CHAPTER 3

Approaches In Bacterial Systematics

1. Introduction

Developments in molecular microbiology now underpin many exciting new methods which have not only been used for the classification and identification of bacteria, but have also provided insight into procaryote evolution. Bacterial systematics has really come of age as it is seen as a fundamental scientific discipline which addresses some of the basic questions facing humankind, such as the extent of microbial diversity and its role in sustainable agriculture (Goodfellow & O'Donnel, 1993).

The three interrelated areas of systematics include classification, nomenclature and identification. Van Berkum & Eardly (1998) describe these areas as follows:

- Classification: the arranging of organisms into taxonomic groups based on similarities.
- Nomenclature: the assignment of names to the taxonomic groups according to international rules.
- Identification: the process of determining whether a new isolate belongs to one of the established and named groups.

The once dull and boring subject area of systematics has developed into an exciting and rapidly developing discipline recognised by diverse members of the scientific community. These include epidemiologists, molecular ecologists and biologists, geochemists, agriculturists, etc. Since systematics is data dependent, many of the advances made within this field are due to the manner in which data is collected and analysed. In many instances multiple traits are analysed to establish the relationships between microorganisms. Consequently, current systematics requires an understanding of microbial chemistry, molecular biology, microbial physiology and data handling procedures (Goodfellow & O'Donnell, 1993).

The integration of multiple traits, termed polyphasic taxonomy by Colwell (1970), arose about 30 years ago. This approach aims to integrate phenotypic, genotypic and phylogenetic information for the delineation of taxa at all levels. Genotypic information is derived from the nucleic acids (DNA and RNA) while phenotypic information is derived from proteins and their function, chemotaxonomic markers, and other expressed features (Vandamme *et al.*,

1996). Information from a variety of molecules may be used within a polyphasic approach. However, the technical complexity in terms of time and labour often determines the sequence of application since not all can be applied to large numbers of isolates. Additionally, it is also important to understand at which level these molecules carry information. An abridged presentation from Vandamme *et al.* (1996) of the level of taxonomic information from the different techniques is given in Fig. 3.1.

The classification of rhizobia based on plant infection tests has been abandoned since the genes coding for nodulation, host specificity and nitrogen fixation are sometimes located on transmissible symbiotic plasmids. The International Subcommittee for the Taxonomy of *Rhizobium* and *Agrobacterium* proposed minimal standards for the description of species of root- and stem-nodulating bacteria (Graham *et al.*, 1991). These include symbiotic performance with selected hosts, cultural and morphological characteristics, DNA-DNA relatedness, rRNA-DNA hybridisation and 16S rDNA analysis, DNA restriction fragment length polymorphisms, and multilocus enzyme electrophoresis. The aim of this section is therefore to discuss some of the major categories of taxonomic techniques required to study bacteria at different levels, their general concept and application in rhizobial taxonomy.

2 Phenotypic Methods

Phenotypic methods comprise all those not directed toward DNA or RNA, and as such also include chemotaxonomic techniques. The term "chemotaxonomy" refers to the application of analytical methods to collect information on various chemical constituents of the cell to classify bacteria. More detail on this approach, classical phenotypic analyses and numerical analyses will be discussed in subsequent sections.

2.1 Classical phenotypic analyses

Classical or traditional phenotypic traits, such as morphological, physiological and biochemical features, form the basis for the formal description of taxa, from species and subspecies up to genus and family level. Very often, highly standardised procedures are required to obtain reproducible results within and between laboratories when considering some of these phenotypic traits. There are also instances where the paucity of these phenotypic characters can cause problems in the description or differentiation of taxa. For

such bacteria, alternative chemotaxonomic or genotypic methods are often required to reliably identify strains (Vandamme *et al.*, 1996).

