Molecular Techniques in the Assessment of Botanical Diversity

ANIELA KARP*, OLE SEBERG† and MARCELLO BUIATTI‡

*IACR-Long Ashton Research Station, Department of Agricultural Sciences, University of Bristol, Long Ashton, Bristol BS18 9AF, UK, †Botanisk Laboratorium, Gøthersgade 140, DK-1123 København K. Denmark and ‡Universita degli Studi di Firenze, Dipartimento di Biologia Animale e Genetica, Via Romana, 17, 50125, Firenze, Italy

Received: 24 February 1996 Accepted: 6 April 1996

‘Majestic is the power of letters when their order’s changed
But atoms bring to bear wider variety,
And so create,
A rich and infinitely wide diversity of things’
Lucretius

A variety of different molecular techniques can be used for the study of botanical diversity. Restriction fragment length polymorphism (RFLP), arbitrary primed DNA, amplified fragment length polymorphism (AFLP), variable number of tandem repeats (VNTR), sequence-tagged simple sequence repeats (SSRs) and polymerase chain reaction (PCR) sequencing are briefly reviewed here. These techniques differ in the way that they resolve genetic differences, in the type of data that they generate and in the taxonomic levels at which they may be most appropriately applied. It is imperative to understand the different ways in which the data from the different molecular techniques can be utilized before embarking upon a programme of applying them to any particular diversity study.

© 1996 Annals of Botany Company

Key words: Biodiversity, RFLP, RAPD, AFLP, VNTR, SSR, molecular data, fingerprinting, genetic distances, phylogeny, conservation.

INTRODUCTION

The term ‘botanical diversity’ conjures up many different images. To some it brings to mind complex natural ecosystems, such as rain forests or grasslands, comprising diverse forms of different species and hybrids. Others may think of living collections of botanic gardens, dried specimens in herbaria or row upon row of containers of seed in gene banks. Examples of this kind present the molecular geneticist with contrasting problems in the assessment of diversity which may be investigated using a variety of different DNA technologies.

Traditional approaches to the measurement of diversity rely upon the ability to resolve differences in morphological characters. The range of characters available may be increased by the use of electron microscopy or biochemical and phytochemical assays. Although these approaches are extremely powerful, it is with the new tidal wave of data emerging from DNA-based techniques that this present review is concerned. An obvious advantage of molecular assays is the immense number of characters that they reveal. It would be erroneous to think, however, that because these characters are in the DNA, they are necessarily superior to those revealed by more traditional methods. In fact, molecular techniques vary in the way that they resolve genetic differences, in the type of data they generate and in the taxonomic levels at which they can be most appropriately applied.

MOLECULAR GENETIC SCREENING TECHNIQUES

In restriction fragment length polymorphism (RFLP), DNA is digested with restriction enzymes, the resultant fragments are separated by gel electrophoresis and blotted onto a filter and then probes are hybridized to the DNA. RFLPs give highly reproducible patterns but variations in fragment lengths between individuals or species can arise either when mutations alter restriction sites, or result in insertions/deletions between them (Burr et al., 1983) (Fig. 1A). Because heterozygotes are distinguishable, RFLPs are co-dominant markers. Their technical limitations are: (a) a good supply of probes is needed and, if heterologous probes are unavailable, cDNA or genomic DNA probes must be developed; (b) the blotting and hybridization steps are time-consuming and difficult to automate and; (c) sufficient quantities (e.g. 10 μg per digestion) of good quality DNA are required and RFLPs are, thus, not applicable where very limited amounts of source material or preserved tissue are available.

