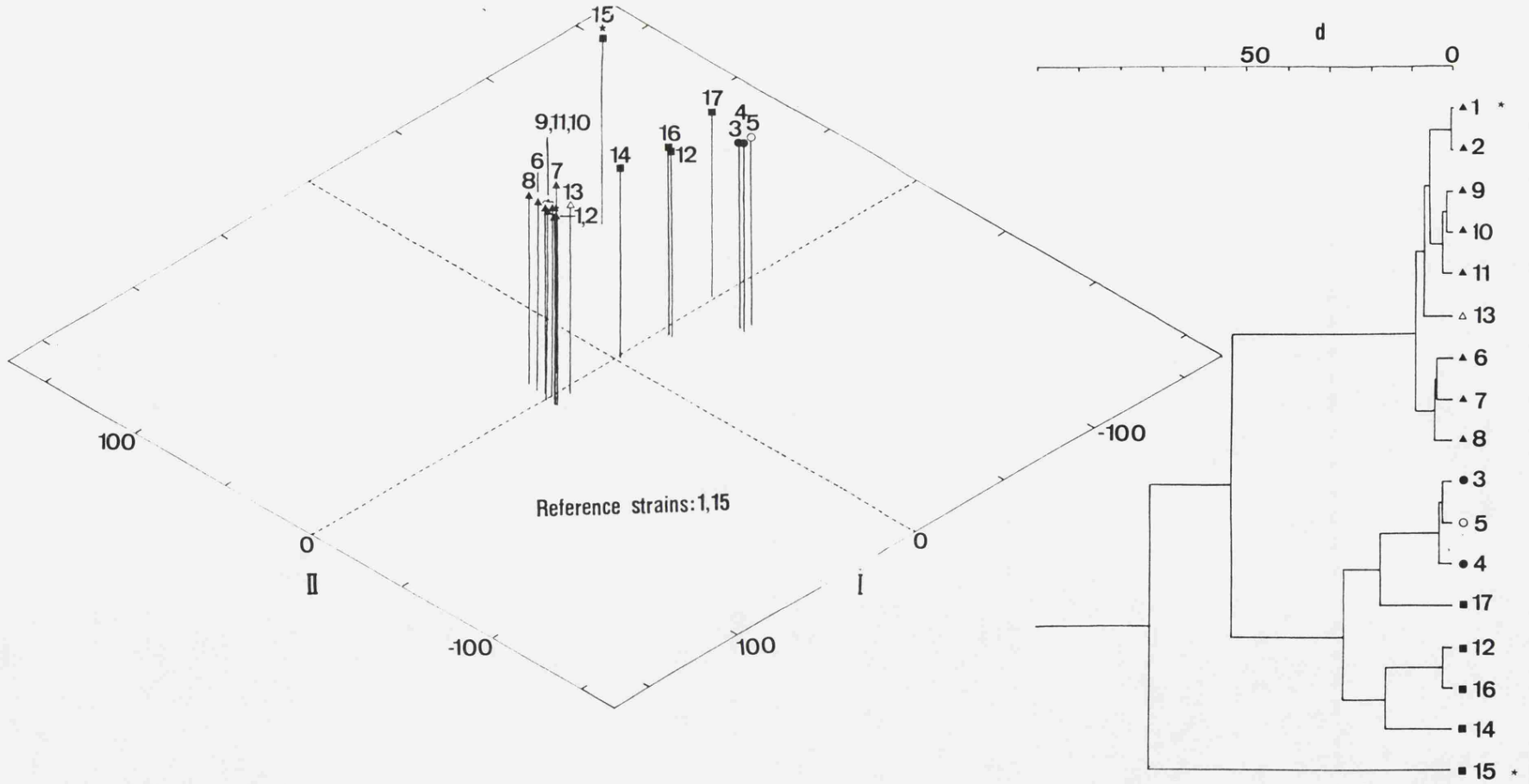


Figure 3.31 Principal Component analysis of Appendix 1 employing strains 1 and 15 as reference strains (marked with asterisks). Other symbols and conventions as Figures 3.27 and 3.28.

Figure 3.32 shows the results from a different pair of reference strains, one (strain 8) from the major cluster and the other (strain 4) from the minor cluster. There is obvious distortion. Only the major cluster is well defined; the minor cluster is dispersed and allied with the pulled-in singletons and the satellites in a loose false cluster; this could be very misleading. The tendency for reference strains to assume peripheral positions (Sneath, 1983) is well shown by strain 4. Strain 2 is now relatively peripheral in the main cluster. Further, the strains of minor cluster 3 and 4 are widely separated relative to the other strains (cf Figure 3.28).

It is not entirely clear why strain 2 has become so peripheral to its own cluster. It is probably due to non-euclidean properties of certain relationships. Strains 2 and 8 appear to be identical when compared directly ($d_{2,8}$ = 0, Appendix 1), yet other values involving them differ considerably. Thus $d_{2,4}$ is 73 % and $d_{8,4}$ is 89 %, which implies that strain 2 is closer to the reference strain 4 than it is to strain 8. Strain 2, therefore, tends to be moved out by its comparative closeness to strain 4 when only similarities involving strains 4 and 8 are available.

Figure 3.33 shows results from choosing two strains, 1 and 13, from the major cluster. These are so close that they are nearly equivalent to one reference strain: they represent, one might say, almost a view from a single point. Consequently the structure is almost one-dimensional (shown also by the low effective dimensionalities; Table 3.9). Strains not of the major cluster are compressed into linear clusters. Strain 14 is now close to the other singletons.

Another choice of two close strains from the major cluster, 1 and 2, gave a similar result. The configuration was again almost linear; there

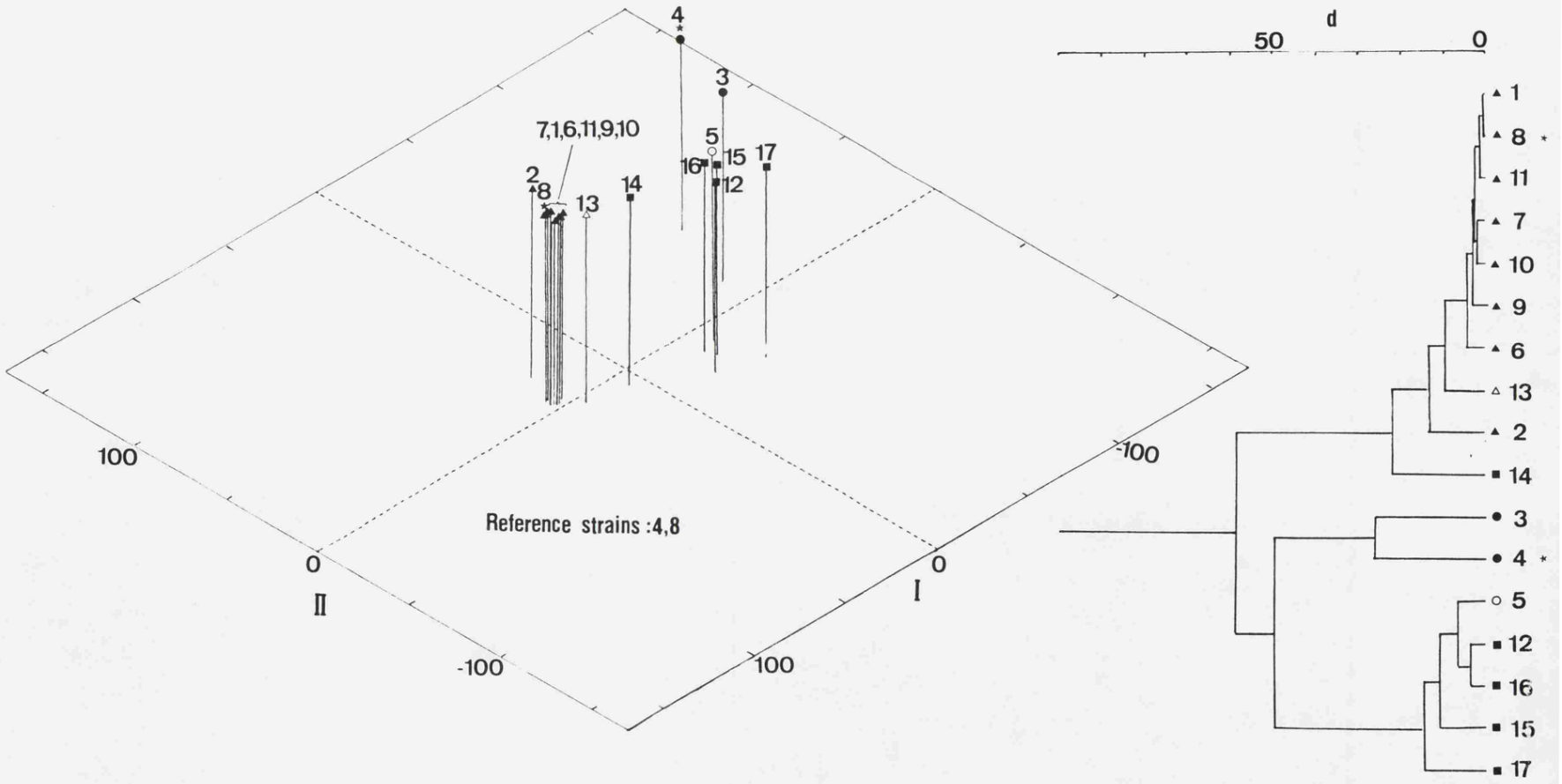


Figure 3.32 Principal Component analysis of Appendix 1 employing strains 4 and 8as reference strains (marked with asterisks). Other symbols and conventions as Figures 3.27 and 3.28.

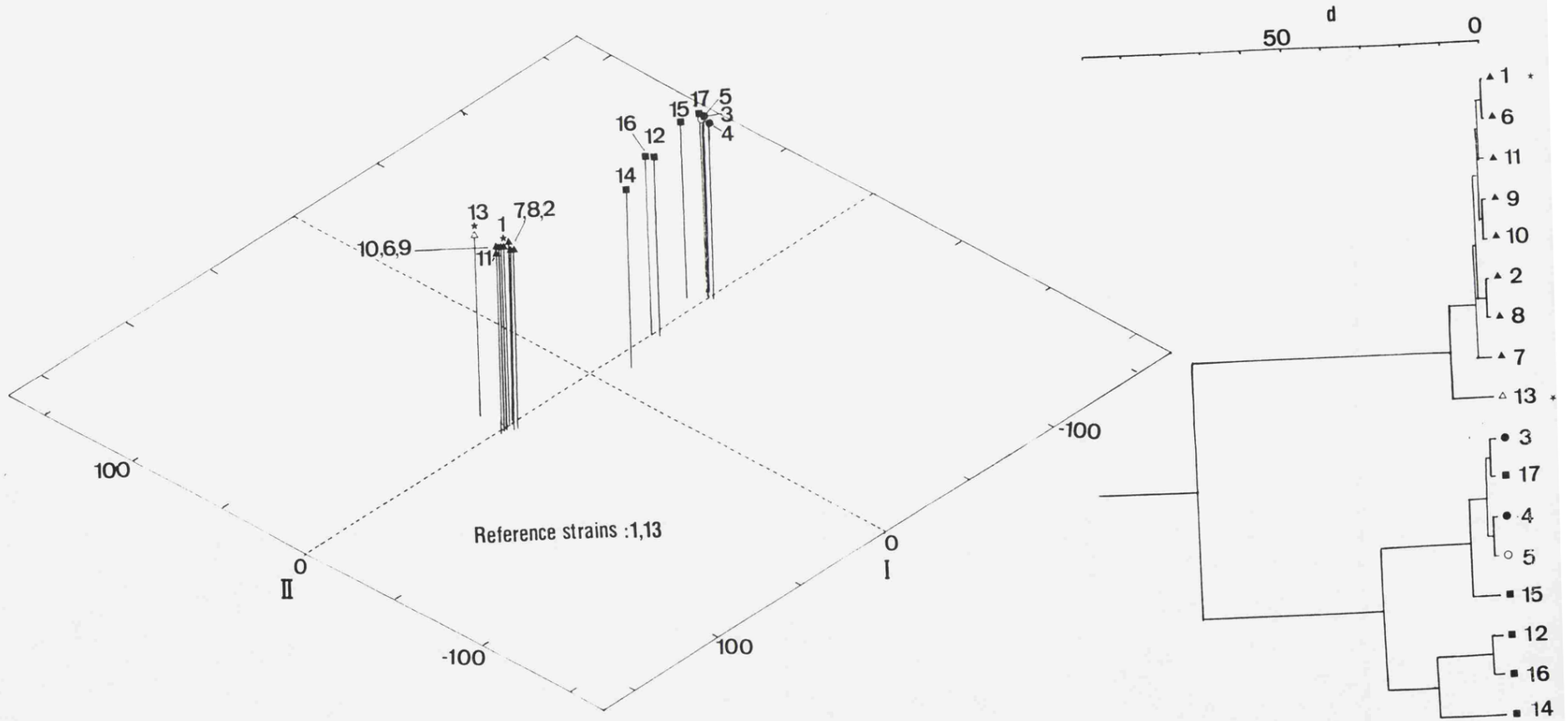


Figure 3.33 Principal Component analysis of Appendix 1 employing strains 1 and 13 as reference strains (marked with asterisks). Other symbols and conventions as Figures 3.27 and 3.28.

was a false cluster composed of the singletons 12 and 16 and another containing the other strains that did not belong to the major cluster. Strain 14 was again near the centroid (as in Figure 3.28). A third such choice, strains 2 and 10, gave similar results, except that strain 13 was pushed further out of the major cluster.

Figure 3.34 shows another choice of three reference strains, one from the major cluster, 1, one satellite, 5, and one singleton, 14. Again there is much distortion, though the major cluster is distinct. Two singletons are grouped with the minor cluster, two are near the centroid, and strain 5 is now very outlying. Strain 8 is pushed out of the major cluster, and I believe the explanation is similar to that for the outlying position of strain 2 in Figure 3.32. Strain 8 is the strain with least dissimilarity to strain 14 (Appendix 1) so that it is drawn out by the latter. Strain 6 shows similar but less marked behaviour, and is again relatively close to strain 14 in Appendix 1.

Figure 3.35 results from choosing 8 reference strains, one from each cluster, one satellite and the five singletons. This might represent a well-balanced choice; the range of variation is spanned, but near duplicates are omitted. The structure is good, though reference strains tend to be peripheral (e.g. strains 3, 5, 1, 15, 17). When this was repeated with omission of strain 17 the structure was little changed (though strain 17 became more central, and strain 12 more peripheral).

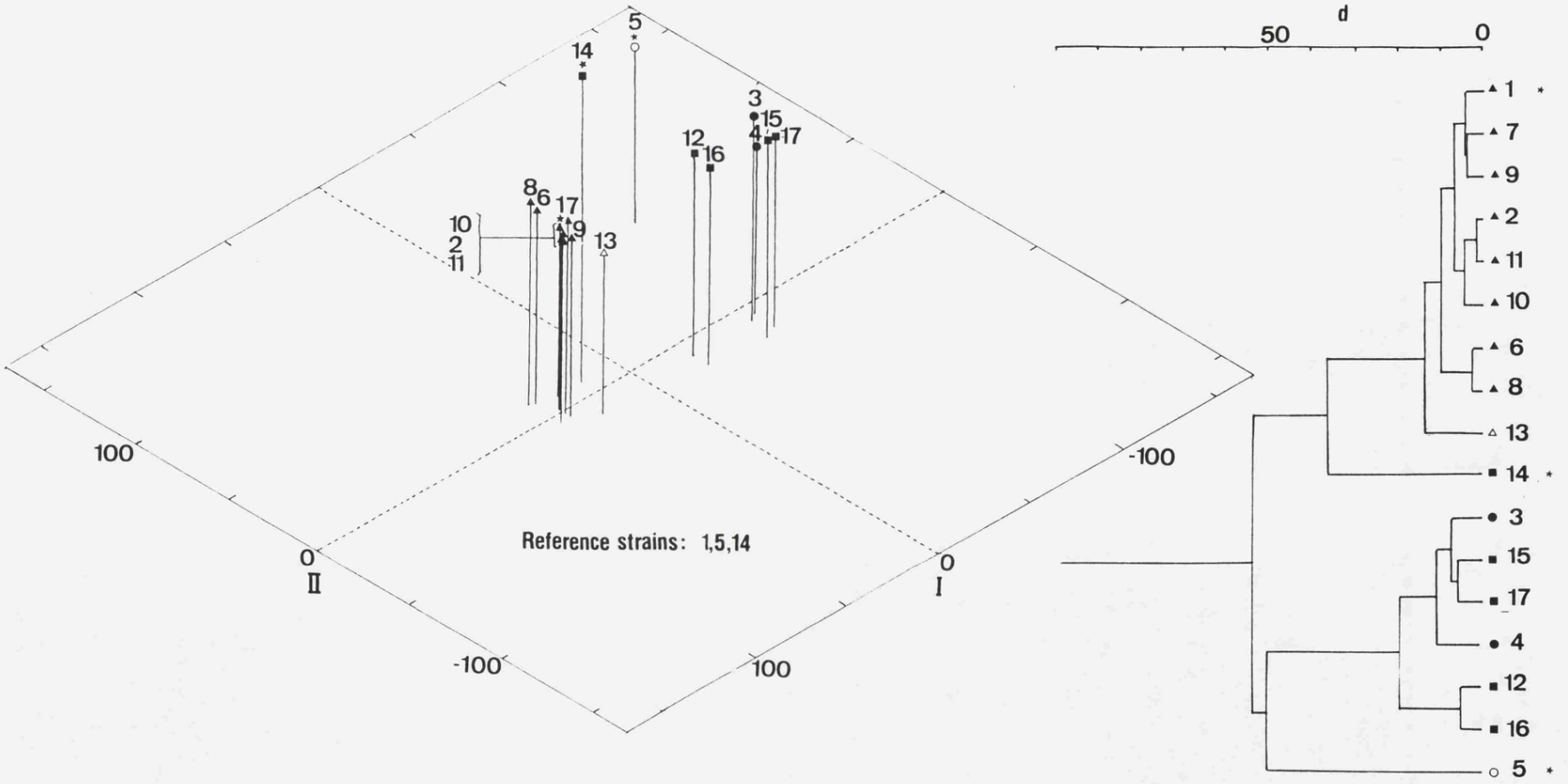


Figure 3.34 Principal Component analysis of Appendix 1 employing strains 1, 5 and 14 as reference strains (marked with asterisks). Other symbols and conventions as Figures 3.27 and 3.28.

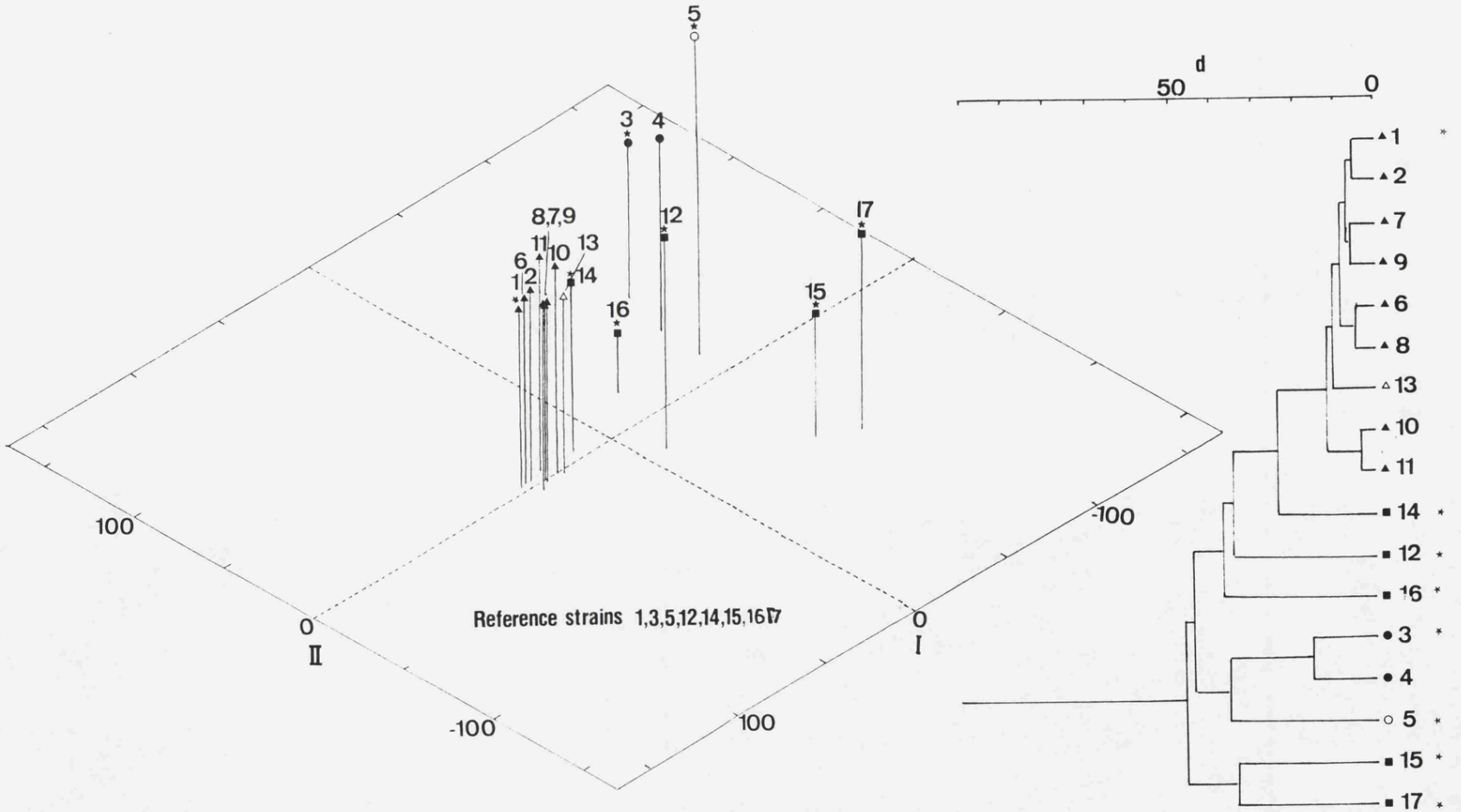


Figure 3.35 Principal Component analysis of Appendix 1 employing strains 1, 3, 5, 12, 14, 15, 16 and 17 as reference strains (marked with asterisks). Other symbols and conventions as Figures 3.27 and 3.28.

Examination of Distortion using a Random Normal Swarm.

A computer program was used to generate distances between 17 points in a three dimensional random normal cluster; the dispersions on the three dimensions were set to those found in the *B. circulans* data (Nakamura and Swezey, 1983*a*) and hence the cluster is flattened (Figure 3.36). The resulting 17 x 17 square matrix was used as input in the Principal Components program.

The cluster is 'spread out' or depicted on a small scale. Points 2 and 6 cluster towards one side of the ordination; 8, 10, 11, 15 as singletons; points 5, 7, 9, 12, 14, 16 as a tight cluster with 1, 3, 13 on the periphery and 4, 17 between these and the singleton, point 8.

To try to maintain the structure using as few reference strains as possible, point 2 from the right hand side of the ordination and point 17 from the opposite extreme of axis I were chosen. The resulting ordination (Figure 3.37) is depicted on the same scale as that of Figure 3.36, the 'true' configuration, for ease of comparison. The reference strains remain at opposing ends of axis I. 2 and 6 are now separated and would not be defined as a cluster. Point 8 is pulled in towards the main body of the ordination, as are points 1 and 4, although they remain satellites of the other points which are compacted into an unrealistic cluster.

Another attempt at establishing the true configuration was made with three points: 1, 2, 14 (Figure 3.38). 2 represents the outlying group, 14 represents the largest group and 1 is a satellite of the large group. The reference strains are situated on the periphery of the ordination. 2 and 6 are again well separated. 4 and 17 remain together and 8 is pulled in towards them. 1 is not grouped with any strains and appears closest to 4 and 17. All other strains are tightly grouped around the centroid.

Figure 3.36 The Principal Component ordination based on data from a 17 x 17 random normal swarm.

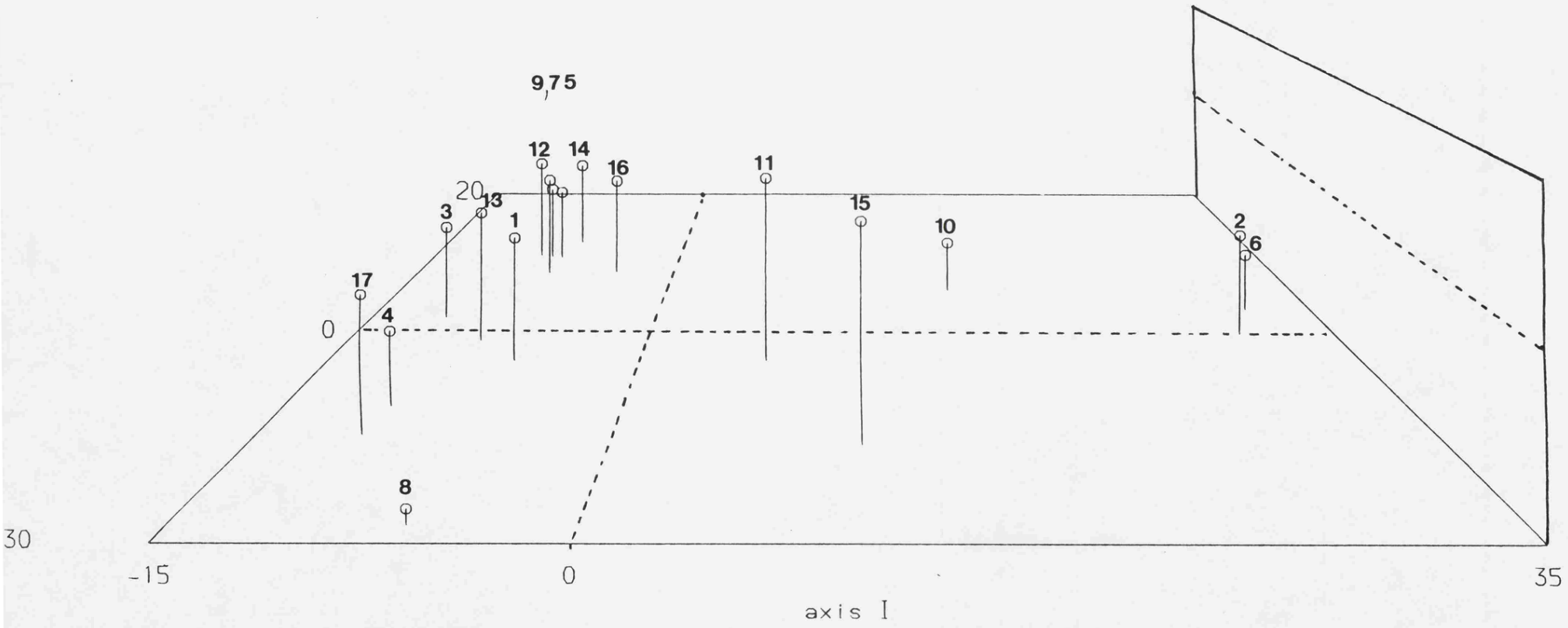


Figure 3.37 The Principal Component ordination based on the 2 x 17 strip matrix derived from the 17 x 17 random normal swarm, using points 2 and 17 as reference points. All points are of the same height on the third axis as there are only two reference points.

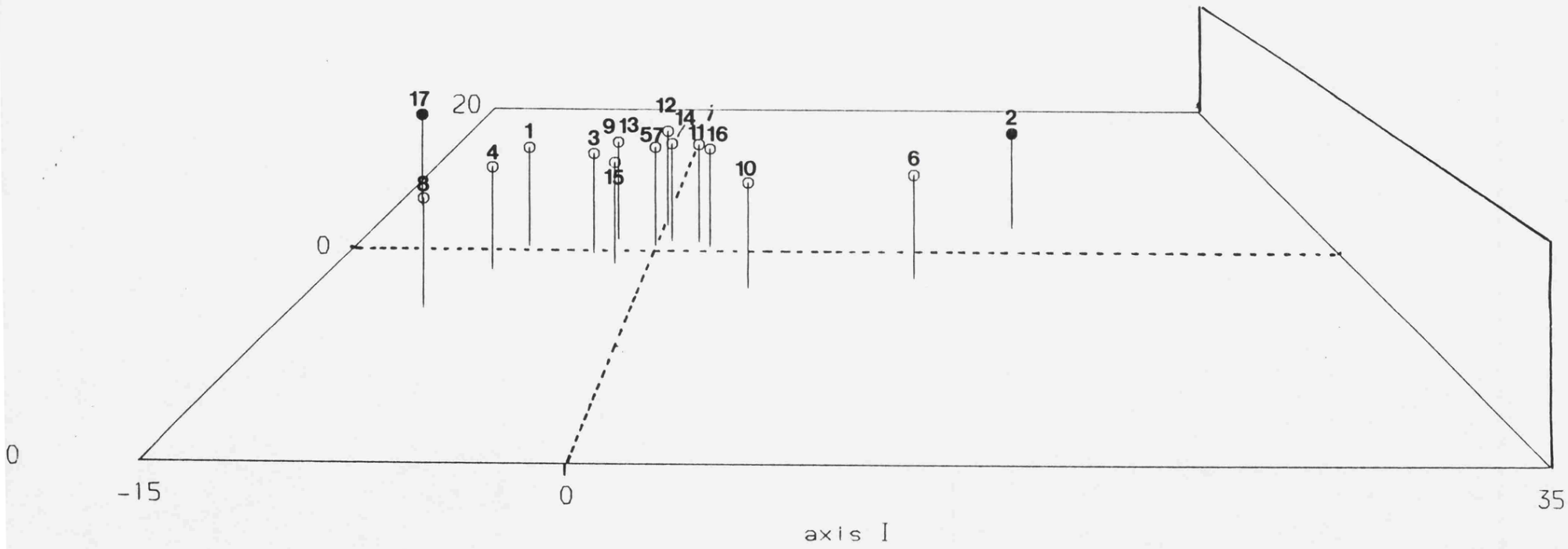
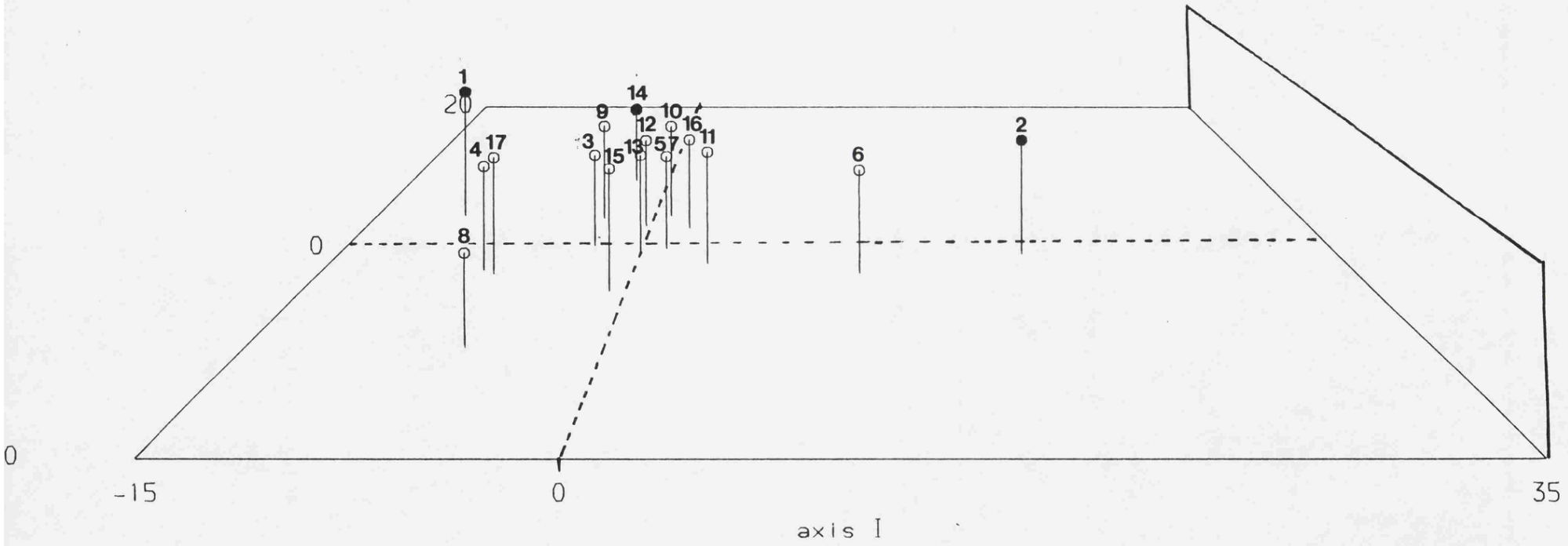


Figure 3.38 The Principal Component ordination based on the 3 x 17 strip matrix derived from the 17 x 17 random normal swarm, using points 1, 2 and 14 as reference points.



3.11 Error from Published Homology Data

Published Standard Deviations.

The results of eight published studies were examined (Table 3.10). The average error, S_E, lies between 3.0 and 8.6 percent; the weighted mean was 5.7 %. Pairing values were plotted against error values (corrected for degrees of freedom). The plot (Figure 3.39) from the data of Potts and Berry (1983) showed greater error with higher percent pairing; this was not always seen, however (see Figure 3.41 below). Mannarelli (1988) used both the filter and optical methods; with the filter method the error increased with an increase in % pairing, from 3.56 at 0-20 % pairing ot 12.67 over 80 % pairing (Figure 3.40).

Error from internal consistency.

Other error estimates are derived from the internal consistency of published data and shown in Tables 3.11-3.14. They are arranged according to the pairing method used, but results are first described according to the methods by which the error was estimated. The techniques, and major groups of bacteria studied, are compared afterwards.

Reciprocal pairs.

Error ranged from 2.26 to 15.4 % (Tables 3.11-3.14). The weighted average was 6.4 %. For five studies, pairing values from labelled strains against several unlabelled strains were compared with the reverse situation (Rocourt *et al.*, 1982; Ezaki *et al.*, 1986; Dent and Williams, 1986*a*, 1986*b*; Johnson and Harich, 1983). The difference in the relative binding does not seem to depend on the labelled strain. The mean and standard deviation of each half of the matrix was found, and a *t*-test of the means showed no significant difference on any of the data sets.

Table 3.10 Error from Published Replications

Technique	Organism	*s.d.	N	Reference
Optical	*Mycobacterium*	5.32	100	Baess (1979)
	" "	3.00	126	Imaeda *et al.* (1982)
	Bacillus	3.97	15	Nakamura (1987*b*)
S1 Nuclease	*Actinobacillus* and *Haemophilus*	6.15	278	Potts and Berry (1983)
	Mycobacterium	8.57	20	McFadden *et al.* (1987)
	" "	4.94	189	DeKesel *et al.* (1987)
	Haemophilus	6.30	228	Potts *et al.* (1986)
Filter	*Butyrivibrio*	6.29	456	Manarelli (1988)

*estimated standard deviation

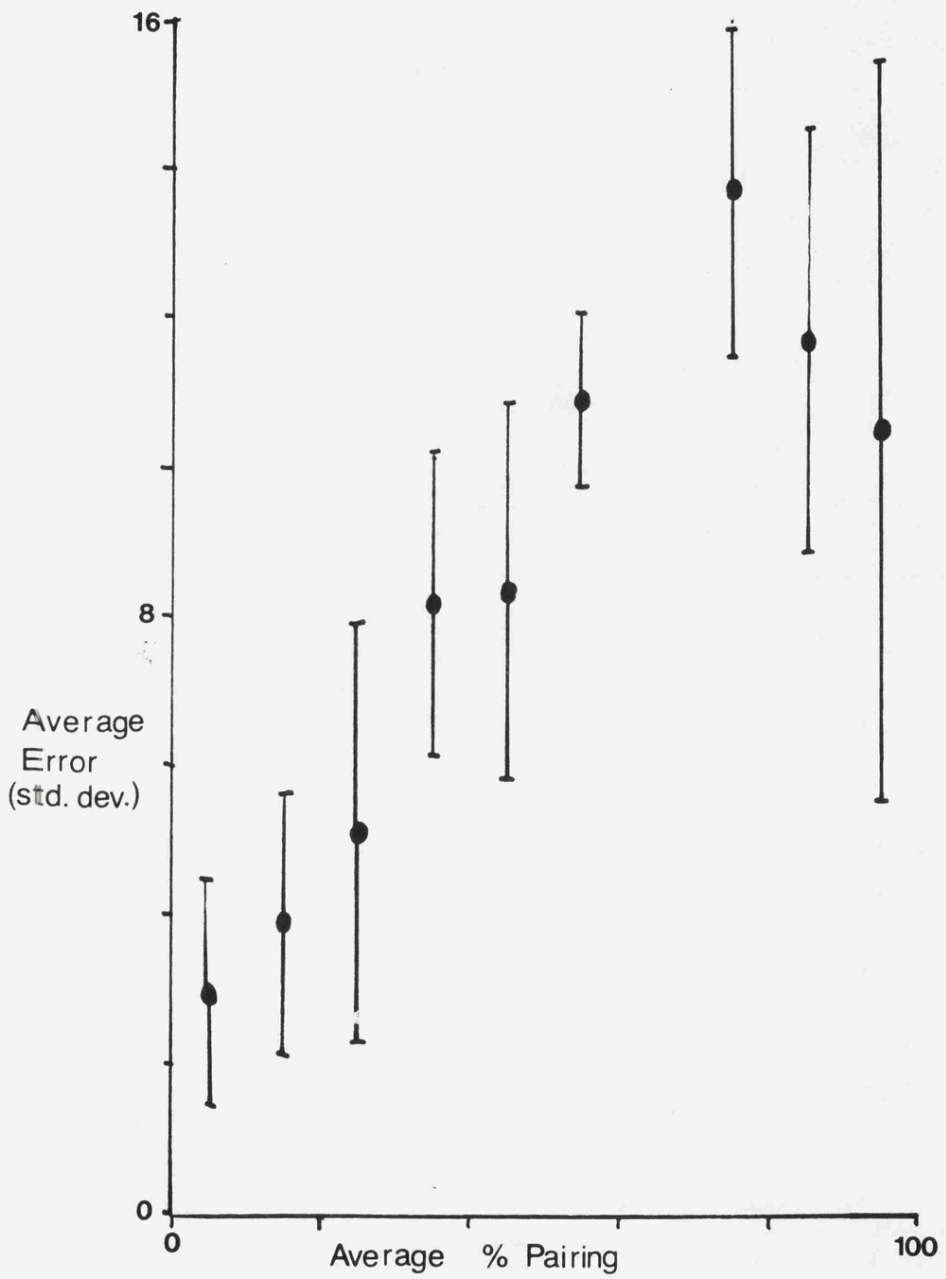


Figure 3.40 Relationship between error (ordinate), expressed as published standard deviations, and average DNA-DNA pairing values (abcissa, data of Mannarelli, 1988).

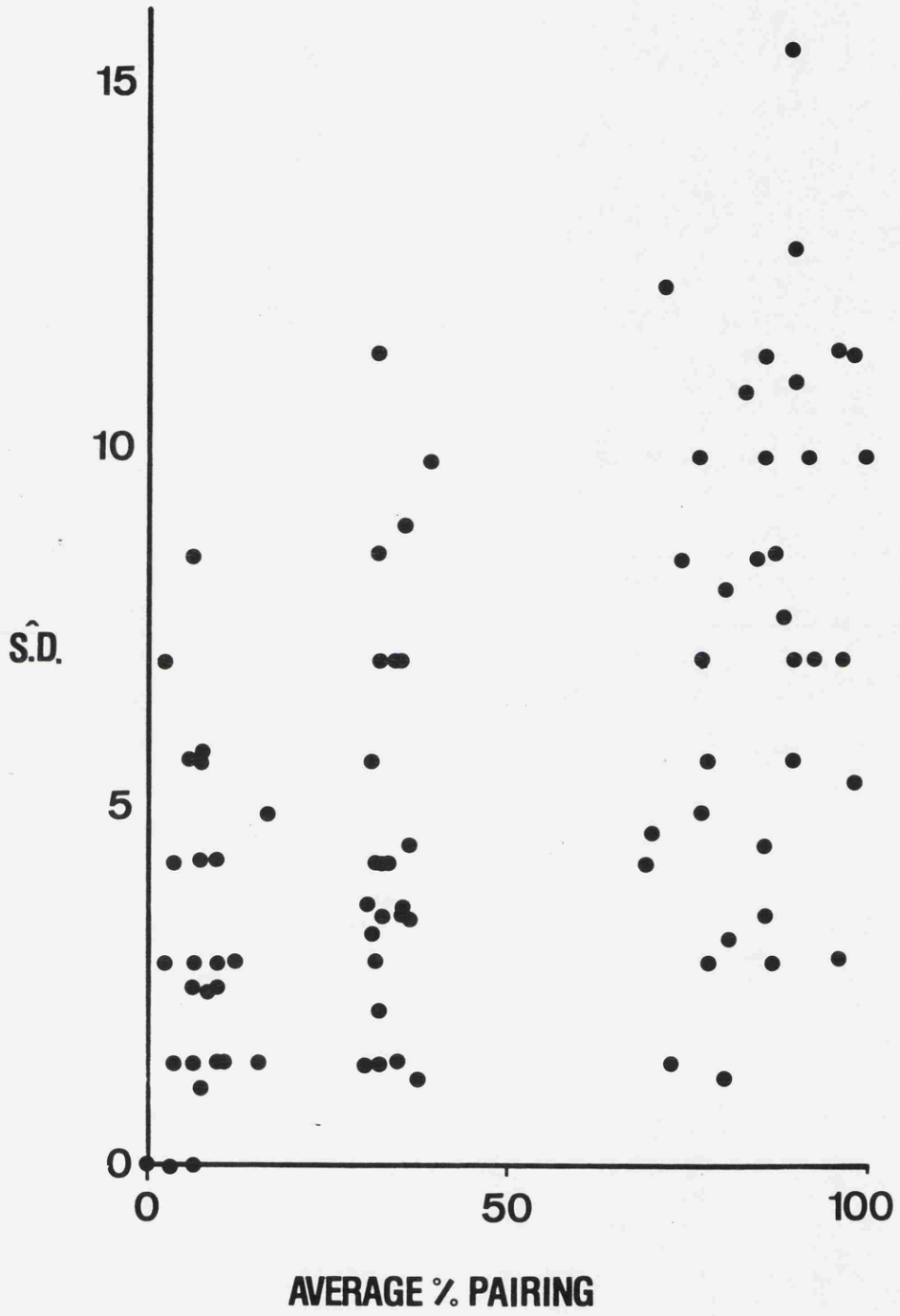


Figure 3.39 Relationship between error (ordinate), expressed as published standard deviations, and average DNA-DNA pairing values (abcissa, data of Potts and Berry, 1983).

Table 3,11 Studies using the Filter Technique

Organism	Average error from reciprocal pairs s,d,	N	Average error from zero-sided triangles s,d,	N	Violating triangles Before taking square -root	After taking square -root	Number of triangles	Reference
Bacteroides	6,21	4	4,15	49	464	59	20387	Johnson and Harich (1986)
Butyrivibrio	3,50	6		0	4	0	452	Manarelli (1988)
Enterococcus	15,40	8	8,46	17	52	21	183	Collins *et al*, (1986)
Streptococcus	10,53	18	8,30	90	122	82	403	Kilpper-Bälz *et al*, (1985)
Streptococcus	9,38	9	10,94	24	35	24	167	Coykendall *et al*, (1987)
Streptococcus	7,12	3	5,96	25	29	22	77	Kilpper-Balz *et al*, (1987)
Actinomyces	4,06	6	2,70	12	34	14	154	Dent and Williams (1986*a*)
Lactobacillus	4,05	17	3,23	21	35	6	435	Dent and williams (1986*b*)
Halophiles	5,45	10		0	0	0	324	Ross and Grant (1985)
Natronococcus and *Natronobacterium*	3,48	6		0	2	0	370	Tindall *et al*, (1984)

*number of cases available to examine

Table 3.12 Studies using the S1 Nuclease Technique

Organism	Average error from reciprocal pairs s.d.	*N	Average error from zero-sided triangles s.d.	N	Violating triangles Before taking square-root	After taking square-root	Number of triangles	Reference
Bacteroides	5.60	21		0	21	0	982	Love *et al.* (1986)
Bacteroides	3.53	15	3.57	20	118	22	430	Love *et al.* (1987*b*)
Fusobacterium	5.92	31	6.25	86	175	85	2682	Love *et al.* (1987*a*)
Haemophilus	3.56	23		0	36	0	408	Morozumi *et al.* (1986)
Pasteurella and *Actinobacillus*	7.84	10	5.48	24	76	24	1226	Escande *et al.* (1984)
Gluconobacter and *Pseudomonas*	3.63	45	5.33	6	70	13	1722	Micales *et al.* (1985)
Selenomonas	3.83	16	5.55	4	24	4	732	Moore *et al.* (1987)
Veillonella	6.40	42	5.05	81	468	112	1447	Johnson and Harich (1983)
Veillonella	3.99	31	6.0'	10	55	7	3909	Mays *et al.* (1982)
Azospirillum	2.26	15		0	0	0	516	Falk *et al.* (1985)
Bifidobacterium	7.23	14		0	5	0	144	Watabe *et al.* (1983)
Hyphomicrobium	4.96	22	4.19	29	30	23	629	Gebers *et al.* (1986)
Methanosarcinaceae	7.43	6		0	9	0	72	Sowers *et al.* (1984)
Streptococcus	10.95	12	8.36	79	125	75	488	Ezaki *et al.* (1986)
Listeria	2.68	9	6.88	100	136	92	1509	Rocourt *et al.* (1982)
Mycobacterium	4.95	6		0	0	0	79	Lévy-Frebault *et al.* (1986)

*number of cases available to examine.