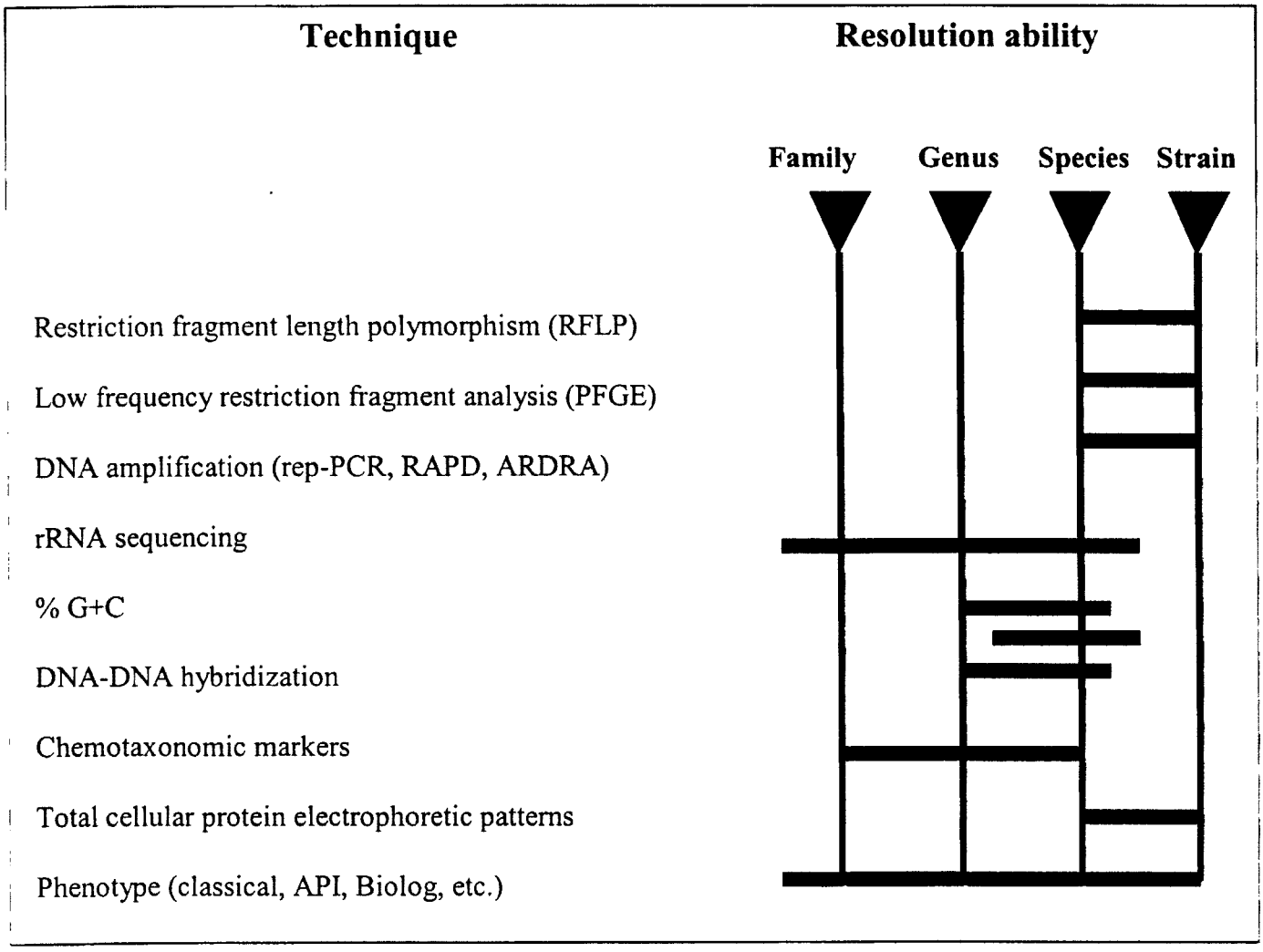

Figure 3.1 Taxonomic resolution of some of the techniques used in polyphasic taxonomy. Abridged from Vandamme *et al.*, 1996.

2.2 Numerical analysis

The primary aim of numerical taxonomy or computer-assisted classification is to assign individual bacterial strains to homogeneous groups using large sets of different types of phenotypic data (Goodfellow & O'Donnell, 1993). Numerical taxonomy arose in parallel with the development of computers and allowed the comparison of large numbers of phenotypic traits for large numbers of strains (Vandamme *et al.*, 1996). The application of

numerical analyses subsequently led to the revision of the pre-1960 classification of many bacterial genera (Goodfellow & O'Donnell, 1993). More recently, it has found application in the taxonomy of rhizobial isolates (see Chapter 4 for more specific references). Additionally, McInroy *et al.* (1999) characterised rhizobia from African acacias and other tropical woody legumes by applying numerical taxonomy based on substrate utilisation patterns.

2.3 Chemosystematics

Chemosystematics is a discipline in which information derived from the whole organism or cell fractions is used to classify, identify or type bacteria. The development of this area of taxonomy is greatly due to the introduction and application of reliable, rapid and sensitive analytical methods, such as chromatography and electrophoresis. Using these tools, traits such as cell wall composition, cellular fatty acids, whole-cell proteins, sugars and amino acids could be used as chemotaxonomic markers. These could be used at all taxonomic levels, although the discriminatory power of these may vary between taxa (Goodfellow & O'Donnell, 1993).

2.4 Multilocus Enzyme Electrophoresis (MLEE)

It is often difficult, and sometimes impossible, to relate phenotypic variation to allelic variation at a specific locus. Hybridisation of total DNA has been widely used to define species limits (Wayne, 1987). However, this application has made little contribution to estimates of the variation within species due to the large experimental error associated with this application arising from the interference of extrachromosomal DNA and variation between laboratories (Selander *et al.*, 1986).

