Chapter 2

Review of Literature
2. Introduction

Information on genetic variation within and among populations is crucial for conservation of endangered species. A species may be considered as a group of individuals organized into populations that share an amalgamation of indicative characters which are not found outside the group. Survival chance of a species is indicated in genetic diversity within the population (Balakrishna, 1999; Tsuda et al., 2009). With greater variation in the genome, a species will be able to adapt to new selection pressure caused by environmental factors. Anthropogenic factors such as cattle grazing, agricultural practices, unscrupulous exploitation for medicinal uses, and natural calamities like flood and wild fire, have a detrimental effect on preserving genetic diversity (Aguilar et al., 2008).

Genes contain information regarding the body plan and traits of an organism. The gene pool, which comprises all the genes of the population, usually occurs in two or more slightly different molecular forms. Each of these forms is called an allele variation. This results in the population as individuals inherit different combinations of these alleles. A gene that exists as two or more alleles in a population is called polymorphic, whereas a monomorphic gene exists as a single allele in a population. By convention, a gene is considered monomorphic when a single allele is found in at least 99% of all instances of a gene. To be polymorphic, a gene has one or more additional alleles that make up at least 1% of the alleles in a population (Hartel and Jones, 1990).

Mutation creates new alleles. On an average, the rate of mutation is between $10^{-5}$ and $10^{-6}$ per gene locus per gamete in each generation. In other words, in a single reproductive season, one gamete in 1,00,000 to 10,00,000 has a new mutation at any
given locus. The allele frequency will stay through the generation if there is no mutation and the population is infinitely large and if it is isolated from other populations of the same species. Additionally mating should be random regarding the alleles, and all individuals should survive and reproduce equally (Hartel and Jones, 1990).

2.1. Molecular markers

Molecular markers have been used for assessing genetic diversity and generating baseline information (Pertoldi et al., 2007; Mirialili et al., 2009). In order to deduct a concrete, reliable data, a marker should meet many criteria; a marker should be inheritable and have the power to discriminate between individuals. It should be easy to generate and interpret. It should be highly polymorphic in nature and frequently distributed throughout the genome. A DNA based marker must have at least two alleles to be useful. Moreover, a marker should be easy to detect and comparable with similar characters (Hillis and Moritz, 1990).

Though there are no such ideal markers existing, three classes of markers based on morphological, biochemical and molecular traits are routinely explored. Morphological characters that are vegetative and floral are highly variable between individuals and often too plastic. Plant phenotypes such as floral features are seasonal in a plant species and can be influenced by changes in the environment such as uncontrolled temperature, drought, and salinity. Though morphological traits are easy to assess, they are not stable and are often reduced to patches of population (Brouat et al., 2001).
Biochemical markers such as allozyme, proteins and phytochemicals are secondary metabolites resulting from expression of genes. They become invaluable markers when there is a difference in the migration pattern of the allelic products during electrophoresis or chromatography. The isoenzymes (Hunter and Markert, 1957) that may not have a direct effect on phenotypic traits are the most explored biochemical markers. Isoenzymes are few in number, sensitive to environmental variations and the study is time consuming. It is also limited to those plant species that produce it in quantifiable amount. Though the small number of loci sampled by the technique limits its analysis, isoenzymes have been extensively used for genetic studies (Quiros and McHale, 1985).

DNA sequence-based molecular markers were introduced in 1970s, and they became the choicest method when PCR technology was invented in 1985. Unlimited number of markers can be developed from the total genome of a species since these markers represent a specific region of DNA within the genome. They are abundant, and exhibit no epistatic effects. DNA markers exhibit a much higher level of allelic variation than morphological and biochemical markers. Many characters are attributed to a suitable molecular marker. It must be polymorphic, co-dominantly inherited, easy and cheap to detect, and reproducible (Weising \em{et al.}, 2005).

### 2.2. The size of the plant genome

Plant nuclear genomes range from 125Mb (bryophytes, crucifers, roses) to more than 70,000 Mb of DNA (angiosperms). It has 5-20% tandem repeats (170–190nt long), mostly occurring in sub-telomeric and centromeric regions. The ribosomal genes (a 10,000nt long 45S rDNA unit which includes the 18S, 5.8S and 28S rRNA genes),
intergenic spacer, and microsatellites also present as tandem repeats in the plant genome, are explored as markers. The length of the core repeat unit, the number of repeat units per locus, and the abundance and distribution of loci distinguish these markers. Tandem repeats are often called ‘satellites’ of DNA. Microsatellites (Variable Number of Tandem Repeats loci), which occur commonly in plants, are used as markers (Rogstad et al., 1988; Sutherland et al., 2008) for studying variation between and within species (Rogstad, 1993; Yamagishi et al., 2008). A single copy DNA generally flanks these motifs and varies in copy numbers between different accessions of a plant. This region is often conserved and can be used to design primers to sequence segments of the genome being studied. Regions of retrotransposons which account for more than 90% of the entire genome are also widely used as molecular markers for DNA diversity studies (Heslop-Harrison, 2000).

The plant genome can also be studied by sequencing one end of the mRNA molecules by a methodology called Expressed Sequence Tags (ESTs). ESTs can be located by hybridization and can be used to classify and compare all the genes in a species. The greatest advantage of ESTs is that large amounts of repetitive DNA need not be sequenced to arrive at a conclusion (Schmidt and Heslop-Harrison, 1998). Some of the nucleic acid fragments that are used as markers occur once in a genome, others are repeated. Many of the repeated sequences used as markers are in the noncoding region and others are part of multigene families. Some of the repeated sequences are interspersed throughout the genome either distributed randomly or in clusters (Rogowsky et al., 1992; Gielly and Taberlet, 1994; Won and Renner, 2005). Moreover, the angiosperm chloroplast that has two copies of a large inverted repeat separated by a short single copy region, is also explored as a molecular marker.
2.3. DNA based molecular markers

DNA based molecular markers have prominence over morphological and biochemical markers mainly because their number is almost unlimited and very reliable. All techniques to identify DNA markers are based on three technologies, namely, DNA-DNA hybridization as in RFLP and AFLP, PCR technique as in RAPD, SNP, ISSR and the method of DNA sequencing. These techniques have been used to characterize the genome and to study genetic variability, genome fingerprinting, genome mapping, gene localization, and population and evolution studies for solving taxonomic ambiguities in plant breeding and in diagnostics (Murphy et al., 2003).