Table 3.13 Studies involving the Optical Technique*

Organism	Average error from zero-sided triangles s.d.	*N	Violating triangles Before taking square-root	After taking square-root	Number of triangles	Reference
Bacteroides	9.09	9	25	23	1755	Tanner *et al.* (1986)
Butyrivibrio		0	175	2	9139	Manarelli (1988)
Haemophilus and *Actinobacillus*		0	7	3	16	Pohl *et al.* (1983)
Pasteurella	7.36	15	64	31	321	Mutters *et al.* (1985)
Cytophaga and *Flavobacterium*		0	0	0	47	Callies and Mannheim (1980)
Thermus	7.31	3	6	3	15	Hensel *et al.* (1986)
Bacillus	5.08	32	32	26	496	Nakamura and Swezey (1983*a*)
Bacillus	4.92	210	270	170	680	Nakamura and Swezey (1983*b*)
Bacillus		0	8	1	90	Nakamura (1987*a*)
Bacillus	4.57	548	857	476	3100	Nakamura (1987*b*)
Mycobacterium	1.87	445	195	191	298	Imaeda (1985)

*for this method no estimate can be made for reciprocal pairs

*number of cases available to examine.

Table 3.14 Other Techniques

Organism and Technique	Average error from reciprocal pairs s.d.	*N	Average error from zero-sided triangles s.d.	N	Violating triangles Before taking square root	After taking square root	Number of triangles	Reference
Bacillus Slot-blot Filter Method	7.46	5	4.69	10	11	8	18	Seldin and Dubnau (1985)
Methylobacterium Multi-blot Filter Method	2.65	14	3.20	34	83	47	719	Hood *et al.* (1987)
Mycoplasma DNA Probes and hydroxyapatite columns	2.92	10		0	1	0	260	Stephens *et al.* (1985)
Rhizobium Hydroxyapatite probe and filter	8.79	14	10.59	41	174	44	768	Wedlock and Jarvis (1986)
Leptospires Hydroxyapatite technique	15.26	48	13.03	42	123	42	1703	Yasuda *et al.* (1987)

*number of cases available to examine

The S1 nuclease and filter methods showed differences in error estimated from reciprocal pairs. To confirm this an analysis of varience (ANOVA) was carried out on six of the most complete studies (Johnson and Harich, 1983, 1986; Dent and Williams, 1986*b*; Gebers *et al.*, 1986; Kilpper-Bälz *et al.*, 1985; Love *et al.*, 1987a). When all six studies were included in ANOVA, the hypothesis that there was no difference in the mean error between studies was rejected ($P < 0.001$). However, this was due entirely to one study, that of Kilpper-Bälz *et al.* (1985) on *Streptococcus*, which had much higher error than the others. When this study was excluded the significance of differences in error between the remaining five studies did not reach $P = 0.2$.

Error from zero sides

Inconsistencies in DNA pairing data from triangles with zero-sides were very common where there were zero-distances, i.e. 100 % pairing values to analyse. Table 3.15 gives typical examples of discrepencies of this sort from the data of Mutters *et al.* (1985).

Error ranged from 1.87 to 13.03 % (Tables 3.11-3.14). The weighted average was 5.0 %. Figure 3.41 shows a plot of standard deviation against average percent pairing pooled from a range of papers (Rocourt *et al.*, 1982; Johnson and Harich, 1983, 1986; Kilpper-Bälz *et al.*, 1985; Mutters *et al.*, 1985; Micales *et al.*, 1985; Dent and Williams, 1986*a*, 1986*b*; Gebers *et al.*, 1986; Tanner *et al.*, 1986).The percent pairing values were divided into 10 bands at 10 % intervals. The average error for each section was plotted against the midpoint of the % pairing band (Figure 3.38). Error seems to remain fairly constant over the range of pairing values, in contrast to Figure 3.39, except for high error in the 80-90 % band. Error is somewhat lower at the extremes, i.e. 0-10 % and 90-100 %

Table 3.15 Inconsistency in percent DNA-DNA pairing.

TRIANGLES WITH ONE ZERO SIDE

Strains			% Difference		
a	b	c	D_{ab}	D_{ac}	D_{bc}
1	5	8	0	8	5
1	5	14	0	27	45
1	3	27	0	78	79
19	22	23	0	8	9
19	21	27	0	57	38

Data of Mutters *et al.* (1985), typical values.

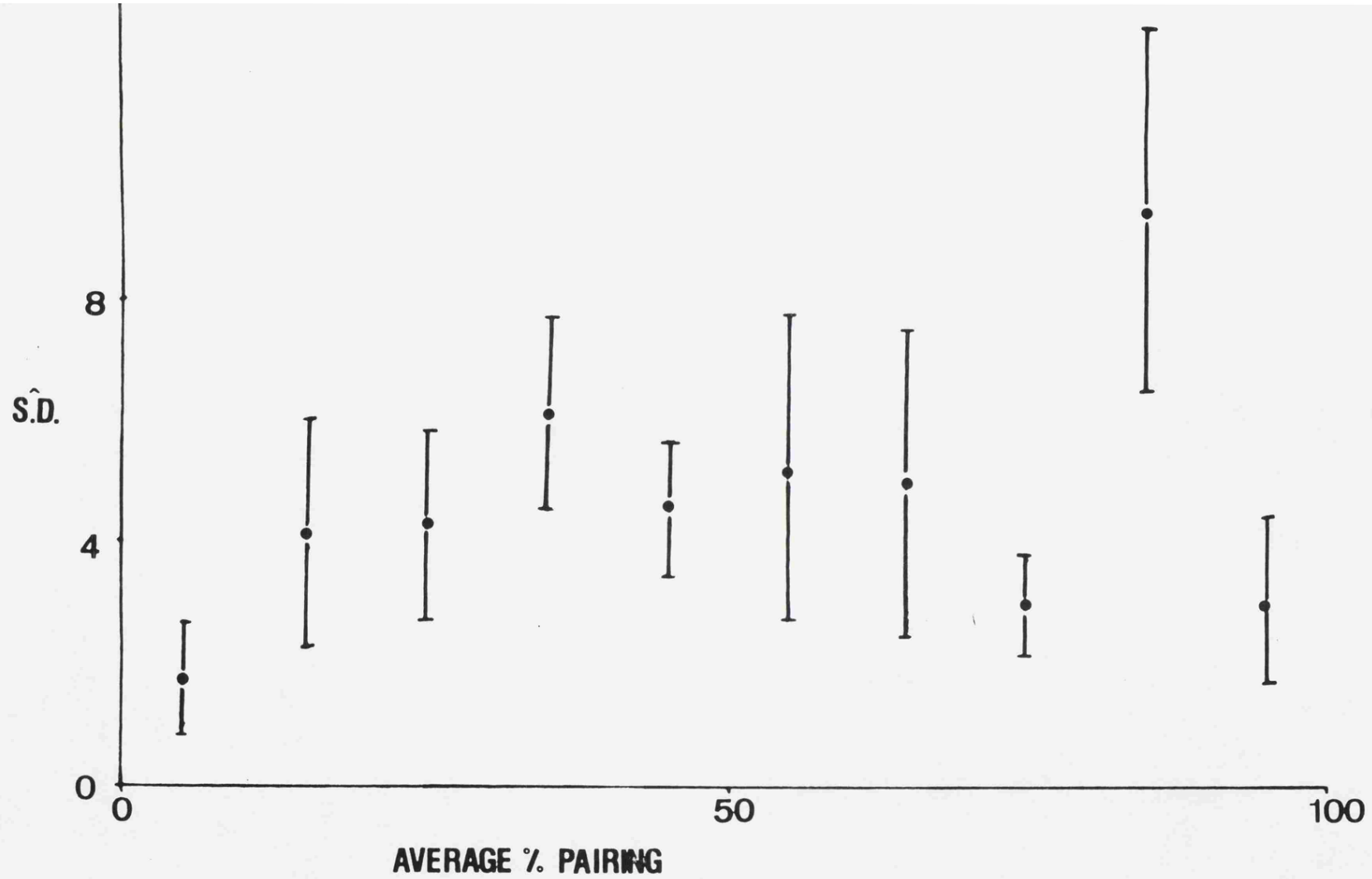


Figure 3.41 Relationship between error from zero-sided triangles and average DNA-DNA pairing. The error is the average s.d. (see section 2.14). Pairing values were divided into 10 bands at 10 % intervals, and the results plotted at the midpoints of the bands, as mean (dots) and one standard deviation above and below (bars). Data are from studies listed in text.

pairing. This is presumably because large error is not compatible with an average close to 0 or 100 %. Thus, if two values give an average of 90 %, the maximum error obtainable will be 14.14 %, which is the standard deviation of homology values 80 and 100 (this is not necessarily so where there are pairing values over 100 %, but there are not enough data for conclusions on methods where such values can occur). These papers were used for an ANOVA to detect significant differences between the studies. Significant differences certainly exist; the hypothesis that there was no difference in the mean error between studies was rejected ($P < 0.001$).

A plot of mean reciprocal pair error against mean zero-sides error, for studies where both could be estimated, gave a reasonable straight line fit:- $Y = 0.31 + 1.07X$ (where X = error from triples with a zero side, Y = reciprocal error) shown in Figure 3.42. Results with less than 10 values in either error method were not used. When the *Lactobacillus* data (Collins *et al.*, 1987) are omitted the line passes close to the origin at almost 45°, and the correlation r is high (over 0.79), so this implies that the error rates obtained from reciprocal pairs and zero sides are consistent and similar in magnitude.

3.11 Triangle Inequalities.

Examples of triangle inequalities in DNA pairing data are shown in Table 3.16.

The proportion of violating triangles (Tables 3.11-3.14) ranged from 0 to 65 %, the average being about 8 % before square-rooting the DNA dissimilarities. Although violating triangles were more frequent in

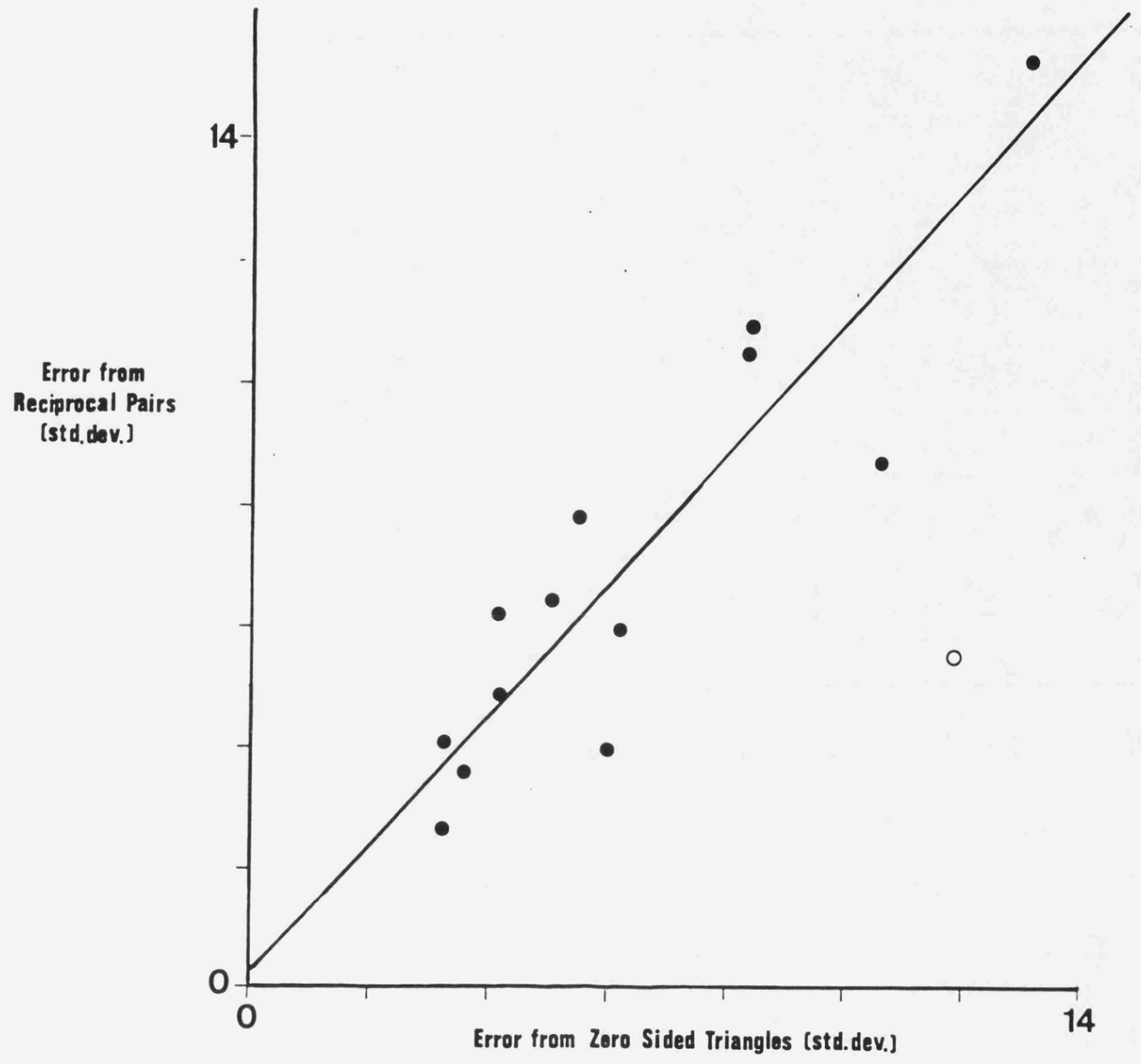


Figure 3.42 Relationship between error estimated from reciprocal pairs and that estimated from zero-sided triangles for studies where both are available. Each circle represents a different study from Tables 3.11-3.14 (see text). The open circle is that of Collins *et al.* (1987). The line of linear regression (excluding open circle) is shown.

Table 3.16 Inconsistency in percent DNA-DNA pairing.

TRIANGLES WITH NO ZERO SIDES

Strains			% Difference			
a	b	c	D_{ab}	D_{ac}	D_{bc}	
1	6	11	2	16	2	*
1	9	14	9	27	27	
1	9	45	9	41	94	*
1	9	52	9	100	76	
19	25	33	22	63	80	
27	30	31	1	20	8	*

* The triangle inequality is not removed by transforming to the square root.

Data of Mutters *et al.* (1985), typical values.

studies with more zero-sided triangles (as expected, see Methods), there were many puzzling features, which are taken up in the Discussion. Six papers were used to detect whether square-rooting significantly reduced the proportion of violating triangles using Fisher's exact method for 2 x 2 tables and combination of probabilities (Conover, 1971; Snedecor, 1956). This gave a probability of 0.01-0.001, indicating a significant improvement from square-rooting.

When all papers in the study were used the average proportion of violating triangles was 3.8 % compared with 8.2 % before square-rooting, and most of the remaining violating triples had a zero side which forces a violation if there is any error, however small. From ten studies (Johnson and Harich, 1983; Dent and Williams, 1986*a*, 1986*b*; Ezaki *et al.*, 1986; Collins *et al.*, 1986*a*, 1987; Love *et al.*, 1986, 1987*a*; Kilpper-Balz *et al.*, 1987; Hood *et al.*, 1987) a total of 1231 violating triples was reduced to 470 by square-rooting the distances; however, 437 of these involved a zero-side. Of the 33 remaining, 20 were from the study of Johnson and Harich (1983). The effect of square-rooting on DNA data was examined by generating the Principal Coordinate ordination of the data of Nakamura and Swezey 1983*a* after square-rooting the complete matrix (Figure 3.43); the results are shown on a much smaller scale than the scale of Figure 3.9. The resulting relationships were very similar to those in Figure 3.9, but the entire configuration is spread out.

Five papers that involved techniques other than those in Tables 3.12-3.13 are shown in Table 3.14. Hood *et al.* (1987) claimed to have an improved multi-blot filter method and the errors were found to be low using both zero sides and reciprocal pairs, but the proportion of violating triangles was fairly high.

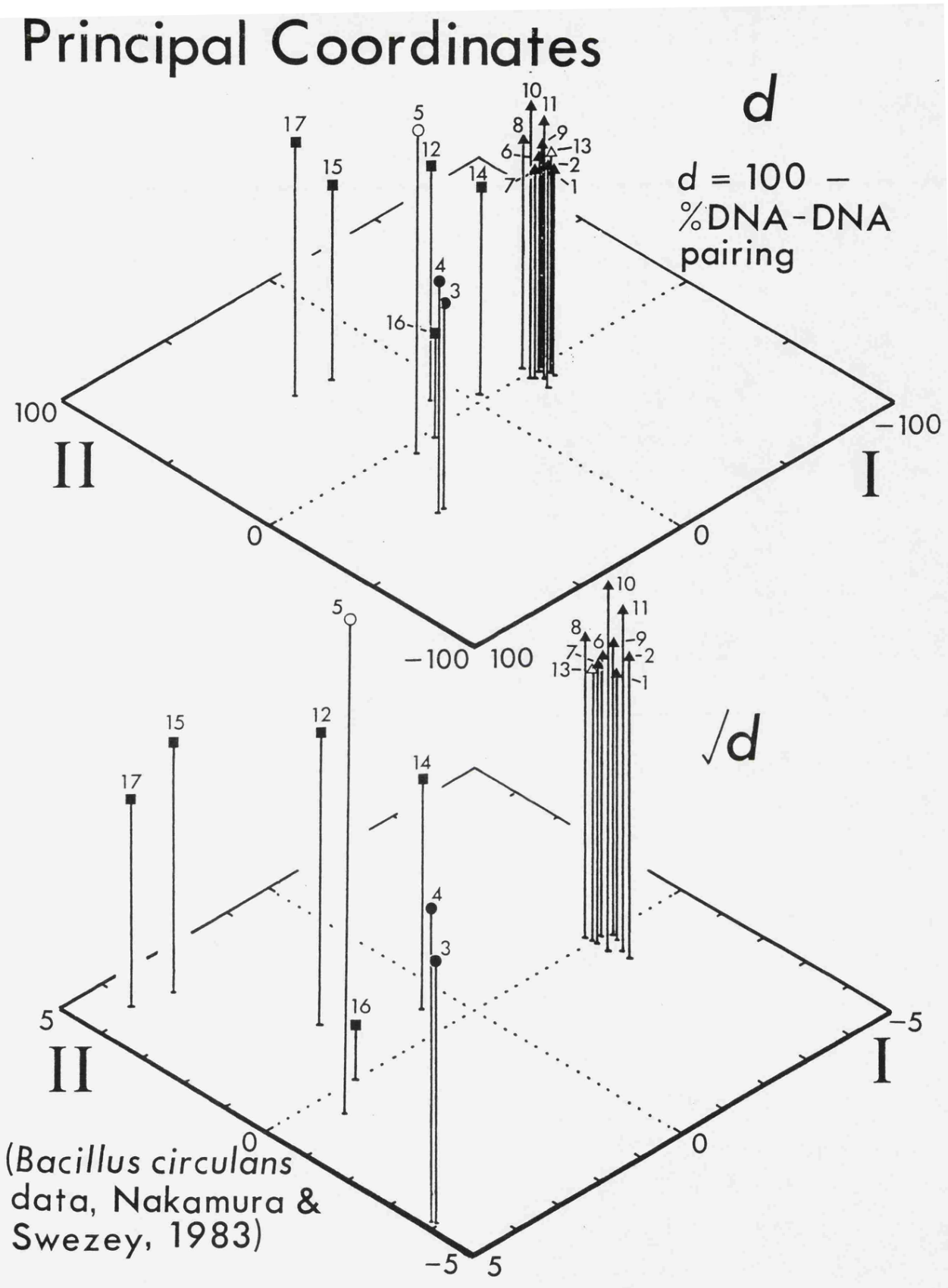


Figure 3.43 The effect of square-rooting on DNA distances in Principal Coordinate analysis. Data are from Nakamura and Swezey (1983*a*) (Appendix 1); all 17 strains are reference strains. The top configuration is a replica of Figure 3.27, included for comparison.

Comparison of techniques and major groups of bacteria.

Results are averaged for each technique and major group of the organisms studied (Table 3.17). An ANOVA on the error from zero-sides was not rejected at $P = 0.05$, that is the zero-side errors did not differ significantly between techniques. An ANOVA on the reciprocal pair error also showed no significant difference between techniques at $P = 0.05$.

Comparison of the error for different major groups of organisms is hampered by the small values of N in many of the cases, but it is noteworthy that the percentage of violating triangles is high in Gram positive groups.

3.12 Error Estimation from *Listeria* Homology *Data*

The standard deviations from replications of *Listeria* pairing data (Appendix 6) is listed in Table 3.18. The average standard deviation was plotted against % average pairing for the ranges of % pairing values given in Table 3.18 (Figure 3.44). Error seemed to be independent of % pairing, although error was higher at very low and high (over 90%) pairing values.

The pairing data was analysed using TRUDNA.PAS (Appendix 2). There were no zero errors, reciprocal pairs or violating triangles.

Table 3.17 Average Error within Methods and Major Groups of Bacteria.

	Technique					
	Optical		S1 Nuclease		Filter	
Reciprocal pairs	Not applicable		s.d.	N	s.d.	N
Gram-negative			4.84	234	6.14	221
Gram-positive			6.86	27	8.14	70
Halophiles			-	-	4.71	16
All Organisms			5.05	318	6.47	312
Zero-sided Triangles	s.d.	N	s.d.	N	s.d.	N
Gram-negative	7.93	27	5.47	231	4.15	49
Gram-positive	4.03	1235	7.53	179	9.11	267
Halophiles	-	-	-	-	-	-
All Organisms	4.11	1262	6.23	439	8.13	328
Triangle Violations	% violating	N	% violating	N	% violating	N
Gram-negative	2.40	11278	7.70	13538	2.25	20839
Gram-positive	30.69	4764	12.57	2076	21.67	1419
Halophiles	-	-	-	-	0.29	694
All organisms	10.83	16057	7.94	16975	3.39	22952
Triangle Violations (after square-rooting)	% violating	N	% violating	N	% violating	N
Gram-negative	0.52	11278	0.20	13538	0.28	20839
Gram-positive	18.50	4664	8.04	2076	11.90	1419
Halophiles					0	694
All organisms	5.80	15957	2.69	16975	0.99	22952

Table 3.18 Error from Replications of *Listeria* Homology Experiments.

Average % Pairing	N*	Standard Deviation	Average Standard deviation	
8.0	2	11.3		
10.5	2	0.6		
11.7	2	2.1	0-20%	
13.6	2	0.2	3.6 ± 3.9	
14.3	2	0.8	n = 14	
15.9	2	4.0		
18.9	2	6.2		
23.4	4	1.7		
24.9	2	2.6		
24.9	3	4.3	>20-30%	
+(25.3	3	3.2)	2.19 ± 1.28	
27.1	2	0.5	n = 14	
27.8	3	1.6		
30.4	3	0.6		
33.0	3	3.2		>20-40%
33.0	2	1.8		2.04 ± 1.47
34.5	2	0.5		n = 45
34.9	2	0.4	>30-40%	
35.4	2	1.8	1.97 ± 1.57	
35.7	2	4.3	n = 31	
35.9	2	0.4		
36.8	3	1.8		
37.3	5	4.4		
38.4	3	1.2		
38.9	2	0.1		

41.8	2	2.6			
42.4	2	1.1			
43.0	3	3.1			
44.1	2	1.0	>40-50%		
44.6	2	3.4	2.05 ± 1.00		
44.7	3	0.9	n = 36		
(42.7	4	4.1)			
(44.2	5	4.9)			
44.9	4	3.4			
(47.8	5	7.1)			
45.5	3	0.8		>40-60%	
46.2	3	1.7		1.88 ± 1.01	
46.6	3	0.7		n = 67	
49.7	9	2.5			
50.3	2	0.6			
51.2	3	0.6			
51.7	3	1.6			
52.2	3	1.8			
52.6	5	3.0	>50-60%		
53.4	4	2.3	1.68 ± 1.00		
53.5	2	0.7	n = 31		
55.0	3	0.6			
56.2	2	0.7			
56.8	2	1.6			
59.7	2	3.5			
62.3	2	2.1	>60-70%		
67.0	2	2.5	2.73 ± 0.56		
69.3	3	3.3	n = 7	>60-80%	
71.9	7	3.2	>70-80% n = 9	2.86 ± 0.51	

				n = 16
73.8	2	2.1		
81.5	3	2.5		
83.4	3	2.6		
84.3	2	3.2		
			>80-90%	
84.9	2	0.1		
			1.82 ± 0.97	
85.1	2	1.5		
			n = 18	
86.2	2	1.9		
89.4	4	1.0		>80%
91.6	2	0.5		2.36 ± 1.95
93.5	2	0.4	>90%	n = 31
94.6	2	7.2	3.12 ± 2.67	
94.9	2	4.2	n = 13	
95.0	3	1.5		
95.5	2	5.7		
(101.3	3	10.9)		

* Number of Replications.

+ Not included in averages; bracketed lines are based on the same set of experiments as the previous line but included an unacceptable run. This is explained in the Discussion.

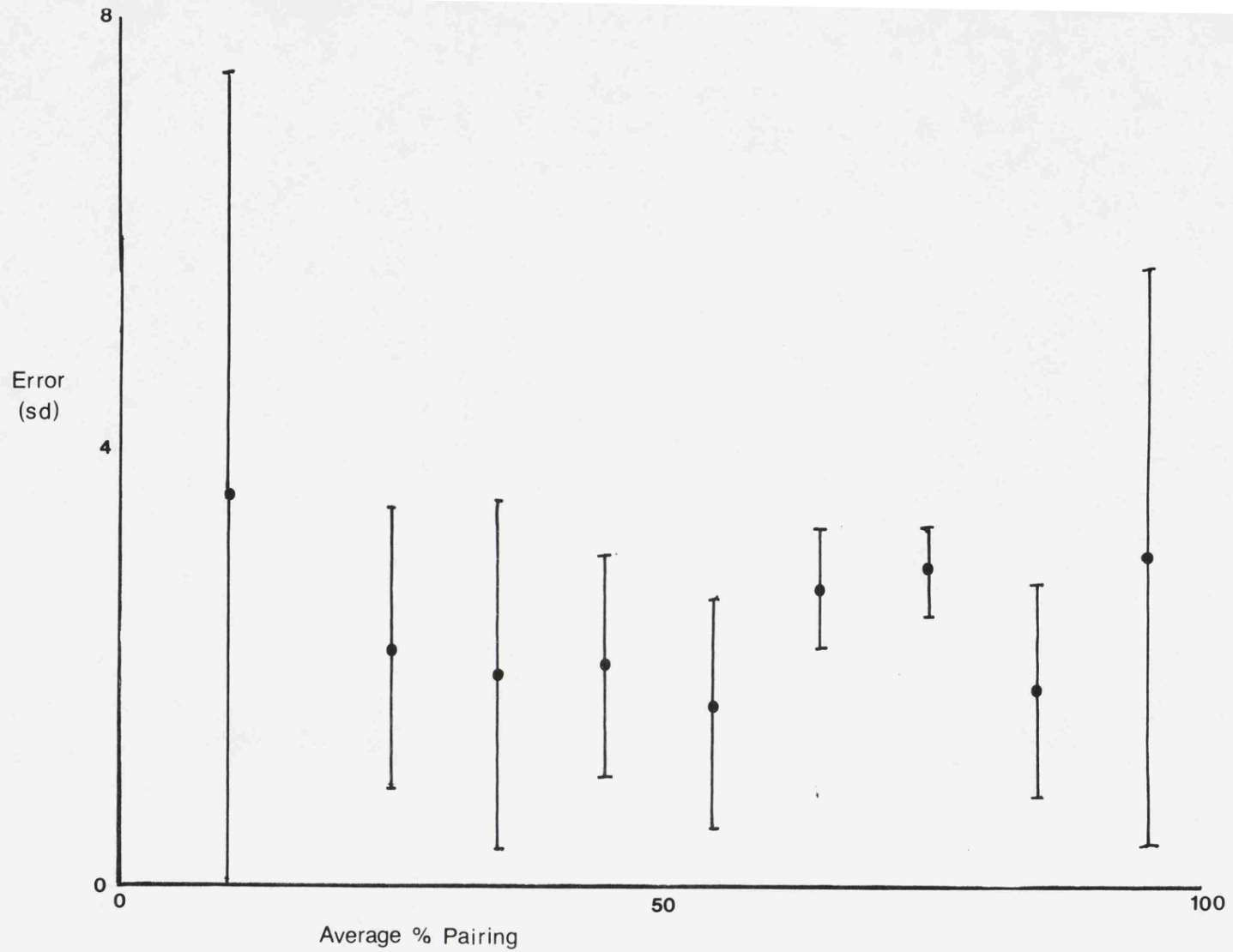


Figure 3.44 Relationship between error from replications and average percent DNA-DNA pairing experiments with *Listeria* strains using the optical technique. The error is the average s.d. (Table 3.18). Pairing values were divided into 10 bands at 10 % intervals, and the results plotted as the midpoints of the bands, as mean (dots) and one standard deviation above and below (bars).

4. DISCUSSION

One of the major past disadvantages of the spectrophotometric technique was the ability to process only one sample at a time, or in some cases three cuvettes at a time. This has been overcome by new technology and % pairing for up to four different pairs of strains may be determined in the same run. This could cut down on one form of experimental error as there is no risk of the environmental conditions varying within an experiment as the three cuvettes involved in each experiment are run over the same time period. However, care must be taken to set the spectrophotometer at the T_{or}; the final temperature may not be the same temperature at which the machine is set. This could be critical if there is more than 5°C difference, as illustrated by the results of stringent and non-stringent conditions (section 3.6). The variation over the cell chamber should also be assessed, especially if % (G+C) determinations are being carried out as a small change in temperature corresponds to a larger change in the derived % (G+C) result (section 3.7). With the PU8700, correction for base composition determinations is advisable; furthermore it was found that a new temperature probe meant the variation had to be re-calibrated. However, in the case of renaturation rate experiments, if the machine is at the T_{or} the homology determination should not be appreciably affected by less than 1°C total variation. If the experiment is carried out in triplicate then the sample positions may be alternated to double check against any discrepancies.

A further advantage of the optical technique is the ability to carry out stringency experiments without having to worry about leaching from filters; however a check for evaporation is necessary at the beginning and end of renaturation. It is also easy to carry out % (G+C) determinations with the same equipment as the homology determinations, and infact these may be done in 2 x SSC at the beginning of an homology experiment as long as a standard is included and evaporation checked for.

Shearing by passage through the French Press was found to be suitable for *Listeria* homology experiments (section 3.2); it was reproducible and produced a fragment size which allowed for easy measurement of the renaturation rate in 2 x SSC. Syringe-shearing did not produce small enough fragments, hence rates were too fast and measurement of the rates from renaturation curves produced by DNA sheared in this way would introduce large errors into the data.

The removal of polysaccharide from DNA preparations was found to be unnecessary for *Listeria* strains (section 3.4). Both methods used to remove polysaccharride reduced the yield of DNA, particularly the CTAB procedure; they also made the DNA purification procedure much longer by the need for extensive dialysis. The difference in the concentration of polysaccharide before and after the 2-methoxyethanol procedure was not significant. Polysaccharide removal may be necessary for some organisms such as *Rhizobium* and some streptococci, as bacteria producing abundant extracellular polysaccharide may also show large error (*Rhizobium*, Table 3.14, Wedlock and Jarvis, 1986). This is possibly due to the interaction between the carbohydrates and DNA (Graves, 1968) whose work implies that the presence of polysaccharides in large amounts would affect renaturation rates using any method of DNA-DNA pairing measurement.

Samples of purified *Listeria* DNA may be stored at -20°C for long periods (over a year) without significantly affecting the degree of binding in renaturation rate experiments; this confirms the results of Crombach, 1974. The samples should be dialysed before use to ensure the correct salt concentration (section 3.5).

The *Listeria grayi* strain, C214, was stored as two subcultures labelled C214*a* and C214*b* over seven years ago by Sarah Ferusu (1980) as part of her study on the taxonomy of *Listeria*. The subcultures have been kept separately at -70°C ever since. The homology between the two 'strains' should therefore be 100 %. However the average value obtained by the optical technique was 95.5 %. Looking at the individual results (Appendix 6) there is a larger than average degree of error in the determinations with these two subcultures. The homology was also determined under stringent conditions (T_{or} + 10°C) to examine the thermal stability of the duplexes formed under optimal conditions (section 3.6). The % pairing was much reduced under stringent temperatures. The range of pairing values varied enormously suggesting a decrease in reliability of results under stringent conditions. The stringency experiments suggest that there has been a change in one or both of the strains over the storage period. It is difficult to believe that extensive genetic changes could have occurred at - 70°C. However, error could possibly account for the discrepancy in the observed degree of binding under optimal conditions. Another pair of closely related strains, JS21 and JS31, which have an average % pairing of 95 % and were isolated from the same source, were analysed under stringent conditions (section 3.6). In this case the % pairing was reduced by 5.4 % to 89.6 % showing that the duplexes formed by renaturation were very stable. A *Listeria monocytogenes* strain, C52, and a *Listeria welshimeri* strain, C1091, with an average homology of 52.6 % under optimal conditions were examined under stringent and non-stringent conditions (section 3.6). Stringent temperatures of 7-10°C above the average T_{or} reduced the degree of binding by up to 16 % and even 4°C above the T_{or} reduced the degree of binding by 10-11 %. This is relevant for experimental error because one can only control temperature in practice by ± 0.5°C; this will be expected to give about 1.25% uncertainty in pairing. The greater the difference in the T_{or}s of two

organisms, whose pairing value is being determined, the more uncertainty introduced into the result. Strains which have a lower DNA homology seem to be more affected by deviation from the T_{or} than closely related strains. Experiments are more difficult to carry out under higher temperatures as more evaporation occurs and bubbles may be introduced into the cuvettes making rate measurements prone error. If experiments are to be done under stringent conditions it would be best to lower the T_{or} of the samples by the addition of formamide to the reaction buffer (Hutton, 1977). Under non-stringent conditions (T_{or} - 10°C) the % pairing between C52 and C1091 increased by 6 % probably due to non-specific base-pairing. The deviation from % pairing under optimal conditions appears to be much larger at T_m-15°C compared to that at T_m-35°C which supports the results of Gillis *et al.* (1970) that showed the degree of binding is scarcely affected over a range of 5°C under the T_{or}.

The standard deviation of base composition determinations was much reduced by using a ramp rate of 0.5°C min^{-1} instead of 1.0°C min^{-1} (section 3.7). The base composition of C644 was closer to the estimation of Ferusu (1980) when it was determined at the slower temperature increase than at a temperature increase of 1°C per minute.

Taxonomy of *Listeria* based on DNA-DNA pairing experiments.

Rocourt *et al.* (1982) clarified the division of *Listeria monocytogenes* into five species based on DNA-DNA homologies using the S1 nuclease technique. However, with the exception of two strains of *L. ivanovii*, only one reference strain was used from each species. The data from the optical technique used here gave a very similar ordination when only one reference strain was used from each species. The choice of a similar number of reference strains, but not spread across the spectrum of relatedness resulted in a slightly different taxonomic structure.

The range of variation of relatedness within the species *L. monocytogenes sensu stricto* was shown to be very large in both studies, suggesting either more than one cluster or a spectrum of relatedness. The latter is presumably the case because even using the complete matrix of DNA distances no clear groups were seen. *L. grayi* and *L. murrayi* however appear definite. C52 shared a higher DNA pairing value with some *L. innocua* strains than with other *L. monocytogenes* strains. These two species are separated on whether or not they produce haemolysis. Rocourt *et al.* (1982) showed greater separation of these two species with DNA homology, but the *L. monocytogenes* strain used as a reference did not seem typical of the species, *L. monocytogenes*. It seems from Figure 3.7, derived from the 16 x 16 matrix, where there were five *L. monocytogenes* reference strains that C52, although derived from the type strain, does not seem to be a particularly typical representative of the species *L. monocytogenes*.

As reflected by the data from numerical taxonomic and other chemical relatedness studies, *Listeria* species are very closely related, especially *L. monocytogenes, L. innocua, L. welshimeri,* and *L. seeligeri*. The DNA pairing data heavily supports this and there seems to be no

strongly defined clusters among these four species. *L. ivanovii* is distinct from these four species; *L. grayi* and *L. murrayi* are also distinct although they may not be separated from each other.

A few strains of *Listeria* were present in both this study and that of Rocourt *et al.* (1982). The pairing values determined in both studies are summarised in Table 4.1. As illustrated in other comparisons of DNA pairing methods, the S1 nuclease technique seems to be more stringent than both the filter and optical techniques. Here, although the resulting ordination gave broadly similar results to that of Rocourt *et al.* (1982), the data from the optical technique was always higher than that of similar or identical pairings in the S1 nuclease data (Appendices 6 and 8, Table 4.2) values were as much as 18 % different between strains included in both studies (Table 4.1).

Several values below 30 % were obtained using a concentration of 75-80$\mu g ml^{-1}$ DNA (Appendix 6). Huss *et al.* (1983) found that values below 30 % were rarely obtained at this DNA concentration and were meaningless with the optical technique. If this is true then *L. grayi* and *L. murrayi* may be less related to the other five species than is suggested by my data.

Table 4.1 Pairing values determined by the optical technique and the S1 nuclease technique (Rocourt *et al.*, 1982).

	C52		C644		C1091		C1090	
	S1	Opt.	S1	Opt.	S1	Opt.	S1	Opt.
C52	100	100						
C644	53	71.9	100	100				
C1091	42	52.6	46/44	49.7	100	100		
C1090	24	52.2	24	50.3	28	46.2	100	100

S1 : data from the S1 nuclease technique (Rocourt *et al.*, 1982)

Opt : data from the optical technique (Appendix 6)

Table 4.2 Range of Relatedness between strains of *Listeria* from two hybridization techniques.

Strain	Strain										
	L. monocytogenes		*L. innocua*		*L. ivanovii*		*L. welshimeri*		*L. seeligeri*		*L. grayi*
	S1[1]	Op[2]	S1	Op	S1	Op	S1	Op	S1	Op	Op
L. monocytogenes	63-98[3]	55-90	45-58		18-49		40-47		21-29		
L. innocua	47-56	45-81	87-99	95	22-45		42-47		19-29		
L. ivanovii	*27-33*	*27-45*	*29-42*	*27-39*	*85-104*	*58-95*	*29-34*		*31-44*		
L. welshimeri	41-46	42-58	38-46	46-71	22-38	24-42	88-96	92	21-30		
L. seeligeri	29-38	40-60	26-34	45-62	34-49	24-43	28-35	43-54	71-89	94	
L. grayi	4-21	12-30	5-26	0-25	3-7	9-36	5-14	7-15	3-4	12-14	96
L. murrayi	4-9	5-26	5-6	4-16	2-3	8-17	2-6	11-24	1-2	13-17	84

[1] ; Rocourt *et al*, 1982 data from the S1 nuclease method,
[2] ; Data from the optical technique,
[3] ; Homologous reactions are not included,

Choice of Reference Strains.

Complete matrices of DNA-DNA pairing values are not common, and the one analysed for distortion of DNA relationships from Nakamura and Swezey (1983*a*) is the largest one I found. Not all of the primary data was published by them, because the values for each of the three replicates are not given separately, and the extent of test reproducibility cannot therefore be directly determined (though the error from zero-sided triangles was found to be about 5.08). It is highly desirable that in such studies the full details be published to allow this to be examined.

Most of the types of distortion observed on phenotypic similarities in a study by Sneath (1983) are seen here. There are no obvious effects peculiar to DNA:DNA data. Choice of strain is a far more important factor than choice of cluster method. There were only minor differences from UPGMA when Single Link or Complete Link clustering was used (Figures 3.7, 3.8).

The most obvious effect is the tendency of outliers to be drawn inwards (even when all strains were used as reference strains Figures 3.7, 3.29). This is more obvious for strains in the loose areas than those in the tight clusters. When all strains are employed as reference strains the clusters may be compressed relative to intercluster distances. This effect (measured by the ratio R of intercluster to intracluster sums of squares) was found in the example of Sneath (1983, Table 4). The *Bacillus* data was not suitable to analyse this as there was only a single well defined cluster.

Reference strains tend to become positioned on the periphery of the configuration. This is not always marked; thus in Figure 3.32 strain 8 is not noticeably peripheral, and the effect is not as constant as in the study of Sneath (1983). Reference strains that belong to a tight cluster

show the effects the least (Figure 3.11, C1090 and C1091; Figure 3.31, strain 1). Reference strains, and strains close to them, tend to disperse. This is more obvious for a loose cluster such as Figure 3.30 than for the tight cluster Figure 3.34 (strain 1) (Also see Figure 3.17).

Swivelling of strains along one major axis is very obvious if reference strains are very close (Figures 3.18, 3.19, 3.33). This is because such a choice is approximately the same as choosing a single reference strain. When a single reference strain is employed the configuration necessarily become linear, because all derived distances behave as if measured from one point (e.g., Sneath, 1983, Figure 10). Further, when reference strains are extremely similar, a considerable portion of the differences between their DNA-DNA pairing values may be due to chance effects of experimental error. Much of the detail in the derived configuration may then depend on these chance effects. Such choices represent, as it were, views from almost a single point in space, or from points of uncertain position.

A single reference strain is thus very unsatisfactory. If the analysis employs principal axis methods this necessarily aligns all the strains along a straight line. If it employs minimal spanning trees this necessarily creates a fan of all points spread round a central reference strain (see Figure 10 in Sneath, 1983). In either instance structure is grossly distorted.

If a reference strain is chosen from each cluster, and others are well spaced, good recovery of structure is obtained (Figures 3.11, 3.21). The problem, of course, is how to make such choices before the clusters are known. The risk of obtaining spurious structure from an unsuitable choice of reference strains is well illustrated by Figures 3.17, 3.18, 3.30, 3.32, 3.33, 3.34). Choosing reference strains based on a previously

show the effects the least (Figure 3.11, C1090 and C1091; Figure 3.31, strain 1). Reference strains, and strains close to them, tend to disperse. This is more obvious for a loose cluster such as Figure 3.30 than for the tight cluster Figure 3.34 (strain 1) (Also see Figure 3.17).

Swivelling of strains along one major axis is very obvious if reference strains are very close (Figures 3.18, 3.19, 3.33). This is because such a choice is approximately the same as choosing a single reference strain. When a single reference strain is employed the configuration necessarily become linear, because all derived distances behave as if measured from one point (e.g., Sneath, 1983, Figure 10). Further, when reference strains are extremely similar, a considerable portion of the differences between their DNA-DNA pairing values may be due to chance effects of experimental error. Much of the detail in the derived configuration may then depend on these chance effects. Such choices represent, as it were, views from almost a single point in space, or from points of uncertain position.

A single reference strain is thus very unsatisfactory. If the analysis employs principal axis methods this necessarily aligns all the strains along a straight line. If it employs minimal spanning trees this necessarily creates a fan of all points spread round a central reference strain (see Figure 10 in Sneath, 1983). In either instance structure is grossly distorted.

If a reference strain is chosen from each cluster, and others are well spaced, good recovery of structure is obtained (Figures 3.11, 3.21). The problem, of course, is how to make such choices before the clusters are known. The risk of obtaining spurious structure from an unsuitable choice of reference strains is well illustrated by Figures 3.17, 3.18, 3.30, 3.32, 3.33, 3.34). Choosing reference strains based on a previously

determined structure may just emphasise that structure, as the reference strains are pushed away from each other, whether it is truly correct or not.

If a reference strain is a singleton, it may compress other strains into a false cluster (Figure 3.31). A similar effect is seen when *L. grayi* and *L. murrayi* are included then removed from the configuration (Figures 3.7, 3.10) and when only the two *L. seeligeri* strains were used as reference strains (Figure 3.18). There is a tendency for reference strains in a tight cluster to push one or two strains out, rather than to simply expand the cluster (for example Figure 3.32). Loose clusters with relatively few members are particularly easily distorted. This was noticed with the random swarm data with, for example, points 1, 3, 13, and to some extent with the *Bacillus* data with strains 3, 4, 5, although this requires some further conformation with other studies. Omission of one reference strain from a good set has minor effects.

Information on the underlying taxonomic structure is necessarily lost when a small number of reference strains are chosen. The new configuration has fewer dimensions than the starting configuration, and it has been noted (Sneath, 1983) that for c reference strains the configuration will have an effective dimensionality of about $c - 0.5$ at the most, because the points will be reflected to one side of a hyperplane of $c - 1$ dimensions (i.e. into one half of a configuration of c dimensions. This reduction of effective dimensionality is accentuated if reference strains are close together, as noted earlier.

One cannot therefore judge the amount of the underlying variation by looking simply at the first few eigenvalues if these are derived from only a few reference strains. All the variation will be in the first c eigenvalues. The fact that, for example, the first two eigenvalues may

recover 99 % of the variation does not ensure that a two-dimensional diagram, based on say three reference strains, will contain almost all the taxonomic structure. Clusters and points that are well separated in the full space may be overlapped in a scatter diagram.