MLEE has long been used as a standard method in eucaryotic population genetics and although not so widely used in procaryotes, a number of reports have verified its applicability in the study of bacteria (Demezas *et al.*, 1991). In the application of the MLEE method microorganisms are characterised by electrophoretic analyses of a range of enzymes, such as glucose-6-phosphate dehydrogenase or malate dehydrogenase, which are important components of metabolic pathways. Due to their essential role in metabolism they are widely distributed across diverse genera. Furthermore, they are chromosomally encoded, present as single copies on the genome and their products are easily identified by staining following

electrophoresis. MLEE method has the advantage over other phenotypic methods in that it provides an unbiased assessment of the population genetic structure since the presence of a specific enzyme electromorphs do not appear to provide a selective advantage for the strains in which they reside (Eardly, 1994).

The electrophoretic variation observed among the enzymes is related to allelic variation in the genes coding for these proteins. Typically, the electrophoretic variation at 15 to 30 enzyme loci are considered and can provide information on the genetic variation within and among species (Eardly, 1994; van Berkum & Eardly, 1998). Genetic distance or relatedness is usually expressed as the proportion of loci at which dissimilar alleles occur (Selander *et al.*, 1986).

The applicability of this technique in the examination of *Rhizobium* species has been described in Eardly (1994). In a recent report describing *Sinorhizobium arboris* and *Sinorhizobium kostiense*, 13 enzyme loci were analysed (Nick *et al.*, 1999b). The *S. arboris* strains revealed six and the *S. kostiense* three distinctive multilocus genotypes. Other earlier reports on rhizobial MLEE studies have indicated that species of *Bradyrhizobium* and *Rhizobium* are extremely diverse in comparison to human bacterial pathogens (Martinez-Romero & Caballero-Mellado, 1996).

3. Genotypic Methods

3.1 Determination of the DNA base ratio (moles percentage G+C)

The determination of the moles guanosine and cytosine is considered one of the classical genotypic methods forming part of the standard description of bacterial taxa. Among procaryotes it ranges between 24% and 76% and within well-defined species and genera the range is not more than 3% and 10%, respectively (Vandamme *et al.*, 1996). Differences in the moles % G+C are taxonomically useful for separating groups, however, similarities in base composition do not necessarily indicate close relationship, since the linear sequence of bases in the DNA molecule is not considered (Rosello-Mora & Amann, 2001). Conversely, organisms with widely different base composition will have few DNA sequences in common and are likely to be distantly related. Most bacterial genera have, however, comparatively narrow ranges of G+C values (Austin & Priest, 1986). It should be noted that estimates of

G+C content must be treated with caution since variation between laboratories have been observed and is therefore always a mean value (Logan, 1994).

The G+C content of DNA may be determined by various methods which exploits the physical, chemical and optical properties of DNA. The most common method is the optical tracking of denaturing of DNA. This denaturation is associated with a higher absorbance (approximately 40%) at 260 nm. The thermal denaturation midpoint (T_m) depends on the DNA base composition and is therefore an important taxonomic feature (Grimont, 1988). This midpoint temperature increases with increased mole % G+C. When T_m is determined the mole % G+C can be determined by an established empirical formula:

$$\text{mole \% G+C} = 2.44\ T_m - 169.3 \text{ (Austin \& Priest, 1986)}$$

Although a less popular method, CsCl gradients can also be used to determine the mole % G+C. In this instance the density of DNA is exploited since DNA density increases linearly with the mole % G+C. The method requires ultracentrifugation for a long period of time and reference DNA of known mol % G+C. Other methods for determining mole % G+C include DNA bromination, comparative absorbance ratio determination and the release of DNA bases by acid hydrolysis and subsequent chromatographical separation (Johnson *et al.*, 1985).

Species of *Rhizobium* usually have G+C values in the range 59 to 64 mole %, while in *Azorhizobium* the range is 66 to 68 mole % and *Bradyrhizobium* strains have an intermediate value of 61 to 65.4 mole % (Graham *et al.*, 1991). *Allorhizobium* (de Lajudie *et al.*, 1998a), *Sinorhizobium* (de Lajudie *et al.*, 1994) and *Mesorhizobium* (Jarvis *et al.*, 1997) have mole % G+C values of 60.1, 60.8 to 65.7 and 59-64, respectively.