2.3.1. Hybridization based technologies

This is based on the presence of complementary strands of nucleic acids. Mismatching in hybridization can be controlled by stringent conditions such as altering temperature, salt concentration, formamide concentration, and the duration of hybridization (Dowling et al., 1990). RFLP is one of the earliest methods used for detecting polymorphism based on hybridization technology (Botstein et al., 1980). In this method the genomic DNA is digested with a restriction enzyme to generate fragments of various lengths as recognition sites are distributed randomly throughout the genome. These fragments are separated by electrophoresis and then made single strand by denaturation. The single strands are fixed on a nitrocellulose membrane and hybridized with a single stranded probe that is species specific to recognize a single locus. They are generally obtained from a cDNA library or a genomic library. Variation in the characteristic pattern of a RFLP digest is caused by base pair deletion, mutation, inversion, translocation and transposition which result in the loss or gain of a recognition site, thereby resulting in a fragment of different length and polymorphism.
In plants, small fragments’ sizes are obtained by using methylation sensitive restriction enzyme, *Pst* I (Swati *et al*., 1999). This produces single or low copy hybridization pattern. The greatest advantage is that the scoring of RFLP as single band is unambiguous and that allelism/co-dominance can easily be recognized (Winter and Kahl, 1995).

RFLP are co-dominant markers. It means a heterozygote can be distinguished from a homozygote. RFLP gives complete genetic information at a single locus. It is the choicest marker in linkage analysis and in breeding, as it is possible to determine if a linked trait is present in a hetero or homozygous state in an individual (Winter and Kahl, 1995). RFLP has been extensively used for genetic studies and plant identification (Gebhardt and Salamini, 1992). Developing RFLP markers is expensive and time consuming. Only few loci are examined on each gel and only one out of several markers provides a polymorphism. Moreover, radioactive probes are required for detection, which makes it hazardous. It cannot be employed for identifying point mutation within the region at which they are detecting polymorphism. RFLP requires suitable probes and large quantity of pure DNA is needed to be digested (Winter and Kahl, 1995).

### 2.3.2. PCR based techniques

Polymerase chain reaction based techniques (Saiki *et al*., 1985) are a better alternative to hybridization techniques for various reasons. Primarily, PCR dependent techniques require less quantity of DNA and sample purity is not a major concern. It is a quick method without the requirement of radioactive labelling. Since primers are employed for amplifications, these can be termed as ‘sequence tagged sites’ (STS). Specificity
and reproducibility of PCR depends on a number of factors such as the concentration and quality of the reaction ingredients, primer design, GC content of the primers, number of cycles, temperature and duration. Many of these disadvantages can be streamlined and an optimum condition can be worked out (Westman and Kresovich, 1997).

Depending on the target sequence information, primers have to be designed in PCR based marker assays. Arbitrary primers can be designed if sequences are not known. In case of sequence tagged sites (STSs) (Olson et al., 1989) presence or absence of the fragment, or the difference in fragment length can be assessed to generate data. This is vital in diploid organisms as the length of the fragment can be assessed as co-dominant alleles.

### 2.3.2.1. Microsatellites

Microsatellites or Single Sequence Length Polymorphisms (SSLPs) are short repeat units of nucleotides present throughout the genome. Though they are not more than 6 bases long, in plants these repeat units per locus may reach up to 50 (Saghai et al., 1994; Kofler et al., 2008). The most common repeats found in plants are (CA)n, (AT)n, (CT)n, (AAT)n, etc. Variation in microsatellites repeat units results from an error in DNA replication and unequal crossing over (Moxon and Wills, 1999; Varshney et al., 2005). A slip in the DNA polymerase when copying the repeat region changes the number of repeats. Variation also results from unequal crossing over, since they produce different repeat length and migrate different distances according to their sizes. Variation between individuals in mono, di or trinucleotide repeats present at a given locus in a single copy sequence is exploited in SSLP study. The repetitive regions are
amplified by priming from the flanking sequences. For developing primers, the target and flanking region sequences have to be known. Therefore, interspersed repeats or known sequences can be used to amplify unknown sequences between the repeats (Rogowsky et al., 1992). Any variation in the number of repeats within the locus is expressed in the length of the amplicon which can be determined in a PAGE (Di Vecchi-Staraz et al., 2009). When many interspersed repeat loci are amplified in one sample, each locus is evaluated separately. When tandem repeat loci are amplified, length polymorphism at a locus is assumed to reflect variations in the number of tandem repeat units. SSLP have a large number of alleles per locus. It assumes that the whole genome is sampled rather than one segment of it. They can be multi-allelic, as each SSLP can have a number of different length variants. The average length of SSLPs allele in plants is less than 300nt and can be amplified by PCR precisely (Condit and Hubbell, 1991; Altintas et al., 2007).

Tandem repeated genes and interspersed sequences are also used as markers (Long and David, 1980). The two internal transcribed spacer regions (ITS1 and ITS2) that separate nuclear ribosomal genes (18S-5.8S-26S) is the most commonly sequenced locus (Alvarez and Wendel, 2003; Carolan et al., 2006) for species level plant systematic studies. These regions and spacers form a unit which is tandem repeated hundreds of times, at several loci in the genome. The individual repeat units are separated by non-transcribed intergenic spacer (IGS) regions at each of these loci. The IGS region also has tandem copies of short sub repeat sequences. These gene clusters are very versatile molecular marker for phylogenetic analysis (Carolan et al., 2006).
Microsatellites are simple, hyper variable, highly reproducible co-dominant markers, showing Mendalian inheritance that can easily detect heterozygosity. They can be automated and are highly informative as they reveal a large number of polymorphisms. As the flanking regions of microsatellites are often conserved, they can be used for designing primers and to construct genomic libraries (Edward et al., 1996; Gimenes et al., 2007). A large number of public SSLP primer pairs are available for exploration and the effective cost per genotype and primer is low. Since it is a PCR-based method, DNA from dried leaves can be used for assay. They are a variant of the RAPD technique, except the high annealing temperature which makes it a more accurate technique. Microsatellites detect variations at individual loci when a combination of loci is examined. They are useful in DNA fingerprinting, population genetics, linkage map construction, genotyping, gene based cloning, diversity study, breeding system exploration, gene flow experiments, cultivar identification, crop improvement programmes, in paternity testing, and phylogeographic studies to assess the distribution pattern (Condit and Hubbell, 1991; Akkaya et al., 1992; Gupta et al., 1996; Varshney et al., 2005; Sharma et al., 2008; Tsuda et al., 2009).