Choice of only two reference strains, even if well spaced, constrains the resulting figuration to a plane. This can be seen by the constant height of points in diagrams such as Figures 3.16, 3.18, 3.19, 3.31, 3.32, 3.33. The risk of overlapping for a plausible statistical model has been shown to be related to the chi-square distribution (Sneath, 1980; 1983). Reduction to two dimensions greatly increases the risk in the chi-square table. If there are several fairly close clusters the danger is considerable, even for three dimensions.

The random swarm study (section 3.10, p.138) backed up the findings from the *Bacillus* data. From the original 17 x 17 ordination (Figure 3.36), points 2 and 6 were picked out as a taxonomic group, however this was not obvious from Figures 3.37 and 3.38. If a study using selected reference strains was carried out (as in Figures 3.37, 3.38) then points 3, 5, 7, 9, 10, 11, 12, 13, 14, 15, 16 would be described as a very closely related taxon. This would be a misconception as seen from the 'true' data in Figure 3.36.

Published Standard Deviations.

A further paper using *Bacteroides* (Tanner *et al.*, 1986) and the optical technique was examined where a small number of replications showed an average error of 0.29 % but this paper is not representative; there were only six replicates and all pairing values were very close to 100 %.

Error seemed to be largely independent of the degree of pairing, although there was some evidence for an increase in error at high homology values in Potts and Berry's (1983) data (Figure 3.39). This is not readily explicable, because in theory the error at least in the optical method will be greatly constrained near 100 % (and near 0 %) and therefore the error here would be reduced. Indeed the error from zero-sided triangles suggests this may be so (Figure 3.41).

Zero-sided triangles

The paper by Nakamura and Swezey (1983*b*) on *Bacillus* sp. had 147 measurable errors of which most were small; the average was raised by a few high errors. Two of the largest errors were based on DNA pairing values of less than 35 %. One explanation is that the purity of DNA or fragment size may vary between the preparations from the strains; these factors are known to affect the degree of binding and the reproducibility of the results (DeLey *et al.*, 1970).

There may be a tendency to round up high values to 100 %. Homology values are generally published as integers, due to the amount of experimental error involved; so many '100 %' pairing values may be, for example, 99.6 %. If pairing determinations could be calculated without any error many of the zero distances would be found to show less than 100 % pairing, This could theoretically affect this method of error

estimation, because such triangles would then be excluded from consideration. I do not however, believe this has caused major bias. If, for example, a pair of strains has a pairing value recorded as 100 % when the true value should be 99 %, this implies that the pairing values to other strains should agree to within one or two percent at the most. The discrepencies observed are commonly far greater than this and the reliability of estimates from zero-sided triangles is supported by other methods of estimation (Table 3.17, Figure 3.42).

It might be thought that if a triangle had a side that was very small, but not zero, one could estimate error if the other two sides differed by a considerable amount. Thus a triangle with sides 1, 10, 30 implies an error close to that estimated from a triangle with sides 0, 10, 30. This, however, requires subtraction of a correlation term (related to quantity 1) from the difference between 10 and 30 and this is not statistically straightforward. The problem was explored by computer simulation and an empirical correction was determined, but the estimates, which were of similar magnitude to those from zero-sided triangles, did not yield further reliable information.

Triangle Inequalities

The proportion of violating triangles fluctuates widely with little relation to average error, method or group of organisms. Thus two studies on *Veillonella* (Mays *et al.*, 1982; Johnson and Harich, 1983) show 1.4 % and 32.3 % of violating triangles respectively, though the error rates are quite typical (Table 3.). Similarly, percentages on *Bacillus* by Nakamura and his colleagues (Nakamura and Swezey, 1983*a*, b; Nakamura, 1987*a*, *b*) vary from 6.5 % to 39.7 % (Table 3.13). High percentages tend to be associated with high error rates though this effect is not marked. Percentages tend to be high in Gram-positive bacteria, particularly for

streptococci and enterococci, but this may not be significant. The expected association between violating triangles and zero-sided triangles has been noted in the results.

The frequent occurrence of triangle inequalities suggests that 100 - % DNA:DNA pairing does not necessarily behave as a eucluidean metric. It is not clear to what extent the inequalities are due to experimental error, but the unexpected features just noted suggest that they are an expression of physicochemical factors that are not yet understood. This would have implications for taxonomic conclusions that are drawn from DNA studies. It seems unlikely that the rather higher error in studies on Gram-positive compared to Gram-negative bacteria would account for the higher proportion of triangles that violate the inequality in the former (Table 3.17). More accurate pairing values are needed to decide this point, but it is possible that DNA pairing differences are inherently non-Eucluidean; if so it would be worth trying the square-root transformation in systematic studies. At present there are few data available for such work, but the effect of square-rooting was examined on the complete matrix of Nakamura and Swezey (1983*a*). The same dendrogram topology was found and very similar principal coordinate relationships (Figure 3.43) as those with unrooted values (Figure 3.29); the main effect was expansion of the scale near the tips of the dendrogram and looser grouping of the tight cluster of strains in the principal coordinate diagram. Because this matrix shows few violations, and a simple taxonomic structure, these findings are not unexpected. The transformation might be of importance in evolutionary reconstructions. Consider a situation where, for strains *a*, *b* and *c*, for example, the implied evolutionary change *a* : *c* between strains *a* and *c* is greater than the sum of the changes *a* : *b* and *b* : *c*. This will prevent evolutionary distances from being additive, and it would have undesirable consequences for alogarithms for phylogenetic reconstruction.

General Factors

The greater error with Gram-positive than Gram-negative (Table 3.15) may well be due to the tough cell wall of the former that makes them so difficult to lyse, which may effect fragment size or purity of the DNA. DNA isolation is difficult for some *Streptococcus* species (Garvie, 1978). This may be why error is particularly high with these organisms. All reciprocal-pair errors with streptococci were much higher than the average for other papers, as were the errors for zero-sided triangles. The proportion of violating triangles was also very high, about one third of all the triangles compared to an average of about 8 % for all papers studied. Analysis of varience carried out on error from reciprocal pairs and error from zero-sides showed no significant differences between the studies on *Streptococcus* and *Enterococcus* at $P < 0.05$.

Error in studies such as those by Gebers *et al.* (1986) and Love *et al.* (1987*a*) may perhaps be high because of the wide % (G+C) range covered; organisms with a % (G+C) difference of more than a few percent will have widely different T_m values and hence different optimal renaturation temperatures; as previously discussed, a temperature of as little as 4°C above the T_{or} may affect the reliability of the results (section 3.6).

Previous comparisons of reassociation techniques showed a straight line relationship between values at high pairing levels, but the relation is curved when extended to levels below 30-40 % pairing (Huss *et al.*, 1983; Bouvet and Grimont, 1986). This behaviour deserves further study. It may be noted, however, that there is considerable scatter about the lines and this scatter must reflect test error. Examination of their figures suggests that the error is 5-8 %, similar to the average found here.

Gibbins and Gregory (1972) found standard deviations with the optical technique (± 4.11 %) to be much higher than those with the filter technique (± 0.87 %) this is the reverse of the findings in this study; it can probably be explained by the vast improvement in optical techniques recently due to instrumentation and hence the ability to run an entire experiment at the same time (i.e. all three cuvettes).

In the study by Huss *et al.* (1983) the magnitude of the DNA:DNA pairing in the filter and optical techniques can be compared directly for values over 30-40 %. The magnitudes for these two methods are about the same. However, the S1 nuclease technique, in Bouvet and Grimont's study (1986), has corresponding pairing values of approximately 20 % less than the filter technique (for values around 30-40 % pairing). This was also found with the data from *Listeria* strains (see Tables 4.1, 4.2). It seems from these studies that the S1 nuclease technique may be more stringent than the other two methods. This difference is not seen in the data from Whiley *et al.* (1988). There were only six comparable values in the latter study, so the question remains open. Differences in the stringency and accuracy of techniques brings into doubt the use for specific cut-off levels for defining a species, sub-species, genus etc. If these must be employed then they should be defined for hybridisations at the T_{or} for each technique. Wayne *et al.* (1987) recently made such recommendationsdefining a species in phylogenetic terms as "strains with approximately 70 % or greater DNA-DNA relatedness and with 5°C or less ΔT_m. Both values must be considered." This would suggest, from both my data and that of Rocourt *et al.* (1982), that *Listeria monocytogenes sensu stricto* consists of more than one genospecies. However, Wayne *et al.* (1987) also recommend that if it is not possible to differentiate strains by any phenotypic properties then they should not be regarded as separate species but as subspecies.

Conclusions and suggestions for further study.

Two- and three-dimensional representations are unsafe for deducing taxonomic structure unless the identity of the strains and taxa are known beforehand, so that points can receive different symbols. If they are so symbolised, the representations are not giving structure *de novo*, but are only confirming previously known structure. It is for this reason that dendrograms are safer than principle axes plots for defining taxonomic groups. In the *Bacillus* sample the structure is not sufficiently complex for there to be much difference between the results of the two methods, but the principle involved is important.

When DNA pairing techniques are used in systematics the error associated with these techniques must affect the inferred relationships. Indeed, all the methods of estimation suggest that the error averages 5-6 %, though varying considerably from study to study. Unless the error is taken into account, incorrect conclusions could easily be drawn, especially if only a few reference strains are used.

It is not yet clear what is the best transformation of DNA-DNA differences but the findings on non-euclidean triangles show that this certainly requires study.

There does not seem to be any distinct 'genospecies' within *L. monocytogenes sensu stricto* but a wide spectrum of homology values. Analysis of strains of *Listeria* using DNA techniques is required involving more strains of *L. innocua* and *L. monocytogenes* to see if any distinct clusters are present. Sequencing of 16srRNA of these two species may reveal whether or not they are two distinct species. As haemolysis is the only phenotypic difference between *L. monocytyogenes* and *L. innocua* strains it would be interesting to see if the gene sequence for haemolysis is present in both species.

5. APPENDICES

Appendix 1 % DNA-DNA dissimilarities from Nakamura and Swezey (1983*a*) for 17 Strains of *Bacillus circulans*.

Strain	Strain 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1,313[1]	0																
2,358	0	0															
3,385	94	95	0														
4,387	90	73	37	0													
5,397	91	82	79	74	0												
6,726	0	2	90	96	76	0											
7,727	5	1	85	88	86	0	0										
8,728	0	0	92	89	75	0	2	0									
9,729	2	0	91	93	90	1	0	3	0								
10,746	3	0	83	90	75	1	0	3	0	0							
11,765	0	3	77	92	82	1	2	0	0	0	0						
12,826	64	71	93	84	84	69	64	73	74	68	68	0					
13,831	10	15	95	96	94	10	13	14	10	9	12	74	0				
14,1108	42	49	74	91	79	37	44	33	49	53	52	65	62	0			
15,1670	89	88	91	94	90	75	72	69	79	81	82	80	85	77	0		
16,1341	65	67	70	80	92	65	68	70	74	92	90	81	70	67	77	0	
17,1353	94	90	96	92	86	97	95	89	92	87	90	86	93	84	67	90	0

[1]The strain serial number is given first followed by the NRRL-NRS number.

Appendix 2 Pascal Program TRUDNA.PAS for the VAX cluster.

Input Files

a) SQTITLE.DAT

If desired, the names of the OTUs may be written to this file. The name or strain number may be up to 18 characters long, and each is entered on a separate line (i.e. *n* rows of names for *n* OTUs.

b) SQVALUES.DAT

Homology data is entered in any order in the form:

```
1 1 100
1 2 90
4 6 43
```

i.e. the numbers of the two OTUs involved separated by a space, followed by the homology value. Each data statement is entered on a separate line. The matrix does not have to be complete and the strain numbers (from 1 to *n*) do not need to be entered in any particular order. The % homology values may be integers or real numbers.

```
(*********************************************)
(*                                           *)
(* Program to Analyse DNA Homology Values    *)
(*                                           *)
(*         Trudy Hartford                    *)
(*          April 1986                       *)
(*********************************************)

PROGRAM  DNADATA(INPUT,OUTPUT,SQVALUE,SQTITLE);

CONST    NMAX=200;  (*MAX NO. OF STRAINS ALLOWED*)

VAR      zeno,N,STRAINA,STRAINB,A,B,C,SMD,MD : INTEGER:
         DATA : ARRAY[1..NMAX,1..NMAX]OF REAL;
         SQVALUE,SQTITLE : TEXT;
         D,P,Q,R,CH,CHI,ZEHI,Y,X,K,PP,QQ,RR,J,NOE : REAL;
         REPLY  : CHAR;
         DATALACK,BEHAVE,MISBEHAVE : INTEGER;
         SUMZE1,SUMPZ,PZ,SUMPE,SUMZE,SUME,BIG,SMALL,MIN,MAX : REAL;

PROCEDURE SPECIESNUMBER;
(* To establish the number of OTUs involved *)
BEGIN
 WRITE('ENTER NUMBER OF OTUS: ');
 READLN(N);
END: (*SPECIESNUMBER*)

PROCEDURE ZEROARRAY;
(* To zero the matrix*)
BEGIN
 FOR STRAINA:= 1 TO N DO
 BEGIN
  FOR STRAINB:= 1 TO N DO
```

```
   BEGIN
   DATA[STRAINA,STRAINB]:= -1:    (* All positions made -1*)
  END;
 END;
END; (*ZEROARRAY*)

PROCEDURE ENTERNAMES;
(* reads species names from file:sqtitle *)
(* used as a check; up to 25 characters allowed *)
VAR  I : INTEGER:
     C1,C2,C3,C4,C5,C6,C7,C8,C9,C10,C11,C12,C13,C14,C15,C16,C17 : CHAR;
     C18,C19,C20,C21,C22,C23,C24,C25 : CHAR;
BEGIN
 I:=0;
 RESET(SQTITLE);
 REPEAT
 I:=I+1;

READLN(SQTITLE,C1,C2,C3,C4,C5,C6,C7,C8,C9,C10,C11,C12,C13,C14,C15,C16,C
17,
 C18,C19,C20,C21,C22,C23,C24,C25);
 WRITELN('OTU ',I:3,' IS:
',C1,C2,C3,C4,C5,C6,C7,C8,C9,C10,C11,C12,C13,C14,
 C15,C16,C17,C18,C19,C20,C21,C22,C23,C24,C25)
 UNTIL EOF(SQTITLE)
 END; (*ENTERNAMES*)

PROCEDURE ENTERDATA;
(*Reads known homology values from a separate file : sqvalue.dat*)
(* and sorts them into the correct position in the matrix*)

BEGIN
 RESET(SQVALUE);
 REPEAT
 READLN(SQVALUE,STRAINA,STRAINB,D);
 DATA[STRAINA,STRAINB]:=D
 UNTIL EOF(SQVALUE)
END;

PROCEDURE COUNTVALUES;
(*Counts proportion of known values*)

VAR  V,VM,VT :  INTEGER;
     VP,VMP  :  REAL;
BEGIN
 V:=0;
 VM:=0;
 FOR STRAINA:=1 TO N DO
 BEGIN
  FOR STRAINB:=1 TO N DO
  BEGIN
   D:=DATA[STRAINA,STRAINB];
```

```
    IF D>=0 THEN
    V:= V + 1
    ELSE VM:= VM + 1;
   END;
  END;
  WRITELN('NUMBER OF KNOWN ENTRIES= ',V);
  WRITELN('NUMBER OF UNKNOWN DATA ENTRIES= ',VM);
  VT:=VM + V;
  VP:=V/VT*100;
  VMP:=100-VP;
  WRITELN;
  WRITELN('THE % OF KNOWN VALUES IS ',VP:2:2);
  WRITELN('THE % OF MISSING VALUES IS ',VMP:2:2);
  WRITELN;
END;

PROCEDURE CONVERTVALUES;
(*Converts values to distances, assumes largest % homology =100*)

BEGIN
 FOR STRAINA:=1 TO N DO
 BEGIN
 FOR STRAINB:=1 TO N DO
  BEGIN
   IF DATA[STRAINA,STRAINB]<=100 THEN BEGIN
    IF DATA[STRAINA,STRAINB]>=0 THEN
    DATA[STRAINA,STRAINB]:=100-DATA[STRAINA,STRAINB];
   END;
  END;
 END;       (*CONVERTVALUES*)

PROCEDURE KNOWNDISTANCES;
(*assumes largest distance = 0, i.e. max % = 100*)

BEGIN
 FOR STRAINA:=1 TO N DO
 BEGIN
  FOR STRAINB:=1 TO N DO
  BEGIN
   IF DATA[STRAINA,STRAINB]>=0 THEN
   WRITELN('DISTANCE FROM ',STRAINA,' TO ',STRAINB,
   ' IS: ',DATA[STRAINA,STRAINB]:2:4);
  END;
 END;
 WRITELN;
END;

PROCEDURE BEHAVIOUR;
(*To find the sets of data points that behave as triangles *)
(*i.e. A+B>=C where C is the longest side of the triangle *)
(* B and C being the other two sides *)
BEGIN
```

```
MAX:=0;
 MIN:=0;
 P:=DATA[C,B];
 Q:=DATA[C,A];
 R:=DATA[B,A];
 IF P>Q THEN MAX:=P ELSE MAX:=Q:
 IF R>MAX THEN MAX:=R;
 IF P<Q THEN MIN:=P ELSE MIN:=Q;
 IF R<MIN THEN MIN:=R;
 IF MIN<0 THEN DATALACK:=DATALACK + 1;
 IF (P+Q+R-MAX) >=MAX THEN BEHAVE:=BEHAVE + 1
 ELSE MISBEHAVE:=MISBEHAVE + 1;
END; (*BEHAVIOUR*)

PROCEDURE BEHAVE2;
BEGIN
 MAX:=0;
 MIN:=0;
 P:=DATA[C,B];
 Q:=DATA[A,C];
 R:=DATA[B,A];
 IF P>Q THEN MAX:=P ELSE MAX:=Q;
 IF R>MAX THEN MAX:=R;
 IF P<Q THEN MIN:=P ELSE MIN:=Q;
 IF R<MIN THEN MIN:=R;
 IF MIN<0 THEN DATALACK:=DATALACK+1;
 IF (P+Q+R-MAX) >=MAX THEN BEHAVE:=BEHAVE + 1
 ELSE MISBEHAVE:=MISBEHAVE + 1;
 IF MIN>=0 THEN BEGIN
  IF (P+Q+R-MAX)<MAX THEN WRITELN(A,B,C)
 END;
END;

PROCEDURE LOOP;
(* the loop required for the eight possible methods of reading a
triangle*)
BEGIN
 MISBEHAVE:=0;
 DATALACK:=0;
 BEHAVE:=0;
 FOR A:=3 TO N DO BEGIN
  FOR B:=2 TO A-1 DO BEGIN
   FOR C:=1 TO B-1 DO BEGIN
    BEHAVIOUR;
    BEHAVE2;
    END;
   END;
  END;
 FOR A:=3 TO N DO BEGIN
  FOR C:=2 TO A-1 DO BEGIN
   FOR B:=1 TO C-1 DO BEGIN
    BEHAVE2;
```

```
    BEHAVIOUR;
    END;
   END;
  END;
FOR C:=3 TO N DO BEGIN
 FOR B:=2 TO C-1 DO BEGIN
  FOR A:=1 TO B-1 DO BEGIN
   BEHAVIOUR;
   END;
  END;
 END;
FOR C:=3 TO N DO BEGIN
 FOR A:=2 TO C-1 DO BEGIN
  FOR B:=1 TO A-1 DO BEGIN
   BEHAVIOUR;
   END;
  END;
 END;
FOR B:=3 TO N DO BEGIN
 FOR C:=2 TO B-1 DO BEGIN
  FOR A:= 1 TO C-1 DO BEGIN
   BEHAVIOUR;
   END;
  END;
 END;
FOR B:=3 TO N DO BEGIN
 FOR A:=2 TO B-1 DO BEGIN
  FOR C:=1 TO A-1 DO BEGIN
   BEHAVIOUR;
   END;
  END;
 END;
 WRITELN('THE NUMBER OF COMPLETE TRIANGLES + ',BEHAVE:2);
 WRITELN('THE NUMBER OF INCOMPLETE TRIPLES + ',DATALACK:2);
 WRITELN('THE NUMBER OF VIOLATING TRIPLES =',(MISBEHAVE-DATALACK):2);
END;

PROCEDURE ORDER;
(* To arrange three distances in numerical order*)

VAR  MID,DE,AV :  REAL;
BEGIN
 IF P>Q THEN MAX:=P ELSE MAX:=Q;
 IF P<Q THEN MID:=P ELSE MID:=Q;
 IF R>MID THEN MID:=R;
 IF R>MAX THEN MID:=MAX;
 IF R>MAX THEN MAX:=R;
 IF (MID+MAX)=0 THEN AV:=0 ELSE AV:= (MID+MAX)/2;
 OE:=2*SQR(MAX-AV);
IF AV=0 THEN CH:=0 ELSE CH:=OE/AV;
END;
```

```
PROCEDURE SQUAREROOT;
(*To squareroot the known distances*)
(*will only squareroot +ve values*)
BEGIN
 FOR STRAINA:=1 TO N DO
 BEGIN
  FOR STRAINB:=1 TO N DO
   BEGIN
    D:= DATA[STRAINA,STRAINB];
    IF D>0 THEN
    DATA[STRAINA,STRAINB]:= (+SQRT(D));
   END;
  END;
 END; (*SQRT*)

PROCEDURE ZEROERROR;

VAR  ZE,W,WE,U,UE,WUE,ZE1,WE1,WUE1,UE1 : REAL;
BEGIN
 Q:=DATA[C,A];
 R:=DATA[B,A];
 IF Q<R THEN MIN:=Q ELSE MIN:=R;
 IF MIN>0 THEN ZE:=(0.5*(Q-R)*(Q-R)) ELSE ZE:=-1;
 IF ZE>=0 THEN ZE1:=SQRT(ABS(ZE));
 (*i.e. ZE is the average variance*)
 IF ZE>= 0 THEN SUMZE:=SUMZE + ZE;
 IF ZE>= 0 THEN SUMZE1:=SUMZE1 + ZE1;
 IF ZE>= 0 THEN ZENO:=ZENO + 1;
 W:=DATA[A,B];
 IF W<Q THEN MIN:=W ELSE MIN:=Q;
 IF MIN>0 THEN WE:=(0.5*(W-Q)*(W-Q)) ELSE WE:=-1; IF WE>=0 THEN
 WE1:=SQRT(ABS(WE));
 IF WE>0 THEN BEGIN
  SUMZE:=SUMZE + WE;
  SUMZE1:=SUMZE1 + WE1;
  ZENO:=ZENO + 1;
 END;
U:=DATA[A,C];
IF U<R THEN MIN:=U ELSE MIN:=R;
IF MIN>0 THEN UE:=(0.5*(U-R)*(U-R)) ELSE UE:=-1;
IF UE>=0 THEN UE1:=SQRT(ABS(UE));
IF UE>ZEHI THEN ZEHI:=WE;
IF UE>=0 THEN BEGIN
  SUMZE:=SUMZE + UE;
  SUMZE1:=SUMZE1 + UE1;
  ZENO:=ZENO + 1;
 END;
IF U<W THEN MIN:=U ELSE MIN:=W;
 IF MIN>0 THEN WUE:=(0.5*(W-U)*(W-U)) ELSE WUE:=-1;
 IF WUE>=0 THEN WUE1:=SQRT(ABS(WUE));
 IF WUE>=0 THEN BEGIN
  SUMZE:=SUMZE + WUE;
```

```
  SUMZE1:=SUMZE1 + WUE1;
  ZENO:=ZENO + 1;
 END;
END; (*ZEROERROR*)

PROCEDURE ERROR;
(*To find experimental error *)
(*where error is equivalent to standard deviation or sum of squares*)
(*corresponds to procedure zeroerror*)
BEGIN
 WRITELN('£££££££ ERROR FROM TRIPLES WITH ZERO DISTANCES £££££££');
 ZENO:=0;
 SUMZE:=0;
 SUMZE1:=0;
 FOR A:=3 TO N DO BEGIN
  FOR B:=2 TO A-1 DO BEGIN
   FOR C:=1 TO B-1 DO BEGIN
    IF DATA[C,B]=0 THEN ZEROERROR;
    IF DATA[B,C]=0 THEN ZEROERROR;
   END;
  END;
 END;
 FOR B:=3 TO N DO BEGIN
  FOR A:=2 TO B-1 DO BEGIN
   FOR C:=1 TO A-1 DO BEGIN
    IF DATA[C,B]=0 THEN ZEROERROR;
    IF DATA[B,C]=0 THEN ZEROERROR;
   END;
  END;
 END;
 FOR C:=3 TO N DO BEGIN
  FOR B:=2 TO C-1 DO BEGIN
   FOR A:=1 TO B-1 DO BEGIN
    IF DATA[C,B]=0 THEN ZEROERROR;
    IF DATA[B,C]=0 THEN ZEROERROR;
   END;
  END;
 END;
 IF ZENO>0 THEN WRITELN('THE AVERAGE VARIENCE WAS ',SUMZE/ZENO:2:2,' N=
',ZENO);
 IF ZENO>0 THEN WRITELN('THE AVERAGE STD DEVN WAS ',SUMZE1/ZENO:2:2);
 END;

PROCEDURE SQERROR;
(* otu a vs otu b should = otu b vs otu a *)

VAR  S2,SAVD : REAL;
     SUMS11,SUMS,DIF,SUMS2,SQE,SQDIF,SP,SQ,RTMNSQDIF : REAL;
     S11,SQE1,S1,S,LO,HIP,HI,SUMS1,NO : REAL;
BEGIN
 WRITELN('£££££££££ ERROR FROM RECIPROCAL PAIRS £££££££££');
 WRITELN('ERROR IS VARIENCE');
```

```
WRITELN;
S:=0;
NO:=0;
SUMS1:=0;
SP:=0;
HIP:=0;
DIF:=0;
SQ:=0;
SUMS11:=0;
SUMS2:=0;
SUMS:=0;
SQDIF:=0;
SAVD:=0;
FOR A:=2 TO N DO
BEGIN
 FOR B:=1 TO A-1 DO
 BEGIN
  IF DATA[A,B]>=0 THEN BEGIN
  IF DATA[B,A]>=0 THEN BEGIN
   P:=DATA[A,B];
   Q:=DATA[B,A];
   SP:=SP + P;
   NO:=NO + 1;
   SQ:=SQ + Q;
   IF P<Q THEN LO:=P ELSE LO:=Q;
   IF P>Q THEN HI:=P ELSE HI:=Q;
   IF LO>=0 THEN DIF:=HI-LO ELSE DIF:=0;
   IF DIF>0 THEN BEGIN
    SQDIF:=SQDIF + SQR(ABS(HI-LO));
    S:=(HI+LO)/2
    SAVD:=SAVD + S;
    S1:=(S-LO)*(S-LO)*2;
    WRITELN('PAIR  ',A,B,' E= ',S1);
    S11:=SQRT(ABS(S1));
    IF S1>HIP THEN HIP:=S1;
    SUMS1:=SUMS1 + S1;
    SUMS11:=SUMS11 + S11;
    IF S>0 THEN SUMS:=SUMS + S;
    END;
   END;
  END;
 END;
END;
WRITELN;
WRITELN('SUM OF AVERAGE DISTANCES =',SAVD:2:2);
WRITELN('SUM OF INTER-DISTANCE VARIANCES = ',SUMS1:2:2);
WRITELN;
IF NO>0 THEN BEGIN
 WRITELN('SUM OF DIFFERENCES =',(SP-SQ):2:3);
 WRITELN('MEAN DIFFERENCE = ',(SP-SQ)/NO:2:3);
 WRITELN('SUM OF SQUARE DIFFERENCES= ',SQDIF:2:3);
 IF SQDIF>0 THEN SQDIF:=(SQDIF/NO) ELSE SQDIF:=0;
```

```
  IF SQDIF>0 THEN RTMNSQDIF:=SQRT(ABS(SQDIF)) ELSE RTMNSQDIF:=0;
  WRITELN('ROOT OF MEAN SQUARE DIFFERENCE = ',RTMNSQDIF:2:3);
  IF HIP>0 THEN WRITELN('THE LARGEST ERROR WAS ',HIP:2:3);
  IF SUMS1>0 THEN BEGIN
   SQE:=SUMS1/NO;
   SQE1:=SUMS11/NO;
   WRITELN('THE AVERAGE VARIENCE ',SQE:2:3);
   WRITELN('THE AV. STD. DEVN = ',SQE1:2:3);
  END;
  WRITELN('NUMBER OF DETERMINATIONS =',NO);
  WRITELN;
 END;
END;

BEGIN
(*main program*)
 WRITELN('PROGRAM TRUDNA TO ESTIMATE ERRORS IN DNA HOMOLOGY DATA');
 WRITELN('THIS VERSION IS FOR SQUARE MATRICES, ALL DATA SHOULD BE IN');
 WRITELN('THE FILE SQVALUE.DAT (AND SQTITLE.DAT IF DESIRED).');
 SPECIESNUMBER;
 ZEROARRAY;
 WRITELN;
 ENTERDATA;
 WRITELN('TO CHECK THE HOMOLOGY BETWEEN TWO OTUS: ');
 REPEAT
 WRITE('SUPPLY OTU NUMBERS: ');
 READLN(STRAIN,NUMBER);
 IF DATA[STRAIN,NUMBER]>=0 THEN
 WRITELN('THE %HOMOLOGY = ',DATA[STRAIN,NUMBER])
 ELSE WRITELN('NO DATA,');
 WRITE('MORE OTUS?');
 READLN(ANSWER)
 UNTIL ANSWER = 'N';
 COUNTVALUES;
 CONVERTVALUES;
 WRITELN;
 WRITELN('£££££ DISTANCE = 100 - %HOMOLOGY £££££');
 WRITELN;
 WRITE('DO YOU WANT A LIST OF KNOWN DISTANCES?');
 READLN(REPLY);
 IF REPLY = 'Y' THEN KNOWNDISTANCES;
 LOOP;
 ERROR;
 WRITELN('£££££ DISTANCE = SQRT(100-%HOMOLOGY £££££);
 WRITELN;
 SQUAREROOT;
 LOOP;
 ERROR;
 SQERROR;
 WRITELN(MISBEHAVE,BEHAVE);
 END.
```

Appendix 3 Variation in Sample Position in the PU800 Spectrophotometer.

	Position Cell	Probe	T_m °C	Ramp Rate (°Cmin^{-1})	ΔT_m^1	ΔT_m^2
E. coli (Sigma) T_m = 77°C	2	4	76.3	1.0	-0.7	0.1
	3	4	77.2	"	0.2	1.0
	5	4	76.2	"	-0.8	0
	7	4	76.4	"	-0.6	0.2
	2	4	76.9	1.0	-0.1	0.9
	3	4	76.2	"	-0.8	0.2
	4	4	76.5	"	-0.5	0.5
	5	4	76.0	"	-1.0	0
	6	4	76.1	"	-0.9	0.1
	7	4	76.8	"	-0.2	0.8
	8	4	77.5	"	0.5	1.5
	2	4	78.0	1.0	1.0	
	3	4	76.6	"	-0.4	
	4	4	77.0	"	0.0	
	6	4	76.6	"	-0.4	
	2	5	76.1	1.0	-0.9	
	3	5	74.9	"	-2.1	
	4	5	75.0	"	-2.0	
	7	5	75.1	"	-1.9	
	8	5	76.4	"	-0.6	
	3	5	72.4	1.0	-4.6	-0.9
	4	5	72.5	"	-4.5	-0.8
	5	5	73.3*	"	-3.7	0
	6	5	71.7	"	-5.3	-1.6
	7	5	72.6	"	-4.4	-0.7
	8	5	73.0	"	-4.0	-0.3
	2	5	76.2	0.5	-0.8	-0.3
	3	5	74.7	"	-2.3	-1.8
	4	5	75.2	"	-1.8	-1.3
	5	5	76.5	"	-0.5	0
	6	5	74.8	"	-2.2	-1.7
	7	5	74.9	"	-2.1	-1.6
	8	5	76.4	"	-0.6	-0.1
	2	5	77.1	0.5	0.1	0.6
	3	5	75.9	"	-1.1	-0.6
	4	5	76.2	"	-0.8	-0.3
	5	5	76.5	"	-0.5	0

7	5	76.3	"	-0.7	-0.2
8	5	76.8	"	-0.2	0.3
2	5	76.4[P]	0.5	-0.6	-0.1
3	5	76.4[P]	"	-0.6	-0.1
4	5	76.7	"	-0.3	0.2
5	5	76.5	"	-0.5	0
6	5	75.8	"	-1.2	0.3
8	5	77.9	"	0.9	1.4
2	5	76.6	0.5	-0.4	0.6
3	5	75.3	"	-1.7	-0.7
4	5	75.8	"	-1.2	-0.2
5	5	76.0	"	-1.0	0
6	5	75.3[P]	"	-1.7	-0.7
7	5	75.8	"	-1.2	-0.2
8	5	76.8	"	-0.2	0.8

[P] : poor curve
[*] : some evaporation occurred.
[1] : T_m - Expected T_m = 77°C (Sigma)
[2] : T_m - T_m of cell position 5.

Appendix 4 : Polysaccharide Removal Experiments.

1. Effect on the T_m.

Strain		Untreated DNA T_m	Untreated DNA T_m^r	Untreated DNA % Hyper-chromicity	Treated DNA T_m	Treated DNA T_m^r	Treated DNA % Hyper-chromicity
C52	i	90.5	n.d.	35.1	89.7	n.d.	n.d.
	ii	90.5	90.1	25.5	91.0	"	22.6
		90.1	90.5	n.d.			
	iii	90.1	87.4	32.4	89.8	87.2	26.6
		89.4	88.1	34.6	89.2	-	31.8
	iv	88.5	86.6	n.d.	90.2	87.8	n.d.
C644	i	89.0	n.d.	n.d.	89.0	n.d.	n.d.
		88.9	n.d.	n.d.	89.5	n.d.	n.d.
		88.4	n.d.	n.d.	90.6	n.d.	n.d.
					90.7	n.d.	n.d.
					89.7	n.d.	n.d.
	ii	88.6	87.4	n.d.	88.6	87.5	n.d.
		87.6	n.d.	n.d.	87.8	n.d.	n.d.
C1090	i	87.4	88.5	n.d.	85.5	86.7	n.d.
		86.9	86.3	23.3	85.5	n.d.	27.5
		86.0	n.d.	25.4			
	ii	87.1	86.9	26.3	89.7	88.8	36.6
					82.6	81.8	32.3
		86.7	87.6	n.d.	88.4	89.0	n.d.
		90.5	91.4	n.d.	86.4	88.7	n.d.
	iii	88.3	88.3	n.d.	89.0	n.d.	n.d.
		89.0	87.7	n.d.	87.0	85.7	n.d.
		88.3	n.d.	n.d.	87.3	n.d.	n.d.
		90.5	n.d.	n.d.	88.8	n.d.	n.d.
		89.0	n.d.	n.d.			
		91.6	91.3	n.d.	89.4	87.9	n.d.
		89.6	88.4	n.d.	90.1	88.5	n.d.
		89.2	88.1	n.d.			

Average T_m

C52	89.9 ± 0.77 (n = 6)	90.0 ± 0.67 (n = 5)
C644	88.5 ± 0.56 (n = 5)	89.4 ± 1.05 (n = 7)
C1090	88.6 ± 1.64 (n = 14)	88.6 ± 2.20 (n =12)

2. Effect on Renaturation Rates.

Strain	Renaturation Rates untreated DNA	treated DNA	Mix[1]	% Pairing
C52	.00371	.00368	.00395	113.8
	.00373	.00365[2]	.00377	104.3
	.00292	.00301		
C644	.00219	.00383		
	.00237	.00392		
	.00210	.00325		
C1090	.00337	.00334	.00304	81.2
	.00323	.00285		
	.00265	.00374	.00322	103.1
	.00402	.00408		
	.00258	.00274		
	.00599	.00487	.00540	99.4
	.00576	.00531	.00547	

Heterologous Pairings:

Strain *a*	treated	Strain *b*	treated	Rate *a*	Rate *b*	Rate *ab*	% Homology	Expected Homology
C52	+	C201	-	.00301	.00315	.00301	95.5	89.4 ± 1.0
				.00344	.00330	.00320	89.9	
C52	-	C1090	+	.00288	.00456	.00334	81.7	52.2 ± 1.8

* : evaporation occurred
T_m^r : T_m of DNA after it has undergone renaturation.
[1] : 50:50 mix of treated and untreated DNA
[2] : 2-methoxyethanol treated, all other samples were treated with CTAB.

Appendix 5 Base Composition Determinations.

Strain	Position of: Cell	Probe	T_m	T_m^{adj}	Ramp	%(G+C)[1]	%(G+C)^{adj}
E. coli	5	5	74.0	74.0	1.0	47.5	53.8
C1091	2	"	65.8	65.0	"	36.7	35.1
"	3	"	65.4	65.5	"	35.9	36.1
"	4	"	65.4	65.4	"	35.9	35.9
C644	6	"	67.0	67.1	"	39.2	39.4
"	7	"	67.0	66.3	"	39.2	37.8
"	8	"	67.5	66.6	"	40.3	38.4
E. coli	5	5	74.6[e]	74.6	1.0	48.8	53.8
C644	2	"	67.1	66.3	"	38.2	36.5
C644[nd*]	3	"	66.0	66.1	"	35.6	36.1
"	7	"	65.9	65.2	"	35.7	34.3
"	8	"	66.5	65.6	"	37.0	35.1
C1091	4	"	66.3[P]	66.3	"	36.5	36.5
"	6	"	65.2	65.3	"	34.2	34.5
E. coli	5	5	75.3	75.3	"	50.2	53.8
C644	2	"	68.8	68.0	"	40.3	38.6
"	3	"	68.1	68.2	"	38.8	39.0
"	8	"	69.3	68.4	"	41.3	39.4
C1091	4	"	67.3[P]	67.3	"	37.2	37.2
"	6	"	68.0[P]	68.1	"	38.6	38.8
E. coli	5	5	72.9	72.9	1.0	45.2	53.8
C644	3	"	65.1	65.2	"	37.6	37.8
"	4	"	65.8	65.8	"	39.0	39.0
"	6	"	65.5	65.6	"	38.4	38.6
"	7	"	65.7	65.0	"	38.8	37.4
"	8	"	66.?	65.3	"	39.9	38.0
E. coli	4	5	75.8	75.8	1.0	51.3	50.4
C644	2	"	69.7	68.9	"	41.1	36.0
"	3	"	69.3	69.4	"	40.3	37.1
"	5	"	69.0	69.0	"	39.7	36.3
"	6	"	69.0[e]	69.1	"	39.7	36.5
"	7	"	69.1	68.3	"	39.9	34.9
"	8	"	71.9	70.9	"	45.7	40.2
E. coli	4	5	75.6	75.6	0.5	(50.8)	50.4
C644	2	"	69.8	69.0	"	38.3	36.7
"	3	"	68.9	69.0	"	36.5	36.7
"	5	"	69.0	69.0	"	36.7	36.7
"	6	"	68.6	68.7	"	35.8	36.0
"	7	"	68.8	68.1	"	36.3	34.8
E. coli	4	5	76.7	76.7	1.0	(53.1)	50.4
C644	2	"	70.3	69.5	"	37.1	35.4
"	3	"	69.7	69.8	"	35.8	36.0
"	5	"	69.7	69.7	"	35.8	35.8
"	6	"	68.5	68.6	"	33.3	33.6
"	7	"	70.1	69.3	"	36.7	35.0

E. coli	5	5	76.7	76.7	1.0	(53.1)	50.4
C644	3	"	70.2	70.3	"	36.9	37.1
"	4	"	69.6	69.6	"	35.6	35.6
"	6	"	69.1	69.2	"	34.6	34.8
E. coli	5	5	77.1	77.1	1.0		53.8
C644	2	"	70.1	69.3	"	39.2	37.6
"	3	"	69.1	69.2	"	37.2	37.4
"	4	"	68.1	68.1	"	35.1	35.1
"	6	"	68.4	68.5	"	35.7	35.9
"	7	"	68.4	67.7	"	35.7	34.2
E. coli	5	5	77.8	77.8	1.0		53.8
C644	3	"	68.6	68.7	"	34.7	34.9
"	6	"	69.5	69.6	"	36.5	36.7
"	7	"	69.2	68.5	"	35.9	34.5
C644	3	5	70.2	70.3	1.0	39.2	
"	4	"	69.6	69.6	"	38.2	
"	6	"	69.1	69.2	"	37.2	
E. coli	5	5	77.1	77.1	1.0		53.8
C644	2	"	70.1	69.3	"	39.2	37.6
"	3	"	69.1	69.2	"	37.2	37.4
"	4	"	69.2	69.2	"	37.4	37.4
"	6	"	68.4	68.5	"	35.7	35.9
"	7	"	68.4	67.7	"	35.7	34.2
E. coli	3	4	78.1	78.2	0.5		50.4
C214*b*	6	"	76.0	76.1	"	46.0	46.0
"	6	"	76.4	76.5	"	46.9	47.1
E. coli	2	4	78.9	78.1	1.0		
C1174	3	"	72.0	72.1	"	39.4	41.3
"	7	"	72.9[e]	72.2	"	41.3	41.5
C214*b*	5	"	75.6	75.6	"	46.9	48.6
E. coli	5	4	77.5	77.5	"		50.4
C1174	3	"	71.9	72.0	"	38.8	39.0
C214*b*	6	"	75.7	75.8	"	46.7	46.9

adj : adjusted for cell position (see 3.1)

[1] : Equation used for calculating %(G+C) =
% (G+C) = 53.8 + 2.08(T_{mx} - T_{mr})
or: % (G+C) = 50.4 + 2.08(T_{mx} - T_{mr})
depending on the *E. coli* standard DNA (Sigma) used.