3.2 DNA-DNA hybridisation studies

Traditional typing methods represent information of no more than 10% of the genome. In contrast, methods involved in determining DNA relatedness represent considerably more information since a larger portion of the genome is examined (Vandamme *et al.*, 1996). A characteristic of DNA and RNA is its ability for hybridisation or reassociation. Under standardised conditions, DNA from different organisms reassociate depending on the similarity of their nucleotide sequences. This allows quantification of the degree of relatedness and is usually expressed as % similarity or homology (Rosello-Mora & Amann,

2001). The percent DNA binding or DNA-DNA hybridisation value is an indirect parameter of the sequence similarity between two entire genomes. Genetically closely related organisms will have more nucleotide sequences in common and therefore a higher degree of nucleotide binding will occur (Vandamme *et al.*, 1996). Based on DNA relatedness, genomic species (or genetic species) is defined as a group of strains showing homology of 70% or more under optimal hybridisation conditions and with 5 °C or less ΔT_m (Wayne *et al.*, 1987). A later description of circumscription of the genomic species was made by Grimont, (1988): strains showing 80% reassociation at optimal temperature with divergence below 5 °C belong to one genomic species and that strains showing less than 60% reassociation and more than 7 °C divergence do not belong to the same genomic species.

Many different procedures have been developed to measure DNA similarity, and can be of two types; immobilised DNA or free solution renaturation. Most of these methods require radioactively labelled reference DNA. In the immobilised assay, single stranded (ss) DNA is immobilised on a nitrocellulose filter, incubated in the presence of labelled DNA from reference organisms and the amount of reassociation is estimated by measuring the radioactivity on the membrane. In this estimation, results from heterologous reactions (involving DNA from different strains) are also included (Grimont, 1988).

In the free solution approach, DNA hybridisation relies on the renaturation rates determined spectrophotometrically at 260 nm. Rates of DNA reassociation between test and reference organism and of each DNA separately are monitored by falls in absorbance at 260 nm (Logan, 1994).

In other methods a small amount of sheared, radiolabelled denatured DNA and larger amounts of sheared unlabelled DNA are mixed in a 1: 500 ratio and allowed to reassociate under optimal conditions. Reassociated fragments are separated from non-reassociated fragments by either hydroxyapatite (HA) chromatography or selective digestion of single stranded fragments by S1 nuclease. Both single and double stranded DNA adsorbs to HA, and can be selectively eluted by raising the buffer molarity. The radioactivity of the single-stranded and double-stranded fragments is measured and the percentage reassociation calculated (Grimont, 1988).

The S1 nuclease procedure is performed under conditions which prevent or reduce the digestion of double-stranded DNA, allowing digestion of nearly all ss-DNA. Half of the hybridisation mixture is treated with S1 nuclease, the remaining double stranded molecules precipitated and the percentage reassociation calculated by comparison with the untreated sample.

Recently, Christensen *et al.* (2001) developed a micro-well DNA-DNA hybridisation assay. This method was aimed at reducing the time and labour used to perform such hybridisations. Briefly, the method entails the covalent binding of mechanically-sheared DNA to the micro-wells and addition of photo-activatable-biotin-labelled reference DNA. The amount of biotin-labelled DNA bound to wells after stringency washes can then be determined by the addition of chemicals, which react with the biotin label, to generate a fluorescent signal.

3.3 DNA-RNA hybridisation

Bacterial cells contain several classes of RNA, including mRNA, rRNA and tRNA. Most comparative RNA hybridisation studies have been performed with either 16S and /or 23S rRNA molecules. Additionally, these two molecules account for 80% of the nucleic acid in a bacterial cell, which can readily be isolated. rRNA is now generally accepted as a target for studying phylogenetic relationships since it is present in all bacteria, functionally constant and consists of conserved and variable regions (Woese, 1987; Stackebrandt & Goebel., 1994).

In DNA-rRNA hybridisations, sequence homology between labelled 16S or 23S rRNA from a reference strain, and the rRNA cistrons within the chromosomal DNA from a second isolate is determined. The usual approach is to expose immobilised, denatured chromosomal DNA to labelled rRNA. Following the removal of unbound labelled rRNA and RNase treatment, the thermal stability of the hybrid is tested by subjecting the filter to a series of temperature increases up to 95 °C. The radioactivity eluted at each step is measured as the hybrids are denatured to give values of thermal stability (Austin & Priest, 1986; Logan, 1994). One drawback of this approach is that the G+C content of rRNA (52-54 mole % range) requires high hybridisation temperatures to be used. Such high temperatures can cause thermal degradation of the rRNA. Using formamide alleviates the problem of RNA degradation since lower temperatures may be used (Johnson, 1985). Furthermore, duplicating hybridisation

conditions between different laboratories is difficult, making correlation of data almost impossible.