2.3.2.2. Random Amplified Polymorphic DNA (RAPD)

Random Amplified Polymorphic DNA (RAPD) is based on the random amplification of sequences using primers that are 10 bases long (Williams et al., 1990). The primer anneals at several priming sites to the template at complementary sequences in both ‘sense’ and ‘antisense’ orientation, and the regions between primers in opposite orientation are amplified. However, geometric amplification occurs only in those regions in which the 3’end of the annealed primers face one another or opposite strands and not more than 3kb apart. In other words, the primer annealing sites must be
inverted repeats. Thus RAPD scans a genome for these small inverted repeats and amplifies intervening DNA sequences of invariable length.

A primer generally produces 1-10 fragments. If there is variation in the nucleotides of different sets of DNAs there will be changes in the priming sites causing presence or absence of bands which are generally polymorphic. The presence of a marker in more than one individual means that the individuals share the same sequence at the primer annealing sites and that these sites are separated by the same number of base pairs. The absence of a fragment in more than one individual does not indicate the same level of homology (Dangi et al., 2004). The failure to amplify may be due to changes in the annealing site, or increased levels of competition from other amplified fragments. Therefore, the presence of polymorphism is far more informative than its absence (Skroch et al., 1993). However, this can lead to mis-scoring since the failure to amplify is interpreted as the absence of an allele. As homozygous bands cannot be distinguished from heterozygous, the presence of a fragment is considered dominant over absence. RAPDs are dominant markers, i.e., profiles are scored for the presence or absence of a single allele. Dominant markers are not affected by the conditions under which the plants are grown or the developmental stage of the tissue from which the DNA is extracted. Though RAPD evaluates many loci simultaneously, it provides very little genetic information about each locus.

One of the greatest concerns about RAPD is its reproducibility. The reproducibility in RAPD using the same target DNA and primers have been investigated in different laboratories. It was found that most RAPD markers are reproducible, with difference between PCR machines accounting for most of the variations (Penner et al., 1993).
This reproducibility problem is usually the case for bands with lower intensity. This may arise from impurities in DNA preparation (Micheli et al., 1994). Other steps such as consistent reaction conditions, thermal profile during amplification and fixed DNA template concentration can ensure reproducibility. Magnesium chloride concentration above 2mM, primer to template concentration ratio, fluctuation in the concentration of reagents in PCR reaction, Taq polymerase above 2U/reaction also affect reproducibility. Annealing temperature below 36°C gives altogether different banding pattern (Penner, 1996). Pipetting error should be minimized. Moreover to ensure the result, the assay should be carried out two or more times and highest possible number of samples should be analysed to eliminate artefacts. A modified method termed high annealing temperature RAPD (HAT-RAPD) (Atienzar et al., 2000) relies on increasing the annealing temperature to 40-46°C and helps in generating reproducible profile.

RAPD has various advantages. It is an easy technique and does not require radioactive labelling as in RFLP, since the separation of bands is done on an agarose gel and visualised by ethidium bromide staining. It amplifies regions where no prior knowledge of sequence exists. This is a much faster and cheaper method than RFLP. Moreover, it uses nanograms of genomic DNA which reduces the damage done to a plant species. This is important if the plant belongs to an endangered species (Winter and Kahl, 1995).

2.3.2.3. Applications of RAPD

RAPD has been employed in various areas due to its simplicity and low cost. Some of the areas include biological control programmes (Wang et al., 2008), genetic mapping (Kiss et al., 1993; Hemmat et al., 1994; Sutherland et al., 2008), breeding programmes,
developing genetic markers linked to a trait of interest (Li et al., 2008) without the
necessity for mapping the entire genome. This has been successful in identifying
markers linked to disease resistance genes in tomato (Martin et al., 1991), lettuce
(Paran et al., 1991), fungi (Ananga et al., 2008), barley (Jansen and Schaffrath, 2009),
testing purity in hybrid seeds (Singh et al., 2007) and genetic diversity in a multitude of
species (Ram et al., 2008; Sharma et al., 2008; Jose et al., 2009; Orabi et al., 2009).

RAPD has also been used in a genome profiling strategy termed Bulk Segregant
Analysis (BSA) (Michelmore et al., 1991). In this method bulked DNA samples from
individuals that have the target trait/gene are compared to bulked DNA samples of
individuals lacking the trait. Markers that are polymorphic between the pools will be
genetically linked to loci determining the trait used to construct the pool. RAPD has
been extensively used in population and evolutionary genetics studies in spite of the
need for large samples of individuals from each population. This is essential if an
accurate estimate of allele and genotype frequencies is needed. Since RAPD is a
dominant marker, the estimated gene frequency is less accurate than those obtained
with co-dominant markers such as Allozyme and RFLP (Bardakci, 2001). Co-migrating
bands derived from the same primer is another problem with RAPD. But as the
population diverges, the co-migrating bands become less homologous. Therefore,
RAPD gives a better result between closely related species (Williams et al., 1993;
Megnegneau et al., 1993; Orabi et al., 2009). Species specific (inter specific) markers
unique to individuals from one species within a genus can be estimated by RAPD
(Arnold et al., 1991).
Economically important traits are polygenic in nature and quantitative. These traits are influenced by various external (environmental) and internal factors. The polygenic loci involved in their expression are termed Quantitative Trait Loci (QTL). In breeding programmes, RFLP markers were used to monitor useful QTL allele transmission from generation to generation (Soller and Backmann, 1983).