Appendix 6 DNA Pairing Data

Bacterial Strain	T_m (°C)	T_{m2}[a] (°C)	%Δ Abs.[b]	T_r[c] (°C)	%Fit[d]	% Homology	n[f]	% Hg	s.d.[h]
C52	-	-	34.8	65.6	-	68.8			
C200	-	-	41.6	"	-				
Mix	-	-	36.1	"	-				
C52	-	-	24.2	66.6	-	65.2			
C200	-	-	32.9	"	-				
Mix	-	-	26.4	"	-		2	67.0	2.5
C52	-	-	30.0	66.1	-	70.6			
C202	-	-	37.0	"	-				
Mix	-	-	-	"	-		1	70.6	0
C52	-	-	30.0	66.1	-	67.1			
C203	-	-	36.3	"	-				
Mix	-	-	-	"	-		1	67.1	0
C52	-	-	-	65.9	98.9	68.3			
C231	-	-	-	"	99.0				
Mix	-	-	-	"	99.8				
C52	90.9	-	-	66.2	99.2	72.9			
C231	91.3	-	33.3	"	99.2				
Mix	91.0	-	34.1	"	99.6				
C52	90.8	-	38.5	66.4	99.4	66.6			
C231	91.3	-	34.3	"	99.6				
Mix	90.7	-	30.6	"	99.2		3	69.3	3.3
C52	91.2	90.7	31.0	63.4	99.1	90.7			
C201	88.2	87.4	28.4	"	99.7				
Mix	89.8	89.0	29.6	"	99.8				
C52	90.5	-	25.5	63.4	99.6	89.2			
C201	88.1	-	24.7	"	99.8				
Mix	89.3	-	26.2	"	99.9				
C52	90.3	90.3	32.0	62.6	99.9	89.4			
C201	89.4	89.7	30.1	"	99.8				
Mix	89.3	-	33.7	"	99.3				
C52	-	-	-	62.6	99.6	88.2			
C201	-	-	-	"	99.6				
Mix	-	-	-	"	99.9		4	89.4	1.0
C52	-	-	33.6	66.1	99.8	55.7			
C228	-	-	33.7	"	99.1				
Mix	-	-	31.4	"	99.4				
C52	-	-	-	66.7	98.6	57.9			
C228	-	-	-	"	99.6				
Mix	-	-	-	"	99.8		2	56.8	1.6
C52	88.9	88.1	29.7	64.3	-	71.5			
C644	91.5	89.4	31.5	"	-				
Mix	89.9	89.5	31.5	"	-				
C52	91.1	-	31.8	63.1	99.5	72.3			
C644	89.6	-	32.3	"	99.6				
Mix	90.2	-	35.3	"	99.8				

C52	89.0	-	-	62.6	99.9	75.1			
C644	89.3	-	-	"	99.9				
Mix	88.8	-	-	"	99.6				
C52	87.5	-	-	62.2	-	67.0			
C644	87.6	-	-	"	-				
Mix	87.2	-	-	"	-				
C52	88.6	-	-	63.6	99.0	69.6			
C644	89.5	-	-	"	98.8				
Mix	87.9	-	-	"	98.3				
C52	86.6	-	-	61.5	-	71.0			
C644	86.5	-	-	"	-				
Mix	86.4	-	-	"	-				
C52	86.1	-	-	64.6	-	76.5			
C644	91.2	-	-	"	-				
Mix	89.3	-	-	"	-		7	71.9	3.2
C52	-	-	29.3	66.6	-	80.6			
C645	-	-	31.9	"	-				
Mix	-	-	-	"	-		1	80.6	0
C52	-	-	-	66.4	-	46.3			
JS31	-	-	-	"	-				
Mix	-	-	-	"	-				
C52	91.2	-	31.4	66.2	99.9	46.2			
JS31	90.8	-	41.1	"	99.9				
Mix	91.8	-	-	"	99.6				
C52	-	-	33.6	66.1	-	47.4			
JS31	-	-	-	"	-				
Mix	-	-	36.5	"	-		3	46.6	0.7
C52	88.9	88.1	28.8	62.3	-	51.7			
C1090	87.1	86.9	28.2	"	-				
Mix	89.0	88.9	31.6	"	-				
C52	87.1	-	30.2	62.6	99.8	54.2			
C1090	86.0	86.5	28.7	"	99.7				
Mix	88.2	88.2	26.6	"	99.1				
C52	89.0	89.6	-	62.3	-	50.6			
C1090	86.7	87.6	-	"	-				
Mix	85.2	87.1	-	"	-		3	52.2	1.8
C52	89.4	88.1	34.6	63.3	-	59.9			
C1090	88.2	88.9	33.9	"	-				
Mix	88.1	87.2	40.6	"	-		4	54.2	4.1
C52	-	-	32.7	66.6	-	57.0			
C1171	-	-	46.2	"	-				
Mix	-	-	33.2	"	-		1	57.0	0
C52	88.7	-	-	62.4	-	52.4			
C1091	86.8	-	-	"	-				
Mix	87.6	-	-	"	-				
C52	-	-	-	63.6	96.1	51.0			
C1091	-	-	-	"	99.0				
Mix	-	-	-	"	99.4				
C52	91.1	-	37.1	62.7	99.8	53.3			
C1091	92.7	-	38.2	"	99.8				
Mix	91.6	-	-	"	99.6				
C52	86.5	-	25.9	62.7	99.2	57.2			
C1091	85.9	-	28.8	"	99.7				
Mix	86.0	-	20.6	"	99.8				

C52	-	-	29.6	65.5	99.7	48.9			
C1091	-	-	34.5	"	99.7				
Mix	-	-	32.7	"	99.7		5	52.6	3.0
C52	91.7	-	33.6	66.3	99.6	44.8			
C1172	90.3	-	38.1	"	99.0				
Mix	90.9	-	35.3	"	99.3				
C52	91.7	-	31.4	65.9	99.8	43.4			
C1172	90.0	-	34.4	"	99.6				
Mix	90.5	-	-	"	99.8		2	44.1	1.0
C52	91.5	-	34.7	66.6	-	45.2			
C1087	90.6	-	24.7	"	94.5				
Mix	91.1	-	29.3	"	95.5				
C52	92.9	-	28.1	66.5	99.7	43.6			
C1087	90.9	-	27.5	"	99.4				
Mix	91.9	-	28.3	"	99.8				
C52	90.2	90.2	36.0	65.4	99.7	45.2			
C1087	89.7	89.6	25.4	"	99.9				
Mix	90.8	91.1	-	"	99.5		3	44.7	0.9
C52	91.3	-	31.5	63.3	99.8	36.7			
C1087	90.3	-	24.3	"	99.1				
Mix	91.5	-	25.9	"	99.2		4	42.7	4.1
C52	90.3	-	-	65.5	94.5	50.2			
C1087	89.8	-	-	"	99.9				
Mix	90.5	-	-	"	99.3		5	44.2	4.9
C52	90.5	-	-	66.3	99.5	30.1			
C663	-	-	-	"	99.4				
Mix	91.9	-	-	"	99.8				
C52	90.9	-	-	66.4	-	29.9			
C663	91.2	-	-	"	-				
Mix	91.8	-	-	"	-				
C52	91.2	-	33.9	66.5	-	31.1			
C663	91.5	-	33.4	"	-				
Mix	-	-	-	"	-		3	30.4	0.6
C52	91.1	-	36.9	66.5	99.6	35.5			
C667	-	-	-	"	98.8				
Mix	-	-	-	"	98.7				
C52	90.7	-	33.5	65.2	99.4	36.0			
C667	90.7	-	35.9	"	99.4				
Mix	90.6	-	34.3	"	98.8				
C52	-	-	34.3	66.5	-	38.8			
C667	-	-	44.6	"	-				
Mix	-	-	-	"	-		3	36.8	1.8
C52	90.7	-	33.5	65.2	99.4	33.7			
C659	92.0	-	36.6	"	99.9				
Mix	90.8	-	-	"	99.9				
C52	90.5	-	-	65.9	98.9	29.6			
C659	90.7	-	-	"	99.5				
Mix	90.2	-	-	"	99.5				
C52	91.8	-	41.1	66.4	99.5	35.8			
C659	91.0	-	25.9	"	99.9				
Mix	91.4	-	-	"	99.6		3	33.0	3.2
C52	90.8	-	38.5	66.4	99.4	32.6			
C666	91.8	-	35.2	"	99.4				
Mix	94.1	-	-	"	99.3				

C52	90.7	-	33.9	66.4	99.5	38.7			
C666	90.7	-	33.1	"	98.6				
Mix	91.6	-	-	"	99.2		2	35.7	4.3
C52	-	-	-	67.2	99.5	11.8			
C214*a*	-	-	-	"	99.0				
Mix	-	-	-	"	99.8		1	11.8	0
C52	90.9	-	42.6	66.8	99.7	22.5			
SLCC7211	92.1	-	39.1	"	99.8				
Mix	91.7	-	-	"	97.4				
C52	91.0	90.6	43.5	66.5	98.8	23.7			
SLCC7211	92.4	91.7	-	"	99.9				
Mix	91.6	91.4	-	"	99.4				
C52	91.3	-	38.3	66.4	99.3	25.7			
SLCC7211	92.4	-	34.0	"	99.8				
Mix	91.8	-	-	"	99.9				
C52	-	-	-	66.4	99.7	21.7			
SLCC7211	-	-	-	"	99.6				
Mix	-	-	-	"	99.5		4	23.4	1.7
C52	93.4	92.8	32.5	67.6	-	21.6			
C1174	93.1	93.3	35.3	"	-				
Mix	93.1	92.6	-	"	-		1	21.6	0
C200	-	-	-		99.6	78.7			
C202	-	-	33.4		99.6				
Mix	-	-	-		99.7				
C200	92.3	-	25.6		-	83.5			
C202	92.6	-	30.8		-				
Mix	92.1	-	-		-				
C200	91.7	-	29.6	67.4	97.0	82.3			
C202	92.5	-	30.2	"	98.6				
Mix	92.7	-	-	"	99.1		3	81.5	2.5
C200	-	-	39.1	66.1	99.3	75.2			
C203	-	-	38.1	"	99.8				
Mix	-	-	36.0	"	99.3		1	75.2	0
C200	92.5	-	35.4	67.1	99.5	75.2			
C231	92.0	-	34.7	"	99.8				
Mix	93.8	-	-	"	99.9				
C200	91.7	-	-		98.9	72.3			
C231	-	-	-		99.7				
Mix	92.0	-	-		99.5		2	73.8	2.1
C200	-	-	36.7	66.1	99.3	71.6			
C228	-	-	32.9	"	99.5				
Mix	-	-	34.8	"	99.1		1	71.6	0
C200	91.7	-	37.5	66.6	99.4	51.2			
C644	92.2	-	32.8	"	99.7				
Mix	92.0	-	22.6	"	99.8		1	51.2	0
C200	-	-	36.8		99.6	55.4			
C645	-	-	29.9		99.6				
Mix	-	-	-		99.7		1	55.4	0
C200	92.5	-	-	67.1	-	55.4			
C1090	92.0	-	-	"	-				
Mix	92.2	-	-	"	-				

C200	91.7	-	-		98.7	54.3			
C1090	92.5	-	-		99.6				
Mix	93.4	-	-		99.3				
C200	92.5	-	35.4	67.1	99.6	55.4			
C1090	92.5	-	-	"	99.5				
Mix	92.9	-	-	"	99.8		3	55.0	0.6
C200	-	-	-		99.4	50.6			
C1171	-	-	-		99.6				
Mix	-	-	-		99.7		1	50.6	0
C200	-	-	37.7	64.3	99.7	57.4			
C1091	-	-	30.0	"	99.3				
Mix	-	-	33.6	"	99.5		1	57.4	0
C200	-	-	38.9	66.6	99.5				
C1172	-	-	31.4	"	99.6				
Mix	-	-	36.7	"	99.5		1	53.4	0
C200	-	-	35.5	65.6	-	34.6			
C1087	-	-	32.7	"	-				
Mix	-	-	35.4	"	-		1	34.6	0
C200	91.7	-	-	66.9	99.0	39.9			
C663	-	-	-	"	99.1				
Mix	92.2	-	-	"	99.6				
C200	92.5	-	-	67.1	99.0	43.6			
C663	92.2	-	28.2	"	99.5				
Mix	94.5	-	-	"	99.7		2	41.8	2.6
C200	-	-	39.1	66.1	99.3	31.0			
C667	-	-	37.7	"	99.6				
Mix	-	-	35.7	"	99.7		1	31.0	0
C200	-	-	37.1	65.6	99.2	30.4			
C659	-	-	-	"	99.8				
Mix	-	-	33.9	"	99.5		1	30.4	0
C200	-	-	37.5		99.4	15.7			
C214*a*	-	-	28.4		99.8				
Mix	-	-	26.6		99.2		1	15.7	0
C200	-	-	32.9	66.6	-	14.9			
C1174	-	-	37.5	"	-				
Mix	-	-	-	"	-		1	14.9	0
C202	92.6	-	-	67.6	-	84.0			
C203	91.6	-	28.3	"	-				
Mix	91.7	-	-	"	-				
C202	92.9	-	32.2	67.4	99.1	86.1			
C203	-	-	-	"	99.6				
Mix	92.0	-	30.6	"	99.6		2	85.1	1.5
C202	92.9	-	32.2	67.6	99.1	79.3			
C231	92.9	-	31.9	"	99.3				
Mix	93.9	-	-	"	99.5		1	79.3	0
C202	-	-	37.6	66.1	99.4	54.8			
C228	-	-	32.9	"	99.5				
Mix	-	-	34.9	"	99.8		1	54.8	0
C202	-	-	35.6	66.6	-	49.7			
C644	-	-	-	"	-				
Mix	-	-	33.0	"	-		1	49.7	0

C202	–	–	37.3	65.9	99.8	19.1			
C645	–	–	30.5	"	99.5				
Mix	–	–	31.0	"	99.9		1	19.1	0
C202	–	–	35.6	66.6	–	39.6			
C1090	–	–	34.1	"	–				
Mix	–	–	32.3	"	–		1	39.6	0
C202	90.2	–	25.6		–	50.3			
C1171	91.6	–	–		–				
Mix	91.0	–	–		–		1	50.3	0
C202	–	–	37.3	65.9	99.8	49.4			
C1091	–	–	35.4	"	99.6				
Mix	–	–	–	"	99.7		1	49.4	0
C202	–	–	33.4		99.6	45.6			
C1172	–	–	32.5		99.6				
Mix	–	–	31.5		99.6		1	45.6	0
C202	–	–	29.7	66.1	–	44.4			
C1087	–	–	28.0	"	–				
Mix	–	–	33.6	"	–		1	44.4	0
C202	92.5	–	30.2	67.4	98.6	33.6			
C663	93.8	–	–	"	99.4				
Mix	92.7	–	–	"	98.8		1	33.6	0
C202	–	–	35.7	66.1	99.8	34.0			
C667	–	–	33.8	"	99.8				
Mix	–	–	33.1	"	99.5		1	34.0	0
C202	–	–	30.6	63.6	99.6	34.3			
C659	–	–	30.5	"	99.3				
Mix	–	–	–	"	99.8		1	34.3	0
C202	–	–	35.5	66.6	–	29.3			
C214*a*	–	–	30.9	"	–				
Mix	–	–	–	"	–		1	29.3	0
C202	–	–	33.8	66.6	–	4.5			
C1174	–	–	33.3	"	–				
Mix	–	–	33.0	"	–		1	4.5	0
C203	91.6	–	28.3	67.6	–	71.4			
C231	92.6	–	30.2	"	–				
Mix	94.9	–	–	"	–		1	71.4	0
C203	91.5	–	34.6	63.6	–	67.1			
C644	92.4	–	30.5	"	–				
Mix	92.2	–	31.4	"	–				
C203	89.4	–	31.1	64.9	99.8	63.7			
C644	90.4	–	33.3	"	99.8				
Mix	90.3	–	–	"	97.4		2	62.3	2.1
C203	–	–	37.2		99.7	49.8/72			
C645	–	–	29.9		99.6				
Mix	–	–	–		99.6				
C203	91.5	–	–	67.0	–	43.7			
C1090	–	–	–	"	–				
Mix	–	–	–	"	–		1	43.7	0

C203	–	–	34.6	63.6	–	40.1			
C1171	–	–	39.6	"	–				
Mix	–	–	–	"	–		1	40.1	0
C203	–	–	37.2	65.6	99.5	55.9			
C1091	–	–	33.9	"	99.8				
Mix	–	–	35.4	"	99.8		1	55.9	0
C203	–	–	36.2	65.9	99.1	52.3			
C1172	–	–	34.2	"	99.2				
Mix	–	–	35.3	"	99.7		1	52.3	0
C203	–	–	31.8	66.1	99.8	29.9			
C1087	–	–	28.0	"	99.8				
Mix	–	–	29.5	"	98.9		1	29.9	0
C203	–	–	23.3	66.7	98.8	34.8			
C663	91.3	–	27.7	"	99.7				
Mix	92.6	–	–	"	99.7				
C203	–	–	–	66.6	99.6	34.1			
C663	–	–	–	"	99.5				
Mix	–	–	–	"	99.8		2	34.5	0.5
C203	–	–	34.1		99.7	36.4			
C667	–	–	31.5		99.7				
Mix	–	–	34.8		98.6		1	36.4	0
C203	–	–	37.2	65.6	99.8	30.4			
C659	–	–	34.9	"	99.4				
Mix	–	–	33.9	"	99.7		1	30.4	0
C203	–	–	36.1	66.6	99.8	23.2			
C214*a*	–	–	30.0	"	99.7				
Mix	–	–	31.1	"	99.8				
C203	–	–	–	67.8	99.8	14.5			
C214*a*	–	–	–	"	99.3				
Mix	–	–	–	"	99.7		2	18.9	6.2
C203	91.7	91.1	26.1	68.3	99.7	11.5			
C214*b*	93.8	93.6	32.6	"	99.7				
Mix	93.3	93.3	–	"	98.5		1	11.5	0
C203	–	–	32.8	66.6	–	26.3			
C1174	–	–	33.3	"	–				
Mix	–	–	31.8	"	–		1	26.3	0
C231	90.4	–	32.1	67.3	99.5	47.0			
C644	89.5	–	–	"	99.8				
Mix	90.6	–	–	"	99.7				
C231	92.5	–	–	67.3	99.9	42.2			
C644	91.6	–	–	"	99.8				
Mix	92.7	–	–	"	99.8		2	44.6	3.4
C231	–	–	32.5	65.9	99.9	54.8			
C645	–	–	28.4	"	99.1				
Mix	–	–	43.3	"	99.9		1	54.8	0
C231	–	–	–	67.1	99.3	59.8			
C1090	–	–	–	"	99.7				
Mix	–	–	–	"	99.2		1	59.8	0

C231	90.9	89.4	-		-	54.0			
C1171	87.5	87.6	-		-				
Mix	87.8	88.0	-		-				
C231	-	-	-	66.9	-	56.2			
C1171	-	-	-	"	-				
Mix	-	-	-	"	-				
C231	92.7	-	-	66.9	97.9	50.8			
C1171	91.4	-	-	"	98.8				
Mix	91.9	-	-	"	99.8				
C231	-	-	-	66.9	99.7	52.5			
C1171	-	-	-	"	99.7				
Mix	-	-	-	"	99.1		4	53.4	2.3
C231	90.0	-	-		-	46.8			
C1171	89.9	-	-		-				
Mix	89.5	-	-		-		5	52.1	3.5
C231	-	-	32.5	65.6	-	42.3			
C1091	-	-	32.8	"	-				
Mix	-	-	31.0	"	-		1	42.3	0
C231	-	-	37.4	65.9	-	50.8			
C1172	-	-	38.7	"	-				
Mix	-	-	37.1	"	-		1	50.8	0
C231	-	-	33.6	65.9	-	38.2			
C1087	-	-	33.6	"	-				
Mix	-	-	33.9	"	-		1	38.2	0
C231	91.8	-	31.8	65.9	99.6	41.2			
C663	90.4	-	27.4	"	99.7				
Mix	90.9	-	-	"	99.7		1	41.2	0
C231	-	-	-	66.6	99.3	27.4			
C667	-	-	-	"	99.6				
Mix	-	-	-	"	99.4				
C231	-	-	29.8	66.1	-	26.7			
C667	-	-	38.6	"	-				
Mix	-	-	-	"	-		2	27.1	0.5
C231	-	-	29.6	66.2	-	21.6*			
C667	-	-	33.4	"	-				
Mix	-	-	-	"	-		3	25.3	3.2
C231	-	-	29.8	66.1	99.4	30.7			
C659	-	-	30.7	"	99.8				
Mix	-	-	-	"	99.6		1	30.7	0
C231	91.3	-	34.3	66.4	99.6	39.0			
C666	91.8	-	35.2	"	99.3				
Mix	92.4	-	-	"	98.8				
C231	91.3	-	33.3	66.2	99.2	37.0			
C666	91.7	-	34.1	"	99.3				
Mix	92.6	-	-	"	99.5				
C231	91.0	91.1	35.6	65.8	99.4	39.1			
C666	90.4	90.4	35.8	"	99.8				
Mix	91.2	91.0	-	"	99.6		3	38.4	1.2
C231	91.2	-	36.4	66.0	99.8	50.6			
C666	91.0	-	35.6	"	99.7				
Mix	92.6	-	-	"	99.9				

C231	90.9	-	34.8	65.8	99.7	17.3*			
C666	90.7	-	35.9	"	99.9				
Mix	91.0	-	-	"	97.5				
C231	-	-	-		99.6	30.1			
C214*a*	-	-	35.9		96.5				
Mix	-	-	24.3		99.6		1	30.1	0
C231	-	-	32.5	66.6	-	25.2			
C1174	-	-	31.5	"	-				
Mix	-	-	29.1	"	-		1	25.2	0
C201	88.3	-	-	63.8	-	79.0			
C644	91.1	-	-	"	-				
Mix	87.6	-	-	"	-		1	79.0	0
C201	88.3	-	-	-	-	25.2			
SLCC7211	87.0	-	-	-	-				
Mix	86.0	-	-	-	-		1	25.2	0
C228	91.6	91.2	33.3	66.9	99.5	51.3			
JS21	91.6	91.3	38.5	"	99.8				
Mix	92.0	91.7	37.1	"	99.9				
C228	91.4	-	33.2	66.5	99.4	51.8			
JS21	91.3	-	-	"	99.6				
Mix	93.1	-	-	"	99.3				
C228	92.2	-	-	66.9	-	50.6			
JS21	91.0	-	-	"	-				
Mix	91.7	-	-	"	-		3	51.2	0.6
C228	91.8	92.8	33.2	66.9	99.5	59.2			
JS21	91.7	-	37.4	"	94.2				
Mix	91.8	91.8	-	"	99.6		4	53.2	4.0
C228	92.9	-	38.1	67.7	98.9	49.8			
JS31	91.8	-	39.1	"	98.8				
Mix	92.2	-	-	"	96.8		1	49.8	0
C228	-	-	-	66.2	99.8	47.9			
C1090	-	-	-	"	98.1				
Mix	-	-	-	"	99.5		1	47.9	0
C228	91.7	-	32.9	66.5	99.8	53.6			
C1091	90.7	-	33.9	"	99.8				
Mix	91.5	-	-	"	99.5				
C228	92.0	-	-	67.0	82.6	51.0			
C1091	91.4	-	-	"	99.1				
Mix	92.4	-	-	"	99.9				
C228	91.5	-	33.2	66.5	99.7	50.6			
C1091	91.1	-	33.3	"	99.7				
Mix	91.5	-	-	"	99.9		3	51.7	1.6
C228	-	-	36.6	66.3	99.1	57.8			
C1172	-	-	32.5	"	99.8				
Mix	-	-	31.5	"	99.4		1	57.8	0
C228	-	-	-	66.2	-	32.9			
C1087	-	-	-	"	-				
Mix	-	-	-	"	-		1	32.9	0
C228	91.4	91.2	33.8	67.4	99.5	9.8			
C214*a*	93.1	93.4	33.5	"	98.8				
Mix	92.7	92.7	40.0	"	99.9		1	9.8	0

C644	-	-	28.4	66.6	99.5	91.9			
C645	-	-	31.9	"	99.3				
Mix	-	-	-	"	99.3				
C644	91.4	-	-	66.7	-	97.8			
C645	-	-	-	"	-				
Mix	91.0	-	-	"	-		2	94.9	4.2
C644	-	-	-	66.6	99.7	49.8			
C1090	-	-	-	"	99.2				
Mix	-	-	-	"	99.7				
C644	91.6	-	34.5	66.6	99.7	50.7			
C1090	91.0	-	-	"	99.1				
Mix	92.0	-	-	"	99.4		2	50.3	0.6
C644	-	-	-	66.6	99.1	49.3			
C1171	-	-	-	"	99.6				
Mix	-	-	-	"	99.5		1	49.3	0
C644	-	-	33.8	64.9	99.8	52.1			
C1091	-	-	27.4	"	99.8				
Mix	-	-	34.0	"	-				
C644	-	-	-	63.6	-	50.5			
C1091	-	-	-	"	-				
Mix	-	-	-	"	-				
C644	-	-	29.1	65.3	-	49.4			
C1091	-	-	25.5	"	-				
Mix	-	-	24.3	"	-				
C644	-	-	21.0	64.4	99.8	53.6			
C1091	-	-	30.1	"	99.4				
Mix	-	-	27.5	"	99.8				
C644	91.3	-	28.8	64.5	99.7	50.6			
C1091	90.2	-	27.7	"	99.4				
Mix	90.3	-	28.0	"	99.6				
C644	-	-	-	63.3	-	49.9			
C1091	-	-	-	"	-				
Mix	-	-	-	"	-				
C644	-	-	29.6	64.9	99.6	44.6			
C1091	-	-	30.3	"	98.8				
Mix	-	-	32.3	"	97.4				
C644	-	-	24.9	63.6	-	48.2			
C1091	-	-	26.8	"	-				
Mix	-	-	27.3	"	-				
C644	-	-	29.8	65.2	99.9	48.8			
C1091	-	-	29.5	"	99.4				
Mix	-	-	24.3	"	99.1		9	49.7	2.5
C644	-	-	33.2	65.8	-	45.4			
C1172	-	-	34.2	"	-				
Mix	-	-	25.7	"	-				
C644	-	-	36.7	65.8	99.9	46.4			
C1172	-	-	38.7	"	99.2				
Mix	-	-	36.8	"	99.9				
C644	-	-	30.9	66.6	99.5	44.8			
C1172	-	-	31.4	"	99.6				
Mix	-	-	32.2	"	99.2		3	45.5	0.8
C644	-	-	36.6	65.6	98.7	26.9			
C1087	-	-	32.7	"	99.6				
Mix	-	-	32.7	"	99.1		1	26.9	0

C644	91.8	-	30.3	67.0	99.8	29.0			
C663	91.5	-	29.4	"	99.6				
Mix	92.5	-	-	"	99.5				
C644	90.5	-	-	65.6	99.8	26.0			
C663	90.6	-	26.4	"	99.8				
Mix	91.0	-	-	"	99.7				
C644	91.7	-	29.2	67.0	99.6	28.5			
C663	91.8	-	28.4	"	99.7				
Mix	92.4	-	-	"	98.5		3	27.8	1.6
C644	-	-	35.7	66.1	-	35.2			
C667	-	-	33.2	"	-				
Mix	-	-	30.1	"	-		1	35.2	0
C644	-	-	33.1		99.7	30.8			
C659	-	-	31.4		98.1				
Mix	-	-	-		99.9		1	30.8	0
C644	-	-	32.8	66.6	99.7	4.3			
C214*a*	-	-	28.4	"	99.8				
Mix	-	-	24.6	"	98.8		1	4.3	0
C644	91.1	-	-		-	25.2			
SLCC7211	87.0	-	-		-				
Mix	86.5	-	-		-		1	25.2	0
C644	-	-	28.3	66.6	-	13.1			
C1174	-	-	34.4	"	-				
Mix	-	-	30.1	"	-				
C644	-	-	26.8	66.6	99.6	18.7			
C1174	-	-	31.5	"	99.8				
Mix	-	-	29.9	"	99.8		2	15.9	4.0
C645	-	-	28.4	65.9	99.1	61.6			
C1090	-	-	33.7	"	99.8				
Mix	-	-	29.4	"	99.4		1	61.6	0
C645	91.7	91.6	37.8	67.0	99.3	49.5			
C1171	91.8	91.7	28.9	"	98.7				
Mix	92.1	91.7	-	"	99.7				
C645	91.5	91.3	-	66.7	99.7	41.3			
C1171	91.4	91.5	28.8	"	99.6				
Mix	92.1	92.1	-	"	99.8				
C645	-	-	-	66.7	-	44.3			
C1171	92.1	-	-	"	-				
Mix	91.4	-	-	"	-				
C645	-	91.0	-	66.7	99.5	44.3			
C1171	92.6	92.4	-	"	99.2				
Mix	92.1	-	-	"	99.7		4	44.9	3.4
C645	-	-	29.3	67.6	99.6	0			
C214*a*	-	-	29.2	"	99.8				
Mix	-	-	30.7	"	99.6		1	0	0
C645	-	-	29.3	67.6	99.6	-3.6			
C1174	-	-	34.4	"	99.7				
Mix	-	-	-	"	99.9		1	-3.6	0
JS21	-	-	37.4	66.8	99.7	93.4			
JS31	-	-	37.0	"	99.5				
Mix	-	-	-	"	99.6				

JS21	92.1	-	35.3	67.7	99.4	96.3			
JS31	91.8	-	39.1	"	98.9				
Mix	94.3	-	-	"	98.9				
JS21	91.0	91.0	-	66.8	99.6	95.2			
JS31	91.3	90.9	-	"	97.9				
Mix	92.0	92.0	-	"	99.3		3	95.0	1.5
JS21	-	-	-	66.6	-	71.1*			
C1091	-	-	-	"	-				
Mix	-	-	-	"	-		1	71.1	0
JS21	-	-	37.7	67.1	-	12.9			
C214*a*	-	-	32.3	"	-				
Mix	-	-	-	"	-		1	12.9	0
JS31	-	-	-	65.8	99.6	57.2			
C1090	-	-	-	"	99.3				
Mix	-	-	-	"	99.8				
JS31	91.3	91.5	36.7	67.3	99.5	62.2			
C1090	92.0	91.7	35.1	"	98.6				
Mix	92.8	92.1	36.2	"	99.5		2	59.7	3.5
JS31	-	-	-	66.2	-	85.1*			
C1091	-	-	-	"	-				
Mix	-	-	-	"	-				
JS31	-	-	36.6	66.3	99.5	56.7			
C1091	-	-	28.1	"	99.3				
Mix	-	-	33.0	"	99.4				
JS31	91.4	-	43.1	66.6	99.4	55.7			
C1091	90.8	-	38.6	"	99.8				
Mix	91.2	-	42.5	"	99.7		2	56.2	0.7
JS31	91.4	-	37.4	66.3	99.3	71.6			
C1172	90.3	-	38.1	"	99.0				
Mix	92.1	-	36.8	"	99.1				
JS31	90.5	-	37.4	66.6	99.8	94.4			
C1172	90.6	-	36.8	"	99.7				
Mix	91.2	-	36.3	"	99.9				
JS31	91.3	-	41.1	66.6	99.7	39.0			
C1087	91.2	-	35.2	"	99.8				
Mix	91.1	-	32.7	"	99.5				
JS31	-	-	36.6	66.3	99.7	38.8			
C1087	-	-	32.5	"	99.5				
Mix	-	-	33.0	"	99.7		2	38.9	0.1
JS31	90.1	91.6	38.5	66.6	-	26.7			
C214*a*	92.4	94.2	37.0	"	-				
Mix	91.5	93.4[e]	36.2	"	-				
JS31	-	-	38.1	67.1	98.9				
C214*a*	-	-	32.3	"	99.8	23.0			
Mix	-	-	34.2	"	99.3		2	24.9	2.6
JS31	89.9	-	32.7	67.2	99.5	7.2			
C1174	93.3	-	31.2	"	99.3				
Mix	92.4	-	31.6	"	99.8		1	7.2	0
C1090	91.3	-	27.8		99.7	93.8			
C1171	92.0	-	14.4		99.5				
Mix	92.6	-	-		99.3				

C1090	91.0	–	33.4	66.6	–	93.2			
C1171	91.1	–	36.1	"	–				
Mix	91.8	–	35.1	"	–		2	93.5	0.4
C1090	91.3	90.2	33.9	65.8	–	48.1			
C1091	91.2	90.5	28.4	"	–				
Mix	91.5	–	28.4	"	–				
C1090	86.6	–	35.6	62.4	–	45.6			
C1091	86.0	–	24.1	"	–				
Mix	86.4	–	–	"	–				
C1090	87.1	–	–	61.2	98.9	44.8			
C1091	85.9	–	–	"	99.7				
Mix	86.8	–	–	"	99.6		3	46.2	1.7
C1090	92.9	–	33.8	66.0	99.5	71.7*			
C1091	90.5	–	32.7	"	99.3				
Mix	92.2	–	–	"	99.4				
C1090	91.0	–	34.0	66.3	–	54.0			
C1172	91.1	–	35.8	"	–				
Mix	91.5	–	36.9	"	–				
C1090	91.4	–	35.5	66.9	–	53.0			
C1172	91.7	–	40.4	"	–				
Mix	92.0	–	35.1	"	–		2	53.5	0.7
C1090	91.3	90.2	33.9	65.8	–	41.1			
C1087	90.0	89.2	25.1	"	–				
Mix	90.5	89.0	27.8	"	–				
C1090	92.0	91.7	35.1	67.3	99.3	33.4			
C1087	91.6	91.0	34.4	"	99.8				
Mix	91.8	91.9	35.8	"	99.4				
C1090	–	–	–	64.3	99.6	39.3			
C1087	–	–	–	"	99.9				
Mix	–	–	–	"	98.2				
C1090	–	–	35.6	63.3	99.9	31.8			
C1087	–	–	24.3	"	99.1				
Mix	–	–	25.9	"	99.2				
C1090	–	–	–	64.3	–	41.0			
C1087	–	–	–	"	–				
Mix	–	–	–	"	–		5	37.3	4.4
C1090	91.1	90.7	34.9	65.9	98.5	39.7			
C663	90.8	90.4	27.6	"	99.8				
Mix	91.2	90.9	–	"	99.7				
C1090	–	–	–	66.7	–	42.6			
C663	–	–	–	"	–				
Mix	91.6	–	–	"	–				
C1090	–	–	–	66.1	–	47.1			
C663	–	–	–	"	–				
Mix	91.6	–	–	"	–				
C1090	91.3	–	35.0	66.4	99.7	42.6			
C633	91.7	–	24.6	"	99.7				
Mix	93.0[e]	–	–	"	99.8		3	43.0	3.1
C1090	91.6	–	34.4	67.2	99.5	25.8*			
C663	93.7	–	31.1	"	99.8				
Mix	91.5	–	–	"	99.7				
C1090	–	–	31.3	65.9	–	38.9			
C667	–	–	32.8	"	–				
Mix	–	–	25.8	"	–		1	38.9	0

C1090	–	–	35.2		99.6	34.1			
C659	–	–	31.4		98.1				
Mix	–	–	32.6		99.7				
C1090	–	–	31.3	65.9	99.6	36.7			
C659	–	–	31.0	"	99.9				
Mix	–	–	29.7	"	99.8		2	35.4	1.8
C1090	91.8	92.5	33.1	66.6	98.9	13.7			
C214*a*	92.4	93.6	32.0	"	99.7				
Mix	91.3	92.5	31.9	"	99.9				
C1090	–	–	–	67.2	98.1	13.4			
C214*a*	–	–	–	"	99.0				
Mix	–	–	–	"	99.7		2	13.6	0.2
C1090	88.8	–	33.9	67.4	99.2	16.9			
C1174	91.3	–	32.1	"	99.3				
Mix	90.3	–	–	"	99.9		1	16.9	0
C1171	–	–	33.7	65.6	–	43.6			
C1091	–	–	32.8	"	–				
Mix	–	–	40.3	"	–		1	43.6	0
C1171	–	–	46.2	66.6	–	42.5			
C1172	–	–	43.9	"	–				
Mix	–	–	46.0	"	–		1	42.5	0
C1171	–	–	38.6	67.1	–	37.2			
C1087	–	–	31.6	"	–				
Mix	–	–	–	"	–		1	37.2	0
C1171	93.3	–	26.1	66.7	–	35.2			
C663	–	–	–	"	–				
Mix	91.6	–	–	"	–				
C1171	–	–	–	66.4	–	36.1			
C663	–	–	–	"	–				
Mix	–	–	–	"	–		2	35.9	0.4
C1171	–	–	31.8	65.9	99.6	38.6			
C667	–	–	33.7	"	99.7				
Mix	–	–	–	"	99.7		1	38.6	0
C1171	–	–	35.2		99.0	24.0			
C659	–	–	31.4		98.1				
Mix	–	–	32.6		99.7		1	24.0	0
C1171	–	–	36.1	66.6	–	10.2			
C214*a*	–	–	31.7	"	–				
Mix	–	–	34.6	"	–				
C1171	93.4	–	37.1	66.6	–	13.1			
C214*a*	93.9	–	33.6	"	–				
Mix	94.0	–	–	"	–		2	11.7	2.1
C1171	–	–	–	66.6	–	12.9			
C1174	–	–	–	"	–				
Mix	–	–	–	"	–		1	12.9	0
C1091	–	–	30.0	64.3	99.1	91.9			
C1172	–	–	35.9	"	99.7				
Mix	–	–	33.6	"	99.6				
C1091	90.4	–	32.7	65.9	99.3	91.2			
C1172	90.5	–	35.2	"	98.4				
Mix	90.5	–	35.3	"	99.6				

C1091	–	–	32.5	66.1	99.2	20.8*			
C1172	–	–	33.2	"	99.0				
Mix	–	–	–	"	99.9		2	91.6	0.5
C1091	90.3	90.2	32.5	66.1	99.2	35.1			
C1087	92.0	91.3	29.8	"	99.4				
Mix	90.5	90.0	–	"	99.3				
C1091	91.2	90.5	28.4	65.8	–	45.4*			
C1087	90.0	89.2	25.1	"	–				
Mix	90.3	89.3	28.7	"	–				
C1091	89.8	–	–	65.3	99.3	48.3*			
C1087	90.8	–	–	"	99.3				
Mix	90.1	–	–	"	97.0				
C1091	–	–	28.0	64.6	99.8	34.6			
C1087	–	–	24.0	"	99.7				
Mix	–	–	–	"	99.6		2	34.9	0.4
C1091	–	–	33.6		99.5	24.3			
C663	–	–	27.4		99.7				
Mix	–	–	31.3		99.7		1	24.3	0
C1091	90.0	–	–	65.5	99.5	31.5			
C667	90.5	–	–	"	99.5				
Mix	90.2	–	–	"	98.9		1	31.5	0
C1091	–	–	34.0	65.6	99.2	29.5			
C659	–	–	–	"	99.8				
Mix	–	–	34.7	"	99.2				
C1091	–	–	–	65.5	–	21.1			
C659	–	–	–	"	–				
Mix	–	–	–	"	–				
C1091	90.0	–	–	65.5	99.5	24.1			
C659	90.5	–	–	"	99.8				
Mix	91.2	–	–	"	99.5		3	24.9	4.3
C1091	93.0	92.2	35.2	67.4	99.8	14.7			
C214*a*	93.1	93.4	31.5	"	98.8				
Mix	92.4	92.4	34.1	"	99.9		1	14.7	0
C1091	92.0[e]	–	22.0	66.3	–	10.9			
SLCC7211	93.1	–	35.9	"	–				
Mix	91.4	–	–	"	–				
C1091	89.0	–	–	66.2	99.8	10.1			
SLCC7211	92.0	–	–	"	97.7				
Mix	91.6	–	–	"	98.2		2	10.5	0.6
C1091	–	–	29.5	64.3	–	24.3			
C1174	–	–	34.8	"	–				
Mix	–	–	30.8	"	–		1	24.3	0
C1172	–	–	36.0	64.3	99.5	13.2*			
C1087	–	–	34.8	"	98.2				
Mix	–	–	30.6	"	–				
C1172	90.9	–	32.8	66.5	99.1	41.6			
C1087	91.4	–	27.5	"	99.4				
Mix	91.9	–	31.0	"	99.9				
C1172	–	–	–	66.6	–	43.2			
C1087	–	–	–	"	–				
Mix	–	–	–	"	–		2	42.4	1.1

C1172	-	-	36.0	64.3	99.5	27.3			
C663	-	-	29.2	"	98.6				
Mix	-	-	29.8	"	98.3		1	27.3	0
C1172	-	-	34.7	66.1	-	36.1			
C667	-	-	33.2	"	-				
Mix	-	-	34.8	"	-		1	36.1	0
C1172	-	-	35.9	65.9	-	32.0			
C659	-	-	35.1	"	-				
Mix	-	-	34.1	"	-		1	32.0	0
C1172	-	-	-	66.6	-	6.7			
C214*a*	-	-	-	"	-				
Mix	-	-	-	"	-		1	6.7	0
C1172	89.3	-	34.2	67.4	-	11.4			
C1174	91.3	-	32.1	"	-				
Mix	89.8	-	-	"	-		1	11.4	0
C1087	-	-	34.8	64.3	99.6	69.4			
C663	-	-	29.2	"	99.2				
Mix	-	-	32.3	"	99.7		1	69.4	0
C1087	-	-	-		-	70.7			
C667	-	-	-		-				
Mix	-	-	-		-		1	70.7	0
C1087	-	-	33.0	63.6	-	84.8			
C659	-	-	30.5	"	-				
Mix	-	-	-	"	-				
C1087	-	-	33.0	65.6	-	87.5			
C659	-	-	35.1	"	-				
Mix	-	-	-	"	-		2	86.2	1.9
C1087	-	-	-	66.2	99.2	8.9			
C214*a*	-	-	-	"	99.9				
Mix	-	-	-	"	99.3		1	8.9	0
C1087	90.3	-	-	66.0	99.5	17.2			
SLCC7211	91.0	-	-	"	99.8				
Mix	90.6	-	-	"	99.8		1	17.2	0
C1087	-	-	31.6	67.1	-	13.7			
C1174	-	-	41.5	"	-				
Mix	-	-	-	"	-				
C1087	91.4	91.0	32.4	67.6	-	14.9			
C1174	93.1	93.3	35.3	"	-				
Mix	92.6	92.5	-	"	-		2	14.3	0.8
C663	91.5	-	33.4	66.5	-	83.8			
C667	92.2	-	28.3	"	-				
Mix	91.9	-	30.1	"	-		1	83.8	0
C663	91.1	-	24.9	66.0	-	58.4			
C659	91.8	-	30.0	"	-				
Mix	91.9	-	-	"	-		1	58.4	0
C663	-	-	-		-	30.4			
C214*a*	-	-	-		-				
Mix	-	-	-		-		1	30.4	0

C663	–	–	27.1		–	9.0			
C1174	–	–	34.8		–				
Mix	–	–	28.6		–		1	9.0	0
C667	–	–	33.8	66.1	99.8	43.6*			
C659	–	–	31.6	"	99.1				
Mix	–	–	30.7	"	99.8				
C667	91.2	–	39.6	65.6	99.5	86.1			
C659	90.7	–	29.6	"	99.9				
Mix	90.4	–	32.8	"	99.5				
C667	90.8	–	–	65.7	99.6	80.9			
C659	90.4	–	–	"	99.8				
Mix	90.7	–	–	"	99.5				
C667	92.0	–	33.5	66.4	–	83.1			
C659	91.0	–	25.9	"	–				
Mix	91.9	–	–	"	–		3	83.4	2.6
C667	90.5	–	–	66.1	99.4	72.6			
C659	90.9	–	–	"	99.5				
Mix	90.6	–	–	"	99.6		4	80.7	5.8
C667	91.0	–	35.6	66.0	99.7	84.9			
C666	90.2	–	34.9	"	99.7				
Mix	91.8	–	–	"	99.3				
C667	90.8	–	35.2	65.9	–	84.8			
C666	91.3	–	–	"	–				
Mix	91.5	–	–	"	–				
C667	90.4	–	34.0	65.9	–	>100*			
C666	90.5	–	37.9	"	–				
Mix	90.9	–	–	"	–		2	84.9	0.1
C667	–	–	33.0	66.6	99.2	28.4			
C214*a*	–	–	30.0	"	99.8				
Mix	–	–	31.0	"	99.8		1	28.4	0
C667	90.7	–	34.4	66.7	99.5	16.0			
SLCC7211	91.3	–	22.2	"	99.4				
Mix	91.4	–	–	"	99.3				
C667	–	–	34.8	66.5	99.4	8.1			
C1174	–	–	30.6	"	99.7				
Mix	–	–	32.7	"	99.7		2	8.0	11.3
C659	92.0	–	27.2	66.4	–	89.5			
C666	90.7	–	33.1	"	–				
Mix	–	–	–	"	–		1	89.5	0
C659	90.9	–	27.3	"	–	99.7*			
C666	90.5	–	37.9	"	–				
Mix	92.6	–	–	"	–		2	94.6	7.2
C659	–	–	–	66.7	99.2	27.5			
C214*a*	91.3	–	22.2	"	99.4				
Mix	92.2	–	21.1	"	99.5		1	27.5	0
C659	90.7	–	–	66.5	99.2	35.7			
SLCC7211	91.3	–	22.2	"	99.2				
Mix	92.2	–	21.1	"	98.9		1	35.7	0
C659	–	–	29.4	66.5	99.5	17.2			
C1174	–	–	30.6	"	99.7				
Mix	–	–	30.2	"	99.1		1	17.2	0

C214*a*	93.2	93.0	32.0	69.0	99.9	91.5			
C214*b*	94.5	94.0	36.8	"	99.5				
Mix	94.1	-	-	"	99.4				
C214*a*	93.5	-	-	68.6	99.0	99.5			
C214*b*	93.4	-	-	"	99.8				
Mix	94.1	-	-	"	99.8		2	95.5	5.7
C214*a*	93.6	-	-	68.9	99.8	113			
C214*b*	93.4	-	-	"	99.9				
Mix	94.4	-	-	"	99.9		3	101.3	10.9
C214*a*	93.0	-	30.7	67.2	99.4	82.0			
C1174	93.3	-	31.2	"	99.2				
Mix	-	-	-	"	99.6				
C214*a*	93.9	-	33.6	66.6	-	86.5			
C1174	93.4	-	41.5	"	-				
Mix	92.2	-	41.2	"	-		2	84.3	3.2

a : %Hyperchromicity, determined from the absorbance at 25°C and that of the sample in the single-stranded form.
b : T_m measured after renaturation.
c : Renaturation temperature.
d : % fit of the linear regression line as determined by MINITAB.
f: number of replicates
g : mean % homology
h : standard deviation.
e : Sample evaporated during the experiment.
* : Poor renaturation rates.

Appendix 7 Programs PCA.BAS. TRU3DA.BAS and TRUGH.FOR to draw 3-dimensional graphs from principal component/coordinate analyses using DNA-DNA Homology matrices on the VAX cluster.

Program PCA.BAS is a modified version of PHS6P14.BAS which was written by P.H.A. Sneath at Leicester University to determine the principal components or principal coordinates of a matrix. The program was altered to output to a file TRUPC.OUT in a form suitable for input to TRU3DA.BAS which transforms the coordinates and outputs them to a file TRUGH.FOR. The latter uses GHOST-80, a graphics package available on the VAX cluster, to produce a 3-dimensional plot of the data.

List of Programs and Languages used.

Program	Language	Author
PCA.BAS	BASIC	P.H.A. Sneath amended by T. Hartford
TRU3DA.BAS	"	T. Hartford (TRUPC.OUT input for TRU3DA)
TRUGH.FOR	FORTRAN	T. Hartford

PCA.BAS

The program, PHS6P14.BAS was amended to output the first three vectors for each OTU to the file TRUPC.OUT in the form of *n* rows, the values being separated by commas and the row culminating in an ampersand. The vectors corresponded to the coordinates for the first three axes of the three dimensional ordination for *n* OTUs.