The development of the polymerase chain reaction (PCR) for amplifying DNA led to a revolution in the applicability...
of molecular methods and a range of new technologies were
developed which can overcome many of the technical
limitations of RFLPs. A subset of the latter involve the use
of a single ‘arbitrary’ primer, (obtainable from commercial
companies) and result in the amplification of several discrete
DNA products (Fig. 1 B). Each product will be derived from
a region of the genome containing two short segments with
sequence similarity to the primer, on opposite strands and
sufficiently close for the amplification to work. The most
common version is RAPD (randomly amplified polymor-
phic DNA) analysis, in which the amplification products are
separated on agarose gels in the presence of ethidium bromide and visualized under ultraviolet light (Williams et
al., 1990). AP-PCR (arbitrary primed PCR) (Welsh and
McClelland, 1990) and DAF (DNA amplification finger-
printing) (Caetano-Anollés, Bassam and Gresshoff, 1991)
differ from RAPDs principally in primer length, primer to
template ratio, the gel matrix used and in the visualization
procedure. The enormous attractions of these arbitrary
priming techniques are: (a) there is no requirement for
DNA probes or sequence information for the design of
specific primers; (b) since the procedure involves no blotting
or hybridizing steps, it is quick, simple and automatatable
and; (c) very small amounts of DNA (10 ng per reaction)
are required. It is absolutely critical, however, to maintain strictly constant PCR reaction conditions in order to achieve reproducible profiles.

The data derived from RAPDs (or AP-PCR and DAF) have their strength in distinguishing individuals, cultivars or accessions, although the difficulty of achieving robust profiles, particularly in RAPDs, makes their reliability for ‘typing’ questionable. The presence or absence of bands can be scored and the data converted into similarity matrices for calculation of genetic distances (Ellsworth, Rittenhouse and Honeycutt, 1993). RAPDs have also been used for population studies, molecular ecology (Hadrys, Balick and Schierwater, 1992) and for taxonomy. In using arbitrary priming procedures, it should be realised, however, that: (a) the markers are dominant and heterozygotes cannot be detected; (b) in the absence of pedigree analysis, the identity of individual bands in the multi-band profiles is not known and there can be uncertainty in assigning markers to specific loci; (c) the presence of a band of apparently identical molecular weight in different individuals cannot be taken as evidence that the two individuals share the same homologous fragment, although this assumption is commonly made and; (d) single bands on the gel can sometimes be comprised of several co-migrating amplification products.

More recently, a new PCR-based technique has been developed, termed amplified fragment length polymorphism (AFLP) (Vos et al., 1995), which is essentially intermediate between RFLPs and PCR. The first step involves restriction digestion of the genomic DNA, which is then followed by selective rounds of PCR amplification of the restricted fragments. The amplified products are radioactively or fluorescently labelled and separated on sequencing gels (Fig. 1D). AFLPs appear to be as reproducible as RFLPs, but they are technically more demanding and require more DNA (1 µg per reaction) than RAPDs. Because of their large genome coverage (on average they give 100 bands per gel compared with 20 for RAPDs), AFLPs are particularly good for mapping and fingerprinting and genetic distances can be calculated between genotypes. They do, however, share many of the limitations, with respect to band homologies and identities as outlined above for RAPDs. Furthermore, AFLPs are a recent technology, about which little new information is continually being revealed regarding the identity and distribution of AFLP bands in the genome. This is clearly related to the selected primers being used, but recent data indicate, for example, that AFLP bands may cluster on genetic maps around the centromeres. In using AFLPs for genetic diversity studies, researchers should thus keep a watchful eye on the latest discussions on this technology.

Another powerful technique for studying diversity utilises hypervariable regions of the genome comprised of tandemly repeated simple sequences. These repeats vary in number (and, hence, length) and are, therefore, generally called VNTRs (variable number of tandem repeats), although the terms ‘microsatellite’ (or simple sequence repeat, SSR) (Fig. 1C) and ‘minisatellite’ are used where the basic repeat unit is around two to eight base pairs in length, or longer, respectively. VNTRs may occur at many sites in the genome. When synthetic oligonucleotide probes containing repeats are hybridized to genomic DNA, multi-band profiles are usually produced. These VNTR fingerprints are generally highly discriminative and are often used to distinguish varieties, or even individuals, and reveal parentage and identity. The multi-locus profiles share many features with RAPDs and AFLPs. Features which are also common to combinations of these methods such as the use of VNTR primers in RAPDs, or SSR primers in AFLPs.