In order to develop methods that have better discriminating power between closely related genotypes variations in RAPD techniques were introduced. These techniques were also aiming at better reproducibility criteria. AP-PCR (Arbitrary Primed PCR) (Welsh and McCelland, 1990) using a single primer of 10-50 bases long and DAF (DNA Amplification Fingerprinting) (Caetano-Anollès et al., 1991) where a 5 base primer is used are techniques evolved with this understanding. DAF involves fragment analysis by PAGE and silver staining. This method is very useful in genetic typing and mapping. The AP-PCR fragments are similar to RAPD and they can be visualized by either ethidium bromide staining or autoradiography. Very often polymorphic fragments are cloned and sequenced to generate reliable markers (Paran and Michelmore, 1993). Based on the sequenced data longer primers can be designed to amplify a specific locus. These are termed Sequence Characterized Amplified Region (SCAR) markers which are usually dominant. They can be converted to a co-dominant marker by digesting them by restriction enzymes and deducing polymorphism by denaturing gradient gel electrophoresis (DGGE) (McCelland et al., 1994). During this process the double stranded DNA is denatured and then separated on a PAGE. Renaturation of the strands leads to distinctive secondary structures of the single stranded DNA. This structure is sequence determined and migrates differentially from
one another through the gel, and the discrimination between DNA fragments is more on the basis of their sequence than on their size.

2.3.2.4. Amplified Fragment Length Polymorphism (AFLP)

AFLP (Vos et al., 1995) is an extremely sensitive, technically demanding method with great stringency conditions. It combines the techniques of RFLP and PCR and produce complex banding pattern of different lengths. This method was developed to detect large number of DNA polymorphisms. AFLP has high resolution and it identifies polymorphism with great accuracy (Powel et al., 1996; Altintas et al., 2007). Therefore this technique has been currently used as a molecular marker to phylogeographic studies to explore geographical history, distribution pattern and genetic structure of many plant species (Toyama and Yahara, 2009; Tsuda et al., 2009). It is also explored to understand sex expression (Roy et al., 2008).

In AFLP the genomic DNA is digested with two restriction enzymes, one a frequent cutter (Eco RI) and the other a rare cutter (Mse I). Three types of fragments are formed from such a digestion: fragments that are cut on either side by a frequent cutter, fragments that are cut on one side by a frequent cutter and the other side by a rare cutter and finally, fragments cut on either side by the rare cutter. These fragments are then made amplifiable in a PCR using an adaptor to both the ends of the restricted sites. The primers used are short, with 6-8 bases. AFLP usually generates 50-100 bands and each band is assumed to come from a different area of the genome. On an average AFLP produces four times as many bands per reaction compared to RAPD. It produces more polymorphic loci per primer than RFLPs, SSRs or RAPDs (Gerber et al., 2000; Gimenes et al., 2007; Sutherland et al., 2008).
AFLP method does not require prior sequence information. AFLP can be used for large-scale population studies since it is based on PCR. Moreover, different genetic loci can be simultaneously analyzed per experiment. AFLP markers are considered to be distributed across the genome as each band is assumed to come from a different area of the plant genome. They are dominant markers as polymorphism is detected as either presence or absence of bands. Dominant markers are less efficient than co-dominant markers for population genetic studies. The greatest advantage is that they can be used for screening large number of cultivars which is beneficial for population studies (Sutherland et al., 2008).

One of the shortcomings of AFLP is poor amplification resulting from a partial initial digestion. Partial digestion occurs if DNA is methylated. Complex band pattern generated are difficult to standardize and score. Similar to RAPD, co-migrating bands may cause confusion in AFLP and accurate comparison between laboratories is almost impossible. However, AFLP is highly reproducible (Sutherland et al., 2008).

2.3.2.5. Single Nucleotide Polymorphisms (SNP)

Single Nucleotide Polymorphism (SNP) occurs when a base is replaced with another base at a particular nucleotide site. Such substitution arises due to point mutation and occurs in coding, noncoding, and intergenic regions between genes. Single nucleotide substitution results in numerous SNPs in every genome. Each SNP can have four alleles as there are four nucleotides. However, most SNPs exist in two alleles, namely, the original sequence and the mutated version (Gupta et al., 2001).
SNPs within a coding sequence will not necessarily change the amino acid sequence of the protein that it produces. This is due to redundancy in the genetic code. SNPs that are not in protein coding regions may still have influence on gene splicing, transcription factor binding, and the sequences of non-coding RNA. SNPs which are more abundant in plants than in human genome are estimated to occur once in every 100-300 nt in a genome (Gupta et al., 2001). For a variation to be considered a SNP it must occur in at least 1% of the population.

Though SNP detection involves huge expenses they are highly reproducible. However, it requires less quantity and high quality of DNA for analysis. SNPs do not always need a gel based system to analyse. They are very reliable in individual genotyping since they are the most abundant marker system. They are used in identifying traits of economical value, marker assisted selection, studying genetic diversity, crop improving, plant breeding, and variety identification. The amazing amount of sequence information will lead to SNP as the preferred marker in the future (Gupta et al., 2001).

2.3.3. Markers based on nucleic acid sequencing

Variation within nucleotides can be detected at very narrow levels by sequencing. Sequencing uses minimal amount of DNA to develop accurate results. Sequencing can also reveal whether the variation is due to substitution or rearrangement. Since this method involves money, labour and time a few loci of importance are sequenced for developing markers. Very specific genes such as phytochrome, heat shock proteins and nuclear rRNA gene families have been sequenced for marker generation (Soltis and Soltis, 1995). The sequence of chloroplast gene, $rbcL$ has been used as a marker in many phylogenetic studies (Clegg and Zurawki, 1992). The chloroplast DNA in plants
has higher mutation rate than the mitochondrial genome and thus can discriminate a species better than the mitochondrial genome (Aguilar et al., 2008). Internal transcribed spacers (ITS1 and ITS2), intergenic spacers, trnL intron (that are variable and polymorphic) (Humby and Zimmer, 1992) are often used for marker generation by sequencing. Though sequencing is considered the ultimate molecular tool in marker research, it has its pitfalls too. As each nucleotide is a separate character the sequence fidelity should be taken care of. So purity of nucleic acid, PCR conditions and Taq DNA polymerase that has proof reading capacity are essential to generate reliable sequencing data (Westman and Kresovich, 1997).