The program also outputs the maximum and minimum *x* and *y* values and the minimum *z* value, where *x*, *y*, *z* are the points on the first three axes : *X*, *Y*, *Z*. These should be noted to decide the most desirable axes sizes to be used on the resulting ordination.

```
100 print "    program phs6p14 for principal component"
110 print "     and principal coordinate analysis"
115 print "    amended for output to file trupc.out."
120 print
125 open "trupc.out" as file £1
127 open "pca.out" as file £2
130 dim a(100,100)
140 dim b(100,100)
150 dim c(100,100)
160 dim m(100)
170 dim p(6,6)
180 dim s(100)
185 dim v(100)
190 let p5 = 1e4
191 let h(1) = 0
192 let h(2) = 0
193 let h(3) = 0
194 let l(1) = 0
195 let l(2) = 0
196 let l(3) = 0
200 print "   data should consist either of:"
210 print " (1)  a matrix of character states for n"
220 print "      characters for t otus, given as the"
230 print "      numbers n, t in data statement 1000"
240 print "      followed by n rows of data statements,"
250 print "      each with t character states at 1001 onwards"
260 print " or "
270 print " (2)  a matrix of t(t+1)/2 dissimilarities"
280 print "      for t otus, including zeros in the diagonal"
290 print "      with the no. t in data statement 2000"
300 print "      followed by the dissimilarities in rows 2001 onwards"
310 print
320 print " if principal components are required enter 0,"
330 print " if principal coordinates are required enter 1,"
400 print
410 input p1
420 if p1 = 0 then 460
430 if p1 = 1 then 460
440 print "incorrect, enter 1 or 0"
450 goto 410
460 if p1 = 1 then 520
470 print
490 print "principal components of character states"
500 print
510 goto 570
520 print
530 print "principal coordinates of dissimilarities"
540 print
560 goto 740
570 print
580 print "if matrix is to be transposed so that rows are treated"
590 print " as otus and columns characters enter 1 else zero"
600 print
620 input p6
630 if p6 = 0 then 670
640 if p6 = 1 then 670
650 print "incorrect, reenter"
660 goto 570
670 print
680 if p6 = 0 then 710
690 print " data transposed"
700 goto 720
```

```
710 print " data not transposed"
720 rem ** continues **
740 goto 2400
999 rem ** Listeria data no c1174 or 214 **
1000 data 14,14
1001 data 100,67,70.6,67.1,70.4,70.5,52.2,57,54.3,44.1,44.4,30.4,36.4,33.5
1002 data 67,100,83,75.2,73.8,53.2,39.8,40.4,52.9,53.4,34.6,41.8,31,30.4
1003 data 70.6,83,100,85.1,79.3,49.7,39.6,45.1,49.4,45.6,44.4,33.6,34,31.4
1004 data 67.1,75.2,85.1,100,71.2,63.7,39.7,40.1,55.9,52.3,29.9,34.5,36.4,30.4
1005 data 70.4,73.8,79.3,71.2,100,44.6,59.8,53.2,42.3,50.8,38.2,41.2,25.3,30.7
1006 data 70.5,53.2,49.7,63.7,44.6,100,50.3,51,50,45.5,26.9,25.9,35.2,70.8
1007 data 52.2,39.8,39.6,39.7,59.8,50.3,100,96.9,46.2,54,34.4,44.4,38.9,36.7
1008 data 57,40.6,45.1,40.1,53.2,51,96.9,100,43.6,40.7,37.2,35.2,38.6,34.8
1009 data 54.3,52.9,49.4,55.9,42.3,50,46.2,43.6,100,91,35.1,21.9,30.1,24.1
1010 data 44.1,53.4,45.6,52.3,50.8,45.5,54,40.7,91,100,41.9,27.3,36.1,32
1011 data 44.4,34.6,44.4,29.9,38.2,26.9,34.4,37.2,35.1,41.9,100,69.4,70.7,87.2
1012 data 30.4,41.8,33.6,34.5,41.2,25.9,44.4,35.2m21.9,27.3,69.4,100,83.8,58.4
1013 data 36.4,31,34,36.4,25.3,35.2,38.9,38.6,30.1,36.1,70.7,83.8,100,85.5
1027 data 33.5,30.4,31.4,30.4,30.7,30.8,36.7,34.8,24.1,32,87.2,58.4,85.5,100
2000 data 17
2001 data 0
2002 data 20.96,0
2003 data 11.69,18.69,0
2004 data 10.36,24.37,9.33,0
2005 data 12.03,16.45,9.2,11.28,0
2006 data 20.24,9.11,18.78,22.92,14.93,0
2007 data 12.08,16.28,9.68,11.96,2.68,14.94,0
2008 data 19.2,28.52,14.42,13.89,15.07,28.62,14.77,0
2009 data 9.12,16.99,9.5,12.65,6.81,16.27,6.12,16.67,0
2010 data 14.24,15.76,15.04,16.99,11.89,9.56,12.02,24.86,11.96,0
2011 data 13.95,13.95,12.95,18.36,14.46,18.03,14.42,22.41,12.49,18.49,0
2012 data 11.12,14.35,7.83,11.86,5.29,15.18,5.32,15.49,6.7,13.38,10.62,0
2013 data 12.93,17.16,6.7,11.78,10.52,19.82,10.77,15.08,10.91,18.23,9.61,7.23,0
2014 data 8.96,15.5,7.57,10.68,6.42,14.22,6.19,17.44,5.65,9.34,13.27,6.43,10.01
2015 data 13.37,19.33,16.61,19.34,17.3,21.88,17.65,25.67,14.69,20.64,8.38,14.78
2016 data 13.21,13.65,7.35,13.52,8.28,14.77,8.79,17.15,9.27,13.56,11.42,6.82,7.
2017 data 8.82,23.41,11.59,11.14,13.19,24.79,13.03,14.3,11.06,20.53,13.56,10.91
2400 rem ** continues **
2420 read n, t
2430 if p6 = 1 then 2460
2440 print "n and t before transposition are"
2450 print n, " and ", t
2460 for i = 1 to n step 1
2470 for j = 1 to t step 1
2480 if p6 = 0 then 2510
2490 read c(j,i)
2500 goto 2520
2510 read c(i,j)
2520 next j
2530 next i
2540 if p6 = 0 then 2580
2550 let n0 = n
2560 let n = t
2570 let t = n0
2580 rem ** exchanges n & t if matrix transposed **
2590 if p1 = 1 then 2620
2600 print " n and t after transposition"
2610 print " are ",n, " and ", t
2620 rem ** overwrites array **
2630 if p1 = 0 then 2710
2640 read t
2650 for j = 1 to t step 1
```

```
2660 for k = 1 to j step 1
2670 read c(j,k)
2680 let c(k,j) = c(j,k)
2690 next k
2700 next j
2710 rem ** data have been read **
2720 rem ** input of option for no. of vectors to print**
2730 print
2740 print " enter no. of vectors to be printed"
2770 input p4
2780 if p4 = 1 then 2850
2790 if p4 < 1 then 2830
2800 if p4 > n then 2830
2810 let p4 = int(p4)
2820 goto 2880
2830 print "number not in range, reprint"
2840 goto 2740
2850 if p4 < 1 then 2830
2860 if p4 > t then 2830
2870 let p4 = int(p4)
2880 print
2890 print " no. of vectors printed is ",p4
2900 rem ** now main program choice **
2910 print
2920 if p1 = 1 then 4560
2930 rem ** calculates column means into row m **
2940 for i = 1 to n step 1
2950 let s1 = 0
2960 for j = 1 to t step 1
2970 let s1 = s1 + c(i,j)
2980 next j
2990 let m(i) = s1/t
3000 next i
3010 print " if principal component analysis is to be"
3020 print " performed on sums of squares and cross-products"
3030 print " enter zero, if on correlations enter 1"
3040 print
3050 print
3060 input p2
3070 if p2 = 0 then 3110
3080 if p2 = 1  then 3110
3090 print " incorrect input,reenter"
3100 goto 3060
3110 if p2 = 1 then 3140
3120 print " uses sums of squares and cross products"
3130 goto 3150
3140 print " uses correlations "
3150 print
3160 for i = 1 to n step 1
3170 for h = 1 to i step 1
3180 let s1 = 0
3190 let s2 = 0
3200 let s3 = 0
3210 let s4 = 0
3220 let s5 = 0
3230 for j = 1 to t step 1
3240 let s2 = s2 + c(h,j)*c(h,j)
3250 let s4 = s4 + c(i,j)*c(i,j)
3260 let s5 = s5 + c(h,j)*c(i,j)
3270 next j
3280 let s0 = s5 - m(h)*m(i)*t
3290 if p2 = 0 then 3390
```

```
3300 let s0 = s0/t
3310 let s2 = (s2/t)-(m(h)*m(h))
3320 let s4 = (s4/t)-(m(i)*m(i))
3330 let x = sqr(abs(s2*s4))
3340 if x > 0 then 3370
3350 let s0 = 0
3360 goto 3390
3370 let s0 = s0/x
3380 rem ** r set to zero for indeterminate values **
3390 let a(h,i) = s0
3400 let a(i,h) = s0
3410 next h
3420 next i
3430 rem ** cross products of correlations in array a **
3440 rem ** option to scale axis **
3450 print " if each column of the eigenvectors is to be "
3460 print " scaled so that its sum of squares is: "
3470 print "    (1) the reciprocal of the square of the axis "
3480 print "        eigenvalue, ENTER 1"
3490 print "    (2) the reciprocal of the axis eigenvalue"
3500 print "       ENTER 2"
3510 print "    (3) unity, ENTER 3 "
3520 print "    (4) the axis eigenvalue, ENTER 4"
3530 print "    (5) the square of the axis eigenvalue,"
3560 print "       ENTER 5"
3570 print
3580 print
3590 input p2
3600 if p2 = 1 then 3670
3610 if p2 = 2 then 3670
3620 if p2 = 3 then 3670
3630 if p2 = 4 then 3670
3640 if p2 = 5 then 3670
3650 print "incorrect entry,try again"
3660 goto 3450
3670 print
3680 print " scaling option is no.",p2
3690 let v2 = (p2 - 3)/2
3700 rem ** eigenvalues and vectors gosub **
3710 let n0 = n
3720 gosub 5730
3730 rem ** now prints eigenvalues **
3740 gosub 5060
3750 rem ** now finds scalars for vectors **
3760 for i = 1 to n step 1
3770 let s2 = 0
3780 for h = 1 to n step 1
3790 let s2 = s2 + b(h,i)*b(h,i)
3800 next h
3810 rem ** sum of sqrs of original vector **
3820 rem ** column in s2 **
3830 let s(i) = s2
3840 for h = 1 to n step 1
3850 let b(h,i) = b(h,i)/sqr(abs(s(i)))
3860 next h
3870 next i
3880 rem ** scalars found **
3890 rem ** sum of squares **
3900 rem ** now prints scaled vectors **
3910 gosub 5350
3920 rem ** subroutine to print weighted coordinates **
3940 for i = 1 to n0 step 1
```

```
3950 let s(i) = abs(a(i,i))
3960 if x > 1e-6 then 3980
3970 let s(i) = 1e-6
3980 next i
3990 print
4000 print £2, "     coordinates of otus on principal axes,"
4010 print £2,
4020 print £2,
4030 print £2, "otus           coordinates on axes"
4040 print £2,
4050 print £2,
4060 for i = 1 to n0 step 4
4070 print £2, " ",
4080 for i1 = i to i + 3 step 1
4090 if i1 > p4 then 4130
4100 print £2, i1,
4110 let p(1,i1-i+1) = 0
4120 let p(2,i1-i+1) = 0
4130 next i1
4140 print £2, " "
4150 print £2,
4160 if i > p4 then 4510
4170 for j = 1 to t step 1
4180 print £2, j,
4181 let l9=0
4190 for i1 = i to i + 3 step 1
4191 let l9 = l9 + 1
4200 if i1 > p4 then 4320
4210 let x = 0
4220 for i2 = 1 to n0 step 1
4230 let x = x + (c(i2,j) - m(i2))*b(i2,i1)
4240 next i2
4270 let x = x*s(i1)^v2
4280 let p(1,i1-i+1) = p(1,i1-i+1) + x
4290 let p(2,i1-i+1) = p(2,i1-i+1) + x*x
4300 print £2, int(p5*x + .5)/p5,
4301 if l9 < 3 then 4306
4302 print £1, int(p5*x + .5)/p5;"&"
4303 if (int(p5*x + .5)/p5) > h(3) then h(3) =(int(p5*x + .5)/p5)
4304 if (int(p5*x + .5)/p5) < l(3) then l(3) =(int(p5*x + .5)/p5)
4305 goto 4320
4306 print £1, int(p5*x + .5)/p5;",";
4307 if l9 >=2  then 4318
4308 if (int(p5*x + .5)/p5) < l(1) then l(1) =(int(p5*x + .5)/p5)
4309 if (int(p5*x + .5)/p5) > h(1) then h(1) =(int(p5*x + .5)/p5)
4310 goto 4320
4318 if (int(p5*x + .5)/p5) < l(2) then l(2) =(int(p5*x + .5)/p5)
4319 if (int(p5*x + .5)/p5) > h(2) then h(2) =(int(p5*x + .5)/p5)
4320 next i1
4325 print £2, " "
4330 next j
4340 print £2,
4350 print £2, "sum",
4360 for i1 = i to i + 3 step 1
4370 if i1 > p4 then 4400
4380 let x = p(1,i1-i+1)
4390 print £2, int(p5*x + .5)/p5,
4400 next i1
4410 print £2, " "
4420 print £2, "sum of"
4430 print £2, "squares",
4440 for i1 = i to i + 3 step 1
```

```
4450 if i1 > p4 then 4480
4460 let x = p(2,i1-i+1)
4470 print £2, int(p5*x + .5)/p5,
4480 next i1
4490 print £2, " "
4500 print £2,
4510 next i
4520 print £2,
4530 print £2,
4540 rem ** end of s/r to print scaled coordinates **
4545 print "xmax=";h(1);" xmin=";l(1)
4546 print "y max =";h(2);" y min =";l(2)
4547 print "zmin =";l(3)
4550 goto 5040
4560 rem ** principal coordinate analysis **
4580 print " number of otus is "; t
4590 for j = 1 to t step 1
4600 for k = 1 to t step 1
4610 let c(j,k) = -c(j,k)*c(j,k)/2
4630 next k
4640 next j
4650 let s2 = 0
4660 for j = 1 to t step 1
4670 let s1 = 0
4680 for k = 1 to t step 1
4690 let s1 = s1 + c(j,k)
4700 next k
4710 let v(j) = s1/t
4720 let s2 = s2 + v(j)
4730 next j
4740 let v0 = s2/t
4750 rem** row and column means in v(j), grand mean in v0 **
4760 for j = 1 to t step 1
4770 for k = 1 to t step 1
4780 let a(j,k) = c(j,k) - v(j) - v(k) + v0
4790 next k
4800 next j
4810 rem ** gower transformation **
4820 rem ** to eigenvalues gosub **
4830 let n0 = t
4840 gosub 5730
4850 rem ** now scales vector col.s **
4860 for k = 1 to t step 1
4870 let s2 = 0
4880 for j = 1 to t step 1
4890 let s2 = s2 + b(j,k)*b(j,k)
4900 next j
4910 if s2 < 1e-6 then 4940
4920 let x = sqr(abs(a(k,k)/s2))
4930 goto 4950
4940 let x = 0
4950 for j = 1 to t step 1
4960 let b(j,k) = b(j,k)*x
4970 next j
4980 next k
4990 rem ** prin. coords. in cols of array b **
5000 rem ** prints eigenvalues **
5010 gosub 5060
5020 rem **prints coord.s **
5030 gosub 5350
5040 stop
5050 rem ** end of run stop **
```

```
5060 rem ** gosub to print eigenvalues **
5070 let s1 = 0
5080 for i = 1 to n0 step 1
5090 let s1 = s1 + a(i,i)
5100 next i
5110 print
5120 print £2,
5130 print £2, "  eigenvalues, percentages of total of ", s1
5140 print £2, "and cumulative percentages are:"
5150 print £2,
5160 print £2, "no, ", "eigenvalue percentage cumulative"
5170 print £2, "                                    percentage"
5180 print £2,
5190 let s0 = 0
5200 for i = 1 to n0 step 1
5210 let s0 = s0 + a(i,i)
5220 print £2, i,
5230 let x = int(a(i,i)*p5 + .5)/p5
5240 print £2, x,
5250 let x = int((a(i,i)/s1)*p5 + .5)/p5
5260 print £2, x*100,
5270 let x = int((s0/s1)*p5 + .5)/p5
5280 print £2, 100*x,
5290 print £2, " "
5300 next i
5310 print £2,
5320 print £2,
5330 return
5340 rem ** end of gosub **
5350 rem ** gosub to print **
5360 rem ** array **
5370 rem ** **
5380 rem ** columns **
5390 rem ** and the **
5400 print £2,
5410 print £2,
5420 if p1 = 1 then 5460
5430 print £2, "     principal components of characters"
5440 print £2, "characters      components"
5450 goto 5480
5460 print £2, "      principal coordinates of otus"
5470 print £2, "otus       coordinates on axes"
5480 print £2,
5490 print £2,
5500 for i = 1 to n0 step 4
5510 print £2, " ",
5520 for i1 = i to i + 3 step 1
5530 if i > p4 then 5550
5540 print £2, i1,
5550 next i1
5560 print £2, " "
5570 print £2,
5580 if i > p4 then 5680
5590 for j = 1 to n0 step 1
5600 print £2, j,
5610 for i1 = i to i + 3 step 1
5620 if i1 > p4 then 5640
5630 print £2, int(p5*b(j,i1) + .5)/p5,
5640 next i1
5650 print £2, " "
5660 next j
5670 print £2,
```

```
5680 next i
5690 print £2,
5700 return
5710 rem ** end of gosub **
5720 rem ** prin. coord.s **
5730 rem ** gosub for **
5740 rem ** of real **
5750 rem ** from davis **
5760 rem ** of matrix **
5770 rem ** matrix b **
5780 rem ** b(j,i) **
5790 rem ** set b to id **
5800 rem ** and final norms **
5810 let a1 = 0
5820 for i = 1 to n0 step 1
5830 for j = 1 to n0 step 1
5840 if i = j then 5870
5850 let b(i,j) = 0
5860 goto 5890
5870 let b(i,j) = 1
5880 let a1 = a1 + a(i,j)*a(i,j)
5890 next j
5900 next i
5910 let a1 = sqr(a1)
5920 let a2 = a1*1e-9/n0
5930 rem ** set indicators and threshold in a3,a4 **
5940 let a3 = a1
5950 let a3 = a3/n0
5960 let a4 = 0
5970 rem ** scan columns for off-diagonal **
5990 for i = 2 to n0 step 1
6000 let i1 = i-1
6010 for j = 1 to i1 step 1
6020 if abs(a(j,i)) - a3 < 0 then 6370
6030 let a4 = 1
6050 let a5 = -a(j,i)
6060 let a6 = (a(j,j) - a(i,i))/2
6070 let a7 = a5/sqr(a5*a5 + a6*a6)
6080 if a6 >= 0 then 6100
6090 let a7 = -a7
6100 let b1 = a7/sqr(2*(1 + sqr(1 - a7*a7)))
6110 let b2 = b1*b1
6120 let b3 = sqr(1 - b2)
6130 let b4 = b3*b3
6150 for k = 1 to n0 step 1
6160 if k = j then 6210
6170 if k = i then 6210
6180 let a0 = a(k,j)
6190 let a(k,j) = a0*b3 - a(k,i)*b1
6200 let a(k,i) = a0*b1 + a(k,i)*b3
6210 let b0 = b(k,j)
6220 let b(k,j) = b0*b3 - b(k,i)*b1
6230 let b(k,i) = b0*b1 + b(k,i)*b3
6240 next k
6250 rem ** continues **
6260 let a8 = 2*a(j,i)*b1*b3
6270 let a0 = a(j,j)
6280 let b0 = a(i,i)
6290 let a(j,j) = a0*b4 + b0*b2 - a8
6300 let a(i,i) = a0*b2 + b0*b4 + a8
6310 let a(j,i) = (a0 - b0)*b1*b3 + a(j,i)*(b4 - b2)
6320 let a(i,j) = a(j,i)
```

```
6330 for k = 1 to n0 step 1
6340 let a(j,k) = a(k,j)
6350 let a(i,k) = a(k,i)
6360 next k
6370 next j
6380 next i
6390 rem ** test for completion **
6400 if a4 > 0 then 5960
6410 if a3 - a2 > 0 then 5950
6420 rem ** now sorts **
6430 for i = 2 to n0 step 1
6440 let j = i
6450 if a(j-1,j-1) - a(j,j) >= 0 then 6560
6460 let a0 = a(j-1,j-1)
6470 let a(j-1,j-1) = a(j,j)
6480 let a(j,j) = a0
6490 for k = 1 to n0 step 1
6500 let a0 = b(k,j-1)
6510 let b(k,j-1) = b(k,j)
6520 let b(k,j) = a0
6530 next k
6540 let j = j - 1
6550 if j - 1 > 0 then 6450
6560 next i
6570 rem ** end of sort **
6590 return
6610 end
```

TRU3DA.BAS

This program generates the amended coordinates for 3-dimensional plotting. Input is read from TRUPC.OUT. The output file, TRUGH.FOR, contains the amended co-ordinates and instructions for the GHOST-80 package. The program allows the size of the axes to be chosen, the maximum total length of each axis being 300 units. The maximum number of OTUs allowed for is 100.

Program TRU3dA.bas

```
10 print " ***** Program tru3dA.bas to generate ***** "
20 print " ***** coordinates for 3-d drawings   ***** "
25 print
30 print "  data input should be in n rows of 3, "
40 print " representing each strain's coordinates on the 3 axes"
50 print " as derived from principal coordinate or component analysis"
60 print "  (see program trupca.bas)  "
70 print
80 print " The name of the output file, to be ran on Ghost 80"
90 print " should be in line 100"
100 open "trugh.for" as file #1
105 open "trupc.out" as file #2
110 dim n(100)
120 dim m(100,3)
130 dim s(100,3)
131 dim f(300)
132 dim g(300)
133 dim h(300)
134 dim p(300)
135 dim q(300)
140 print "Enter size of x axis"
142 print "(i.e. total length in units including negative scale)"
145 input p
150 rem ** p is the size of the x axis **
155 print " Enter total size of the y axis"
160 input q
200 print " enter n, no. of otus/strains"
201 input n
305 print n,"otus"
330 mat input #2, m(n,3)
360 rem ** data read **
```

```
370 print " Enter maximum minus value for x, y, zwithoutentering - sign"
380 input f,g,h
400 for i = 1 to n step 1
405 let x = m(i,1) + f
410 let y = m(i,2) + g
415 let z = m(i,3) + h
440 if x > (p/2) then 600
450 if x < (p/2) then 480
460 if x = (p/2) then 470
470 let s(i,1) = x
475 goto 510
480 let a = q/((p/4)-(x/2))
485 let a = (y/a) + x
490 let s(i,1) = a
500 rem ** a is now new x coordinate **
510 let b = q - (y/2)
520 let c = (z*b)/q
521 rem ** b = z max **
530 let c = c + y
540 let s(i,3) = c
550 let s(i,2) = y
560 rem ** z is now adjusted to read on the y axis **
570 rem ** coordinates to plot are (a,y) and (a,c) **
580 rem ** i.e. x,y and x,z then join them up **
590 goto 800
600 let a = q/((x/2) - (p/4))
605 let a = y/a
607 let a = x - a
610 let s(i,1) = a
620 let b = q - (y/2)
630 let z = (z*b)/q
640 let c = z + y
650 let s(i,3) = c
660 let s(i,2) = y
670 goto 800
800 next i
850 print #1, " program trugh3d.for to draw 3-D plots "
860 print #1, " Put & in 6th space at begining of each row "
870 print #1, " which is the continuation of a data statement "
880 print #1, " Shorten lines by removing some spaces if necessary "
900 print #1, "      real x(";n;"),y1(";n;"),y2(";n;")
910 print #1, "      data x /";s(1,1);";"
911 for i = 2 to n-1
912 print #1,s(i,1);",";
913 next i
914 print #1,s(n,1);"/,"
920 print #1, "      & y1 /";s(1,2);","
921 for i = 2 to n-1
922 print #1,s(i,2);","
923 next i
924 print #1,s(n,2);"/,"
930 print #1, "      £ y2 /"s(1,3);","
931 for i = 2 to n-1
```

```
932 print #1,s(i,3);",";
935 next i
936 print #1,s(n,3);"/"
940 print #1, "      call paper (1)"
941 print #1, "      call pspace(0.1,0.9,0.1,0.9)"
942 print #1, "      call map (0.0,";p;".0,0.0,200.0)"
944 print #1, "      call positn (0.0,0.0)"
945 print #1, "      call join (";(p/4);c;".0)"
946 print #1, "      call join ("(p-(p/4));q;".0)"
947 print #1, "      call join (";p;".0,0.0)"
948 print #1, "      call join (0.0,0.0)"
949 print #1, "      call plotcs ("(p/2):",150.0,'ENTER TITLE HERE')"
950 print #1, "      call pcscen (";(p/2);",-10.0,'axis I')"
951 print #1, "      call pcscen (0.0,-5.0,'-";f;"')"
952 print #1, "      call pcscen (";f;",-5.0,'0')"
953 print #1, "      call pcscen (";p;",-5.0,'";(p-f);"')"
954 print #1, "      call pcscen (-5.0,0.0,'-";g;"')"
955 print #1, "      call pcscen (";(g*((p/4)/q))-1;",";0;")"
956 print #1, "      call pcscen (";q;",";(p/4)-1;",'";(q-g);"')"
957 print #1, "      do 100 I = 1,";n;",1"
958 print #1, "      call positn (x(I),y1(I))"
959 print #1, "      call join (x(I),y2(I))"
960 print #1, "      call pcscen (x(I),y2(I),'o')"
961 print #1, "  100 continue"
987 print #1, "      call grend"
988 print #1, "      stop"
989 print #1, "      end"
990 stop
999 end
```

TRUGH.FOR

This program uses GHOST-80 to draw three dimensional plots. TGhe input data are in data statements at the begining of the program. This needs editing before use by adding an ampersand into the sixth space at the begining of each data row which does not have a data statement.

OTUs are marked, by default, with a symbol 'o'. OTUs may be differentiated by the addition of different characters in the program. Alterations are made in lines 957-961 of TRU3DA.BAS or in the corresponding lines of TRUGH.FOR by repeating these five lines using different characters (i.e. 'o' is replaced by *, +, x etc.) and

specifying the strain numbers in the first of these three lines as shown in the example below.

```
      real x( 49 ),y1( 49 ),y2( 49 )
      data x / 51.1811,28.9394,41.1958,35.6022,61.8183,
     &  80.7654,59.1489,27.3824,32.8495,39.0699,57.5389,86.4772,
     &  51.3704,26.5693,35.7658,36.1384,61.7921,79.1289,63.6187,
     &  26.2711,33.2358,43.7389,32.9577,42.461,82.2981,30.551,
     &  29.6147,44.3146,35.3543,23.9803,106.342,34.5768,29.3993,
     &  43.2762,25.6913,25.8977,117.212,30.887,30.9914,33.2681,
     &  25.9671,24.1272,114.753,32.4329,44.5858,57.6501,32.873,
     &  27.6587,110.523 /,
     & y1 / 111.435,73.0207,56.068,90.982,57.4135,31.4413,
     &  134.334,74.1203,62.8406,88.4415,53.4206,23.4021,106.909,
     &  74.4351,55.7453,93.3417,55.185,29.0855,132.181,71.6659,
     &  58.2945,82.5621,50.2288,46.3193,111.677,69.2333,60.2359,
     &  80.7175,51.9394,40.5542,109.244,67.5161,61.8416,86.7246,
     &  85.0727,14.8164,69.912,78.0086,72.5646,57.2229,79.4465,
     &  9.1643,84.7749,87.5393,92.1695,101.879,91.1263,17.9055,
     &  45.9151 /,
     & y2 / 152.546,141.621,101.334,126.634,146.097,51.6673,
     &  167.608,134.334,105.11,126.21,142.024,46.4689,143.655,
     &  135.101,96.9656,133.251,144.773,60.2098,164.797,133.718,
     &  96.9839,124.43,143.314,55.2463,155.54,139.981,103.525,
     &  123.464,142.443,75.4385,141.15,138.643,104.918,131.698,
     &  142.172,87.5409,143.294,141.385,111.439,125.993,135.263,
     &  85.5529,135.105,140.557,117.142,145.315,141.33,90.4088,
     &  127.491 /
      call paper (1)
      call pspace(0.1,0.9,0.1,0.9)
      call map (0.0,140.0,0.0,200.0)
      call positn (0.0,0.0)
      call joiln ( 35.0,140.0)
      call joiln ( 105.0,140.0)
      call joiln ( 140.0,0.0)
      call joiln (0.0,0.0)
      call plotcs ( 70.0,200.0,'Complete data')
      call pcscen (70.0,-10.0,'axis I')
      call pcscen (0.0,-5.0,'-40')
      call pcscen ( 40.0,-5.0,'0')
      call pcscen ( 140.0,-5.0,'100')
      call pcscen (-5.0,0.0,'-70')
      call pcscen ( 10.5,70.0,'0')
      call pcscen ( 34.0,140.0,'70')
      do 100 I = 1,49,1
      call positn (x(I),y1(I))
      call joiln (x(I),y2(I))
      call pcscen (x(I),y2(I),'o')
  100 continue
      call grend
      stop
      end
```

Appendix 8. DNA relatedness among *Listeria* strains : Rocourt *et al.* 1982.

Source of unlabelled DNA		% Homology with labelled DNA from:					
SLCC[1] number	Serotype	SLCC 5329	SLCC 2479	SLCC 3769	SLCC 3379	SLCC 5334	SLCC 3990
L. monocytogenes							
5329	1/2a	100	41	32	53	47	25
2371	1/2a	98	41	22	50	41	24
53	1/2a	63	39	22	53	42	24
30	1	95	37	28	45	41	25
1044	1	84	39	29	46	45	23
3939	1/2c	96	40	27	50	44	26
2373	3a	94	41	28	46	40	22
2540	3b	75	41	31	46	44	25
3993	3b	83	38	21	49	45	24
2479	3c	84	31	29	46	42	29
4210	3c	91	40	25	49	43	22
2374	4a	72	37	26	54	46	23
788	4a	76	39	21	52	46	23
2375	4b	79	39	23	53	45	22
1382	4b	77	39	23	47	44	26
5510	4b	75	41	27	58	42	22
2376	4c	71	38	24	57	45	21
3737	4c	70	37	18	48	42	24
2377	4d	70	41	31	49	43	25
2378	4e	68	49	25	50	41	24
1745	4e	89	32	22	53	47	26
2482	"7"	83	47	29	56	44	26
L. ivanovii							
2379	5	33	100	100	42	34	44
3769	5	31	102	100	30	32	39
3887	5	28	101	104	31	31	44
5378	5	27	93	91	29	29	32
5379	5	33	92	103	29	29	32
5380	5	29	98	85	31	29	31
L. innocua							
3379	6a	54	40	23	100	44	24
4275	6a	54	45	28	99	47	28
4883	6a	53	33	25	91	45	28
5375	6a	47	36	28	89	42	19
3423	6b	52	35	22	87	46	26
5290	6b	56	40	29	92	45	28
5337	6b	47	43	24	89	42	28
4482	-	55	32	29	93	47	29
2745	4ab	53	36	25	92	45	25

L. welshimeri							
5334	6a	46	38	28	46	100	28
3809	6a	41	29	22	38	88	21
3810	6a	41	27	24	39	86	22
5328	6b	44	27	30	42	96	30
L. seeligeri							
3954	1/2b	38	34	40	34	34	73
4115	4c	35	49	43	34	30	78
3754	4d	36	41	47	32	35	89
3616	6b	35	38	42	33	35	71
3678	-	29	43	40	26	28	85
4109	6b	32	45	44	28	32	87
L. grayi							
2080		4	7	3	5	5	3
5330		21	7	5	16	14	4
L. murrayi							
4425		4	2	3	6	2	2
4426		9	3	2	6	4	1
4427		8	2	3	5	6	1

[1]Strain number of the Special *Listeria* Culture Collection.

Appendix 9 Abbreviations

CTAB : cetyltrimethyl ammonium bromide

DNA : deoxyribonucleic acid

EDTA : ethylenediaminotetraacetic acid

HA : hydroxyapatite

HCl : hydrochloric acid

n : number of replications

OTU : operational taxonomic unit

RNA : ribonucleic acid

SSC : standard saline citrate

s.d. : standard deviation

T_m : melting temperature

T_{or} : optimal reassociation temperature

% H : % homology

5. REFERENCES

Anon 1983. Editorial: *Listeria* infections in farm animals. Veterinary Record 112 314.

Anon 1986. Communicable disease quarterly. Public Health Laboratory Service, Microbiology Digest 3 35-37.

Applequist, J. 1967. *In*: Conformation of Biopolymers. p.403. G.M. Ramachandran (ed.). Academic Press : New York.

Audurier, A., Pardon, P., Marly, J. and Lantier, F. 1980. Experimental infection of mice with *Listeria monocytogenes* and *Listeria innocua*. Annales de Microbiologie 131B 47-57.

Baess, I. 1979. Deoxyribonucleic acid relatedness among species of slowly-growing Mycobacteria. Acta Pathologica Microbiologica Scandinavica Section B87 221-226.

Beaven, G. H., Holiday, E. R. and Johnson, E. A. 1955. Optical properties of nucleic acids and thier components. *In* : The Nucleic Acids, volume 1, pp. 493-545, E. Chargaff and J.N. Davidson (ed.s) Academic Press, New York.

Bellamy, A. R. and Ralph, R. K. 1968. Recovery and purification of nucleic acids by means of cetyltrimethylammonium bromide 104 156-160.

Bergey's Manual of Systematic Bacteriology 9th edition Volume 2, 1984. N. R. Krieg and J. G. Holt (ed.s). Williams and Wilkins, Baltimore.

Bernardi, G. 1965. Chromatography of nucleic acids on hydroxyapatite. Nature 206 779-783.

Bolton, E. T. and McCarthy, B. J. 1962. A general method for the isolation of RNA complementary to DNA. Proceedings of the National Academy of Science, U.S.A. 48 1390-1397.

Bonner, J., Kung, G. and Bekher, I. 1967. A method for hybridization of nucleic acid molecules at low temperature. Biochemistry 6 3650-3653.

Bonner, T. I., Brenner, D. J., Neufeld, B. R. and Britten, R. J. 1973. Reduction in the rate of DNA reassociation by sequence divergence. Journal Molecular Biology 81 123-135.

Bouvet, P. J. M. and Grimont, P. A. D. 1986. Taxonomy of the genus *Acinetobacter* with the recognition of *Acinetobacter baumanii* sp. nov., *Acinetobacter haemolyticus* sp. nov., *Acinetobacter johnsonii* sp. nov., and *Acinetobacter junii* sp. nov. and emended descriptions of *Acinetobacter calcoaceticus* and *Acinetobacter lwoffi.* International Journal of Systematic Bacteriology 36 228-240.

Bradley, S. G. 1972. Relationships among Mycobacteria and Nocardiae based upon Deoxyribonucleic acid reassociation. Journal of Bacteriology 113 645-651.

Brenner, D. J. 1973. Deoxyribonucleic acid reassociation in the taxonomy of enteric bacteria. International Journal of Systematic Bacteriology 23 298-307.

Brenner, D. J. and Cowie, D. B. 1968. Thermal stability of *Escherichia coli - Salmonella typhimurium* DNA duplexes. Journal of Bacteriology 95 2258-2262.

Brenner, D. J., Martin, M. A. and Hoyer, B. H. 1967. Deoxyribonucleic acid homologies among some bacteria. Journal of Bacteriology 94 486-487.

Brenner, D. J., Fanning, G. R., Johnson, K. E., Citarella, R. V. and Falkow, S. 1969. Polynucleotide sequence relationships among members of the *Enterobacteriaceae.* Journal of Bacteriology 98 637-650.

Brenner, D. J., Fanning, G. R., Skerman, F. J. and Falkow, S. 1972*a*. Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. Journal of Bacteriology 109 953-965.

Brenner, D. J., Fanning, G. R. and Steigerwalt, A. G. 1972*b*. Deoxyribonucleic acid relatedness among species of *Erwinia* and between *Erwinia* species and other enterobacteria. Journal of Bacteriology 110 12-17.

Brenner, D. J., Farmer, J. S., Fanning, G. R., Steigerwalt, A. G., Klykken, Paal, Wathen, H. G., Hickman, F. W. and Ewing, W. H. 1978. Deoxyribonucleic acid relatedness of *Proteus* and *Providencia* species. International Journal of Systematic Bacteriology 28 269-282.

Bridge, P. D. and Sneath, P. H. A. 1983. Numerical Taxonomy of *Streptococcus.* Journal of General Microbiology 129 565-597.

Britten, R. J. 1969. Reassociation of non-repeated DNA. Carnegie Institute Washington Yearbook 67 330-335.

Britten, R. J. and Kohne, D. E. 1966. Nucleotide sequence repetition in DNA. *In* : Carnegie Institute, Washington Yearbook, 65 78-106.

Britten, R. J., Graham, D. E. and Neufeld, B. R. 1974. Analysis of repeating DNA sequences by reassociation. Methods in Enzymology 29 363-405.

Callies, E. and Mannheim, W. 1980. Deoxyribonucleic acid relatedness of some menaquinone producing *Flavobacterium* and *Cytophaga* strains. Antonie van Leeuwenhoek 46 41-49.

Caster, W. O. 1951. Variability in the Beckman Spectrophotometer. Analytical Chemistry 23 1229.

Chargaff, E. 1955. *In* : The Nucleic Acids Volume 1, pp 307-371 (E. Chargaff and J.N. Davidson, eds.). Academic Press, New York.

Citarella, R. V. and Colwell, R. R. 1970. Polyphasic taxonomy of the genus *Vibrio* : Polynucleotide sequence relationships among selected *Vibrio* species. Journal of Bacteriology 104 434-442.

Collins, M. D., Farrow. J. A. E. and Jones, D. 1986. *Enterococcus mundtii* species nova. International Journal of Systematic Bacteriology 36 8 - 12.

Collins, M. D., Farrow, J. A. E., Phillips, B. A., Ferusu, S. and Jones, D. 1987. Classification of *Lactobacillus divergens, Lactobacillus pisicola* and some catalase-negative, asporogenous, rod-shaped bacteria from poultry in a new genus, *Carnobacterium.* International Journal of Systematic Bacteriology 37 310-316.

Conover, W. J. 1971. *In* : Practical Nonparametric Statistics, 163, John Whiley, New York.

Coykendall, A. L. and Munzenmaier, A. J. 1978. DNA base sequence studies on glucan-producing and glucan negative strains of *Streptococcus mitior.* International Journal of Systematic Bacteriology 28 511-515.

Coykendall, A. L., Setterfield, J. and Slots, J. 1983. DNA relatedness among *Actinobacillus actinomycetemcomitans, Haemophilus aphrophilus* and other *Actinobacillus* sp.s. International Journal of Systematic Bacteriology 33 422-424.

Coykendall, A. L., Wesbecher, P. M. and Gustafson, K. B. 1987. "*Streptococcus milleri*", *Streptococcus constellatus* and *Streptococcus intermedius* are later synonyms of *Streptococcus anginosus.* International Journal of Systematic Bacteriology 37 222-228.

Cristifolini, G. 1980. Interpretation and analysis of serological data. *In* : Chemosystematics : Principles and Practice. F.A. Bisby, J.G. Vaught, C.A. Wright (ed.s). Academic Press, London and New York.

Crombach, W. H. J. 1972. DNA base composition of soil arthrobacters and other coryneforms from cheese and seafish. Antonie van Leeuwenhoek 38 105-120.

Crombach, W. H. J. 1973. Deep-freezing of bacterial DNA for thermal denaturation and hybridization experiments. Antonie van Leeuwenhoek 39 249-255.

Crombach, W. H. J. 1974. Thermal stability of homologous and heterologous bacterial DNA duplexes. Antonie van Leeuwenhoek 40 133-144.

Crosa, J. H., Brenner, D. J. and Falkow, S. 1973. Use of a single strand specific nuclease for analysis of bacterial and plasmid deoxyribonucleic acid homo- and hetero duplexes. Journal of Bacteriology 115 904-911.

DeKesel, M., Coene, M., Portaels, F. and Cocito, C. 1987. Analysis of deoxyribonucleic acid from armadillo-derived Mycobacteria. International Journal of Systematic Bacteriology 37 317-322.

DeLey, J. 1969. Compositional nucleotide distribution and the theoretical prediction of homology in bacterial DNA. Journal of Theoretical Biology 22 89-116.

DeLey, J. 1970. Reexamination of the association between melting point, buoyant density and chemical base composition of DNA. Journal of Bacteriology 101 738-754.

DeLey, J. 1971. Hybridization of DNA p.311-329. *In* : Methods in Microbiology 5A. Norris, J.R. and Ribbons, D.W. (ed.s). Academic Press (London).

DeLey, J. and Friedman, S. 1964. Deoxyribonucleic acid hybrids of acetic acid bacteria. Journal of Bacteriology 88 937-945.

DeLey, J. and Friedman, S. 1965. Similarity of *Xanthomonas* and *Pseudomonas* deoxyribonucleic acids. Journal of Bacteriology 89 1306-1309.

DeLey, J., Cattoir, H. and Reynaerts, A. 1970. The quantitative measurements of DNA hybridization from renaturation rates. European Journal of Biochemistry 12 133-142.

DeLey, J. and Tijtgat, R. 1970. Evaluation of membrane filter methods for DNA-DNA hybridization. Antonie van Leeuwenhoek 36 461-474.

DeLey, J., Tijtgat, R., DeSmedt, J. and Michiels, M. 1973. Thermal stability of DNA-DNA hybrids within the genus *Agrobacterium*. Journal of General Microbiology 78 241-252.

Denhardt, D. T. 1966. A membrane filter technique for the detection of complementary DNA. Biochemical and Biophysical Research Communications 23 641-646.

Dent, V. E. and Williams, R. A. D. 1986*a*. *Actinomyces slackii* species nov. from dental plaque of dairy cattle. International Journal of Systematic Bacteriology 36 392-395.

Dent, V. E. and Williams, R. A. D. 1986*b*. DNA reassociation between strains of *Lactobacillus animalis, Lactobacillus odontolyticus* and strains resembling *Lactobacillus acidophilus* isolated from animals. International Journal of Systematic Bacteriology 36 481-482.

Doty, P., Marmur, J., Eigner, J. and Schildkraut, C. 1960. Strand separation and specific recombination in deoxyribonucleic acids : physical and chemical studies. Proceedings of the National Academy of Science U.S.A. 46 461-476.

Durst, J. and Berencsi, G. 1975. New selective media of rivanol content for *Listeria monocytogenes*. Zentrablatt fur Bakteriologie, Mikrobiologie und Hygiene Abt 1 Orig. A. 232 410-411.

Escande, F., Grimont, F., Grimont, P.A.D. and Bercouvier, H. 1984. Deoxyribonucleic acid relatedness among strains of *Actinobacillus* spp. and *Pasteurella ureae*. International Journal of Systematic Bacteriology 34 309-315.

Ezaki, T., Facklam, R., Takeuchi, N. and Yabuchi, E. 1986. Genetic relatedness between the type strain of *Streptococcus anginosus* and minute colony-forming B-haemolytic streptococci carrying different Lancefield grouping antigens. International Journal of Systematic Bacteriology 36 345-347.

Falk, E. C., Johnson, J. L., Baldini. V. L. D., Dobereiner, J. and Krieg, N. R. 1985. Deoxyribonucleic acid and ribonucleic acid homology studies of the genera *Azospirrillum* and *Conglomeromonas*. International Journal of Systematic Bacteriology 36 80-85.

Falkow, S., Rownd, R. and Baron, L. S. 1962. Genetic homology between *E. coli* K12 and *Salmonella*. Journal of Bacteriology 84 1303-1312.

Fenlon, D. R., 1985. Wild birds and sileage as reservoirs of *Listeria* in the agricultural environment. Journal of Applied Bacteriology 59 537-543.

Ferusu, S. 1980. Taxonomic Studies on *Listeria, Erysipelothrix* and atypical Lactobacilli. PhD Thesis, Leicester University.

Fisher, R. A. and Yates, F. 1957. Statistical tables for biological, agricultural and medical research. Fifth edition. Oliver and Boyd : Edinburgh.

Fredericq, E., Oth, A. and Fontaine, F. 1961. The ultraviolet spectrum of Deoxyribonucleic acids and their constituents. Journal of Molecular Biology 3 11-17.

Friedman, S. and DeLey, J. 1965. "Genetic Species" concept in *Xanthomonas*. Journal of Bacteriology 89 95-100.

Garvie, E. I. 1978. *Streptococcus raffindaetis* Orla-Jensen and Harven a Group N Streptococcus found in raw milk. International Journal of Systematic Bacteriology 28 190-193.

Gebers, R., Martens, B., Wehmeyer, U. and Hirsch, P. 1986. DNA homologies of *Hyphomicrobium* species, *Hyphomonas* species and other hyphal, budding bacteria. International Journal of Systematic Bacteriology 36 241-245.

Gibbins, A. M. and Gregory, K. F. 1972. Relatedness among *Rhizobium* and *Agrobacterium* species determined by three methods of nucleic acid hybridisation. Journal of Bacteriology 111 129-141.

Gillespie, D. and Gillespie, D. 1971. Ribonucleic acid - Deoxyribonucleic acid hybridization in aqueous solutions and in solutions containing formamide. Biochemistry Journal 125 481-487.

Gillespie, D. and Spiegelman, S. 1965. A quantitative assay for DNA-RNA hybrids with DNA immobilised on a membrane. Journal of Molecular Biology 12 829-842.