In 1988 Grimont described the relationship between DNA-DNA and DNA-RNA reassociation. It was clear that these two methods covered different domains of relatedness; DNA-DNA hybridisation being useful for species delineation where DNA-RNA lacks accuracy. On the other hand, DNA-DNA hybridisation is insufficiently accurate when organisms are distantly related, whereas DNA-RNA reassociation is fully able to determine these distant relationships.

3.4 DNA based typing techniques

DNA-directed typing methods have improved substantially during the last few years. Their suitability as a taxonomic tool includes their universal applicability, reproducibility and is relatively easy to perform.

First-generation DNA-based typing methods included restriction fragment length polymorphism (RFLP) analyses of the whole genome and plasmid DNA. The former method entails the digestion of genomic DNA with restriction endonucleases, electrophoresis and visualisation of the DNA fragments. In general, these patterns are very complex and difficult to compare. The complexity may be resolved by selecting low frequency cutting restriction enzymes or southern blotting with specific probes. Since the fragments generated with low frequency cutting enzymes are usually very large, pulsed–field gel electrophoresis (PFGE) is employed instead of conventional agarose electrophoresis (Vandamme *et al.*, 1996). Studies by Gordillo *et al.* (1993) and Tenover *et al.* (1995) have shown PFGE to be a highly discriminatory typing method. Alternatively, these complex DNA patterns may be transferred to membranes and hybridised with a specific labelled probe. Typically rRNA probes are used, which may be 16S rRNA, 23S rRNA or both (Vandamme *et al.*, 1996).

The introduction of the polymerase chain reaction (PCR) methodology (Mullis & Faloona, 1987; Saiki *et al.*, 1988) has led to the development of numerous typing methods. These PCR-based techniques have attracted much interest because of their universal applicability, simplicity and speed with which it may be carried out (Vandamme *et al.*, 1996). Subsequently, the approach to determine relationships based on restriction endonuclease sites,

was extended to target specific genomic regions amplified by PCR. Both whole genome- and amplified gene RFLP analysis have several limitations. In both instances sequence divergence is estimated based on information from only a small portion of the genome and may not be representative of the sequence divergence across the entire genome. If southern hybridisation is performed, the applied probes may react poorly, if at all, with distantly related lineages. Furthermore, fingerprint patterns may also be influenced the presence of polymorphic insertion sequences (van Berkum & Eardly, 1998).

3.4.1 DNA fingerprinting patterns generated by PCR.

(a) Random whole-genome analysis

Whole genome analysis relies on the presence of repetitive elements, which are targeted by a number of PCR-based techniques. These include RAPD, ERIC-, BOX- and REP-PCR, and the analysis of restriction enzyme sites on the genome by PFGE and AFLP. All of these techniques recognise random sites on the genome, which cannot be predicted without the whole genome sequence.

(i) Random amplified polymorphic DNA assay (RAPD)

The RADP assay, also known as arbitrary primed PCR, was first described by Williams *et al.* (1990) and Welsh & McClelland (1990). In this technique short random sequence primers are used to initiate amplification of random regions on the bacterial genome. Since the number and site of these random priming sites differ for different strains of a bacterial species, separation of such amplification products, using agarose gel electrophoresis, will result in a banding pattern characteristic of a particular strain. However, in most cases the sequences of the RAPD primers and reaction parameters, which generate the best DNA banding pattern, need to be determined empirically (Olive & Bean, 1999). According to Vila *et al.* (1996), the RAPD assay is more discriminatory than RFLP of either the 16S rRNA genes or the 16S-23S ITS region, but less discriminatory than REP-PCR. The major drawback associated with this technique is its sensitivity; slight changes in reaction conditions and reagents make standardisation extremely difficult (Olive & Bean, 1999).

(ii) Analyses of interspersed repetitive elements (rep-PCR)

Bacterial genomes contain various repetitive DNA elements: the repetitive extragenic palindromic (REP) element, the enterobacterial repetitive intergenic consensus sequence (ERIC) and BOX elements located within the intergenic regions. The exact function of these highly repeated and conserved elements remains unknown. Although their involvement in various cellular functions, such as mRNA stabilisation and homologous recombination have been suggested, no single function has emerged explaining their conserved and ubiquitous nature (de Bruijn, 1992).

The REP sequence consists of a 33 bp sequence which is found in approximately 500 copies dispersed around the chromosome of *Eschericia coli* and *Salmonella typhimurium* (Gilson *et al.*, 1984; Stern *et al.*, 1984). The 126 bp ERIC sequences are present in many copies on the genome of many enterobacteria; to date there is no evidence of it being present outside gram-negative enterobacterial species (Versalovic *et al.*, 1991; Hulton *et al.*, 1991). The BOX element is an inverted repeat element initially found in *Streptococcus pneumoniae*. They are mosaic repetitive elements composed of various combinations of three subunit sequences referred to as boxA, boxB and boxC (Martin *et al.*, 1992). Additionally, these BOX elements have now been found in a number of other bacterial species and have no sequence similarity to the other two repetitive elements (Olive & Bean, 1999).