2.4. Chloroplast DNA

Chloroplast DNA (cpDNA) has an undisputable place as a marker in phylogeny and in population genetics. It is used to differentiate between closely related species. cpDNA is a circular molecule, is maternally inherited in angiosperms and lacks recombination capacity. It is much more stable than the mitochondrial genome and is highly conserved in size and structure (Shinozaki et al., 1986). It has two long inverted repeats which separate a large single copy region. The genes and their arrangements in cpDNA are conserved, which helps in designing universal primers to study phylogenetic and population genetics (Brouat et al., 2001).

The genome of the chloroplast has three functional categories: protein coding regions, introns, and intergenic spacers. The introns and intergenic spacers do not encode proteins and are referred to as noncoding regions. The highly variable regions of the noncoding regions are normally examined in genetic analysis, phylogeography studies, molecular systematics (Muir and Filatov, 2007), population genetics, and in species
identification by generating barcodes (Gielly and Taberlet, 1994). Gielly and Taberlet used cpDNA to resolve phylogenetic relationships when the sequences for the rbcL gene showed too little variation (Gielly and Taberlet, 1994).

Many of the amplified fragments in cpDNA are small which can be directly sequenced. If the PCR product is long the reaction must be followed by restriction digestion. The degree of universality in these cpDNA primers within the plant kingdom vary yet they are conserved enough to amplify any land plants and many algae (Taberlet et al., 1991). Since there are many copies of cpDNA per cell as compared to nuclear DNA sequences, the amplification is efficient and rapid. Therefore, this is a more attractive method than the traditional RFLP and RAPD for genetic marker identification (Collins et al., 2003).

2.4.1. DNA Barcodes

A DNA barcode uses genes to identify a species. It can be developed at various sites in a plant such as nuclear, mitochondrial and chloroplast sequences. However, the designed barcode should be in a region of the genome which is variable but conserved enough to design primers that can amplify short regions of 100-150bp. Since cpDNA evolves faster than mitochondria and shows considerable mutation rate, it is considered a better choice for developing a barcode for angiosperms (Clegg, 1993; Taberlet et al., 1991).

There are many controversies existing over the value of DNA barcoding. Taxonomists consider that the traditional morphology based identification of a plant species would diminish and result in incorrect species identification as cpDNA relies solely on genetic
divergence (Kress et al., 2005). Moreover, taxonomy of science is based on a detailed understanding of morphology, physiology and behavioral attributes (Balakrishna, 1999; Ebach and Holdrege, 2005; Arvind et al., 2007; Pandey, 2007) and barcoding generates information, not knowledge. Species identification should be based on multigene phenotype rather than a single gene sequence (Moritz and Cicero, 2004). Though DNA barcoding provides rapid species identification, its accuracy relies on PCR technology by using a standardized DNA region as a tag (Hebert and Gregory, 2005). The following are some of the criteria for an ideal DNA barcoding system:

1. It should be standardized with the same DNA region as far as possible for different taxonomic groups
2. The target DNA region should contain enough phylogenetic information to easily assign species to its taxonomic group (genus, family, etc) and reveal significant species level genetic variability and divergence
3. It should be extremely robust with highly conserved priming sites for developing universal primers, and highly reliable DNA amplifications and sequencing
4. The target DNA region should be short enough to allow DNA extraction and amplification of degraded DNA (Taberlet et al., 2006).

Chloroplasts code for 30 tRNA genes in angiosperms in order to synthesize proteins within the chloroplasts (Sugiura, 1992; Wakasugi, 1997). These tRNAs can read all codons using the wobble mechanism.
Genetic diversity can be analysed by using the highly variable trnL-F part of the chloroplast genome. This region consists of tRNA genes for trnT (UGU), trnL (UAA) and trnF (GAA) (Fig. 4). They are arranged in tandem and separated by noncoding spacer regions. It is located in the large single copy region approximately 8kb downstream of rbcL. The trnL gene contains a group I intron positioned between the U and the A of the UAA anticodon loop. The position of this intron, interrupting the anticodon of the tRNA-Leu (UAA) gene (U-intron-AA), is conserved from cyanobacteria to plant chloroplast (Kuhsel et al., 1990). This renders the intron in the tRNA Leu (UAA) gene a suitable candidate for developing barcodes. The trnL intron was the first group I intron described in chloroplast DNA and also the first one described to interrupt a tRNA gene (Bonnard et al., 1985; Simon et al., 2003). The gene encoding tRNA Leu (UAA)’s group I intron can be folded into a normal secondary structure (only with minor length variations in unpaired loop regions). Group I introns are mobile RNA enzymes (ribozymes), which are able to catalyse their own removal from pre-RNA. It also encodes conserved primary and secondary structures required for autocatalysis. Therefore, it not only has catalytic ability but also capacity for information storage.

Fig. 4 Positions and directions of universal primers (c and d) used to amplify the chloroplast trnL (UAA) intron region (Taberlet et al., 2006). c = Forward primer, d = Reverse primer
In plants, the \textit{trnL} intron usually shows sequence conservation in the regions flanking both \textit{trnL} exons (i.e, \textit{trnL} (UAA) 5’ exon and \textit{trnL} (UAA) 3’ exon, whereas the central part is highly variable (i.e \textit{trnL} intron). It is about 550pb sequence in \textit{Tylophora} species. Within the intergenic spacer, which is between \textit{trnL} (UAA) 3’exon and \textit{trnF} (GAA) exon, no secondary structural elements have been found that could serve as splicing points. This indicates that \textit{trnL} and \textit{trnF} are probably co-transcribed (Bonnard \textit{et al}., 1985). The structure has complementary regions that form nine stem-loop structures (P1-P9). There are three gaps in the \textit{trnL} introns of which two belong to the P8 stem and loop, and one is in the P9 element. Analyzing the secondary structure in the spacer region and the \textit{trnL} intron P8 loop are useful in studying phylogenetic relationships. However, to establish phylogenetic relationship at a generic level more variable regions in cpDNA, such as \textit{rps4}, \textit{trnT} and \textit{trnF}, should be explored further (Taberlet \textit{et al}., 2006).