Gillis, M., DeLey, J. and DeCleene, M. 1970. The determination of molecular weight of bacterial genome DNA from renaturation rates. European Journal of Biochemistry 12 143-153.

Gitter, M. 1986. A changing pattern of ovine listeriosis in Great Britain. *In* : Proceedings of the 9th International Symposium on Problems of Listeriosis, pp 294-299. Ed. Courtieu, A.L., Espaze, E.P. and Reynaud, A.E. Nantes: Universite de Nantes.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53 325-338.

Graves, I. L. 1968. Interaction of heat denatured HeLa cell DNA with synthetic and natural polysaccharides. Biopolymers 6 1573-1578.

Gray, M. L. 1957. A rapid method for the detection of colonies of *Listeria monocytogenes*. Zentralblat Bakteriol. Orig. 169 373-377.

Gray, M. L., Stafseth, H. J., Thorp Jr., F., Sholl, L. B. and Riley Jr., W. F. 1948. A new technique for isolating listerellae from the bovine brain. Journal of Bacteriology 55 471-476.

Gray, M. L. and Killinger, A. H. 1966. *Listeria monocytogenes* and listeric infections. Bacteriological Reviews 30 309-382.

Grimont, P. A. D. and Popoff, M. Y. 1980. Use of principal component analysis in the interpretation of deoxyribonucleic acid relatedness. Current Microbiology 4 337-342.

Grimont, P. A. D., Popoff, M. Y., Grimont, F., Coynault, C. and Lemelin, M. 1980. Reproducibility and correlation study of three deoxyribonucleic acid hybridization procedures. Current Microbiology 4 325-330.

Gronstol, H. and Aspoy, E. 1977. A new selective medium for the isolation of *Listeria monocytogenes*. Nord. Vet-Med. 29 446-451.

Hartford, T. and Sneath, P. H. A. 1988. Distortion of taxonomic structure from DNA relationships due to different choice of reference strains. Systematic and Applied Microbiology 10 241-250.

Hartford, T. and Sneath, P. H. A. 1990. *A Review* : Experimental error in DNA-DNA pairing : a survey of the literature. Journal of Applied Bacteriology 68 527-542.

Hastings, R. B. and Kirby, K. S. 1966. The nucleic acids of *Drosophila melanogaster*. Biochemistry Journal 100 532-539.

Hensel, R., Denharter, W., Kandler, O., Kroppenstedt, R. M. and Stackebrandt, E. 1986. Chemotaxonomic and molecular-genetic studies of the genus *Thermus* : Evidence for a phylogenetic relationship of *Thermus aquaticus* and *Thermus ruber* to the genus *Deinococcus.* International Journal of Systematic Bacteriology 36 444-453.

Hildebrand, D. C., Huisman, O. C. and Schroth, M. N. 1984. Use of DNA hybridization values to construct three-dimensional models of fluorescent pseudomonad relationships. Canadian Journal of Microbiology 30 306-315.

Hood, D. W., Dow, C. S. and Green, P. N. 1987. DNA:DNA hybridization studies on the pink pigmented facultative methylotrophs. Journal of General Microbiology 133 709-720.

Hope, K. 1968. Methods of multivariate analysis. University of London Press, London.

Hoyer, B. H., Bolton, E. T. and McCarthy, B. J. 1964. A molecular approach to the systematics of higher organisms. Science 144 959-968.

Hoyer, B. H. and King, J. R. 1969. Deoxyribnonucleic acid sequence losses in a stable streptococcal form. Journal of Bacteriology 97 1516-1517.

Huss, V. A. R., Festl, H. and Schleirer, K. H. 1983. Studies on the spectrophotometric determination of DNA hybridization from renaturation rates. Systematic and Applied Microbiology 4 184-192.

Hutton, J. R. 1977. The effects of formamide and urea on the thermal stability and renaturation kinetics of DNA. Nucleic Acids Research 4 3537-3555.

Imaeda, T. 1985. Deoxyribonucleic acid relatedness among selected strains of *Mycobacterium tuberculosis, Mycobacterium bovis, Mycobacterium bovis* BCG, *Mycobacterium microti* and *Mycobacterium africanum.* International Journal of Systematic Bacteriology 35 147-150.

Imaeda, T., Barksdale, L. and Kirchheimer, W. F. 1982. Deoxyribonucleic acid of *Mycobacterium lepraemurium* : Its genome size, base ratio and homology with those of other Mycobacteria. International Journal of Systematic Bacteriology 32 456-458.

Ivanov, I. 1957. La Listeriose chez les ovins et les caprins. Bull. Off. Int. Epizoot. 48 571-583.

Ivanov, I. 1975. Establishment of non-motile strains of *Listeria monocytogenes* type 5. pp. 18-26. *In* : Woodbine, M. (ed.), Problems of listeriosis. Leicester : Leicester University Press.

Johnson, J. L. 1973. Use of nucleic acid homologies in the taxonomy of anaerobic bacteria. International Journal of Systematic Bacteriology 23 308-315.

Johnson, J. L. 1985. Determination of DNA base composition. *In* : Methods in Microbiology, Volume 18 pp. 1-32, (G. Gottschalk ed.). Academic Press, Inc. (London) Ltd., London.

Johnson, J. L. and Harich, B. 1983. Ribosomal ribonucleic acid homology among species of the genus *Veilonella* Prevot. International Journal of Systematic Bacteriology 33 760-764.

Johnson, J. L. and Harich, B. 1986. Ribosomal ribonucleic acid homology among species of the genus *Bacteroides*. International Journal of Systematic Bacteriology 36 71-79.

Johnson, J. L. and Ordal, E. J. 1968. Deoxyribonucleic acid homology in bacterial taxonomy : Effect of incubation temperature on reaction specificity. Journal of Bacteriology 95 893-900.

Jones, D. 1975. The taxonomic position of *Listeria*. *In* : Problems of Listeriosis. Woodbine, M. (ed.) Leicester University Press.

Jones, D. 1986. The genus *Erypsipelothrix*. *In* : Bergey's Manual of Systematic Bacteriology, Volume 2 pp. 1245-1249. (P. H. A. Sneath, N. S. Mair and M. E. Sharpe, ed.s.). Baltimore : Williams and Wilkins.

Kalb, V. F. and Bernlohr, R. W. 1977. A new spectrophotometric assay for protein in cell extracts. Analytical Biochemistry 82 362-371.

Khan, M. A., Seaman, A. and Woodbine, M. 1973. The pathogenicity of *Listeria monocytogenes*. Zentrablatt fur Bakteriologie, Mikrobiologie und Hygiene Abt 1 Orig. A. 224 362-375.

Kilpper-Bälz, R., Wenzig, P. and Schleifer, K. H. 1985. Classification of Viridans streptococci as *S. oralis*. International Journal of Systematic Bacteriology 35 482-458.

Kilpper-Balz, R. and Schleifer, K. H. 1987. *Streptococcus suis* sp. nov., nom. rev. International Journal of Systematic Bacteriology 37 160-162.

Kingsbury, D. T., Fanning, G. R., Johnson, K. E. and Brenner, D. J. 1969. Thermal stability of interspecies *Neisseria* DNA duplexes. Journal of General Microbiology 55 201-208.

Kohne, D. E. 1968. Taxonomic Applications of DNA Hybridisation techniques. *In :* Chemotaxonomy and Serotaxonomy pp.117-130. Systematics Association Special Volume No.2 J.G. Hawkes (ed.) 1968.

Kourilsky, P., Leidner, J. and Tremblay, G. Y. 1971. DNA-DNA Hybridisation on filters at low temperature in the prescence of formamide or urea. Biochemie 53 1111-1114.

Kramer, P. A. and Jones, D. 1969. Media selective for *Listeria monocytogenes*. Journal of Applied Bacteriology 32 381-394.

Lachance, M. 1980. A simple method for determination of DNA relatedness by thermal elution in hydroxyapatite microcolumns. International Journal of Systematic Microbiology 30 433-436.

Laird, C. D., McConaughy, B. L. and McCarthy, B. J. 1969. On the rate of fixation of nucleotide substitutions on evolution. Nature 224 149.

Legault-Demare, J., Desseaux, B., Hayman, T., Seror, S. and Ress, G. P. 1967. Studies on hybrid molecules of nucleic acids. I. DNA-DNA hybrids on nitrocellulose filters. Biochemical and Biophysical Research Communications 28 550-557.

Levy-Frebault, V., Grimont, F., Grimont, P. A. D. and David, H. L. 1986. DNA relatedness study of the *Mycobacterium fortuitum* - *Mycobacterium chelonae* complex. International Journal of Systematic Bacteriology 36 458-460.

Love, D. N., Johnson, J. L., Jones, R. F., Bailey, M. and Calverley, A. 1986. *Bacteroides tectum* species nova and characters of other non-pigmented *Bacteroides* isolates from soft-tissue infections from cats and dogs. International Journal of Systematic Bacteriology 36 123-128.

Love, D. N., Cato, E. P., Johnson, J. L., Jones, R. F. and Bailey, M. 1987*a*. DNA hybridization among strains of *Fusobacterium* isolated from soft tissue infections of cats: Comparison with human and animal Type strains from oral and other sites. International Journal of Systematic Bacteriology 37 23-26.

Love, D. N., Johnson, J. L., Jones, R. F. and Calverley, A. 1987*b*. *Bacteroides salivosus* sp. nov., an assacharolytic, black-pigmented species from cats. International Journal of Systematic Bacteriology 37 307-309.

Mackaness, G. B. 1971. Resistance to intracellular infection. Journal of Infectious Diseases 123 439-454. International Journal of Systematic Bacteriology 38 340-347.

Mandel, M., Igambi, L., Bergendahl, J., Dodson Jr., M. L. and Scheltgen, E. 1970. Correlation of melting temperature and caesium chloride buoyant density of bacterial deoxyribonucleic acid. Journal of Bacteriology 101 333-338.

Mannarelli, B. M. 1988. Deoxyribonucleic acid relatedness among strains of the species *Butyrivibrio fibrisolvens.* International Journal of Systematic Bacteriology 38 340-347.

Marmur, J. 1961. A procedure for the isolation of deoxyribonucleic acid from microorganisms. Journal of Molecular Biology 3 208-218.

Marmur, J. and Doty, P. 1961. Thermal renaturation of deoxyribonucleic acids. Journal of Molecular Biology 3 585-594.

Marmur, J. and Doty, P. 1962. Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. Journal of Molecular Biology 5 109-118.

Marmur, J., Schildkraut, C. L. and Doty, P. 1961. The reversible denaturation of DNA and its use in studies of nucleic acid homologies and the biological relatedness of microorganisms. Journal Chim. Phys. Physicochem. Biol. 58 945-955.

Marmur, J., Rownd, R. and Schildkraut, C. L. 1963. Denaturation and Renaturation of Deoxyribonucleic acids. Progress in Nucleic Acid Research in Molecular Biology. 1 231-300.

Marsh, J. L. and McCarthy, B. J. 1974. Effect of reaction conditions on the reassociation of divergent deoxyribonucleic acid sequences. Biochemistry 13 3382-3388.

Martin, M. A. and Hoyer, B. H. 1966. Thermal stabilities and specificities of reannealed animal deoxyribonucleic acids. Biochemistry 3 2706-2713.

Mays, T. D., Holdeman, L. V., Moore, W. E. C., Rogosa, M. and Johnson, J. L. 1982. Taxonomy of the genus *Veillonella* Prevot. International Journal of Systematic Bacteriology 32 28-36.

McCarthy, B. J. 1967. Arrangement of base sequences in deoxyribonucleic acid. Bacteriology Reviews 31 215-229.

McCarthy, B. J. and Bolton, E. T. 1963. An approach to the measurement of genetic relatedness among organisms. Proceedings of the National Academy of Science U.S.A. 50 156-164.

McConaughy, B. L., Laird, C. D., McCarthy, B. J. 1967. Nucleic acid reassociation in formamide. Biochemistry 8 3289-3295.

McFadden, J. J., Butcher, P. D., Chiodini, R. J. and Hermon-Taylor, J. 1987. Determination of genome size and DNA homology between an unclassified *Mycobacterium* species isolated from patients with Crohn's disease and other Mycobacteria. Journal of General Microbiology 133 211-214.

McLauchlin, J. 1987. A review : *Listeria monocytogenes*, recent advances in the taxonomy and epidemiology of listeriosis in humans. Journal of Applied Bacteriology 63 1-11.

Micales, B. K., Johnson, J. L. and Claus, G. W. 1985. Deoxyribonucleic acid homologies among organisms in the genus *Gluconobacter*. International Journal of Systematic Bacteriology 35 79-85.

Miyazawa, Y. and Thomas, Jr. C. A. 1965. Composition of short segments of DNA molecules. Journal of Molecular Biology 11 223-237.

Moore, L. V. H., Johnson, J. L. and Moore, W. E. C. 1987. *Selenomonas noxia* sp. nov., *Selelomonas fluggei* sp. nov., *Selelomonas infelix* sp. nov., *Selenomonas dianae* sp.nov. and *Selenomonas artemedis* sp. nov., from human gingival crevice. International Journal of Systematic Bacteriology 37 271-280.

Morozumi, T., Urs, P., Braun, R. and Nicolet, J. 1986. DNA relatedness among strains of *Haemophilus parasuis* and other *Haemophilus* spp. of swine origin. International Journal of Systematic Bacteriology 36 17-19.

Murray, E. G. D., Webb, R. A. and Swann, M. B. R. 1926. A disease of rabbits characterised by large mononuclear leukocytosis caused by a hitherto undescribed bacillus *Bacterium monocytogenes* (sp. n.). Journal of Pathology 29 407-439.

Mutters, R., Ihm, P., Pohl, S., Fredriksen, W. and Mannheim, W. 1985. Reclassification of the genus *Pasteurella* Trevisan 1887 on the basis of DNA homology, with proposals for the new species *Pasteurella dogmatis, Pastuerella canis, Pasteurella anatis* and *Pasteurella langaa.* International Journal of Systematic Bacteriology 35 309-322.

Nakamura, L. K. 1987*a*. *Bacillus alginolyticus* sp. nov. and *Bacillus chrondroitinus* sp. nov.. Two alginate-degrading species. International Journal of Systematic Bacteriology 37 284-286.

Nakamura, L. K. 1987*b*. *Bacillus polymyxa* (Prazmowski) Mace 1889 deoxyribonucleic acid relatedness and base composition. International Journal of Systematic Bacteriology 37 391-397.

Nakamura, L. K. and Swezey, J. 1983*a*. Taxonomy of *Bacillus circulans* Jordan 1890: base composition and reassociation of deoxyribonucleic acid. International Journal of Systematic Bacteriology 33 46-52.

Nakamura, L. K. and Swezey, J. 1983*b*. Deoxyribonucleic acid relatedness of *Bacillus circulans* Jordan 1890 strains. International Journal of Systematic Bacteriology 33 703-708.

Nygaard, A. P. and Hall, B. D. 1963. A method for the detection of RNA-DNA complexes. Biochemical and Biophysical Research Communications 12 98-104.

Ogasawara-Fujita, N. and Sakaguchi, K. 1976. Classification of micrococci on the basis of DNA homology. Journal of General Microbiology 94 97-106.

Okanishi, M. and Gregory, K. F. 1970. Methods for the determination of deoxyribonucleic acid homologies in *Streptomyces*. Journal of Bacteriology 104 1086-1094.

Owen, R. J., Hill, L. R. and Lapage, S. P. 1969. Determination of DNA base compositions from melting profiles in dilute buffers. Biopolymers 7 503-516.

Owen, R. J. and Lapage, S. P. 1976. The thermal denaturation of partly purified bacterial DNA and its taxonomic applications. Journal of Applied Bacteriology 41 335-340.

Owen, R. J. and Snell, J. S. S. 1976. Deoxyribonucleic acid reassociation in the classification of Flavobacteria. Journal of General Microbiology 93 89-102.

Pace, B. and Pace, N. R. 1971. Gene dosage for 5s ribosomal ribonucleic acid in *Escherichia coli* and *Bacillus megaterium*. Journal of Bacteriology 105 142-149.

Pirie, J. H. H. 1927. The genus *Listerella* Pirie. Science 91 383.

Pirie, J. H. H. 1940. *Listeria* : change of name for a genus of bacteria. Nature 145 264.

Pohl, S., Bertschinger, H. U., Fredriksen, W. and Mannheim, W. 1983. Transfer of *Haemophilus pleuropneumoniae* and the *Pasteurella haemolytica* - like organism causing porcine necrotic pleuropneumonia to the genus *Actinobacillus* (*Actinobacillus pleuropneumoniae* comb. nov.) on the basis of phenotypic and deoxyribonucleic acid relatedness. International Journal of Systematic Bacteriology 33 510-514.

Popoff, M. and Coynault, C. 1980. Use of DEAE-cellulose filters in the S1 nuclease methjod for bacterial deoxyribonucleic acid hybridization. Annales de Microbiologie (Institute Pasteur) 131A 151-155.

Potts, T. V. and Berry, E. M. 1983. Deoxyribonucleic acid - deoxyribonucleic acid hybridization analysis of *Actinobacillus actinomycetemcomitans* and *Haemophilus aphrophilus*. International Journal of Systematic Bacteriology 33 765-771.

Potts, T. V., Mitra, T., O'Keefe. T., Zambon, J. J. and Genco, R. J. 1986. Realtionships among isolated of oral haemophili as determined by DNA-DNA hybridization. Archives in Microbiology 145 136-141.

Ralovich, B., Ferray, A., Mero, E., Malovics, H. and Szazados, I. 1971. New selective medium for isolation of *Listeria monocytogenes*. Zentrablatt fur Bakteriologie, Mikrobiologie und Hygiene Abt 1 Orig. A. 216 88-91.

Rocourt, J. and Seeliger, H. P. R. 1985. Distribution des especes du genre *Listeria*. Zentrablatt fur Bakteriologie, Mikrobiologie und Hygiene Abt 1 Orig. A Medizinische Mikrobiologie Infektionskrankheiten und Parasitologie (Stuttgart) 259 317-330.

Rocourt, J., Grimont, F., Grimont, P. A. D. and Seeliger, H. P. R. 1982. DNA relatedness among serovars of *Listeria monocytogenes* sensu lato. Current Microbiology 7 383-388.

Rocourt, J., Schrettenbrunner, A. and Seeliger, H. P. R. 1983. Differentiation biochimique des groupes genomiques de *Listeria monocytogenes* (*senso lato*). Annuals de Microbiologie (Paris) 134A 65-67.

Rocourt, J., Hof, H., Schrettenbrunner, A., Malinvenni, R. and Bille, J. 1986. Meningite purulente aigue a la *Listeria seeligeri* chez un adulte immunocompetent. Schwiz. med. Wschr. 116 248-251.

Rocourt, J., Wehmeyer, U. and Stackebrandt, E. 1987*a*. Transfer of *Listeria denitrificans* to a new genus *Jonesia* gen. nov. as *Jonesia denitrificans* comb. nov. International Journal of Systematic Bacteriology 37 266-270.

Rocourt, J., Wehmeyer, U., Cossart, P. and Stackebrandt, E. 1987*b*. Proposal to retain *Listeria murrayi* and *Listeria grayi* in the genus *Listeria*. International Journal of Systematic Bacteriology 37 298-300.

Rogel, M., Brendle, J. J., Haapala, D. K. and Alexander, A. D. 1970. Nucleic acid similarities among *Pseudomonas pseudomallei, Pseudomonas multivorans, Actinobacillus mallei*. Journal of Bacteriology 101 827-835.

Ross, H. N. M. and Grant, W. D. 1985. Nucleic ascid studies on halophilic archaebacteria. Journal of General Microbiology 131 165-173.

Ruhland, G. and Fiedler, F. 1987. Occurence and biochemistry of lipotechoic acids in the genus *Listeria*. Systematic and Applied Microbiology 9 40-46.

Saltzberg, S., Levi, Z., Aboud, M. and Goldberger, A. 1977. Isolation and charecterisation of DNA-DNA and DNA-RNA hybrid molecules formed in solution. Biochemistry 16 25-29.

Schildkraut, C. L. and Lifson, S. 1965. Dependence of the melting temperature of DNA on salt concentration. Biopolymers 3 195-208.

Schildkraut, C. L., Marmur, J., Doty, P. 1961. The formation of hybrid DNA molecules and their use in studies of DNA homologies. Journal of Molecular Biology 3 595-617.

Schmeckpeper, B. J. and Smith, D. 1972. Use of formamide in nucleic acid reassociation. Biochemistry 11 1319-1326.

Seeliger, H. P. R. 1981. Apathogene Listerien : *Listeria innocua* sp. n. (Seeliger und Schoofs, 1977). Zentrablatt fur Bakteriologie und Hygiene. Abt. I Orig. A 249 487-493.

Seeliger, H. P. R. and Hohne, K. 1979. Serotyping of *Listeria monocytogenes* and related species, pp. 31-49. *In* : Bergan, T., Norris, J.R. (ed.s), Methods in microbiology, vol. 13. New York, Toronto, Sydney, San Francisco : Academic Press.

Seeliger, H. P. R. and Jones, D. 1986. Genus *Listeria* Pirie 1940, pp. 1235-1245. *In* : Bergey's Manual of Systematic Bacteriology Volume 2. P. H. A. Sneath, N. S. Mair, M.E. Sharpe (ed.s). The Williams and Wilkins Company, Baltimore.

Seidler, R. J. and Mandel, M. 1971. Quantitative aspects of deoxyribonucleic acid renaturation: Base composition, state of chromosome replication and polynucleotide homologies. Journal of Bacteriology 106 608-614.

Seldin, L. and Dubnau, D. 1985. DNA homology in *Bacillus* species. International Journal of Systematic Bacteriology 35 151.

Selin, Y. M., Harich, B. and Johnson, J. L. 1983. Preparation of labeled nucleic acids (nick translation and iodination) for DNA homology and rRNA hybridization experiments. Current Microbiology 8 127-132.

Shahamat, M., Seaman, A. and Woodbine, M. 1980*a*. Influence of NaCl, pH and temperature on the inhibitory activity of sodium nitrite on *Listeria monocytogenes*. *In*: Survival in Extremes of Environment, pp. 227-237. Ed. Gould, G.W. and Cavy, E.L., London Academic Press.

Shahamat, M., Seaman, A. and Woodbine, M. 1980*b*. Survival of *Listeria monocytogenes* in high salt concentrations. Zentrablatt fur Bakteriologie, Mikrobiologie und Hygiene Abt 1 Orig. A Medizinische, Mikrobiologie, Infektionskrankheiten und Parasitologie 246 506-511.

Skerman, V. B. D., McGowan, V. and Sneath, P. H. A. 1980. Approved lists of Bacterial Names. International Journal of Systematic Bacteriology 30 225-420.

Sneath, P. H. A. 1957. The application of computer to taxonomy. Journal of General Microbiology 17 201-226.

Sneath, P. H. A. 1972. Computer Taxonomy *In* : Methods in Microbiology 7A pp. 29-98, (J.R. Norris and D.W. Ribbons, ed.s) Academic Press, London and New York.

Sneath, P. H. A. 1978. Classification of micro-organisms. *In*: Essays in Microbiology. J.R. Norris and M.H. Richmond (ed.s) : John Wiley : Chichester.

Sneath, P. H. A. 1980. The probability that distinct clusters will be unrecognised in low dimensional ordinations. Classification Society Bulletin 4 22-43.

Sneath, P. H. A. 1983. Distortions of taxonomic structure from incomplete data on a restricted set of reference strains. Journal of General Microbiology 129 1045-1073.

Sneath, P. H. A and Sokal, R. R. 1973. Numerical taxonomy : The principles and practice of numerical classification. W. H. Freeman, San Fransisco.

Sneath, P. H. A. and Stevens M. 1985. A numeric taxonomical study of *Actinobacillus, Pasteurella* and *Yersinia*. Journal of General Microbiology 131 2711-2738.

Snedecor, G. W. 1956. Statistical methods applied to experiments in agriculture and biology. 5th ed. Ames, Iowa: Iowa State University Press.

Sokal, R. R. and Michener, C. D. 1958. A statistical method for evaluating systematic relationships. Kansas University Science Bulletin 38 1409-1438.

Sowers, K. R., Johnson, J. L. and Ferry, J. G. 1984. Phylogenetic relationships among the methylotrophic methane producing bacteria and emendation of the family Methanosarcinaceae. International Journal of Systematic Bacteriology 34 444-450.

Staley, T. E. and Colwell, R. R. 1973*a*. Application of molecular genetics and numerical taxonomy to the classification of bacteria. Annual Review of Ecology and Systematics 273-300.

Staley, T. E. and Colwell, R. R. 1973*b*. Deoxyribonucleic acid reassociation among members of the genus *Vibrio*. International Journal of Systematic Bacteriology 23 316-332.

Stephens, E. B., Robinson, I. M. and Barile, M. F. 1985. Nucleic acid relationships among the anaerobic mycoplasmas. Journal of General Microbiology 131 1223-1227.

Stuart, M. R. and Pease, P. E. 1972. A numerical study on the relationships of *Listeria* and *Erypsipelothrix*. Journal of General Microbiology 73 551-565.

Stuart, S. E. and Welshimer, H. J. 1973. Intrageneric relatedness of *Listeria* Pirie. International Journal of Systematic Bacteriology 23 8-14.

Stuart, S. E. and Welshimer, H. J. 1974. Taxonomic reexamination of *Listeria* Pirie and transfer of *Listeria grayi* and *Listeria murrayi* to a new genus *Murraya*. International Journal of Systematic Bacteriology 24 177-185.

Subriana, J. A. 1966. Kinetics of Renaturation of denatured DNA II: products of the reaction. Biopolymers 4 189.

Subirana, J. A. and Doty, P. 1966. Kinetics of renaturation of denatured DNA. I. Spectrophotometric results. Biopolymers 4 171-187.

Tanner, A. C. R., Listgarten, M. A., Ebersole, J. L. and Strzempko, M. N. 1986. *Bacteroides forsythus* sp. nov., a slow-growing, fusiform *Bacteroides* sp. from the human oral cavity. International Journal of Systematic Bacteriology 36 213-221.

Tindall, B. J., Ross, H. N. M. and Grant, W. D. 1984. *Natronobacterium* gen. nov. and *Natronococcus* gen. nov., two new genera of haloalkaliphilic archaebacteria. Systematic and Applied Microbiology 5 41-57.

Umbreit, W. W. and Burris, R. H. 1964. Manometric and chemical estimation of metabolites and enzyme systems, p.120. *In* : Manometric Techniques, 4th edition. Umbreit, W, W., Burris, R. H. and Stauffer, J. F. (ed.s). Burgess Publishing Company, Minneapolis.

Warnaar, S. O. and Cohen, J. A. 1966. A quantitative assay for DNA-DNA hybrids using membrane filters. Biochemical and Biophysical Research Communications 24 554-558.

Watabe, J., Benno, Y. and Mitsuoka, T. 1983. *Bifidobacterium gallinarum* sp. nov. : a new species isolated from the ceca of chickens. International Journal of Systematic Bacteriology 33 127-132.

Wayne, L. G., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O., Krichevsky, M. I., Moore, L. H., Moore, W. E. C., Murray, R. G. E., Stackebrandt, E., Starr, M. P. and Truper, H. G. 1987. Report of the ad hoc committee on reconciliation of approaches to systematics. International Journal of Systematic Bacteriology 37 463-464.

Wedlock, D. N. and Jarvis, B. D. W. 1986. DNA homologies between *Rhizobium fredii*, rhizobia that nodulate *Galega* sp., and other *Rhizobium* and *Bradyrhizobium* species. International Journal of Systematic Bacteriology 36 550-558.

Welshimer, H. J. and Meredith, A. L. 1971. *Listeria murrayi* sp. n. : A nitrate reducing mannitol fermenting *Listeria*. International Journal of Systematic Bacteriology 21 3-7.

Wetmur, J. G. 1976. Hybridization and renaturation kinetics of nucleic acids. Annual Review of Biophysics and Bioengineering 5 337-361.

Wetmur, J. G. and Davidson, N. 1968. Kinetics of renaturation of DNA. Journal of Molecular Biology 31 349-370.

Whiley, R. A., Russell, R. R. B., Hardie, J. M. and Beighton, D. 1988. *Streptococcus downei* sp. nov. for strains previously described as *Streptococcus mutans* serotype h. International Journal of Systematic Bacteriology 38 25-29.

Wilkinson, B. J. and Jones, D. 1975. Some serological studies on *Listeria* and possibly related bacteria. *In*: Problems of Listeriosis, pp. 251-261. M. Woodbine (ed.) Leicester University Press.

Wilkinson, B. J. and Jones, D. 1977. A numerical taxonomic survey of *Listeria* and related bacteria. Journal of General Microbiology 98 399-421.

Woese, C. R., Pribula, L. D., Fox, G. E. and Zablen, L. B. 1975. The nucleotide sequence of the 5s ribosomal RNA from a Photobacterium. Journal of Molecular Evolution 5 35-46.

Yasuda, P. H., Steigerwalt, A. G., Sulzer, K. R., Kaufmann, A. E., Rogers, F. and Brenner, D. J. 1987. DNA relatedness between serogroups and srovars in the family *Leptospiraceae* with proposals for seven *Leptospira* species. International Journal of Systematic Bacteriology 37 407-415.

Reprint from

SYSTEMATIC AND APPLIED MICROBIOLOGY

formerly Zentralblatt für Bakteriologie,
Mikrobiologie und Hygiene
I. Abt. Originale C

SEMPER BONIS ARTIBUS

Gustav Fischer Verlag
Stuttgart · New York

System. Appl. Microbiol. 10, 241–250 (1988)

Distortion of Taxonomic Structure from DNA Relationships due to Different Choice of Reference Strains

TRUDY HARTFORD and P. H. A. SNEATH

Department of Microbiology, Leicester University, Leicester LE1 7RH, U.K.

Received September 10, 1987

Summary

Most studies using DNA-DNA pairing employ a restricted set of reference strains which are compared with a larger series of strains. The resulting matrix of relationships is thus incomplete, and from these the underlying taxonmic structure must be reconstructed. An analysis was made on a published matrix of complete DNA relationships between 17 strains of *Bacillus circulans*. The data represent two distinct clusters and a number of outlying strains. It is shown that one form of principal component analysis is equivalent to principal coordinate analysis of derived distances. Three-dimensional diagrams, together with dendrograms from UPGMA cluster analysis, were compared when incomplete matrices were used, due to different choices of reference strains. Great distortion in apparent taxonomic structure can result unless reference strains are widely spaced and representative of the clusters that are present.

Key words: DNA relationships – Clustering – Principal components – Principal coordinates – Distortion of relationships – Choice of strains – Taxonomic structure – Incomplete matrices.

Introduction

A major problem with DNA-DNA pairing in systematics is the cost and effort of obtaining a complete matrix of values between all pairs of strains. It is usual to employ only a few strains as reference strains, and to compare all other strains to this restricted set. This strategy gives, in effect, a few strips of DNA-DNA values in an incomplete inter-strain matrix. The problem is to recover the underlying taxonomic structure, i.e. to recover a structure that is as similar as possible to the structure one would obtain from a complete matrix between all pairs of strains.

The problem has been discussed by *Cristofolini* (1980), *Grimont* and *Popoff* (1980), *Sneath* (1980; 1983) and *Mütters* et al. (1985). Three strategies have been commonly used: (a) to define one taxonomic group at a time; (b) to construct a network of closest neighbours; and (c) to derive a new complete matrix from the incomplete relationships and analyse this. The last method may lead to an explicit derived matrix of resemblances between all the strains (*Sneath*, 1980; 1983), or to principal component analysis in which a derived matrix is implicit (*Grimont* and *Popoff*, 1980).

Non-standard abbreviation: UPGMA, Unweighted Pair Group Method with Averages.

It has been shown (*Sneath*, 1983) that the choice of reference strains makes a large difference to the taxonomic structure that is recovered from derived matrices. In this paper we illustrate this with a complete DNA-DNA pairing matrix (*Nakamura* and *Swezey*, 1983) as part of a re-examination of the reliability of DNA techniques for taxonomy. Effects of the choice of reference strains on the other methods, (a) and (b), for recovering structure, and the influence of experimental errors, will be considered elsewhere.

Materials and Methods

The symbolism in *Sneath* (1983) has been used here. It is assumed that a complete matrix of DNA-DNA pairing values is available for t strains, and that any replicates have been averaged to give a single pairing value between a pair of strains j and k. It is further assumed that for those methods in which values of j versus k may be different from values of k versus j such reciprocal pairs have been averaged. This allows construction of a symmetrical $t \times t$ matrix (which is more convenient here than the usual lower triangular matrix). The values are then represented as distance between strains, d_{jk}, by appropriate transformation. Values in the principal diagonal are set to zero. Then c reference strains

are chosen and only the tc values representing the c columns are retained. The remaining values are treated as unknown.

The derived matrices were obtained either by principal component analysis of the DNA-DNA pairing values or by a single cycle of iteration of formula (1) in *Sneath* (1983). It is shown below that the two methods are algebraically identical when principal components are obtained in one particular way (*Gower*, 1966), and this way was employed here. The formula for derived distances is

$$d^*_{jk} = \left[\frac{1}{c} \sum_{r=1}^{r=c} (d_{rj} - d_{rk})^2 \right]^{\frac{1}{2}}$$

where r is a reference strain, but in this paper the summed squares were not divided by c as shown above, so as to retain the algebraic relations with principal components. The only effect, however, is to introduce a constant scaling factor of $1/\sqrt{c}$ that affects all relationships alike.

The taxonomic structure was then represented in two ways. The first is three-dimensional ordination from the first three principal axes of principal component analysis. The second is a dendrogram from UPGMA cluster analysis (*Sneath* and *Sokal*, 1973, p. 230) of the derived distances, d^*_{jk}. The ordination gives the most convenient visual representation of salient features. The dendrogram gives more reliable information, because it is based on the distances in the full space of c dimensions (not simply in the first three dimensions): it is, however, less easy to interpret by eye. In the present study, nevertheless, the discrepancies between the two representations were small, and are only mentioned where appropriate.

Taxonomic structure cannot be satisfactorily represented if the number of dimensions is reduced too much. A suitable measure of the effective dimensionality of the derived configurations is therefore needed. If points lie in a straight line the dimensionality is 1. This is true even if the points are embedded in a space of many dimensions. If they lie almost in a straight line, but show small displacements from it in numerous dimensions, the points cannot be represented exactly in one dimension. The effective dimensionality, n' however, is only a little greater than 1, and it might, for example, be 1.13.

The measure of n' is $1/\Sigma p_i^2$, *where* p_i is the proportion $\lambda_i/\Sigma\lambda_i$, where λ_i values are the non-negative eigenvalues from principal component or principal coordinate analysis (*Sneath*, 1983). A simpler formula is for n' is $(\Sigma\lambda_i)^2/\Sigma\lambda^2_i$. It is necessary to exclude negative eigenvalues because these represent "imaginary" or "non-euclidean" dimensions. Then n' cannot be more than the lesser of the number of characters n and $t-1$; it is maximal for a hyperspherical configuration.

When a model of lower dimensionality is prepared this removes some of the variation. The effective dimensionality is therefore calculated as m', where summation is over only the m non-negative eigenvalues of the m axes in the model.

Grimont and *Popoff* (1980) and *Rocourt* et al. (1982) have employed principal component analysis of DNA pairing values to obtain taxonomic structure from data on reference strains, whereas *Sneath* (1983) employed principal coordinate analysis (*Gower*, 1966) of euclidean distances, d^*_{jk}, between strains. The equivalence of principal coordinates with one form of principal components is illustrated in Table 1. Strains 1 and 3 are reference strains, with hypothetical DNA percent dissimilarity values as shown in Table 1a. It should be noted that in Table 1a reciprocal distances are not identical, and also that the triangle inequality does not hold for all cases. Thus the sum of distance between 1 and 2 and 1 and 3 is either 23 or 28, depending on whether 11% or 17% is chosen to represent the distance from 1 to 3. The distance from 2 to 3 is far greater than either 23 or 28 at 41, so

Table 1. Hypothetical DNA-DNA dissimilarities to illustrate principal coordinate and principal component analysis (see text)

		Strains	Strains 1	2	3	4
Original data	(a)	1	0		17	
		2	12		41	
		3	11		0	
		4	18		38	
		Mean	10.25		24	

		Strains	Strains 1	2	3	4
Distance between strains	(b)	1	0			
		2	26.8328	0		
		3	20.2485	41.0122	0	
		4	27.6586	6.7082	38.6394	0

		Strains	Axes 1	2	3	4
Principal coordinates	(c)	1	8.8419	8.7111	0	0
		2	−17.0190	1.5347	0	0
		3	23.4069	5.3552	0	0
		4	−15.2298	−4.9105	0	0
		Sum	0	0	0	0
		Sum of squares	1147.6583	131.0912	0	0

		Old variates	New variates 1	2
Principal components from sums of squares and products		1	.1925	.9813
		2	.9813	−.1925
	(d)	λ	1147.66	131.09

		Strains	Axes 1	2
Coordinates from components in (d) scaled to eigenvalues		1	−8.8419	−8.7111
		2	17.0190	−1.5547
		3	−23.4069	5.3552
	(e)	4	15.2298	4.9105
		Sum	0	0
		Sum of squares	1147.6583	131.0912

		Strains	Axes 1	2
Coordinates from (d) scaled to unity on each axis		1	−.2610	−.7608
		2	.5024	−.1358
	(f)	3	−.6909	.4677
		4	.4496	.4289
		Sum	0	0
		Sum of squares	1.0	1.0

		Old variates	New variates 1	2
Principal components from correlations		1	.7071	−.7071
		2	.7071	.7071
	(g)	λ	1.4436	.5564

		Strains	Axes 1	2
Coordinates from (g) scaled to unity on each axis		1	−.4230	.1086
		2	.4598	.5097
	(h)	3	−.5702	−.9272
		4	.5334	.2089
		Sum	0	0
		Sum of squares	1.0	1.0

the points 1, 2 and 3 cannot be represented as a triangle in euclidean space. However, such features are not uncommon in DNA data, and the analyses show that they can be accommdated by principal axis methods.

Distances between strains are shown in Table 1b. For example $d^*_{1,2} = [(0-12)^2 + (17+41)^2]^{\frac{1}{2}} = 26.8328$. On analysing Table 1b by principal coordinates one obtains a new distance matrix scaled in the manner given by *Gower* (1966), and this matrix has four eigenvalues; $\lambda_1 = 1147.66$, $\lambda_2 = 131.09$, and the other two are zero. On scaling the eigenvectors of this new matrix so that the sum of squares of each column equals the corresponding eigenvalue, one obtaines the coordinates in Table 1c. These coordinates represent a rigid rotation about the centroid of points representing the strains. Note, however, that the positive and negative ends of the axes are arbitary, because this information is lost when calculating interstrain distances. Thus the configuration may appear reflected about the centroid when compared with that from principal components (Table 1e).

If one performs principal component analysis on Table 1a using sums of squares and crossproducts, the same eigenvalues are obtained. Scaling the eigenvectors so that the sum of squares of each column is unity gives the principal component matrix Table 1d. This represents a rotation matrix such that if one centres the values of Table 1a by subtracting column means, and then matrix multiplies by Table 1d one obtains the coordinates in Table 1e. For example, strain 1 on axis 1 has the coordinate (0–10.25) × .1925 + (17–24) × .9813 = 8.8419, and on axis 2 (0–10.25) × .9813 + (17–24) × –.1925 = –.8711. It can be seen that these coordinates are the same (within machine accuracy) as those in Table 1c, except for change of sign as mentioned above. It was this form of principal components that was used in this paper.

However, if other forms of principal components are used the resulting configurations can be very different (*Hope*, 1968). One common practice is to scale each principal axis so that its sum of square is unity. If this is done, the coordinates become those in Table 1f: the resulting plots or models show equal variance on each principal axis, and, for example, a mainly linear configuration can be turned into a mainly circular or spherical one.

Another variant of principal components employs correlations in place of sums of squares and cross products. Correlations do not yield a rigid rotation, because the relations are distorted before rotation takes place, so that the final coordinates bear no simple relation to the starting configuration. Table 1g shows the principal components from correlations after scaling so that sums of squares are unity. The resulting coordinates are shown in Table 1h, and are obviously very different from Tables 1e and 1f.

It should be emphasized that only Table 1c and 1e represent the data of Table 1a in the manner that is normally desired for taxonomy.

The matrix of DNA-DNA values was that between 17 strains of *Bacillus circulans (Nakamura* and *Swezey,* 1983). Their percent homologies were converted into "DNA distances"by subtracting from 100 and are shown as a square matrix (Table 2). These authors averaged three replicates to obtain each value, but since they used the thermal reassociation method there is no difference between reciprocal pairs.

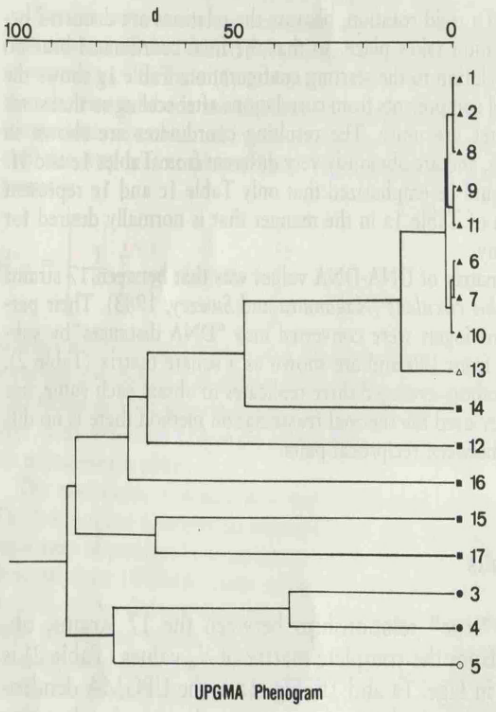
Results

The "true" relationships between the 17 strains, obtained from the complete matrix of d_{jk} values (Table 2) is shown in Figs. 1a and 1b. Fig. 1a is the UPGMA dendrogram, and is the best representation, because it takes into account the distances in the full space of $t - 1 = 16$ euclidean dimensions. Single Linkage cluster analysis gave almost the same results. Fig. 1b, the three-dimensional model from principal coordinates, only represents the first three of the 16 dimensions, and therefore neglects some of the information, but it does allow an easier appreciation of salient relations than Fig. 1a. In this instance it gives broadly the same information.

Strains 1, 2, 6, 7, 8, 9, 10, 11 and 13 form a major cluster. The first eight are a tight, cluster, whereas strain 13 is a satellite of the cluster, lying some distance away. Strains 3 and 4 form a minor, looser, cluster, and strain 5 is a satellite of this. Strains 12, 14, 15, 16 and 17 are

Table 2. Percent DNA-DNA dissimilarities from *Nakamura* and *Swezey* (1983) for 17 strains of *Bacillus circulans*

Strain serial no. and NRRL-NRS no.		Strain 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	313	0	0	94	90	91	0	5	0	2	3	0	64	10	42	89	65	94
2	358	0	0	95	73	82	2	1	0	0	0	3	71	15	49	88	67	90
3	385	94	95	0	37	79	90	85	92	91	83	77	93	95	74	91	70	96
4	387	90	73	37	0	74	96	88	89	93	90	92	84	96	91	94	80	92
5	397	91	82	79	74	0	76	86	75	90	75	82	84	94	79	90	92	86
6	726	0	2	90	96	76	0	0	0	1	1	1	69	10	37	75	65	97
7	727	5	1	85	88	86	0	0	2	0	0	2	64	13	44	72	68	95
8	728	0	0	92	89	75	0	2	0	3	4	0	73	14	33	69	70	89
9	729	2	0	91	93	90	1	0	3	0	0	0	74	10	49	79	74	92
0	746	3	0	83	90	75	1	0	4	0	0	0	68	9	53	81	92	87
1	765	0	3	77	92	82	1	2	0	0	0	0	68	12	52	82	90	90
2	826	64	71	93	84	84	69	64	73	74	68	68	0	74	65	80	81	86
3	831	10	15	95	96	94	10	13	14	10	9	12	74	0	62	85	70	93
4	1108	42	49	74	91	79	37	44	33	49	53	52	65	62	0	77	67	84
5	1670	89	88	91	94	90	75	72	69	79	81	82	80	85	77	0	77	67
6	1341	65	67	70	80	92	65	68	70	74	92	90	81	70	67	77	0	90
7	1353	94	90	96	92	86	97	95	89	92	87	90	86	93	84	67	90	0



outlying singletons. Of these, 17 is the most outlying. The percentage of variation accounted for by the first three dimensions is 77.9 (Table 3). This is rather high for taxonomic structures; values of about 50% are more usual (*Bridge* and *Sneath*, 1983; *Sneath*, 1983; *Sneath* and *Stevens*, 1985), though these refer to complex taxonomies from phenotypic analysis, not from DNA data. A few negative eigenvalues were present (because the distances in Table 2 are not completely euclidean), but they only totalled 12.3%. The effective dimensionality, n', is only 4.5, a good deal less than the nominal dimensionality of 16, and this phenomenon has been noted before (*Sneath*, 1983). Reduction to three dimensions (Fig. 1b) reduces the effective dimensionality to $m' = 2.4$ (Table 3).