Comparative sequence analyses of the REP and ERIC elements have led to the development of oligonucleotides targeting these regions (Versalovic *et al.*, 1991). These were employed to detect the presence of REP- and ERIC-like sequences in a number of eubacteria. Surprisingly these elements were found in a large variety of bacterial genera, preferable gram-negative bacteria as described by de Bruijn (1992). A pilot study by the same author showed the presence of these elements in four genera of the family *Rhizobiaceae*. The suitability of this method for classification purposes was also investigated. The results showed that these elements were highly conserved within this group and could indeed be used to distinguish closely related rhizobia. Subsequent studies by others, such as Vinuesa *et al.* (1998) led to similar conclusion. Another study of 51 fast-growing Sudanese and Kenyan rhizobial isolates by Nick *et al.* (1999a) found that rep-PCR fingerprinting could be used as a first method to rapidly classify rhizobial strains of unknown taxonomic status.

These repetitive elements have superior discriminatory abilities over those of restriction analysis of the 16S rRNA and 16S-23S spacer region, MLEE and biochemical characterisations (Olive & Bean, 1999).

(b) Specific gene variation

Both single-locus and multilocus approaches have been used for molecular typing of bacterial isolates. The single-locus approach includes highly variable genes usually implicated in causing disease such as the neurotoxins of *Clostridium botulinum*. In contrast, the multilocus approaches include multi-locus sequence typing, the analysis of multi-gene families such as *rrn* operons and tRNA genes (Gürtler & Mayall, 2001). In 1996 Gürtler & Stanisch reported on the content and order (5' to 3') of the *rrn* operon as being: 16S rRNA, spacer, 23S rRNA, spacer and 5S rRNA sequences. This arrangement is universal for most bacteria with exceptions reported by Gürtler (1999). Some of these genomic regions, other novel regions and their specific application in polyphasic taxonomy will be discussed in later sections.

(i) The 16S or small subunit ribosomal RNA (ssu-rRNA) gene

The sequencing of the ssu rRNA gene has led to a better understanding of the relationship among the bacteria and may be considered the most useful and most used molecular chronometer (Woese, 1987). The usefulness of 16S rRNA sequences for classification purposes was summarised by Woese *et al.* (1985):

- The rRNA molecule is part of a large molecular complex central to the function of the cell, making transfer between species impossible. Phylogeny based on rRNA is therefore reflective of the true phylogeny of the whole organism.
- It is functionally constant to a degree rare among macromolecules. Consequently "non-chronometric" changes in sequence (i.e. those that are selected and so would distort phylogenetic analysis) occur rarely, making rRNA a particularly accurate molecular chronometer.
- The molecule is large and contains many functionally defined domains. Thus, the rare evolutionary redesign of one such domain (which involves non-chronometric, selective changes) amount to a smaller perturbation only on the molecule as a whole, which is not the case for smaller molecules, such as cytochrome *c* or the 5S rRNA.

Additionally, the ssu-rRNA molecules have both conserved and variable regions making them suitable for use in the analyses of closely related and more distantly related organisms.

However, the reliability of using the ssu rRNA gene alone for estimating phylogenies, has been questioned due to the conserved nature of this molecule. The applicability of the species definition, based on the ssu rRNA similarity in relation to DNA-DNA reassociation values, was investigated by Stackebrandt & Goebel (1994). These authors concluded that ssu rRNA sequence data are most reliable at similarity values lower than 97%, whereas at values higher than 97% DNA-DNA hybridisation provides a more reliable estimate. The 16S rRNA nucleotide similarity values (Table 3.1) among members of the *Rhizobiaceae* range from 88% to 96.3%

Young & Huakka (1996) highlighted some of the limitations of the ssu rRNA as a phylogenetic and taxonomic tool when they considered the diversity and phylogeny of rhizobia. The sources of the limitations were illustrated by the following cases:

- **Recombination among ssu sequences**

  The first 300 bases of *S. meliloti* are significantly more similar to that of *R. leguminosarum* than to the *M. loti* sequence which is more likely due to recombination events than to parallel evolution.

- **Heterogeneity within species**

  Certain genotypes of *R. leguminosarum* have been found to have a 73-nucleotide insertion in the first loop of the ssu rRNA.

- **Heterogeneity within genomes**

  Most bacteria have several copies of the rRNA genes. The cloning and sequencing of PCR-amplified ssu rRNA of *Sinorhizobium saheli* revealed two different sequences in the first stem-loop structure.