Although VNTRs may be present at several sites in the genome, at each of these sites the flanking sequences may well be unique. If individual VNTRs are cloned, primers to the flanking regions can be designed, thereby converting the VNTR into a sequence-tagged site (Fig. 1C). Minisatellites are more difficult to use in this way, due to their size. Retrieval of SSRs from new species has hitherto not been easy, due to their relative low abundance in plant genomes, but several methods for highly efficient isolation of SSRs are now available (Edwards et al., 1996). The highly mutable nature of SSR loci increases the possibility of having many allelic variants at each locus. Over 30 different alleles have been reported in some instances. Agarose gels can be used to distinguish alleles differing by several base pairs (i.e. several repeats) in length, but polyacrylamide gels are necessary if differences of a single repeat length are to be resolved. Sequence-tagged SSRs give highly reproducible profiles and are co-dominant.

The most precise and informative molecular data on diversity is obtained from sequencing known targets in the genome. Sufficient sequence information must be known initially in order to design primer pairs that will lead to amplification of only the target sequence. How precise this information has to be depends on how conserved the target sequence is. The amplified product can be compared on an agarose gel to the corresponding product from another individual, but only those differences in length that result from many base pair changes will be detected. A number of gel systems, such as TGE (thermal gradient gel electrophoresis) (Riesner et al., 1992), DGGE (denaturing gradient gel electrophoresis), single strand conformational polymorphism (SSCP) (Hayashi, 1992) and heteroduplex formation (White et al., 1992) provide sensitive detection assays of sequence differences without the need for complete sequencing, but they do require highly controllable conditions. In another alternative, termed PCR-RFLP, or CAPS, the amplified product is digested with a specific restriction enzyme and the products directly visualized on the agarose gel after ethidium bromide staining (Akopyanz et al., 1992). This simple approach is most informative when the restriction sites are mapped, rather than simply detected as RFLPs, and will not, of course, resolve all differences. Potentially, only sequencing the fragment, either manually, or using an automated DNA sequencer, will resolve all the possible differences between the samples. Sequences derived from protein-coding genes are usually straightforward to align and the interpretation of the data poses few problems. For non-protein-coding sequences (e.g. rDNA), however, one of the most difficult and least understood components of sequence analysis—sequence alignment—has to be confronted and data interpretations can, therefore, become very difficult. The PCR-sequencing approach is applicable.
to extremely small samples, down to a single pollen grain or tiny fragments and DNA can even be extracted from herbarium specimens or fossils (Herrmann and Hummel, 1992).

Plants possess three different genomes and, therefore, three potential sources of sequences for a PCR-targeted approach. The chloroplast genome (cpDNA) is uniparentally (often maternally) inherited in plants. It is highly abundant in leaves and therefore amenable to isolation in large quantities. The entire chloroplast DNA sequence is known for three species (a liverwort, tobacco and rice) and large quantities. The entire chlororplast DNA sequence is abundantly in leaves and therefore amenable to isolation in

**INTERPRETING AND ANALYSING DATA FROM DIFFERENT MOLECULAR SCREENING TECHNIQUES**

It imperative to understand the different ways that data from different molecular techniques can be utilized before embarking upon a programme of applying any one of the techniques to a diversity study (Hillis and Moritz, 1990; Soltis, Soltis and Doyle, 1992; Avise, 1994; Weir, 1996). The most important issue, by far, is whether one is working at or below the species level, i.e. whether the exemplars used in the analysis usually exchange genes or not. Molecular data are collected in two fundamentally different ways, viz. as discrete characters or as continuous characters. Whereas discrete data (e.g. sequence data) relates to individual specimens/species, continuous data (e.g. DNA-DNA-hybridization data) are quantitative pairwise comparisons between specimens or species. Moreover, discrete data can always be transformed into continuous data (distance measures), but the converse is not true. Data gathered by the methodologies described here are all discrete, although they are also of two different qualities. Shared bands (or restriction sites) are scored as presence/absence data, while sequence data are scored as nucleotides (only five permissible values of A,T,C,G and U). Some techniques, like TGGE, DGGE and SSR, are only short-cuts, by-passing sequencing.