The tRNA-Leu (\textit{trnL}) intron has many drawbacks. It has a low resolution compared with several other noncoding chloroplast regions. It has very low discriminating capacity for closely related species which is not highly suitable for phylogenetic studies. The low resolution of tRNA-Leu (\textit{trnL}) intron is therefore linked to a lower intraspecific variation and compared with other noncoding regions of chloroplast DNA (Shaw \textit{et al}., 2005).

In spite of the low resolution it has many advantages. tRNA-Leu (\textit{trnL}) intron is used for phylogenetic studies among closely related genera and species (Gielly and Taberlet, 1996) because universal primers are available for this region (Taberlet \textit{et al}., 1991). The evolution of tRNA-Leu (\textit{trnL}) intron has been analyzed in detail and is well
understood (Quandt and Stech, 2005). The number of tRNA-Leu (\textit{trn}L) intron sequences available in databases is already very high. This is the most available non-coding chloroplast DNA sequences in the database. Furthermore, this region is the only group I intron in chloroplast DNA (Palmer, 1991) and has a conserved secondary structure with alteration of conserved and variable regions (Quandt \textit{et al.}, 2004). Therefore, robust primers can be designed for amplifying short variable region in between the tRNA-Leu (\textit{trn}L) intron. Such short regions are the P6 loop of 10-143bp and P8 loop present within the \textit{trn}L region (Won and Renner, 2005). Primers developed for this region are highly conserved from Bryophytes to Angiosperms. Studies show the entire tRNA-Leu (\textit{trn}L) intron and the P6 loop amplified allow a much greater identification of species in comparison to \textit{psbB}–\textit{psbH}, \textit{rpoB}–\textit{trn}C (GCA), \textit{rps}16 intron, \textit{trn}D (GUC)–\textit{trn}T (GGU), \textit{trn}H (GUG)–\textit{prb}A and \textit{trn}S (UGA)–\textit{trn}M (CAU) primers. Another advantage of tRNA-Leu (\textit{trn}L) intron is that it does not represent the most variable non-coding region of chloroplast DNA. The tRNA-Leu (\textit{trn}L) intron shows polymorphic phylogeny in various \textit{Tylophora} species (Liede \textit{et al.}, 2002).

2.5. Applications of molecular markers

Looking at the history of the development of molecular markers one can appreciate the advancement of markers in the last two decades. It has evolved from an expensive, time consuming, elaborate technique to an automated, accurate, reproducible and reliable method. Markers also broadened their application from breeding programmes to localising genes, marker assisted selection, evaluating genetic diversity, solving taxonomic disputes, assessing genetic drift, phylogenetic analysis and mapping (Condit and Hubbell, 1991; Gupta \textit{et al.}, 1996; DeWalt and Hamrick, 2004; Varshney \textit{et al.}, 2005; Sharma \textit{et al.}, 2008; Tsuda \textit{et al.}, 2009; Verma \textit{et al.}, 2009). They have
generated a vast array of information in the recent past that helps in understanding and thus formulating policies for establishing a balanced ecosystem.

(1) Assessment of genetic variability and characterization of germplasm

Genetic variation has been reduced drastically in the cultivated species due to micropropagation, domestication of species and continuous selection pressure for specific traits (Li et al., 2008). Therefore it is important to study the composition of the existing germplasm and compare it with the initial and with related species. This information can help in understanding phylogenetic relationship (Belaj et al., 2007; Orabi et al., 2009), identifying useful traits and assessing genetic variability. Plant breeding programmes must incorporate sufficient genetic diversity. This allows the production of new varieties that have enhanced crop production and resistance to biotic and abiotic factors. Superior genotypes are developed from a cross by the measurement of genetic similarity or genetic distance between parents. Such studies provide information not only about the phylogenetic relationship, but also about the possibility of finding new, economically important alleles. Quantification of genetic diversity within closely related cultivars is essential for genotyping. Microsatellites that are repetitive short tandem repeats are ideal choice for such genotyping as they represent a single locus.

Various molecular techniques have been used for genetic variability studies. AFLP has been used with great success in those species where identifying molecular markers has been extremely difficult. Similarly ISSR markers were used for diversity studies of rice (Blair et al., 1999) and wheat (Nagaoka and Ogihara, 1997; Altintas et al., 2007).
(2) Marker assisted selection

Identifying alleles of economic importance such as pathogen resistance, stress tolerance and high yielding variety are vital in any crop improvement programme. These markers have been used for locating and modifying Quantitative Trait Loci (QTL) (Babu et al., 2004; Noel et al., 2007) to introduce new characters in existing germplasm. Markers associated with 40 traits of economic importance have been reported in wheat (Gupta et al., 1999) which helps in segregating a plant by a simple PCR early in its growth. Thus marker assisted selection helps in eliminating unfavourable alleles in the early stage of plant development. RAPD has been used to identify markers linked to disease resistance genes in cereals (Adamblondon et al., 1994) and other economically important crop species.

(3) Genome mapping

Plant breeding programmes rely on selecting the correct alleles. Plant breeders have been using the information of plant genome mapping for generating improved varieties in breeding programmes. They are able to identify the useful alleles and breeding can be done to generate plants with valuable alleles (Babu et al. 2004; Kumar, 2008). RFLP has been used in genome mapping. Once a framework map is generated, various other markers are generated by different techniques to be located in the map. Microsatellite markers that follow Mendelian inheritance are used for index mapping which provide a reference point for specific regions of the genome. Microsatellite maps have been developed for linkage groups in rice, soyabean, potato etc. (Carolan et al., 2006).

RAPD has been used in genome mapping of several plants species including conifers. It is extremely difficult to map conifers due to their large genome size (30-40 x10⁹nt)
and high proportion of repetitive DNA (Penner, 1996). RAPD has been used for mapping where RFLP has not been able to detect polymorphism. RAPD has often been converted to SCAR markers and used as diagnostic markers.