Keswani & Whitman (2001) investigated the relationship between 16S rRNA sequence similarity (*S*) and the extent of DNA hybridisation (*D*) in procaryotes. These authors found that it was possible to accurately estimate the distribution of *D* from *S* when the presence of nonultrametric rRNA sequences and differences between genera or families were controlled. The relationship between *D* and *S* varied between procaryotic taxa but was not significantly

different between procaryotic domains. The extent of DNA hybridisation also changes more rapidly for closely related organisms than for distantly related organisms. Therefore ranges of *D* assigned for intergeneric relationships should be interpreted with caution. The nonultrametric property of some rRNA led to lower *S* values than expected from the value of *D*. For such taxa, 16S rRNA sequence similarity was a poor indicator of the relatedness for closely related strains. Therefore, the ultrametric property of rRNA sequences should be tested before making taxonomic or phylogenetic conclusions based upon *S*.

(ii) The large subunit (lsu) or 23S gene

The lsu rRNA gene is used to a lesser extent for phylogenetic inferences. The topology of the lsu rRNA phylogenetic tree has been reported to be similar to that of the tree based on the ssu rRNA gene sequences (Ludwig *et al.*, 1995). However, this region is almost twice the size of the ssu rRNA molecule and contains highly variable stem-loop structures or intervening sequences which may be useful for classification and identification purposes (van Berkum & Eardly, 1998). Tesfaye *et al.* (1997) compared the partial 23S rDNA sequences from *Rhizobium* species. Comparative sequence analysis identified divergent regions that appeared to include characteristic species- or strain-specific sequences. This led to a later report (Tesfaye & Holl, 1998) of 23S-specific primers, which could differentiate between different effectiveness groups within the *Rhizobium-Trifolium* cross-inoculation group.

(iii) 16S-23S Intergenic spacer regions

The 16S-23S Intergenic spacer (IGS) region is not well conserved and exhibits a large degree of sequence, length and frequency variation (Massol-Deya *et al.*, 1995). This is particularly true for members of the α-subclass of the *Proteobacteria*, in which its length varies from 800-1200 nucleotides (Normand *et al.*, 1996). Comparing it to other genes of the *rrn* operon, the IGS region is at least twice as variable as the 23S rRNA gene and four times as variable as the 16S rRNA gene (Grundmann *et al.*, 2000). A detailed study by Gürtler (1999) showed that the IGS region was composed of highly conserved blocks of sequences while others were more variable. A more detailed analysis of these variations led the authors to conclude that these variable regions might have arisen due to recombination or selection events. Due to this variation, analysis of this region results in a higher level of discrimination, particularly between closely related species, and has been used for the identification purposes of a number

of bacteria. Comparative sequence analyses of the IGS region were able to provide high resolution of the relatedness between members of the genus *Nitrobacter* (Grundmann *et al.*, 2000). Similarly, an IGS-RFLP study by Laguerre *et al.* (1996) was able to indicate intraspecific polymorphism between different biovars of *Rhizobium leguminosarum*.

Table 3.1. Small subunit rRNA nucleotide similarity values among the genera of the *Rhizobiaceae*.

	Mesorhizobium	*Sinorhizobium*	*Agrobacterium*	*Rhizobium*	*Azorhizobium*	*Bradyrhizobium*
Allorhizobium (X67221)	93.1	94.2	96.3	93.5	88	86.4
Mesorhizobium (X67229)		95.9	93.6	94.4	91.4	88.1
Sinorhizobium (X67222)			95.3	96	91.2	88.3
Agrobacterium (D14500)				94.5	90.6	88
Rhizobium (U29386)					91.4	88.9
Azorhizobium (X67221)						90.3

The ssu rRNA sequences used were *M. loti*, *S. meliloti*, *Agrobacterium tumefaciens*, *R leguminosarum* bv. *vicae*, *Allorhizobium caulinodans* and *B. japonicum*. Adapted from van Berkum & Eardly (1998). Comparative values for *Allorhizobium undicola* were obtained from de Lajudie *et al.*, (1998a). GenBank accession numbers are indicated in parenthesis.

(iv) Other genomic regions used as phylogenetic markers.

Several other molecules have been used as suitable phylogenetic markers such as cytochromes, ferredoxins and azurins, ATPase, elongation factor and heat shock proteins. However, these are not as easy to handle as ribosomal RNA. The reasons are:

- The degeneracy of the genetic code prevents the formation of conserved regions necessary for the development of PCR amplification primers that could cover the full range of prokaryotic diversity; material to be sequenced must be obtained following time-consuming cloning steps.
- Not all of these proteins are ubiquitously distributed, restricting analyses to only partial phylogenetic trees.

- The occurrence of protein families, which make it difficult to decide whether proteins are truly homologous or paralogous (Stackebrandt & Rainey, 1995).