At the intraspecific level, many problems, such as studies of mating systems, levels of heterozygosity and paternity testing, require discrete genetic markers and information from many individuals at many loci. Studies of geographic variation, gene flow, and hybridization also require discrete genetic markers of this type, but is best done with a more detailed knowledge of the inheritance of the markers, e.g. whether or not they are maternally or biparentally inherited (see, for example, Swoford and Olson, 1990; Queller, Strassmann and Hughes, 1993; Avise, 1994; Weir, 1996). The interpretive tools used at this level range from traditional probabilistic methods (e.g. given two similar banding patterns, what is the chance that they are the results of chance alone) to population genetics (e.g. to be meaningful the data are usually turned into allele frequencies), (Hartl and Clark, 1989).

At the above species level, data are used for the reconstruction of phylogenies and, although it is difficult to find justification for turning discrete data into distances, in principle, both continuous and discrete data may be employed. In the analysis of distance data a matrix of pairwise distances is subjected to a clustering algorithm, like UPGMA (unweighted pair group method using arithmetic averages), or neighbour-joining. Alternatively, the analysis of sequence data can be done using either parsimony based methods or maximum likelihood methods. The aim of parsimony methods is to produce the tree, or group of trees, that require the least amount of change to explain the observed data, while the aim of maximum likelihood
methods is to produce the tree that gives the highest probability of a data set being derived from it. For excellent reviews on this subject see Swofford and Olson (1990) or Forey et al. (1992).

The apparently sharp division between the intra- and interspecific levels is not, however, absolute. Algorithms have been developed that attempt to estimate phylogeny from gene frequency data (Swofford and Berlocher, 1987) and the presence or absence of an SSR, for example, is information of potential phylogenetic relevance. Moreover, gene phylogenies, as opposed to species phylogenies, can be made using standard phylogenetic methods provided that the gene in question is located in an unparentally inherited non-recombinating genome, e.g. cpDNA or mtDNA (see, e.g. Avise, 1994).

In general, the major theoretical drawback of methods that use the presence or absence of bands as data, is the fact that the characters are not independent. An RFLP-band that has disappeared (e.g. due to the evolution of a new restriction site) may, or may not, depending on the size of the new fragments, be detected and scored as two new bands. This has only limited consequences for recognition of patterns (e.g. in fingerprinting) but may have consequences for determination of gene frequencies, or in reconstructions of phylogeny. Additionally, there is a pronounced asymmetry in the probability of losing and gaining bands (Bremer, 1991; Albert, Mishler and Chase, 1992; Backeljau et al., 1996).

Seemingly, sequence data are straightforward to understand. In a set of sequences, the characters are the nucleotides sharing positional homology, viz. occurring at the same position. Positional homology is, however, not always easy to interpret. If the aim is to study species phylogeny rather than gene phylogeny, the most important condition that must be met is that the sequence must be orthologous, not paralogous, but even if that is the case, the presence of insertions/deletions may make the assignment of positional homology (alignment) highly ambiguous. The major drawbacks of distance methods, as they are used in reconstruction of phylogenies, are the facts that all sequence data are probably non-additive and that it remains unclear what the distances mean biologically. The main problem related to the use of maximum likelihood methods is the relatively limited number of taxa that they can effectively handle. Furthermore, the underlying evolutionary models may be too simple.

Sequences (and restriction site analyses) are the only molecular markers that contain a comprehensible record of their own history. Hence, appropriate analyses based on sequence data (or restriction site data) can provide hypotheses on the relationships between the different genotypic categories (or species) that they class together. Some would argue that this particular molecular approach is the only appropriate method for taxonomic studies, being superior even to all traditional approaches, but molecular data suffer the same problems (viz. problems of homology) as morphological data (see Patterson, 1988; Sanderson and Donoghue, 1989). The advantage that they offer is the almost infinite amount of data that can be obtained.

