# DNA BARCODING OF *VERBENACEAE* PLANT SPECIES USING ITS2 AND *RBCL* GENES

IMRAN BUKHARI

THIS THESIS IS SUBMITTED IN FULLFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR OF BIOTECHNOLOGY (HONOURS)

FACULTY OF HEALTH AND LIFE SCIENCES
INTI INTERNATIONAL UNIVERSITY
PUTRA NILAI, MALAYSIA

2016

# NON-PLAGIARISM DECLARATION

By this letter I declare that I have written this thesis completely by myself, and that I have used no other sources or resources than the ones mentioned.

I have indicated all quotes and citations that were literally taken from publications, or that were in close accordance with the meaning of those publications, as such. All sources and other resources used are stated in the references.

Moreover I have not handed in a thesis similar in contents elsewhere.

In case of proof that the thesis has not been constructed in accordance with this declaration, the Faculty of Health & Life Sciences has the right to consider the research proposal as a deliberate act that has been aimed at making correct judgment of the candidate's expertise, insights and skills impossible.

I acknowledge that the assessor of this item may, for the purpose of assessing this item,

- reproduce this assessment item and provide a copy to another member of the University; and/or,
- communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).

In case of plagiarism the examiner has the right to fail me and take action as prescribed by the rules regarding Academic Misconduct practiced by INTI International University.

| | |
|---|---|
| Name | Signature |
| I.D.Number | Date |

# ABSTRACT

The Verbenaceae refers to a family of vervain which are shrubs or trees and composed of 100 genera and about 2600 species around the world which usually grow in sub-tropical and tropical areas. Due to its wide availability there have been difficulties in differentiating the verbenaceae species based on their morphologies. DNA barcoding while becoming a common method of species identification, has had problems in its use such as the lack of consensus on gene loci and their utility across taxons, as well as methods of analysis. Hence this study tests ITS2 and *rbcl* for their ability in identification of Verbenacea. The three species used in this experiment are *Duranta repens, Lantana camara* and *Stachytarpheta jamaicensis*. The potential barcodes were amplified using polymerase chain reaction followed by sequencing. The sequences were characterized, then similarity (BLAST) and distance analysis (Barcode gap and tree topologies were carried out). The negative barcode gap using *rbcl* suggests that it is not suitable for identification, although tree topology based on *rbcl* enabled identification. Conversely, for ITS2 there were barcode gaps between all three species, but the tree topology, especially with combined experimental and downloaded data, did not allow clear identification. Increasing sample size and including samples from different geographic areas, as well as testing other loci is suggested and only the combined effort of scientist around the world will be able to build a comprehensive database, making DNA barcoding a feasible method of species identification.

## Table of contents

CHAPTER

# LIST OF TABLES

## LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AFLP | Amplified fragment length polymorphism |
| BLAST | Basic local Alignment search tool |
| bp | base pair |
| CBOL | Consortium for barcode of life |
| COI | Cytochrome c I oxidase |
| DNA | Deoxyribonucleic acid |
| EDTA | Ethylene diamine tetraacetic acid |
| ITS2 | Internal transcribed spacer 2 |
| *matK* | Maturase K |
| MEGA6 | Molecular Evolutionary genetics analysis 6 |
| NJ | Neighbor joining |
| PCR | Polymerase chain reaction |
| RAPD | Random amplified polymorphic DNA |
| RFLP | Restriction fragment length polymorphism |
| RT-PCR | Real time polymerase chain reaction |
| TAE | Tris-acetate-EDTA |
| TE | Tris/EDTA |
| TBE | Tris-Borate-EDTA |

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

DNA barcoding technique was first proposed by Hebert and colleagues (Jeanson et al., 2011). DNA barcoding refers to screening of specific universal DNA sequence to document and identify living organisms and contribute towards finding new species (Jeanson et al., 2011). It supports identification, as there are insufficient taxonomists (Carvalho et al., 2007) and it can even be used in identification when the whole plant especially the parts used in morphological identification are not available (Hartvig et al., 2015).

Plant identification is important as there is a dire need to conduct further research on biodiversity to obtain information about what are the different species as extinction of species is occurring very quickly (Lande et al., 1988). This will help conservation planning. Identification will also ensure food safety as where the plant material is directly consumed as food or used as herbal cures, accidental or intentional contamination by other less-valuable plants species has been reported (Mattia et al., 2010).

However, amongst the problems in DNA barcoding is that barcodes are not available for all plant species (Janzen et al., 2009), and sampling has generally been restricted to specific geographic areas (Janzen et al., 2009) This will result in the barcoding gap being easily detectable and species delimitation clear (Bergsten et al, 2012). However, such unambiguous species identification may not be the case when a greater number of closely related species and samples from different geographic regions are included, as it is expected that there might be smaller interspecific variation and greater intraspecific variation (Hamrick, Godt & Sherman-Boyles, 1992).

This suggests a need to carry out further sampling in different geographic areas, as well as extend the sampling to include species currently not having barcode sequences in the database. Additionally, while there exists reports that some of the barcodes are universally useful for plants and *rbcl* and *matk* were the proposed genes for barcoding plants (CBOL, 2009), Huang et al., (2015) for example has reported poor ability of these genes in identification. Hence it is still important to screen for suitable barcodes for any new species (Ledford., 2008).