A few of these, not so widely used genomic regions, will be discussed briefly in the following sections:

❖ **Glutamine Synthetase II as a novel taxonomic marker.**

Glutamine synthetases (GS) are key enzymes in nitrogen metabolism and are ubiquitous, well-conserved enzymes. Together with glutamate synthase, GS is responsible for ammonium assimilation. Different GS's (designated GSI and GSII) differ with regard to their primary and tertiary structure and seem to have resulted from the duplication events. Very few prokaryotes have two GS isoenzymes, however, *Rhizobium* and *Agrobacterium* are among such genera. A proteomic study (isoelectric focussing and SDS-PAGE), genetic characterisation (southern blotting, PCR-RFLP) and immunochemical characterisation of the GS of various *Rhizobium tropici* and *Rhizobium etli* strains revealed the usefulness of GSII as a novel taxonomic marker. The GSII results support the separation of *R. etli* and *R. tropici* as *bona fide* species (Taboada *et al.*, 1996). An expansive study by Turner & Young (2000) showed that the GSI and GSII phylogeny of rhizobia are incongruent, while GSI phylogeny closely resembles that of 16S rRNA phylogeny.

❖ **Sigma factors of the σ^{70} family**

The sigma factor protein is a dissociable subunit of the eubacterial RNA polymerase holoenzyme conferring the promoter specificity required for transcription initiation. Two broad families of sigma factors have been identified: the σ^{70} and σ^{54}. The former (σ^{70}) is further subdivided into three groups:

- Group 1: the primary sigma factors, present in all known eubacteria and are essential for cell viability;
- Group 2: similar in sequence to the primary sigma factors, include stationary-phase-specific sigma factor RpoS, considered dispensable for cell growth.
- Group 3: varies considerably in sequence from the first two groups, although it is considered dispensable for the cell, this group includes functional groupings such as heat shock and sporulation sigma factors.

Some of the ideal phylogenetic characteristics of Group 1 sigma factors include: it is ubiquitously distributed among eubacteria, essential for cell viability, contains highly conserved structural features and a large number of sequences are already available. Sequence analyses of Group 1 sigma factors of various eubacteria showed similar resolution as obtained with other molecules, such as the ssu rRNA (Gruber *et al.*, 1997).

❖ **The elongation factor Tu and ATP-synthetase β-subunit genes**

Eubacteria, mitochondria and chloroplasts contain a proton-translocating ATP-synthase (ATPase) complex containing two portions: F_1 and F_0. F_0 is intrinsic to the membrane forming a proton channel. The other (F_1) is an extrinsic membrane protein complex composed of five subunits. The β-subunit contains the catalytic site of the complex and its primary structure is highly conserved and is an ideal macromolecule for deducing phylogenetic relationships (Amann *et al.*, 1988).

The elongation factor (Tu) is also considered a suitable phylogenetic marker. These molecules play a central role in the translation machinery that is different from the tasks of the rRNAs, though functionally connected. Consequently, elongation factor phylogenies are only partly useful for testing the validity of rRNA-based phylogenies.

Since the ATPases are functionally independent of the components of the translation apparatus, such phylogenies are useful to evaluate rRNA and elongation factor phylogenies. A comparative study by Ludwig *et al.* (1993) found that elongation factor data were in good agreement with rRNA trees, although the elongation factor trees showed reduced resolution at certain phylogenetic levels. On the other hand, the discriminatory power of the ATPase β-subunit was less than that of the elongation factor.

❖ **Transfer RNA's as genotypic fingerprints of eubacteria.**

Höfle (1990) described a new method for rapid genotypic classification of bacteria based on the high-resolution gel electrophoresis of low molecular weight (LMW) RNA fraction of bacterial strains. This fraction is composed of the total pool of transfer RNA and the 5S rRNA. Such an electrophoresis profile consists of bands belonging to different size ranges: 5S rRNA (110-131 nucleotides), class 2 tRNAs (82-96 nt) and class 1 tRNAs (72-79 nt). A

study of members of the family *Rhizobiaceae* by Velazquez *et al.* (1998) showed their LMW RNA profiles to be consistent with established taxonomic classification. The 5S profiles gave sufficient discrimination between the different genera of the *Rhizobiaceae*, while class 2 tRNA profile were species specific.

4. Conclusion

Considerable advances have been made in the field of bacterial systematics. To a large extent this is due to the development of molecular techniques. As discussed, each of these techniques has its specific level of discrimination and, as such, these techniques should not be applied on an exclusive basis, but rather complement each other. The choice of technique will, to a large extent, be dictated by the aims of a particular study; where diversity is the primary question one or two of these techniques will be sufficient. On the contrary, the work of a bacterial systematisist will be more extended to generate a consensus taxonomy.