It is against the configurations of Figs. 1a, b that the others are to be judged. It may be noted that strains 1 and 9 both derive from ATCC 4516, and the differences in values for these two in Table 2 are probably due to the experimental error of estimating DNA pairing. We are less confident that experimental error completely accounts for the differences between values for strains 6 and 13, which both derive from Ford 26, because our unpublished analyses of literature results suggest that the differences between 6 and 13 are too large for this.

The results from principal coordinate analysis of d^* coefficients using all 17 strains as reference strains is

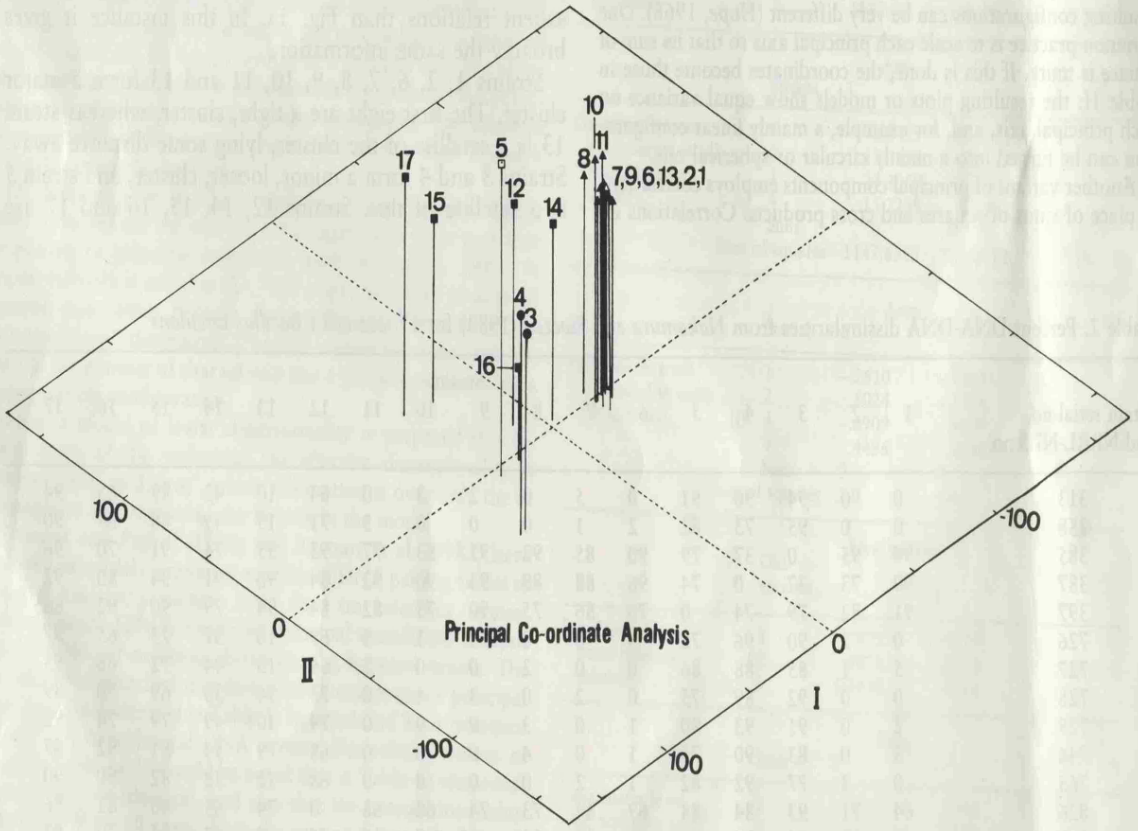


Fig. 1. (a) Dendrogram from UPGMA clustering of percent DNA-DNA dissimilarities of table 2 treated as distances, *d*.
(b) Three-dimensional representation of relationships from principal coordinates analysis of percent DNA-DNA dissimilarities of table 2. First two principal axes are horizontal, third axis is vertical (baseplate at –90).
Symbols: solid triangles members of main cluster; open triangle, satellite of main cluster; solid circles, members of minor cluster; open circle, satellite of minor cluster; solid squares, outlying singletons.

Table 3. The eigenvalues of the first three axes of the figures, together with the effective dimensionality of the configuration, n', and that for the three-dimensional representation, m'

Fig. No.	c	λ_I	λ_{II}	λ_{III}	Percent variation in first three axes	n'	m'
1	NA	16,688	7,362	5,392	77.9	4.51	2.40
2	17	239,490	19,234	10,847	91.3	1.54	1.26
3	3	18,203	6,972	1,997	100	1.92	1.92
4	2	27,856	6,713	0	100	1.46	1.46
5	2	28,101	6,010	0	100	1.41	1.41
6	2	51,171	380	0	100	1.01	1.01
7	3	32,567	6,440	3,666	100	1.63	1.63
8	8	36,354	15,267	9,985	75.3	3.80	2.29

NA not applicable

shown in Fig. 2. The taxonomic structure is essentially correct, and the distortion is small. Within the major cluster strain 8 is now closer and strain 13 relatively a little less close. The effective dimensionality has been reduced, and consequently there is more variation in the first three axes (91.3%, Table 3). The arbitrary reflection on axes I and II can be seen by comparing Fig. 2 with Fig 1b (see Methods).

Fig. 3. shows the results from three reference strains, the members of the minor clusters, 3 and 4, and its satellite, 5. There is bizarre distortion. The minor cluster has greatly expanded, and all the remaining strains, including singletons, have been compressed into an apparently tight but false group near the centroid. This behaviour is particularly significant, because a choice such as this could easily occur if the first strains examined happened to be from a loose cluster.

When strains 3 and 4 were employed, without strain 5, the results were similar: the two strains of the minor cluster became widely separated and all other strains (including strain 5) were in one compact group.

Fig. 4. shows the results from two reference strains, one from the major cluster, 1, and a singleton, 15. The structure is remarkably good: both clusters are easily recognized and the other strains are placed appropriately. Because $c = 2$ all the variation is in the first two axes, and the points are all at a constant height above the baseplate.

There is notable compression of the loose minor cluster together with its satellite strain 5. The reference strain 15 is now very peripheral. The singletons 12 and 16 are close,

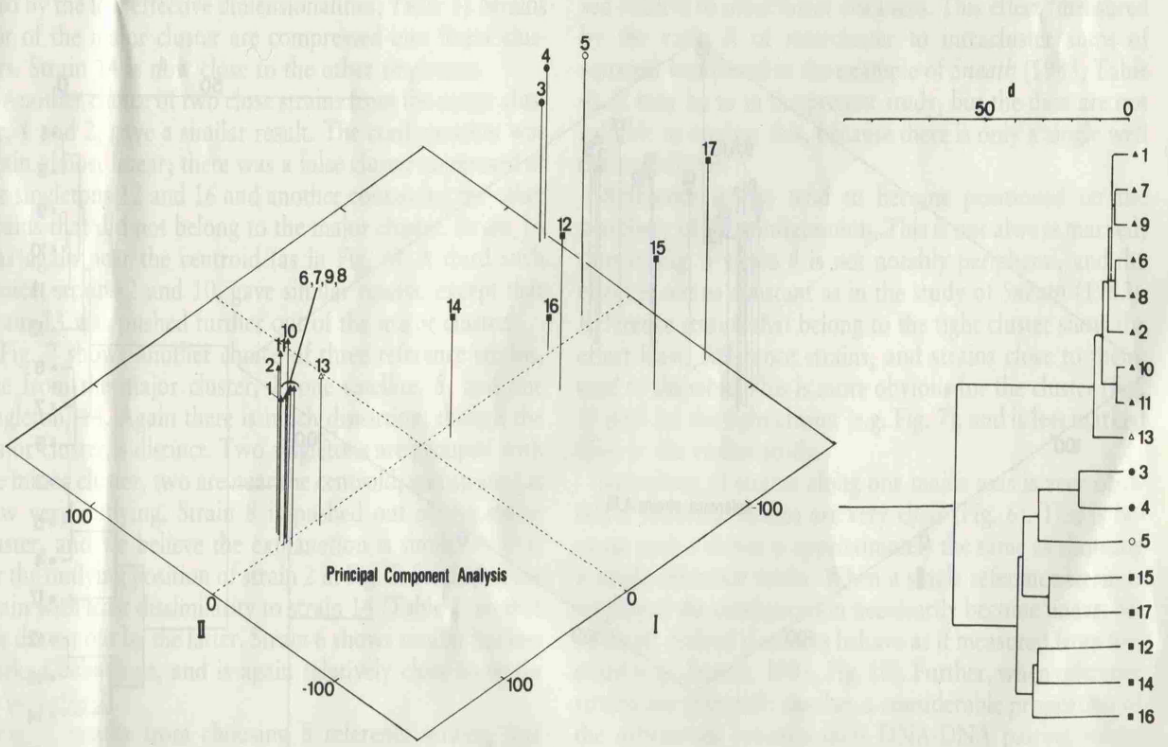


Fig. 2. Princpal component analysis of percent DNA-DNA dissimilarities of table 2 and UPGMA dendrogram of derived distances when all 17 strains are reference strains (i.e. analyses of complete matrix of table 2). Third axis vertical, with baseplate at –90. The distance in the dendrogram are derived distances d^* (see text). Other symbols as in Fig. 1.

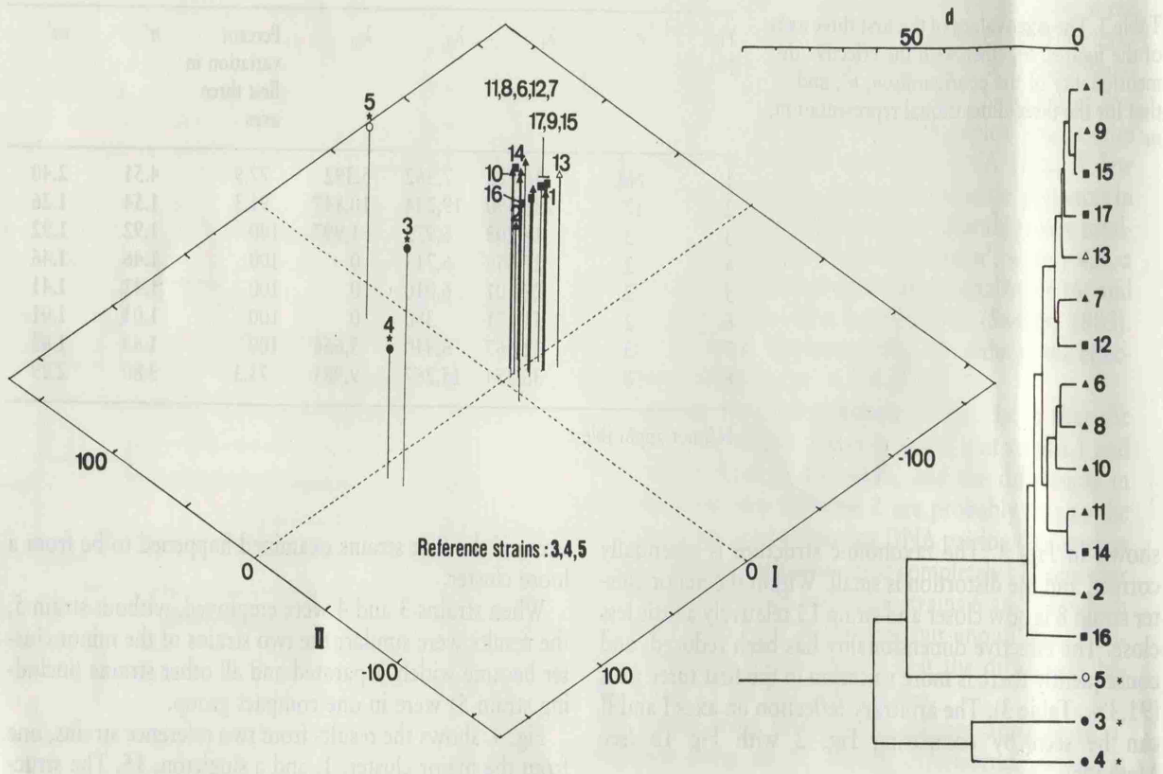


Fig. 3. Principal component analysis of table 2 employing strains 3, 4 and 5 as reference strains (marked with asterisks), and corresponding UPGMA dendrogram from derived distances. Other symbols and conventions as in Figs. 1 and 2.

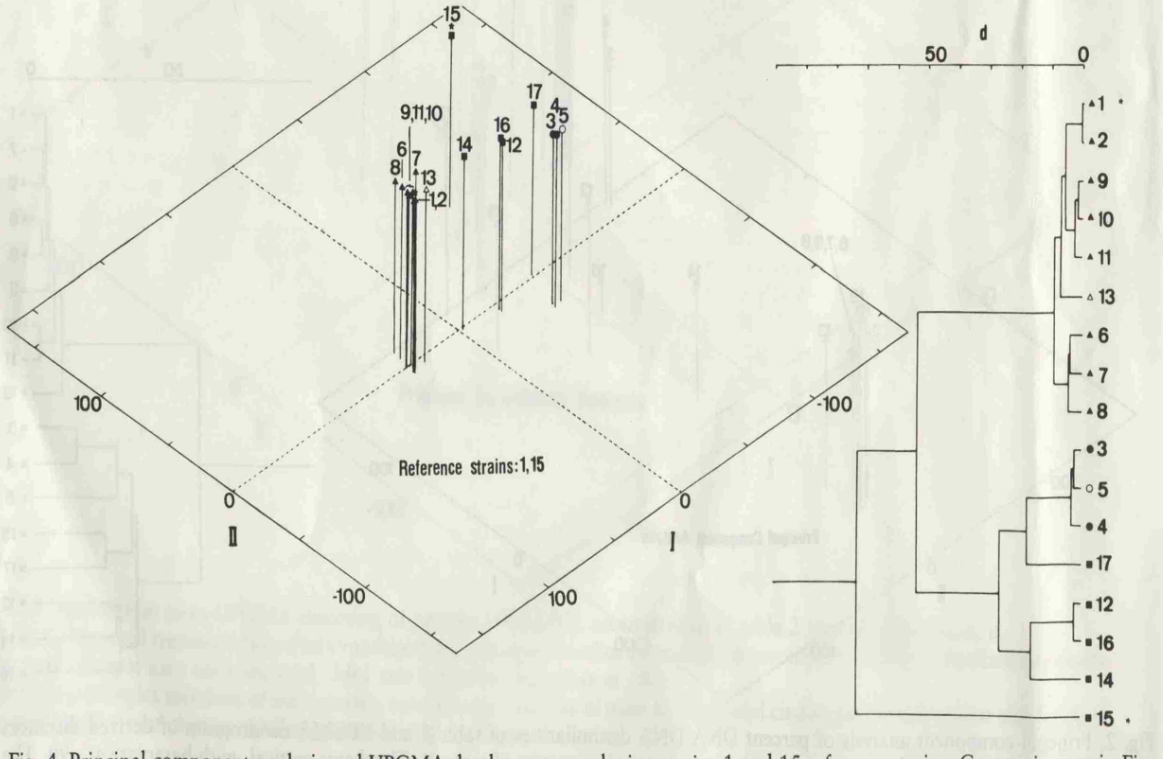


Fig. 4. Principal component analysis and UPGMA dendrogram employing strains 1 and 15 reference strains. Conventions as in Figs. 1–3.

giving the false impression that they form the nucleus of a cluster.

Another similar choice, strain 10 from the major cluster and the singleton 17, resulted again in compression of the other singletons and the minor cluster into one group; in this instance one might easily be misled into thinking that those singletons belonged to the minor cluster.

Fig. 5 shows the results from a different pair of reference strains, one (strain 8) from the major cluster and the other (strain 4) from the minor cluster. There is obvious distortion. Only the major cluster is well defined; the minor cluster is dispersed and allied with the pulled-in singletons and the satellites in a loose false cluster; this could be very misleading. The tendency for reference strains to assume peripheral positions (*Sneath,* 1983) is well shown by strain 4. Strain 2 is now relatively peripheral in the main cluster. Further, the strains of the minor cluster 3 and 4, are widely separated relative to the other strains (cf. Fig. 1).

It is not entirely clear why strain 2 has become so peripheral to its own cluster. We believe it is probably due to non-euclidean properties of certain relationships. Strains 2 and 8 appear to be identical when compared directly ($d_{2,8} = 0$, Table 2), yet other values involving them differ considerably. Thus $d_{2,4}$ is 73% and $d_{8,4}$ is 89%, which implies that strain 2 is closer to the reference strain 4 than it is to strain 8. Strain 2, therefore, tends to be moved out by its comparative closeness to strain 4 when only similarities involving strains 4 and 8 are available.

Fig. 6 shows results from choosing two strains, 1 and 13, from the major cluster. These are so close that they are nearly equivalent to one reference strain: they represent, one might say, almost a view from a single point. Consequently the structure is almost one-dimensional (shown also by the low effective dimensionalities; Table 3). Strains not of the major cluster are compressed into linear clusters. Strain 14 is now close to the other singletons.

Another choice of two close strains from the major cluster, 1 and 2, gave a similar result. The configuration was again almost linear; there was a false cluster composed of the singletons 12 and 16 and another containing the other strains that did not belong to the major cluster. Strain 14 was again near the centroid (as in Fig. 6). A third such choice, strains 2 and 10, gave similar results, except that strain 13 was pushed further out of the major cluster.

Fig. 7 shows another choice of three reference strains, one from the major cluster, 1, one satellite, 5, and one singleton, 14. Again there is much distortion, though the major cluster is distinct. Two singletons are grouped with the minor cluster, two are near the centroid, and strain 5 is now very outlying. Strain 8 is pushed out of the major cluster, and we believe the explanation is similar to that for the outlying position of strain 2 in Fig. 5. Strain 8 is the strain with least dissimilarity to strain 14 (Table 2) so that it is drawn out by the latter. Strain 6 shows similar but less marked behaviour, and is again relatively close to strain 14 in Table 2.

Fig. 8 results from choosing 8 reference strains, one from each cluster, one satellite and the five singletons. This might represent a well-balanced choice; the range of variation is spanned, but near-duplicates are omitted. The structure is good, though reference strains tend to be peripheral (e.g. strains 3, 5, 1, 15, 17). When this was repeated with omission of strain 17 the structure was little changed (though strain 17 became more central, and strain 12 more peripheral).

Discussion

Complete matrices of DNA-DNA pairing values are not common, and the one we have analysed from *Nakamura* and *Swezey* (1983) is the largest we have found. Not all the primary data, however, has been published by them, because the values for each of the three replicates are not given separately, and the extent of test reproducibility cannot therefore be determined. It is highly desirable that in such studies the full details should be published to allow this to be examined. We hope to present elsewhere an analysis of the experimental reproducibility of DNA methods.

Most of the types of distortion observed in an earlier study on phenotypic similarities (*Sneath,* 1983) are seen here. There are no obvious effects peculiar to DNA-DNA data, though more experience is clearly needed. Choice of strains is a far more important factor than choice of cluster method: we found only minor differences from UPGMA when Single Link or Complete Link clustering was used.

The most obvious effect is the tendency of outliers to be drawn inwards (even when all strains are used as reference strains, although this effect was not prominent in the present study). This is more obvious for strains in the loose areas than those in the tight clusters. When all strains are employed as reference strains the clusters may be compressed relative to intercluster distances. This effect (measured by the ratio R of intercluster to intracluster sums of squares) was found in the example of *Sneath* (1983, Table 4). It may be so in the present study, but the data are not suitable to analyse this, because there is only a single well defined cluster.

Reference strains tend to become positioned on the periphery of the configuration. This is not always marked; thus in Fig. 5 strain 8 is not notably peripheral, and the effect is not as constant as in the study of *Sneath* (1983). Reference strains that belong to the tight cluster show the effect least. Reference strains, and strains close to them, tend to disperse. This is more obvious for the cluster (Fig. 3) than for the tight cluster (e.g. Fig. 7), and is less marked than in the earlier study.

Swivelling of strains along one major axis is very obvious if reference strains are very close (Fig. 6). This is because such a choice is approximately the same as choosing a single reference strain. When a single reference strain is employed the configuration necessarily become linear, because all derived distances behave as if measured from one point (e.g., *Sneath,* 1983, Fig. 10). Further, when reference strains are extremely similar, a considerable proportion of the differences between their DNA-DNA pairing values may be due to chance effects of experimental error. Much of the detail in the derived configuration may then depend on these chance effects. Such choices represent, as it were,

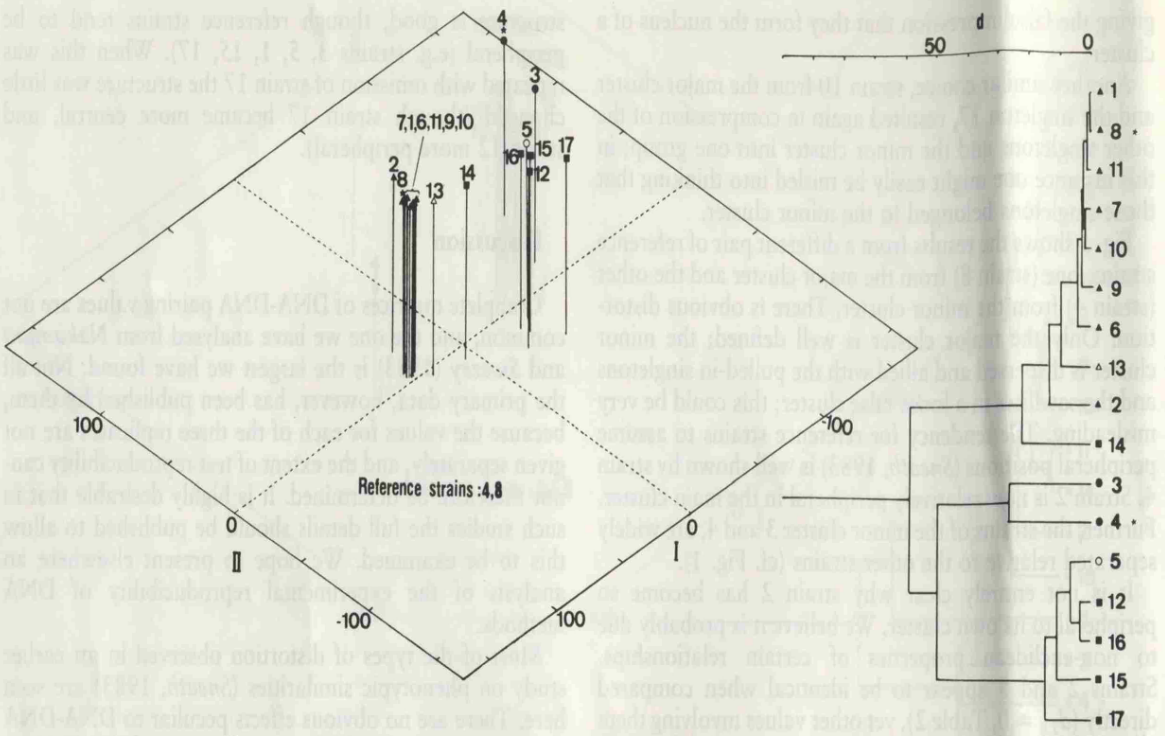


Fig. 5. Principal component analysis and UPGMA dendrogram employing strains 4 and 8 as reference strains. Conventions as in Figs. 1–3.

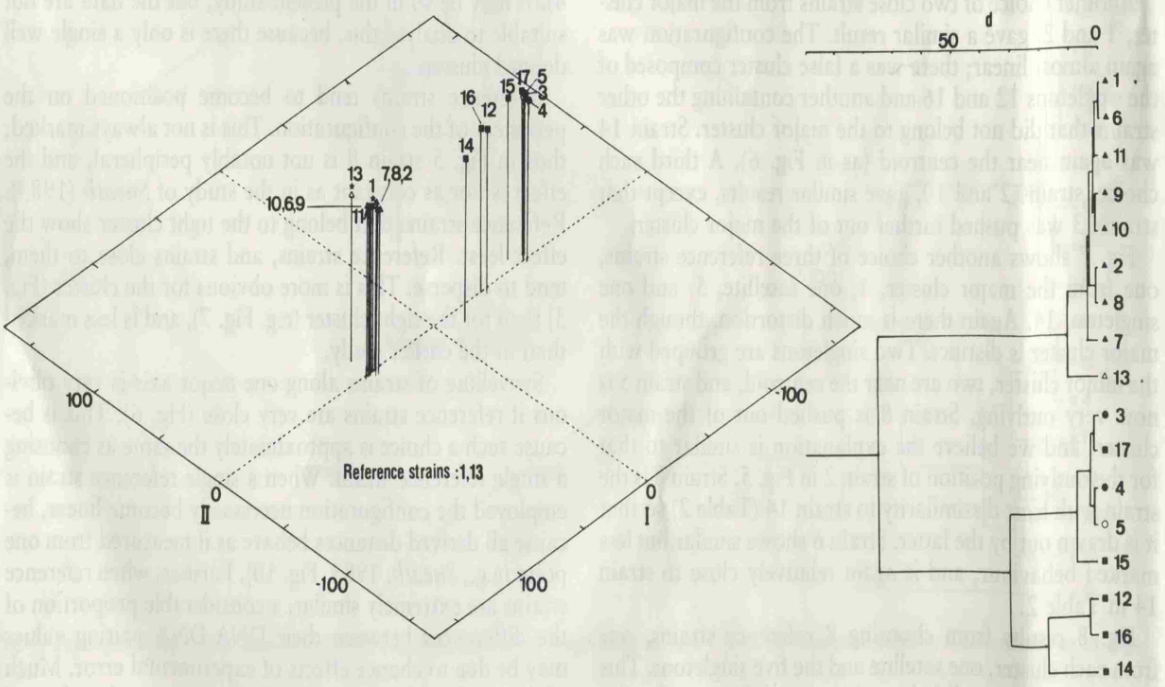


Fig. 6. Principal component analysis and UPGMA dendrogram employing strains 1 and 13 as reference strains. Conventions as in Figs. 1–3.

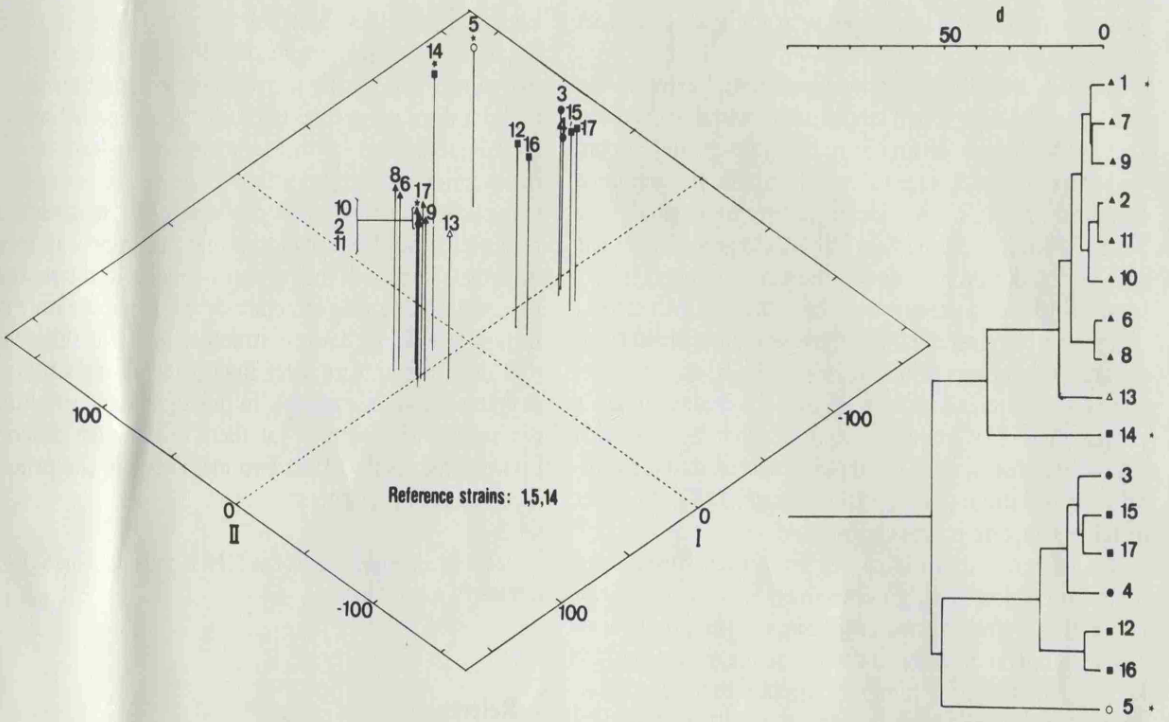


Fig. 7. Principoal component analysis and UPGMA dendrogram employing strains 1, 5 and 14 as reference strains. Conventions as in Figs. 1–3.

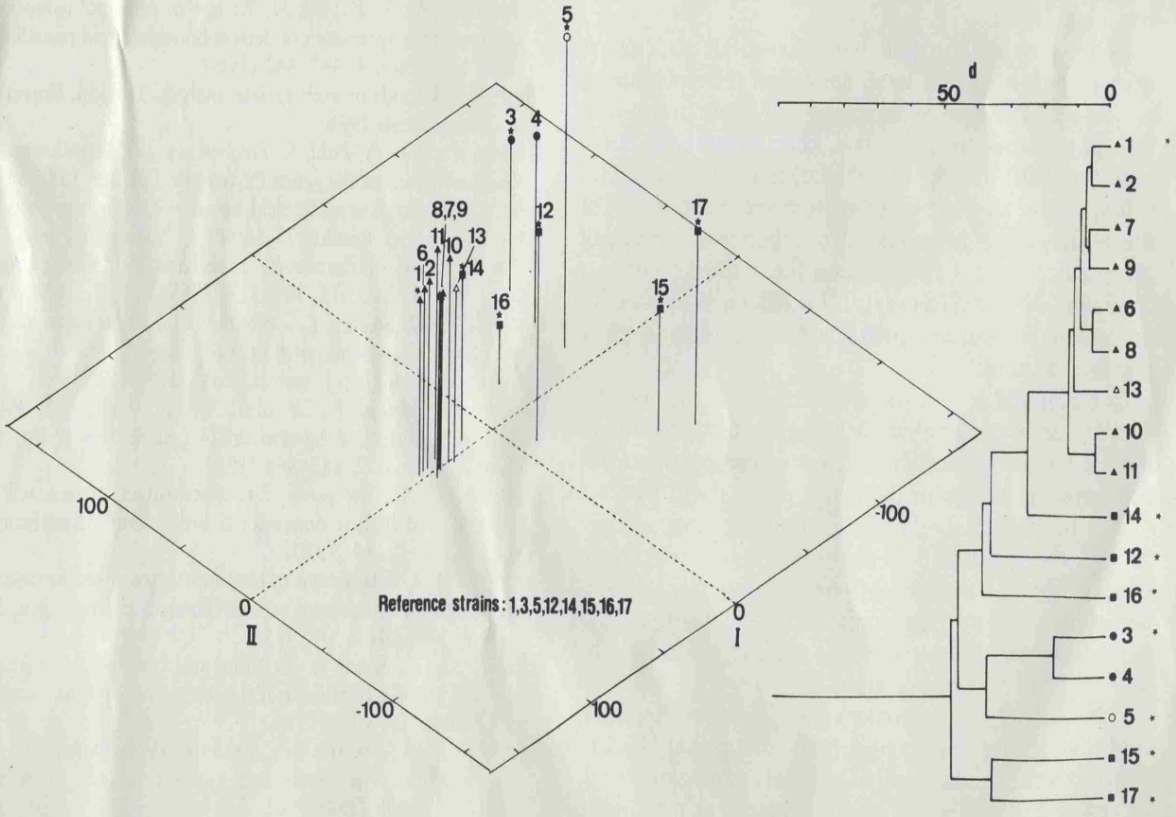


Fig. 8. Principal component analysis and UPGMA dendrogram employing strains 1, 3, 5, 12, 14, 15, 16 and 17 as reference strains. Conventions as in Figs. 1–3.

views from almost a single point in the space, or from points of uncertain position.

There is undoubtedly some experimental error in the DNA data, because if two strains are identical (e.g. strains 1 and 2 with zero dissimilarity in Table 2) they should have the same percentage values when both are compared to a third strain. This is evidently not so; for example 1 to 5 is 91% but 2 to 5 is 82% in Table 2; in general columns such as 1 and 2 do not contain the same percentages. We hope elsewhere to present ways of analysing such errors.

A single reference strain is thus very unsatisfactory. If the analysis employs principal axis methods, as in the present study, this necessarily aligns all strains along a straight line. If it employs minimal spanning trees this necessarily creates a fan of all points spread round a central reference strain (see Fig. 10 in *Sneath,* 1983). In either instance structure is grossly distorted.

If a reference strain is chosen from each cluster, and others are well-spaced, good recovery of structure is obtained (Fig. 8 and *Sneath,* 1983, Fig. 6). The problem, of course, is how to make such choices before the clusters are known. The risk of obtaining spurious structure from an unsuitable choice of reference strains is well illustrated by Figs. 3, 5, 6 and 7.

Additional points from the present study are as follows. If a reference strains is a singleton, it may compress other strains into a false cluster (Fig. 4). There is a tendency for a reference strain in a tight cluster to push one or two strains out, rather than simply to expand the cluster (e.g. Fig. 5). Loose clusters with few members are particularly easily distorted (this finding requires confirmation with other studies). Omission of one reference strain from a good set has minor effects.

Information on the underlying taxonomic structure is necessarily lost when a small number of reference strains are chosen. The new configuration has fewer dimensions than the starting configuration, and it has been noted (*Sneath,* 1980) that for c reference strains the configuration will have an effective dimensionality of about $c - \frac{1}{2}$ at the most, because the points will be reflected to one side of a hyperplane of $c - 1$ dimensions (i.e. into one half of a configuration of c dimensions. This reduction of effective dimensionality is accentuated if reference strains are close together, as noted earlier.

One cannot therefore judge the amount of the underlying varition by looking simply at the first few eigenvalues if these are derived from only a few reference strains. All the variation will be in the first c eigenvalues. The fact that, for example, the first two eigenvalues may recover 99% of the variation does not ensure that a two-dimensional diagram, based on say 3 reference strains, will contain almost all the taxonomic structure. Clusters and points that are well separated in the full space may be overlapped in a scatter-diagram.

Choice of only two reference strains, even if well spaced, constrains the resulting configuration to a plane. This can be seen by the constant height of points in diagrams such as Figs. 4 and 5. The risk of overlapping for a plausible statistical model has been shown to be related to the chi-square distribution (*Sneath,* 1980; 1983). Reduction to two dimensions greatly increases the risk in the chisquare table. If there are several fairly close clusters the danger is considerable, even for three dimensions. Indeed two- and three-dimensional representations are unsafe for deducing taxonomic structure unless the identity of the strains and taxa are known beforehand, so that the points can receive different symbols. If they are so symbolized, the representations are not giving structure *de novo,* but are only confirming previously known structure. It is for this reason that dendrograms are safer than principle axes plots for defining taxonomic groups. In this sample the structure is not sufficiently complex for there to be much difference between the results of the two methods, but the principle involved is important.

Acknowledgements. One of us (TH) is grateful to the SERC for a research scholarship.

References

Bridge, P. D., Sneath, P. H. A.: Numerical taxonomy of *Streptococcus.* J. gen. Microbiol. *129,* 565–597 (1983)

Christofolini, G.: Interpretation and analysis of serological data. In: Chemosystematics: principles and practice (*F. A. Bisby, J. G. Vaughn, C. A. Wright,* eds.). London and New York, Academic Press 1980

Gower, J. C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika *53,* 325–338 (1966)

Grimont, P. A. D., Popoff, M. Y.: Use of principal component analysis in interpretation of deoxyribonucleic acid relatedness. Curr. Microbiol. *4,* 337–342 (1980)

Hope, K.: Methods of multivariate analysis. London, University of London Press 1968

Mütters, R., Ihm, P., Pohl, S., Frederiksen, W., Mannheim, W.: Reclassification of the genus *Pasteurella* Trevisan 1887 on the basis of deoxyribonucleic acid homology, with proposals for the new species *Pasteurella dagmatis, Pasteurella canis, Pasteurella stomatis, Pasteurella anatis* and *Pasteurella langaa.* Int. J. system. Bact. *35,* 309–322 (1985)

Nakamra, L. K., Swezey, J.: Taxonomy of *Bacillus circulans* Jordan 1890: base composition and reassociation of deoxyribonucleic acid. Int. J. system. Bact. *33,* 46–52 (1983)

Rocourt, J., Grimont, F., Grimont, P. A. D., Seeliger, H. P. R.: DNA relatedness among serovars of *Listeria monocytogenes.* Curr. Microbiol. *7,* 383–388 (1982)

Sneath, P. H. A.: The probability that distinct clusters will be unrecognized in low dimensional ordinations. Classification Soc. Bull. *4,* 22–43 (1980)

Sneath, P. H. A.: Distortions of taxonomic structure from incomplete data on a restricted set of reference strains. J. gen. Microbiol. *129,* 1045–1073 (1983)

Sneath, P. H. A., Sokal, R. R.: Numerical taxonomy: the principles and practice of numerical classification. San Francisco, W. H. Freeman 1973

Sneath, P. H. A., Stevens, M.: A numerical taxonomy study of *Actinobacillus, Pasteurella* and *Yersinia.* J. gen. Microbiol. *131,* 2711–2738 (1985)

P. H. A. Sneath, Dept. of Microbiology, Leicester University, Leicester LE1 7RH, U.K.

Journal of Applied Bacteriology 1990, **68**, 527–542

A Review

Experimental error in DNA–DNA pairing: a survey of the literature

T. HARTFORD & P.H.A. SNEATH* *Department of Microbiology, Leicester University, Leicester LE1 7RH, UK*

Accepted 18 November 1989

1. Introduction

It is important with any laboratory method to know its experimental accuracy, both in order to interpret existing data and as a first step to studying the factors that determine accuracy. No substantial survey of the literature has been published for DNA–DNA pairing, and we present a first such survey. Results of experimental work will be presented elsewhere.

The amount of error associated with nucleic acid pairing studies is rarely taken into account; even when reproducibility is recorded, the effects of the error on the conclusions drawn from work are often not discussed. Yet DNA–DNA pairing has, in principle, great potential for determining accurate relationships because of its low sampling error. The precision with which relationships can be estimated depends (in part) on the number of nucleotide differences observed. If one observes 10 differences between two short nucleotide sequences of two organisms one does not expect exactly 10 from another pair of short sequences. Also, obviously one cannot in this case estimate relationship accurately to one part in a hundred. If DNA–DNA pairing reflects differences between whole genomes (representing many thousands of nucleotide differences) as is believed, and reflects them faithfully, then in principle it could yield accuracy of one part in a hundred or better. This potential is greatly diminished by experimental error.

Few papers state both the number of replications and standard deviation associated with each result. The study of Potts & Berry (1983) is one of the few which lists these. Also, conclusions are

* Corresponding author.

frequently drawn from incomplete sets of data, and this leads to difficulties in obtaining with confidence the underlying taxonomic structure, particularly when there are only a few reference strains (Hartford & Sneath 1988).

Pairing values of greater than 100% (that is, greater than the hybridization of homologous DNA) are often published; values as high as 115% have been published, especially in radiolabelling methods. Theoretically these should not exist and they point to some form of inconsistency. Similarly, low values such as 0% cannot be taken at face value as percent sequence similarity, because four alternative nucleotides will give about 25% of random matching. Indeed, little is known of the details of DNA reassociation and the many factors that influence it. Some light on these may be shed by considering the metric properties of DNA–DNA dissimilarities, by examining triangle inequalities as shown later.

Comparison between values from different methods of nucleic acid pairing has been discussed by several authors (Grimont *et al.* 1980; Huss *et al.* 1983; Bouvet & Grimont 1986). In this paper we consider the evidence from a selection of published work which we believe to be representative of the field, and examine experimental error within each method, together with some comments on differences between methods.

2. Methods

2.1 PUBLISHED STANDARD DEVIATIONS

The average error corrected for degrees of freedom, as a standard deviation s_E, was obtained as follows: individual standard deviations, s_i, and number of replicates, n on which s_i was based, were tabulated; then

$$s_E = \sqrt{[\sum s_i^2(n_i - 1)/\sum(n_i - 1)]}$$

In a few instances it was evident from internal evidence that the published s_i values had not been corrected for degrees of freedom, so we then recalculated them.

2.2 RECIPROCAL PAIRS

For methods involving radiolabelling techniques (membrane filter or S1 nuclease techniques) a square matrix is often published, where, for strains a and b, the corresponding values i.e. (a versus b and b versus a) are not duplicates, but where first strain a was the labelled nucleic acid, and then strain b. Theoretically the relation of a:b should equal that for b:a but this is not always so. Error was calculated as the standard deviation between the reciprocal pair. In spectrophotometric techniques this error does not arise as the experiment for measuring pairing between a:b and b:a is identical. The standard deviation s for such a pair of reciprocal values, $X_{a:b}$ and $X_{b:a}$, has 1 degree of freedom and

$$s = \sqrt{[[(X_{a:b} - \text{mean})^2 + (X_{b:a} - \text{mean})^2]/1]}$$

which reduces to

$$\sqrt{[(X_{a:b} - X_{b:a})^2/2]}$$

The average s_E for m such pairs is $\sqrt{(\sum s^2/m)}$.

2.3 USE OF TRIANGLES

This method involves looking at all possible combinations of each of three strains in turn. For convenience the X values are converted to dissimilarities, e.g. 90% DNA pairing corresponds to 10% dissimilarity or a 'distance' of 10. Thus any three strains can be represented as apices of a triangle with the lengths of the sides corresponding to the distances between strains. With a square matrix

(not using the optical method) there will be eight possible triangles for three strains a, b and c, assuming it is a complete matrix:

	Matrix of values		
	a	b	c
a	0	27	30
b	25	0	42
c	35	40	0

Triangles

1. $X_{a:b}, X_{a:c}, X_{b:c}$
2. $X_{a:b}, X_{a:c}, X_{c:b}$
3. $X_{a:b}, X_{c:a}, X_{c:b}$
4. $X_{a:b}, X_{c:a}, X_{b:c}$
5. $X_{b:a}, X_{c:a}, X_{b:c}$
6. $X_{b:a}, X_{c:a}, X_{c:b}$
7. $X_{b:a}, X_{a:c}, X_{c:b}$
8. $X_{b:a}, X_{a:c}, X_{b:c}$

Such data permit two kinds of analysis in addition to study of reciprocal pairs. The first is to estimate test error from triangles with one zero side. The second is to determine whether the values satisfy the triangle inequality (the triangle hypothesis, see below); if so, they have properties consistent with a Euclidean metric, and are therefore well suited to spatial geometric representations of taxonomic structure.

Most published tables of DNA pairing data are very incomplete, consisting of only a limited number of the possible strain comparisons. Therefore a computer program was used to list the complete triangles or triples (i.e. the cases where three sides were reported) and then to determine the number where the triangle inequality was violated and to compute errors.

2.3.1 *Triangles with zero sides*

Strains which appear identical within the sensitivity of the experiment will form a triangle with one side zero. Theoretically the other two sides should be equal, but they are frequently unequal and the discrepancy may be used as a measure of error. Error was determined as $\sqrt{[(X_{a:c} - X_{b:c})^2/2]}$ for a triangle with $X_{a:b} = 0$, and averaged as for reciprocal pairs. Analysis of variance (ANOVA) was used to examine the amount of variation between and within comparable sets of data.

2.3.2 *Triangle inequalities*

The triangle hypothesis was studied by counting the proportion of triples which do not satisfy the triangle inequality. Thus, if $X_{a:b}$ is 25 and $X_{a:c}$ is 25 then $X_{b:c}$ cannot be greater than 50, i.e. the largest side of the triangle must be equal to or less than the sum of the other two sides. Any triangle with a zero side will violate the triangle inequality if any error is present (though perhaps only to a small extent) because then the other two sides will not be exactly equal.

We noted that square-rooting all distances generally reduced the number of violating triangles. Significance of this reduction was tested by chi-square with Yate's correction, and the probability determined for one degree of freedom (except where any value in the 2 × 2 table was less than 5, when Fisher's Exact method was used; Conover 1971).

For complete triangles the numbers of non-violating and violating triangles, before and after square-rooting are tabulated as:

	Before taking square-root	After taking square-root	
Violating	a	b	(a + b)
Not violating	c	d	(c + d)
	(a + c)	(b + d)	

$$\chi^2 = n([ad - bc] - n/2)/(a + b)(c + d)(a + c)(b + d)$$

Table 1. Error from published replications

Technique	Organism	$\widehat{s.d.}$*	N†	Reference
Optical	*Mycobacterium*	5·32	100	Baess (1979)
	Mycobacterium	3·00	126	Imaeda *et al.* (1982)
	Bacillus	3·97	15	Nakamura (1987b)
S1 Nuclease	*Actinobacillus* and *Haemophilus*	6·15	278	Potts & Berry (1983)
	Mycobacterium	8·57	20	McFadden *et al.* (1987)
	Mycobacterium	4·94	189	DeKesel *et al.* (1987)
	Haemophilus	6·30	228	Potts *et al.* (1986)

* Estimated standard deviation.
† Total degrees of freedom, i.e. $\sum(n_i - 1)$; see text.

3. Results

3.1 PUBLISHED STANDARD DEVIATIONS

Seven papers were examined (Table 1). Average error, s_E, lies between 3·0 and 8·6%; the weighted mean was 5·6%.