Mapping helps in tagging traits such as yield, drought resistance, stress tolerance, disease resistance, seed quality, etc., which are important for breeding. Mapped genes can also be used for identifying other useful genes in the new genotype generated in a hybrid programme. RFLP has been used for gene tagging and locating quantitative trait loci.

(4) Phylogenetic studies

Biologists have great interest in understanding species divergence and diversity associated with species formation. This is important to understand genetic variation. Until the arrival of DNA based markers, such phylogenetic studies were based on morphological, biochemical and geographical data. These markers can often be proved incorrect when a detailed analysis based on nucleic acids is carried out. RFLP has been used in such phylogenetic studies.

Transposon elements *tos1-1, tos2-1, and tos3-1* are found in rice. These transposon elements have been used to assess genetic differences among different rice cultivars across ecotypes (Fukuchi *et al.*, 1993). Techniques such as RFLP, microsatellites, AFLP and RAPD have been used to understand genetic variation and phylogenetic relationships among these rice cultivars. Gene pool variation and classification of germplasm have been achieved with the help of such studies. Though these markers give valuable information, expressed sequence tags (ESTs) are currently explored for
phylogenetic and evolutionary studies. Since phylogenetic studies require one to estimate the evolution of functional genes (Mason-Gamer et al., 1998), a trend towards exploring ESTs is understandable.

Having reviewed molecular markers at length, it is important to evaluate the same in *Tylophora* species in order to understand and assess the genetic diversity.

Three species of *Tylophora* are reported from various regions of Gujarat (Shah, 1978; D’Cruz, 2003). Ethnomedicinal demand and reduced forest cover have nearly wiped out the population of at least two of these species, namely, *Tylophora rotundifolia* and *Tylophora fasciculata* from the wild (D’Cruz, 2003). However, *Tylophora indica* is reported to be endangered from certain regions of India (Faisal et al., 2007). A detailed analysis of the genetic diversity based on molecular markers is vital for devising conservation and propagation strategies of these endangered species.

Tissue culture is the choicest method for large scale propagation of any endangered species due to speed, minimum space requirement, and multiplication of superior clones. Sharma and Chandel (Sharma and Chandel, 1992) established a protocol for auxiliary bud induction and shoot multiplication from *Tylophora indica*. Plant regeneration through somatic embryogenesis has also been explored by Jayanti and Mandal (Jayanti and Mandal, 2001). It was observed that plants derived through somatic embryogenesis enhanced the production of secondary metabolites.

Genetic variations in the clonally propagated plants were examined based on the expression of allozymes (Parani and Parida, 1997) and by RAPD (Parani and Parida,
1997; Jayanti and Mandal, 2001). These works assessed the genetic content and its distribution in different populations of *Tylophora indica*. They observed that the studied population was not expressing any significant structural polymorphism and the distribution of genetic content in the micropropagated plantlets were fairly uniform and generated plantlets of true-to-type nature.

In order to meet the need for the large scale propagation as well as for commercial production of tylophorine, Bera and Roy explored micropropagation by multiple bud formation from mature leaf explants without callus intervention (Bera and Roy, 1993). Though they could achieve 72% success in field trials of the micropropagated plants, there was no further report about their commercialization and wide spread cultivation. Faisal *et al.* developed a protocol for high frequency shoot organogenesis and plant establishment from stem derived callus of *Tylophora indica* (Faisal *et al.*, 2007). However, no such work has been reported about the other two species of *Tylophora*.

### 2.6. High Performance Thin Layer Chromatography (HPTLC)

Biochemical markers have their own significance and importance in chemical fingerprinting. Allozymes were the choicest biochemical markers in plants due to various strengths. However, modern and sensitive technologies for identifying markers based on gene expression, such as GCMS, HPLC, NMR, IR, and HPTLC, have replaced allozymes. A plant during its life span produces various phytoactive compounds as secondary metabolites for its own growth and survival. Identification and characterization of these active principles can be used in generating a species specific fingerprint (Anandjiwala *et al.*, 2007). Generating more than one class of molecular markers to estimate genetic diversity is highly warranted in endangered
Review of Literature

species (Esselman et al., 1999; Joshi et al., 2004). A genotypic marker in comparison to a phenotypic marker does not reveal the content, the active principle or the chemical constituent. Therefore, the current trend on chemotype-based fingerprint to support the genotype based molecular markers helps in the proper structuring of a species beyond any level of ambiguity.

HPTLC, a planar chromatography, has many advantages. It is fast, flexible, cheap and highly reproducible. As the technology is highly automated, the results are very reliable. Fractions from the generated chemical pattern can be purified and used for further analysis. Moreover, the choices of solvent system are innumerable and one has the freedom of modifying the compositions. Many samples can run parallel, which makes the technique fast. It has the possibility of multiple evaluation of the plate with different parameters because all fractions of the sample are stored on the plate. Moreover, HPTLC is a quantitative determination procedure.

HPTLC has a very low differentiation capacity as multitudes of active principles are generally present in all the individuals of a species (Anandjiwala et al., 2006). HPTLC has very limited alleles and loci available for study. Besides, only those gene products that are soluble in water, methanol or chloroform are used in chemoprofiling. Discrepancy may arise from the differences between tissues sampled and the environmental conditions under which sampling are done. Variation in chemical profile is highly subjected to extrinsic factors such as cultivation, harvesting, drying and storage conditions (Joshi et al., 2004). Therefore the marker based on secondary metabolites should be able to discriminate one species from another species, one accession from other accessions. Such markers may not necessarily be a therapeutically
important active principle. Any gene product that is neutral and unaffected by environmental changes can function as a phytochemical marker.

Considerable work has been carried out to correlate DNA markers with phytochemical compositions among closely related species. (Sugimoto et al., 1999; Baum et al., 2001; Yang et al., 2001; Fico et al., 2003; Li et al., 2003; Joshi et al., 2004). Merging of these profiles will certainly help in developing a comprehensive understanding of a species.