Frequency data from markers such as arbitrary primed DNA, AFLPs and SSRs only provide the means to classify individuals into nominal genotypic categories. This distinction is important to grasp for population studies, particularly when the diversity data are used as a basis for making decisions about conservation. In principle, sequencing will allow the determination of which gene sequences, in samples taken from within or between populations, are the most closely related and hence share a most recent common ancestor. For such genealogical relationships (which may be separate from the genealogy of the individuals carrying the genes) the influences of genetic factors, such as mutation, are independent from demographic factors such as population size, whereas, in the case of markers that provide only frequency information, these factors are confounded (for an excellent review on this subject, see Milligan, Leebens-Mack and Strand, 1994). This difference is of particular relevance to conservation, where demography (the description and prediction of population growth and age structure) is considered to be as important, if not more important, than genetic factors (Lande, 1988).

In using molecular techniques, it is important to recognise that all molecular diversity data are subject to experimental errors which differ depending on the technique. For example, assays dependent on restriction enzyme digestion are more robust than those that do not (although partial digestion may occur with phenolic and polysaccharide contamination); PCR reactions using single primers are more prone to irreproducibility than those using two primers; and so on. Many of the techniques are very new and there are few guidelines on the infinite number of ways in which unexpected results can arise. Reproducibility is the only easy way of assessing the quality of the data. Assessment of reproducibility is most evident when the data are in the form of a sequence (although, even here, it is possible to be misled) and very unclear when it comes to variations in band intensity. Although the probability of differences due to chance (or error) being generated can sometimes be estimated theoretically (e.g. the probability of Taq polymerase making an error in sequencing, see Koop et al., 1993) it is difficult to take account of all possible experimental variables. Most investigators use careful controls and may try to determine whether unexpected bands appear consistently or not, but few have time to determine why they appear. In practice, all researchers use their own judgement as to what is, or is not, significant. One only needs to observe how two investigators score the same RAPD gel to see evidence of this.

A further qualifying issue concerns the interpretation given to molecular data. At the intra-specific level, in particular, the rationale behind the development of molecular tools for screening diversity has generally been the search for polymorphic, ‘easy-to-handle’ markers, to overcome the main problems of detecting sufficient polymorphisms and of tackling large sample sizes. Molecular markers are thus largely used as arbitrary indicators of diversity. Often the markers are used without knowledge of whether they are located in single or repeated sequences, in coding or non-coding regions nor even, sometimes, whether they are of nuclear or organellar origin. The markers will also be heterogeneous in the frequencies of their different variations and, thus, their rates of evolution, all of which
introduces a potential bias in the interpretation of the variability observed. Whether this variation can be taken as reflecting that which influences future adaptation or individual fitness is debatable (Milligan et al., 1994). We have only questionable evidence as to whether the markers are neutral or not. Few molecular markers for fitness are known and there is little conception of what fitness means and little knowledge of how to measure it adequately (Bachmann, 1994).

As our knowledge of different marker systems advances, so our interpretation of their biological significance changes. The neutrality of mitochondrial markers has recently been questioned (Ballard and Kreitman, 1995). VNTR variation, which arises from changes in repeat numbers derived from slippage and/or unequal crossing-over has also been considered ‘neutral’. A series of human genetic diseases, such as the fragile X syndrome, have, however, now been attributed to changes in the number of SSRs (Sutherland et al., 1994). We wish to point also to long-range correlations in eukaryotes have been shown to exist, comprised of homopurine/homopyrimidine GC/AT containing stretches of variable length (Liò, Russo and Buatti, 1994) of SSR nature (Liò et al. unpubl. res.) suggesting that genome variation may also be under ‘internal constraints’.

CONCLUDING REMARKS

Molecular techniques do have much to offer for the assessment of botanical diversity, but potential users should recognise that this is an emerging, rapidly evolving field in which the technology has advanced faster than scientific understanding. Many exciting opportunities lie before those who are interpreting the data as we finally bridge the gap between the marker technology and Biological Sciences Research Council of the United Kingdom.

LITERATURE CITED