Pairing values were plotted against error values (corrected for degrees of freedom). The plot (Fig. 1) from the data of Potts & Berry (1983) showed greater error with higher per cent pairing; this was not always seen, however (see Fig. 2).

3.2 ERROR FROM INTERNAL CONSISTENCY

Other error estimates are derived from the internal consistency of published data and shown in Tables 2–5. They are arranged according to the pairing technique used, but results are first described according to the methods by which the error was estimated. The techniques, and major groups of bacteria studied, are compared afterwards.

3.3 RECIPROCAL PAIRS

Error ranged from 2·26 to 15·4% (Tables 2, 3, 5). The weighted average was 6·4%. For five studies, pairing values from labelled strains against several unlabelled strains were compared with the reverse situation (Rocourt *et al.* 1982; Ezaki *et al.* 1986; Dent & Williams 1986a, b; Johnson & Harich 1983). The difference in the relative binding does not seem to depend on the labelled strain. The mean and standard deviation of each half of the matrix were found, and a *t*-test of the means showed no significant difference on any of the data sets.

The S1 nuclease and filter methods showed differences in error rate estimated from reciprocal pairs. To confirm this an analysis of variance (ANOVA) was carried out on six of the most complete studies (Johnson & Harich 1983, 1986; Dent & Williams 1986b; Gebers *et al.* 1986; Kilpper-Bälz *et al.* 1985; Love *et al.* 1987a). When all six studies were included in ANOVA, the hypothesis that there was no difference in the mean error between studies was rejected ($P < 0{\cdot}001$). However, this was due entirely to one study, that of Kilpper-Bälz *et al.* (1985) on *Streptococcus*, which had much higher error than the others. When this study was excluded the significance of differences in error between the remaining five studies did not reach $P = 0{\cdot}2$.

3.4 ERROR FROM ZERO SIDES

Error ranged from 1·87 to 13·03% (Tables 2–5). The weighted average was 5·0%. Figure 2 shows a plot of standard deviation against average per cent pairing pooled from a range of papers (Rocourt *et al.* 1982; Johnson & Harich 1983, 1986; Mutters *et al.* 1985; Micales *et al.* 1985; Kilpper-Bälz *et al.*

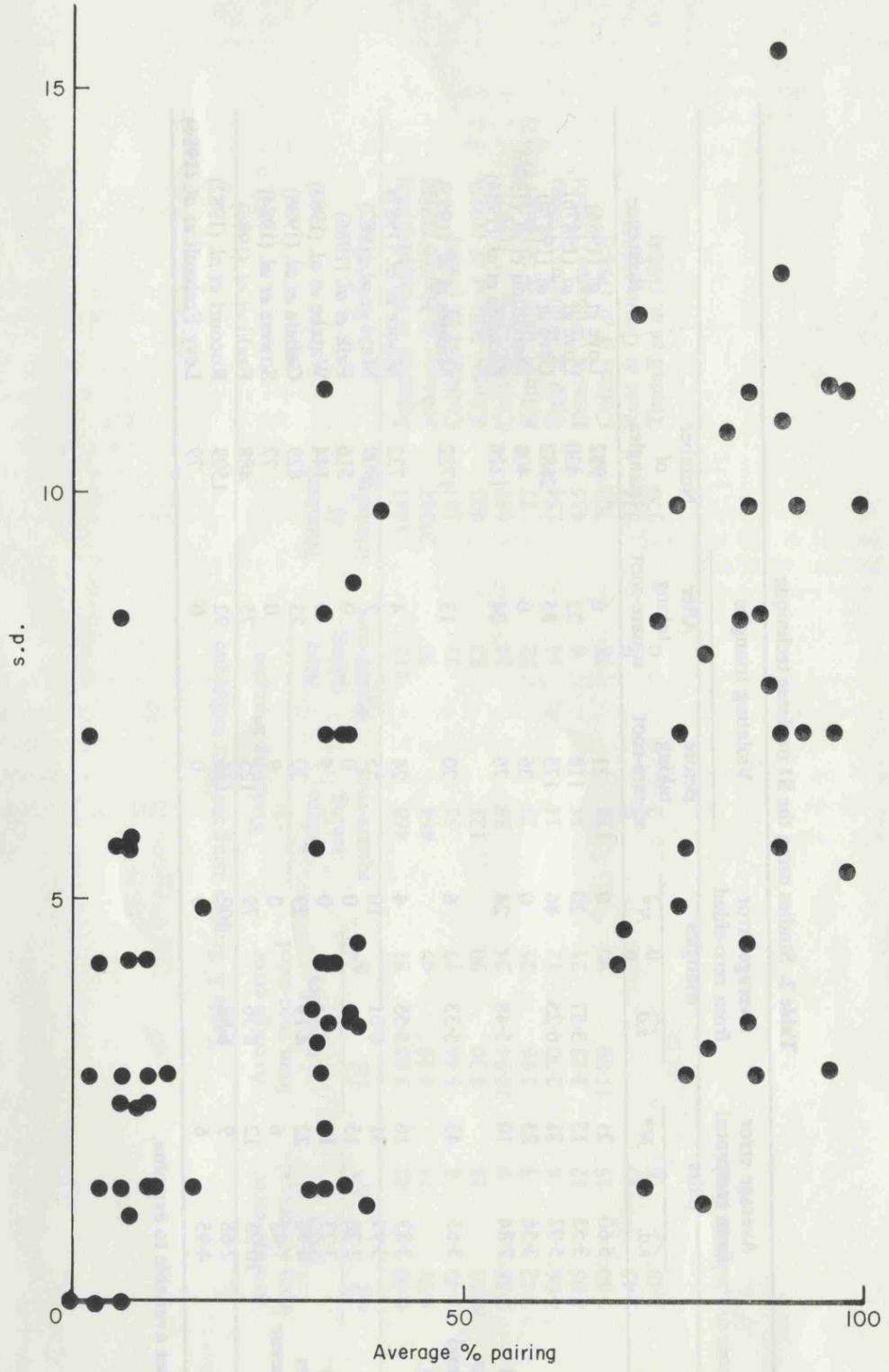


Fig. 1. Relationship between error (ordinate), expressed as published standard deviations, and average DNA–DNA pairing values (abscissa, data of Potts & Berry 1983).

1985; Dent & Williams 1986a, b; Gebers *et al.* 1986; Tanner *et al.* 1986). For each of the 10 papers looked at in detail the per cent pairing values were divided into 10 bands at 10% intervals. The average error for each section was plotted against the midpoint of the % pairing band. Error seems to remain fairly constant over the range of pairing values, in contrast to Fig. 1, except for high error in the 80–90% band. Error is somewhat lower at the extremes, i.e. 0–10% and 90–100% pairing. This is presumably because large error is not compatible with an average close to 0 or 100%. Thus, if two

Table 2. Studies using the S1 nuclease technique

Organism	Average error from reciprocal pairs		Average error from zero-sided triangles		Violating triangles		Number of triangles	Reference
	$\widehat{s.d.}$	N^*	$\widehat{s.d.}$	N^*	Before taking square-root	After taking square-root		
Bacteroides	5·60	21		0	21	0	982	Love *et al.* (1986)
Bacteroides	3·53	15	3·57	20	118	22	430	Love *et al.* (1987b)
Fusobacterium	5·92	31	6·25	86	175	85	2682	Love *et al.* (1987a)
Haemophilus	3·56	23		0	36	0	408	Morozumi *et al.* (1986)
Pasteurella and *Actinobacillus*	7·84	10	5·48	24	76	24	1226	Escande *et al.* (1984)
Gluconobacter and *Pseudomonas*	3·63	45	5·33	6	70	13	1722	Micales *et al.* (1985)
Selenomonas	3·83	16	5·55	4	24	4	732	Moore *et al.* (1987)
Veillonella	3·99	31	6·01	10	55	7	3909	Mays *et al.* (1982)
Azospirillum	2·26	15		0	0	0	516	Falk *et al.* (1986)
Bifidobacterium	7·23	14		0	5	0	144	Watabe *et al.* (1983)
Hyphomicrobium	4·96	22	4·19	29	30	23	629	Gebers *et al.* (1986)
Methanosarcinaceae	7·43	6		0	9	0	72	Sowers *et al.* (1984)
Streptococcus	10·95	12	8·36	79	125	75	488	Ezaki *et al.* (1986)
Listeria	2·68	9	6·88	100	136	92	1509	Rocourt *et al.* (1982)
Mycobacterium	4·95	6		0	0	0	79	Levy-Frebault *et al.* (1986)

* Number of cases available to examine.

Table 3. Studies using the filter technique

Organism	Average error from reciprocal pairs		Average error from zero-sided triangles		Violating triangles		Number of triangles	Reference
	$\widehat{\text{s.d.}}$	N*	$\widehat{\text{s.d.}}$	N*	Before taking square-root	After taking square-root		
Veillonella	6·40	42	5·05	81	468	112	1447	Johnson & Harich (1983)
Bacteroides	6·21	215	4·15	49	464	59	20387	Johnson & Harich (1986)
Enterococcus	15·40	8	8·46	17	52	21	183	Collins *et al.* (1986)
Streptococcus	10·53	18	8·30	90	122	82	403	Kilpper-Bälz *et al.* (1985)
Streptococcus	9·38	9	10·94	24	35	24	167	Coykendall *et al.* (1987)
Streptococcus	7·12	3	5·96	25	29	22	77	Kilpper-Bälz & Schleifer (1987)
Actinomyces	4·06	6	2·70	12	34	14	154	Dent & Williams (1986a)
Lactobacillus	4·05	17	3·23	21	35	6	435	Dent & Williams (1986b)
Lactobacillus	5·50	15	11·80	90	128	78	702	Collins *et al.* (1987)
Halophiles	5·45	10		0	0	0	324	Ross & Grant (1985)
Natronococcus and *Natronobacterium*	3·48	6		0	2	0	370	Tindall *et al.* (1984)

* The number of cases available to examine.

Table 4. Studies involving the optical technique*

Organism	Average error from zero-sided triangles		Violating triangles		Number of triangles	Reference
	$\widehat{s.d.}$	N†	Before taking square-root	After taking square-root		
Bacteroides	9·09	9	25	23	1755	Tanner *et al.* (1986)
Haemophilus and *Actinobacillus*		0	7	3	16	Pohl *et al.* (1983)
Pasteurella	7·36	15	64	31	321	Mutters *et al.* (1985)
Cytophaga and *Flavobacterium*		0	0	0	47	Callies & Mannheim (1980)
Thermus	7·31	3	6	3	15	Hensel *et al.* (1986)
Bacillus	5·08	32	32	26	496	Nakamura & Swezey (1983a)
Bacillus	4·92	210	270	170	680	Nakamura & Swezey (1983b)
Bacillus		0	8	1	90	Nakamura (1987a)
Bacillus	4·57	548	857	476	3100	Nakamura (1987b)
Mycobacterium	1·87	445‡	195	191	298	Imaeda (1985)

* For this method no estimate can be made for reciprocal pairs.
† Number of cases available to examine.
‡ There may be several cases per triangle, see Methods.

values give an average of 90%, the maximum error obtainable will be 14·14%, which is the standard deviation of homology values 80 and 100 (this is not necessarily so where there are pairing values over 100%, but there are not enough data for conclusions on methods where such values can occur).

The same 10 papers were used for an ANOVA to detect significant differences between the studies. Significant differences certainly exist; the hypothesis that there was no difference in the mean error between studies was rejected ($P < 0{\cdot}001$).

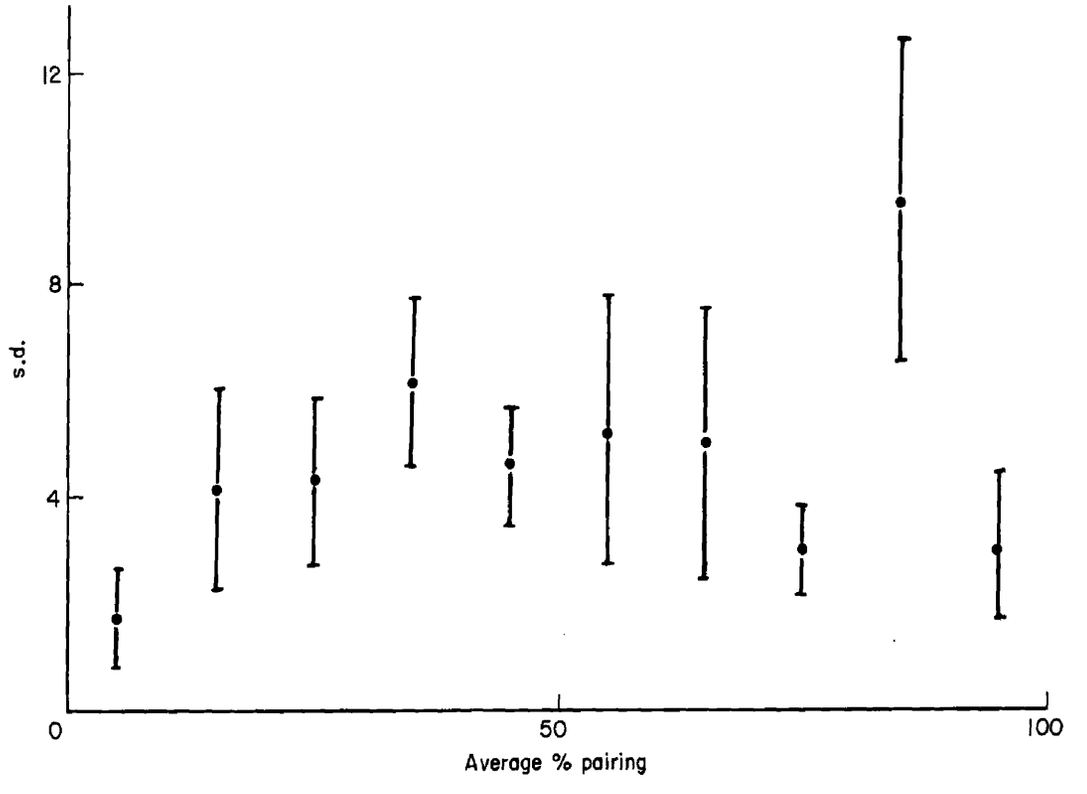


Fig. 2. Relationship between error from zero-sided triangles and average DNA–DNA pairing. The error is the average s.d. (see Methods). Pairing values were divided into 10 bands at 10% intervals, and the results plotted at the midpoints of the bands, as mean (dots) and one standard deviation above and below (bars). Data are from studies listed in text.

Table 5. Other techniques

Organism and technique	Average error from reciprocal pairs		Average error from zero-sided triangles		Violating triangles		Number of triangles	Reference
	$\widehat{s.d.}$	N^*	$\widehat{s.d.}$	N^*	Before taking square-root	After taking square-root		
Bacillus Slot-blot Filter method	7·46	5	4·69	10	11	8	218	Seldin & Dubnau (1985)
Methylobacterium Multi-blot Filter method	2·65	14	3·20	34	83	47	719	Hood *et al.* (1987)
Mycoplasma DNA probes and hydroxyapatite columns	2·92	10		0	1	0	260	Stephens *et al.* (1985)
Rhizobium Hydroxyapatite probe and filter	8·79	14	10·59	41	174	44	768	Wedlock & Jarvis (1986)
Leptospires Hydroxyapatite technique	15·26	48	13·03	42	123	42	1703	Yasuda *et al.* (1987)

* Number of cases available to examine.

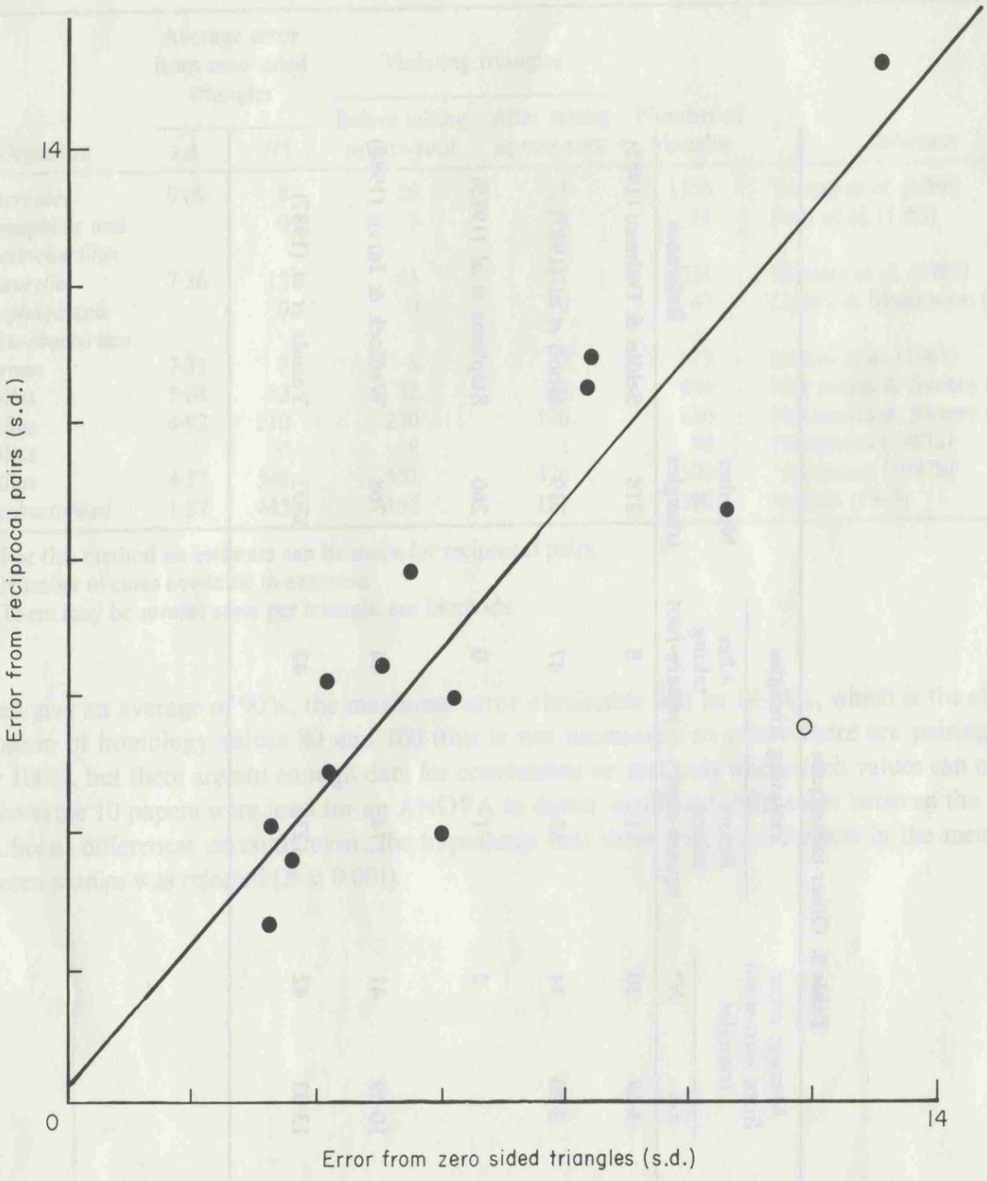


Fig. 3. Relationship between error estimated from reciprocal pairs and that estimated from zero-sided triangles for studies where both are available. Each circle represents a different study from Tables 2, 3 and 5 (see text). The open circle is that of Collins *et al.* (1987). The line of linear regression (excluding open circle) is shown.

A plot of mean reciprocal pair error against mean zero-sides error, for studies where both could be estimated, gave a reasonable straight line fit: $Y = 0{\cdot}31 + 1{\cdot}07X$ (where X = error from triples with a zero-side, Y = reciprocal error) shown in Fig. 3. Results with less than 10 values in either error method were not used. When the *Lactobacillus* data (Collins *et al.* 1987) are omitted the line passes close to the origin at almost 45°, and the correlation r is high (over 0·79), so this implies that the error rates obtained from reciprocal pairs and zero sides are consistent and similar in magnitude.

3.5 TRIANGLE INEQUALITIES

The proportion of violating triangles (Tables 2–5) ranged from 0 to 65%, the average being about 8% before square-rooting. Although violating triangles were more frequent in studies with more zero-sided triangles (as expected, see Methods 2.2), there were many puzzling features, which are taken up in the Discussion. Six papers were used to detect whether square-rooting significantly reduced the proportion of violating triangles using Fisher's exact method for 2 × 2 tables and combination of

Table 6. Variation of average error and triangle violations with method and major group of bacteria

	Technique					
	Optical (Table 4)		S1 Nuclease (Table 2)		Filter (Table 3)	
			$\widehat{s.d.}$	N*	$\widehat{s.d.}$	N
Reciprocal pairs	Not applicable					
Gram-negative			4·50	192	6·24	257
Gram-positive			6·86	27	8·14	70
Halophiles			—	—	4·71	16
All organisms			4·84	276	6·51	349
Zero-sided triangles	$\widehat{s.d.}$	N	$\widehat{s.d.}$	N	$\widehat{s.d.}$	N
Gram-negative	7·93	27	5·89	150	4·71	130
Gram-positive	4·03	1235	7·53	179	9·11	267
Halophiles	—	—	—	—	—	—
All organisms	4·11	1262	6·49	358	7·52	409
Triangle violations	% violating	N	% violating	N	% violating	N
Gram-negative	4·49	2139	4·76	12091	4·27	21834
Gram-positive	30·69	4764	12·57	2076	20·39	1967
Halophiles	—	—	—	—	0·29	694
All organisms	22·61	6918	5·67	15528	5·55	24649

* Number of cases available to examine.

probabilities (Conover 1971; Snedecor 1956). This gave a probability of 0·01–0·001, indicating a significant improvement from square-rooting.

When all papers in the study were used the average proportion of violating triangles was 3·8% compared with 8·2% before square-rooting, and most of the remaining violating triples had a zero side which forces a violation if there is any error however small. From 10 studies (Johnson & Harich 1983; Dent & Williams 1986a, b; Ezaki *et al.* 1986; Collins *et al.* 1986, 1987; Love *et al.* 1986, 1987a; Kilpper-Bälz & Schleifer 1987; Hood *et al.* 1987) a total of 1231 violating triples was reduced to 470 by square-rooting the distances; however, 437 of these involved a zero-side. Of the 33 remaining, 20 were from the study of Johnson & Harich (1983).

Five papers that involved techniques other than those in Tables 2–4 are shown in Table 5. Hood *et al.* (1987) claimed to have an improved multi-blot filter method and we found the errors were low using both zero sides and reciprocal pairs, but the proportion of violating triangles was fairly high.

3.6 COMPARISON OF TECHNIQUES AND MAJOR GROUPS OF BACTERIA

Results are averaged for each technique and major group of the organisms studied (Table 6). An ANOVA on the error from zero-sides was not rejected at $P = 0{\cdot}05$, that is the zero-side errors did not differ significantly between techniques. An ANOVA on the reciprocal pair error also showed no significant difference between techniques at $P = 0{\cdot}05$.

Comparison of the error for different major groups of organisms is hampered by the small values of N in many of the cases, but it is noteworthy that the percentage of violating triangles is high in Gram positive groups. These points are taken up in the Discussion.

4. Discussion

4.1 PUBLISHED STANDARD DEVIATIONS

A further paper using *Bacteroides* (Tanner *et al.* 1986) and the optical technique was examined where a small number of replications showed an average error of 0·29% but this paper is not representative; there were only six replicates and all pairing values were very close to 100%.

Error seemed to be largely independent of the degree of pairing, although there was some evidence for an increase in error at high homology values in Potts & Berry's (1983) data (Fig. 1). This is not readily explicable, because in theory the error at least in the optical method will be greatly constrained near 100% (and near 0%) and therefore the error here would be reduced. Indeed the error from zero-sided triangles suggests this may be so (Fig. 2).

4.2 ZERO-SIDED TRIANGLES

The paper by Nakamura & Swezey (1983b) on *Bacillus* sp. had 147 measurable errors of which most were small; the average was raised by a few high errors. Two of the largest errors were based on DNA pairing values of less than 35%. One explanation is that purity of DNA or fragment size may vary between the preparations from the strains; these factors are known to affect the degree of binding and the reproducibility of the results (DeLey *et al.* 1970).

There may be a tendency to round up high values to 100%. If pairing determinations could be calculated without error many of the zero distances would be found to show less than 100% pairing. This could theoretically affect this method of error estimation, because such triangles would then be excluded from consideration. We do not believe, however, that this has caused major bias. If, for example, a pair of strains has a pairing value recorded as 100% when the true value should be 99%, this implies that the pairing values to other strains should agree to within one or two per cent at the most. The discrepancies we observed are commonly far greater than this and the reliability of estimates from zero-sided triangles is supported by other methods of estimation (Table 1, Fig. 3).

It might be thought that if a triangle had a side that was very small, but not zero, one could estimate error if the other two sides differed by a considerable amount. Thus a triangle with sides 1, 10, 30 implies an error close to that estimated from a triangle with sides 0, 10, 30. This, however, requires subtraction of a correction term (related to the quantity 1) from the difference between 10 and 30, and this is not statistically straightforward. We explored the problem by computer simulation, and concluded that although one could work out an empirical correction, these new estimates (which were of similar magnitude to those from zero-sided triangles) did not yield further reliable information.

4.3 TRIANGLE INEQUALITIES

The proportion of violating triangles fluctuates widely, with little relation to average error, method or group of organisms. Thus two studies on *Veillonella* (Mays *et al.* 1982; Johnson & Harich 1983) show 1·4% and 32·3% of violating triangles respectively, though the error rates are quite typical (Tables 2, 3). Similarly, percentages on *Bacillus* by Nakamura and his colleagues (Nakamura & Swezey 1983a, b; Nakamura 1987a, b) vary from 6·5% to 39·7% (Table 4). High percentages tend to be associated with high error rates though this effect is not marked. Percentages tend to be high in Gram-positive bacteria, particularly for streptococci and enterococci, but this may not be significant. The expected association between violating triangles and zero-sided triangles has been noted in Results.

The frequent occurrence of triangle inequalities suggests that 100 – % DNA–DNA pairing does not necessarily behave as a Euclidean metric. It is not clear to what extent the inequalities are due to experimental error, but the unexpected features just noted suggest that they are an expression of physicochemical factors that are not yet understood. This would have implications for taxonomic conclusions that are drawn from DNA studies. It seems unlikely that the rather higher error in studies on Gram-positive compared with Gram-negative bacteria would account for the higher proportion of triangles that violate the inequality in the former (Table 6). More accurate pairing values are needed to decide this point, but it is possible that DNA pairing differences are inherently non-Euclidean; if so it would be worth trying the square root transformation in systematic studies. At present there are few data available for such work, but we have examined the effect of square rooting on the complete matrix of Nakamura & Swezey (1983a) and found the same dendrogram topology,

and very similar principal co-ordinate relationships, as those with unrooted values in Hartford & Sneath (1988, Fig. 1); the main effect was expansion of the scale near the tips of the dendrogram and looser grouping of the tight cluster of strains in the principal co-ordinate diagram. Because this matrix shows few violations, and a simple taxonomic structure, these findings are not unexpected. The transformation might be important in evolutionary reconstructions. Consider a situation where, for strains a, b and c, for example, the implied evolutionary change a:c between strains a and c is greater than the sum of the changes a:b and b:c. This will prevent evolutionary distances from being additive, and it would have undesirable consequences for algorithms for phylogenetic reconstruction.

4.4 GENERAL FACTORS

The greater error in Gram-positive than Gram-negative bacteria (Table 6) may well be due to the tough cell wall of the former that makes them more difficult to lyse, which may affect fragment size or purity of DNA. DNA isolation is difficult for some *Streptococcus* species (Garvie 1978). This may be why error is particularly high with these organisms. All reciprocal-pair errors with streptococci were much higher than the average for other papers, as were the errors from zero-sided triangles. The proportion of violating triangles was also very high, about one third of all triangles compared to an average of about 8% for all papers studied. Analysis of variance carried out on error from reciprocal pairs and error from zero-sides showed no significant differences between the studies on *Streptococcus* and *Enterococcus* at $P < 0{\cdot}05$.

Bacteria producing abundant extracellular polysaccharide may also show large error (*Rhizobium*, Table 5, Wedlock & Jarvis 1986), possibly due to the interaction between carbohydrates and DNA (Graves 1968).

Error in studies such as those by Gebers *et al.* (1986) and Love *et al.* (1987a) may perhaps be high because of the wide %(G + C) range covered; organisms with a %(G + C) difference of more than a few per cent will have widely different T_m values and hence different optimal renaturation temperatures.

Previous comparisons of reassociation techniques showed a straight line relationship between values at high pairing levels, but the relation is curved when extended to levels below 30–40% pairing (Huss *et al.* 1983; Bouvet & Grimont 1986). This behaviour deserves further study. It may be noted, however, that there is considerable scatter about the lines, and this scatter must reflect test error. Examination of their figures suggests that the error is 5–8%, similar to the average found here.

In the study by Huss *et al.* (1983) the magnitude of the DNA–DNA pairing in the filter and optical techniques can be compared directly for values over 30–40%. The magnitudes for these two methods are about the same. However, the S1 nuclease technique, in the study of Bouvet & Grimont (1986), has corresponding pairing values of approximately 20% less than the filter technique (for values above 30–40% pairing). It seems from these studies that the S1 nuclease technique may be more stringent than the other two methods. This difference is not seen in data from Whiley *et al.* (1988). There were only six comparable pairings in the latter study, however, so the question remains open.

When DNA pairing techniques are used in systematics the error associated with these techniques must affect the inferred relationships. Indeed, all the methods of estimation suggest that the error averages 5–6%, though varying considerably from study to study. Unless the error is taken into account, incorrect conclusions could easily be drawn, especially if only a few reference strains are used.

It is difficult at present to make recommendations on the most repeatable methods. Study on the effects of specific factors such as DNA purity, fragment size and reassociation temperature is clearly needed, and accurate control of temperature and other parameters is important. The regular use of standard DNA preparations might be valuable. Similarly, it is not yet clear what is the best transformation of DNA–DNA differences. One useful approach would be to compare transformed values with the numbers of differences in rRNA sequence comparisons and seek for transformations that gave similar proportional distances in triangles from rRNA and DNA.

5. References

BAESS, I. 1979 Deoxyribonucleic acid relatedness among species of slowly growing mycobacteria. *Acta Pathologica Microbiologica Scandinavica B* **87**, 221–226.

BOUVET, P.J.M. & GRIMONT, P.A.D. 1986 Taxonomy of the genus *Acinetobacter* with the recognition of *Acinetobacter baumannii* sp.nov., *Acinetobacter haemolyticus* sp.nov., *Acinetobacter johnsonii* sp.nov., and *Acinetobacter junii* sp.nov. and emended descriptions of *Acinetobacter calcoaceticus* and *Acetinobacter lwoffii*. *International Journal of Systematic Bacteriology* **36**, 228–240.

CALLIES, E. & MANNHEIM, W. 1980 Deoxyribonucleic acid relatedness of some menaquinone-producing *Flavobacterium* and *Cytophaga* strains. *Antonie van Leeuwenhoek* **46**, 41–49.

COLLINS, M.D., FARROW, J.A.E. & JONES, D. 1986 *Enterococcus mundtii* sp. nov. *International Journal of Systematic Bacteriology* **36**, 8–12.

COLLINS, M.D., FARROW, J.A.E., PHILLIPS, B.A., FERUSU, S. & JONES, D. 1987 Classification of *Lactobacillus divergens*, *Lactobacillus pisicola* and some catalase-negative, asporogenous, rod-shaped bacteria from poultry in a new genus, *Carnobacterium*. *International Journal of Systematic Bacteriology* **37**, 310–316.

CONOVER, W.J. 1971 *Practical Nonparametric Statistics*, New York: John Wiley.

COYKENDALL, A.L., WESBECHER, P.M. & GUSTAFSON, K.B. 1987 *'Streptococcus milleri', Streptococcus constellatus* and *Streptococcus intermedius* are later synonyms of *Streptococcus anginosus*. *International Journal of Systematic Bacteriology* **37**, 222–228.

DEKESEL, M., COENE, M., PORTAELS, F. & COCITO, C. 1987 Analysis of deoxyribonucleic acids from armadillo-derived mycobacteria. *International Journal of Systematic Bacteriology* **37**, 317–322.

DELEY, J., CATTOIR, H. & REQUAERTS, A. 1970. The quantitative measurement of DNA hybridisation from renaturation rates. *European Journal of Biochemistry* **12**, 133–142.

DENT, V.E. & WILLIAMS, R.A.D. 1986a. *Actinomyces slackii* sp. nov. from dental plaque of dairy cattle. *International Journal of Systematic Bacteriology* **36**, 392–395.

DENT, V.E. & WILLIAMS, R.A.D. 1986b Deoxyribonucleic acid reassociation between strains of *Lactobacillus animalis, Lactobacillus odontolyticus* and strains resembling *Lactobacillus acidophilus* isolated from animals. *International Journal of Systematic Bacteriology* **36**, 481–482.

ESCANDE, F., GRIMONT, F., GRIMONT, P.A.D. & BERCOUVIR, H. 1984 Deoxyribonucleic acid relatedness among strains of *Actinobacillus* spp and *Pasteurella ureae*. *International Journal of Systematic Bacteriology* **34**, 309–315.

EZAKI, T., FACKLAM, R., TAKEUCHI, N. & YABUACHI, E. 1986 Genetic relatedness between the type strain of *Streptococcus anginosus* and minute-colony-forming β haemolytic streptococci carrying different Lancefield grouping antigens. *International Journal of Systematic Bacteriology* **36**, 345–347.

FALK, E.C., JOHNSON, J.L., BALDANI, V.L.D., DOBEREINER, J. & KRIEG, N.R. 1986. Deoxyribonucleic acid and RNA homology studies of the genera *Azospirillum* and *Conglomeromonas*. *International Journal of Systematic Bacteriology* **36**, 80–85.

GARVIE, E.I. 1978 *Streptococcus raffinolactis* Orla-Jensen and Hansen, a group N streptococcus found in raw milk. *International Journal of Systematic Bacteriology* **28**, 190–193.

GEBERS, R., MARTENS, B., WEHMEGER, U. & HIRSCH, P. 1986 Deoxyribonucleic acid homologies of *Hyphomicrobium* spp., *Hyphomonas* spp., and other hyphal, budding bacteria. *International Journal of Systematic Bacteriology* **36**, 241–245.

GRAVES, I.L. 1968 Interactions of heat-denatured HeLa cell DNA acid with synthetic and natural polysaccharides. *Biopolymers* **6**, 1573–1578.

GRIMONT, P.A.D., POPOFF, M.Y., GRIMONT, F., COYNAULT, C. & LEMELIN, M. 1980 Reproducibility and correlation study of three deoxyribonucleic acid hybridization procedures. *Curr. Microbiol.* **4**, 325–330.

HARTFORD, T. & SNEATH, P.H.A. 1988. Distortion of taxonomic structure from DNA relationships due to different choice of reference strains. *Systematic and Applied Microbiology* **10**, 241–250.

HENSEL, R., DEMHARTER, W., KANDLER, O., KROPPENSTEDT, R.M. & STACKEBRANDT, E. 1986. Chemotaxonomic and molecular-genetic studies of the genus *Thermus:* evidence for a phylogenetic relationship of *Thermus aquaticus* and *Thermus ruber* to the genus *Deinococcus*. *International Journal of Systematic Bacteriology* **36**, 444–453.

HOOD, D.W., DOW, C.S. & GREEN, P.N. 1987 DNA:DNA hybridisation studies on the pink-pigmented faculative methylotrophs. *Journal of General Microbiology* **133**, 709–720.

HUSS, V.A.R., FESTL, H. & SCHLEIFER, K.H. 1983 Studies on the spectrophotometric determination of DNA hybridization from renaturation rates. *Systematic and Applied Microbiology* **4**, 184–192.

IMAEDA, T. 1985 Deoxyribonucleic acid relatedness among selected strains of *Mycobacterium tuberculosis, Mycobacterium bovis, Mycobacterium bovis* BCG, *Mycobacterium microti* and *Mycobacterium africanum*. *International Journal of Systematic Bacteriology* **35**, 147–150.

IMAEDA, T., BARKSDALE, L. & KIRCHEIMER, W.F. 1982 Deoxyribonucleic acid of *Mycobacterium lepraemurium:* its genome size, base ratio and homology with those of other mycobacteria. *International Journal of Systematic Bacteriology* **32**, 456–458.

JOHNSON, J.L. & HARICH, B. 1983 Ribosomal ribonucleic acid homology among species of the genus *Veillonella* Prévot. *International Journal of Systematic Bacteriology* **33**, 760–764.

JOHNSON, J.L. & HARICH, B. 1986 Ribosomal ribonucleic acid homology among species of the genus *Bacteroides*. *International Journal of Systematic Bacteriology* **36**, 71–79.

KILPPER-BÄLZ, R. & SCHLEIFER, K.H. 1987 *Streptococcus suis* sp. nov., nom. rev. *International Journal of Systematic Bacteriology* **37**, 160–162.

KILPPER-BÄLZ, R., WENZIG, P. & SCHLEIFER, K.H.

1985 Molecular Relationships and Classification of viridans streptococci as *Streptococcus oralis* and Emended description of *Streptococcus oralis*. *International Journal of Systematic Bacteriology* **35**, 482–488.

LEVY-FREBAULT, V., GRIMONT, F., GRIMONT, P.A.D. & DAVID, H.L. 1986 Deoxyribonucleic acid relatedness study of the *Mycobacterium fortuitum–Mycobacterium chelonae* complex. *International Journal of Systematic Bacteriology* **36**, 458–460.

LOVE, D.N., JOHNSON, J.L., JONES, R.F., BAILEY, M. & CALVERLEY, A. 1986 *Bacteroides tectum* sp. nov. and characteristics of other non-pigmented *Bacteroides* isolates from soft-tissue infections from cats and dogs. *International Journal of Systematic Bacteriology* **36**, 123–128.

LOVE, D.N., CATO, E.P., JOHNSON, J.L., JONES, R.F. & BAILEY, M. 1987a DNA hybridisation among strains of fusobacteria isolated from soft tissue infections of cats: comparison with human and animal type strains from oral and other sites. *International Journal of Systematic Bacteriology* **37**, 23–26.

LOVE, D.N., JOHNSON, J.L., JONES, R.F. & CALVERLEY, A. 1987b *Bacteroides salivosus* sp. nov., an asaccharolytic, black-pigmented species from cats. *International Journal of Systematic Bacteriology* **37**, 307–309.

MCFADDEN, J.J., BUTCHER, P.D., CHIODINI, R.J. & HERMON-TAYLOR, J. 1987 Determination of genome size and DNA homology between an unclassified *Mycobacterium* species isolated from patients with Crohn's disease and other mycobacteria. *Journal of General Microbiology* **133**, 211–214.

MAYS, T.D., HOLDEMAN, L.V., MOORE, W.E.C., ROGOSA, M. & JOHNSON, J.L. 1982 Taxonomy of the genus *Veillonella* Prévot. *International Journal of Systematic Bacteriology* **32**, 28–36.

MICALES, B.K., JOHNSON, J.L. & CLAUS, G.W. 1985 Deoxyribonucleic acid homologies among organisms in the Genus *Gluconobacter*. *International Journal of Systematic Bacteriology* **35**, 79–85.

MOORE, L.V.H., JOHNSON, J.L. & MOORE, W.E.C. 1987 *Selenomonas noxia* sp. nov., *Selenomonas flueggei* sp. nov., *Selenomonas infelix* sp. nov., *Selenomonas dianae* sp. nov. and *Selenomonas artemidis* sp. nov. from human gingival crevice. *International Journal of Systematic Bacteriology* **37**, 271–280.

MOROZUMI, T., URS, P., BRAUN, R. & NICOLET, J. 1986 Deoxyribonucleic acid relatedness among strains of *Haemophilus parasuis* and other *Haemophilus* spp. of swine origin. *International Journal of Systematic Bacteriology* **36**, 17–19.

MUTTERS, R., IHM, P., POHL, S., FREDERIKSEN, W. & MANNHEIM, W. 1985 Reclassification of the genus *Pasteurella* Trevisan 1887 on the basis of deoxyribonucleic acid homology, with proposals for the new species *Pasteurella dagmatis*, *Pasteurella canis*, *Pasteurella stomatis*, *Pasteurella anatis*, *Pasteurella langaa*. *International Journal of Systematic Bacteriology* **35**, 309–322.

NAKAMURA, L.K. 1987a *Bacillus alginolyticus* sp. nov., and *Bacillus chrondroitinus* sp. nov. two alginate-degrading species. *International Journal of Systematic Bacteriology* **37**, 284–286.

NAKAMURA, L.K. 1987b *Bacillus polymyxa* (Prazmowski) Macé 1889 deoxyribonucleic acid relatedness and base composition. *International Journal of Systematic Bacteriology* **37**, 391–397.

NAKAMURA, L.K. & SWEZEY, J. 1983a Taxonomy of *Bacillus circulans* Jordan: composition and reassociation of deoxyribonucleic acid. *International Journal of Systematic Bacteriology* **33**, 46–52.

NAKAMURA, L.K. & SWEZEY, J. 1983b Deoxyribonucleic acid relatedness of *Bacillus circulans* Jordan 1890 strains. *International Journal of Systematic Bacteriology* **33**, 703–708.

POHL, S., BERTSCHINGER, H.V., FREDERIKSEN, W. & MANNHEIM, W. 1983 Transfer of *Haemophilus pleuropneumoniae* and the *Pasteurella haemolytica*-like organism causing porcine necrotic pleuropneumonia to the genus *Actinobacillus* (*Actinobacillus pleuropneumoniae* comb. nov.) on the basis of phenotypic and deoxyribonucleic acid relatedness. *International Journal of Systematic Bacteriology* **33**, 510–514.

POTTS, T.V. & BERRY, E.M. 1983 Deoxyribonucleic acid-deoxyribonucleic acid hybridization analysis of *Actinobacillus actinomycetemcomitans* and *Haemophilus aphrophilus*. *International Journal of Systematic Bacteriology* **33**, 765–771.

POTTS, T.V., MITRA, T., O'KEEFE, T., ZAMBON, J.J. & GENCO, R.J. 1986 Relationships among isolates of oral haemophili as determined by DNA–DNA hybridization. *Archives of Microbiology* **145**, 136–141.

ROCOURT, J., GRIMONT, F., GRIMONT, P.A.D. & SEELIGER, H.P.R. 1982 DNA relatedness among serovars of *Listeria monocytogenes sensu lato*. *Current Microbiology* **7**, 383–388.

ROSS, H.N.M. & GRANT, W.D. 1985 Nucleic acid studies on halophilic archaebacteria. *Journal of General Microbiology* **131**, 165–173.

SELDIN, L. & DUBNAU, D. 1985 Deoxyribonucleic acid homology among *Bacillus polymyxa*, *Bacillus macerans*, *Bacillus azotofixans*, and other nitrogen-fixing *Bacillus* strains. *International Journal of Systematic Bacteriology* **35**, 151–154.

SNEDECOR, G.W. 1956 *Statistical Methods Applied to Experiments in Agriculture and Biology* 5th edn. Ames, Iowa: Iowa State University Press.

SOWERS, K.R., JOHNSON, J.L. & FERRY, J.G. 1984 Phylogenetic relationships among the methylotrophic methane-producing bacteria and emendation of the family *Methanosarcinaceae*. *International Journal of Systematic Bacteriology* **34**, 444–450.

STEPHENS, E.B., ROBINSON, I.M. & BARILE, M.F. 1985 Nucleic acid relationships among the anaerobic mycoplasmas. *Journal of General Microbiology* **131**, 1223–1227.

TANNER, A.C.R., LISTGARTEN, M.A., EBERSOLE, J.L. & STREZEMPKO, M.N. 1986 *Bacteroides forsythus* sp.nov., a slow-growing, fusiform *Bacteroides* sp. from the human oral cavity. *International Journal of Systematic Bacteriology* **36**, 213–221.

TINDALL, B.J., ROSS, H.N.M. & GRANT, W.D. 1984. *Natronobacterium* gen. nov. and *Natronococcus* gen.

nov., two new genera of haloalkaliphilic archaebacteria. *Systematic and Applied Microbiology* **5**, 41–57.

WATABE, J., BENNO, Y. & MITSUOKA, T. 1983. *Bifidobacterium gallinarum* sp.nov.: a new species isolated from the ceca of chickens. *International Journal of Systematic Bacteriology* **33**, 127–132.

WEDLOCK, D.N. & JARVIS, B.D.W. 1986 DNA homologies between *Rhizobium fredii*, rhizobia that nodulate *Galega* spp., and other *Rhizobium* and *Bradyrhizobium* species. *International Journal of Systematic Bacteriology* **36**, 550–558.

WHILEY, R.A., RUSSELL, R.R.B., HARDIE, J.M. & BEIGHTON, D. 1988 *Streptococcus downei* sp. nov. for strains previously described as *Streptococcus mutans* serotype h. *International Journal of Systematic Bacteriology* **38**, 25–29.

YASUDA, P.H., STEIGERWALT, A.G., SULZER, K.R., KAUFMANN, A.F., ROGERS, F. & BRENNER, D.J. 1987 Deoxyribonucleic acid relatedness between serogroups and serovars in the family *Leptospiraceae* with proposals for seven new *Leptospira* species. *International Journal of Systematic Bacteriology* **37**, 407–415.