HPTLC uses a stationary phase which is a thin layer (0.25-2.0mm) of silica on a metal foil or a glass plate. The sample is applied with the help of an automated applicator as a thin streak. Sample is applied by spraying with the help of nitrogen gas. Since the mass distribution is uniform over the full range of the bands, densitometric estimation can be done by scanning. The plate is then developed in a saturated chamber containing the developing solvents. The separation of compounds may be based on adsorption, partition, chiral, ion exchange or molecular exclusion principles. The separation rates of compounds are based on their distribution coefficients between the mobile phase and the stationary phase. Once the mobile phase reaches the front end of the plate, retardation factor (Rf) is calculated as,

\[ Rf = \frac{\text{Distance moved by the analyte from the origin}}{\text{Distance moved by the mobile phase front from the origin}} \]

Densitometric estimation can be done by scanning in the spectral range from 190-800nm.
A plant may produce innumerable bioactive compounds (Cowan, 1999) such as phenolics, terpenoids, alkaloids, saponins and lectins. In most cases, these substances serve to defend the plant against predation by microorganisms, insects, and herbivorous animals. Certain terpenoids give plants their odour. Quinones and tannins are responsible for plant pigments. Terpenoids offers flavour to the plant. Some secondary metabolites can be pharmacologically active and antimicrobial agents (Hostettmann, 1999). *Tylophora* species is reported to have high levels of alkaloids, which are heterocyclic nitrogen containing compounds (0.2-0.46%) (Viswanathan and Pai, 1985; D’Cruz, 2003; Chandrasekhar *et al*., 2006; Mujeeb *et al*., 2009).

Due to its medicinal importance, attempts have been made to isolate and identify the active principles of various *Tylophora* species. Govindachari *et al*. isolated two alkaloids (Tylophorine and Tylophorinine) from *Tylophora indica* which are potential anticancer agents with anti-inflammatory properties. Structural analysis showed that these compounds possess a phenathroidolizidine skeleton (Govindachari, 2002). In order to meet the pharmacological need to design anticancer drugs these were chemically synthesised in large quantities. Studies established that tylophorine exists in two enantiomeric forms and the natural tylophorine was probably a racemic mixture in the ratio of 2:3 (Nagarajan, 2008).

To widen the future of drug designing based on the *Tylophora* alkaloids, other species of *Tylophora* such as *Tylophora dalzellii* and *Tylophora crebriflora* were explored. The details of these alkaloids and their extraction procedure have been patented by Rao (Rao, K.V. US patent no: 3,497,593). Tylophorines from plant extracts were explored for their anti-tumor activity particularly in leukemia in human derived tumour cells.
(Gellert and Rudzats, 1964). In an effort to increase their potency several structural analogs of tylophorines were synthesised. One such analog DCB-3503 had great potency against human derived tumour \textit{in vitro} and \textit{in vivo}. Further, DCB-3503 was found to be modulating cell cycle regulatory proteins and NF-B signalling pathway (Baker and Cheng, 2006). Wu \textit{et al.} proposed that tylophorine retard S-phase progression and arrest the cells at G1 phase in carcinoma cells. It was down regulating cyclin A2 which thus arrests the G1 phase of cell cycle in carcinoma cells (Wu \textit{et al.}, 2009). Structural derivatives from \textit{Tylophora sylvatica} were tested for anti allergic properties and found to be effective in suppressing allergic agents from basophils (Gnabre \textit{et al.}, 1994). However, these studies showed that the chemically synthesised compounds were not as potent as the natural products. It may be because these were structural analogs and may not be functional analogs (Gao \textit{et al.}, 2008). Udupa \textit{et al.} (Udupa \textit{et al.}, 1991) found that alcoholic extracts of \textit{Tylophora indica} have significant effects on the functional activities of the adrenal glands. It reduced plasma steroid, cholesterol and vitamin C levels in experimental rats.

Preliminary toxicology studies showed that consumption of \textit{Tylophora indica} extracts (>200mg/kg body wt) may produce side effects like drowsiness, giddiness, loss of taste for salt, mouth pain, stomach upset, temporary nausea and vomiting. Tightness in throat or chest, chest pain, skin rashes, or itchy or swollen skin may occur in some cases (Gupta \textit{et al.}, 1979; Malathi and Gomez, 2008). The methanolic extract induces cytotoxicity, inhibits protein synthesis and induces cell apoptosis in experimental rats. This extract was apparently safe in smaller doses to produce a therapeutic effect (Malathi and Gomez, 2008).
Chaudhuri *et al.* (Chaudhuri *et al.*, 2006) genetically transformed *Tylophora indica* with *Agrobacterium rhizogenes* to increase the production of tylophorine. Various other pharmacological derivatives such as tylophorinine, tylophorinidine, tylophoriside A a novel steroid were also isolated from other *Tylophora* species. (Huang *et al.*, 2002). However no such extractions of ethnomedicinally active substances have been reported from *Tylophora rotundifolia* and *Tylophora fasciculata*. Hence it opens the possibility of exploring these plants for various active principles. These plants have not been scrutinized until now as they are unavailable to the public for investigation. Therefore, the need of the hour is to propagate and make them widely available for further phytochemical studies.

2.7. Interpretation for the present study

There is a challenge for studying plant genetic variation despite having many advanced molecular techniques available to assess genetic variation. This begins with choosing the most appropriate technique to address a research hypothesis. Most methods currently in use are an upgradation of pre-existing technologies for assessing plant genetic variation.

The level of polymorphism that different methodologies reveal is also important. If it reveals too little variation, then it may not be possible to discriminate taxa. If the variation found is too high, then the relationship between the taxa is concealed. This results in inaccurate prediction of relationships.

Choosing an appropriate marker assay is also important. There are several considerations before opting for a marker assay, namely, what is the hypothesis we are
interested in checking, what should be the level of resolution needed in the marker, can the observed result be related to the characteristics of the taxa being studied, etc.

Future marker assays must be able to detect genetic variation with accuracy. They should be reproducible across laboratories without any ambiguity. They should be user friendly, precise, low cost and potentially automated. Understanding the plant genome calls for evaluating many individuals from many taxa. It means techniques must be developed to assess numerous samples in nanolitre volume within a short period of time.