This study will focus on the geographic region of Malaysia particularly Nilai where there is as far as we know limited studies on DNA barcoding plants. My project will be part of a larger project, and based on time and economic constraints I will be obtaining the DNA sequences of 2 loci *(rbcl* and ITS2) for 3 member of the *Verbenacea* family (*Duranta repens, Lantana camara and Stachytarpheta jamaicensis*), plants known for their medicinal properties (Rahmatullah et al., 2011). For example *Duranta repens* has inhibitory effects on malaria and its extracts have insecticidal and antifeedant properties (Anis et al., 2002), *Stachytarpheta jamaicensis* is used in treating ulcers in stomach and asthma (Becker et al., 2004), *Lantana camara* in folk medicine to reduce flu, fever, stomach ache and high blood pressure when mixed with tea (Ghisalberti et al., 2010). *Duranta repens* otherwise known as *Duranta erecta and Lantana camara* have *rbcl* and ITS2 sequences in the database whereas *Stachytarpheta jamaicensis* only has *rbcl* and no ITS2 moreover the samples found in the database are not from Malaysia.

## 1.2 Aims:

Sequences of ITS2 and *rbcl* will be obtained from the plant specimens to validate the use of these two loci in their ability to identify the species *Duranta repens, Stachytarpheta jamaicensis and Lantana camara* found in Nilai, Malaysia and also the effect on identification when locally obtained sequences are analysed together with sequences downloaded from the database.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Species identification

The task of an accurate identification of new plant species is not easy because of the presence of a morphologically similar species and a limited or cluttered taxonomic source and history.

Studying DNA sequences differences is an alternate technique in identifying species whereby genetic differences on a molecular level is studied to identify species and some of these examples include Restriction fragment length polymorphism (RFLP), Amplified fragment length polymorphism (AFLP), Random amplified polymorphic DNA (RAPD), and Real time polymerase chain reaction (RT-PCR) (National centre of biotechnology Institute, 2014).

DNA barcode on comparison provides a standard operating procedure for species identification while the other DNA based methods have been developed on specific genera or species (Novak et al., 2008). DNA barcoding can be carried out with high reproducibility but procedures such as RAPD produces results that are highly dependent on the laboratory in which the procedure is carried out partly due to gel analysis result interpretation which could vary amongst researchers and is highly affected by concentration of DNA and parameters of PCR (Mbwana et al., 2014).

Identification of species of animals was done by means of using a mitochondrial gene termed as Cytochrome c Oxidase I and it was able to identify 200 species of lepidopterans. It has also helped scientists in terms of identifying fish (Ward et al., 2005), birds (Kerr et al., 2007) and animals. This system has proven to be highly reliable and cost effective (Hebert et al., 2003).

While identification amongst animals has been made easy via *Cytochrome c Oxidase I* in plants low substitution rates of mitochondrial DNA have resulted in a search for alternative loci for barcoding the leading candidates for this whereby four coding regions which are *rbcl, matK, rpoB* and *rpoCL* and 3 non-coding spacers *atpF-atpH, trnH-psbA* and *psbK-psbI* where different groups have proposed a different combination of these loci as there barcodes for plants but a consensus has not emerged (Janzen et al., 2009).

The Consortium for barcode of life(CBOL Plant Working Group, 2009), proposed the genetic markers maturase K (*matK)* and RuBisCo large subunit (*rbcl)* to be the standard markers for plants. The advantage of using *rbcl* in detecting most land plants is easy amplification (Jawdat et al., 2013). However, the criteria of detecting most land plants using *rbcl* has been challenged by extensive results and this is primarily due to low variation (Chase et al., 2005) *matK* on the other hand provides high variation, but is difficult to amplify in certain plants (Fineschi et al.,2005; Robertson et al., 2010). So in the search for a useful barcode for plants, another marker that has widespread support to be used is the second internal transcribed spacer of nuclear ribosomal DNA, because the ability in identification using ITS2 has reached a level of 92.7% in plant species(Chen et al., 2010).

Characteristics that make ITS2 a viable DNA barcode is universality and short sequence of 200 base pairs which helps in successful amplification of the loci in land plants (Yao et al., 2010) and makes sequencing easier. Other criteria making ITS2 an ideal choice is that ITS2 secondary structure variation improves accuracy of species delimitation (Yao et al., 2010).

### 2.2 Problems incurred in DNA barcode analysis

Quality control of sequence data has sometimes been ignored which leads to incorrect barcoding results (Mayol & Rossello., 2001). Non-phylogenetic signal from sequence data has multiple sources (Philippe, Delsuc, Brinkmann, & Lartillot, 2005). Among them (i) low quality sequence (Little, 2010) (i) incorrect identification of orthologs, (Buhay, 2009) (iii) substitution model violations (Lemmon & Moriarty 2004; Ripplinger & Sullivan 2008).

This non-historical noise has to be reduced to improved validity of barcoding. Firstly noise is reduced by ensuring quality of sequences used in analysis meets the Standards for Barcode records (Hanner, 2009; Little, 2010). Secondly selection of the orthologous may be undertaken for example by comparing base composition of the sequence, looking for absence of stop codons in coding sequence, annotating motifs within the sequence and using secondary structures when available.

Another problem recognized is the use of suboptimal substitution models (Fregin et al. 2011), the use of the optimal substitution model. There are different kinds of models that are used as shown in Table 2.1

Table 2.1 Models of nucleotide substitution and its criteria.

| Model | Criteria |
|---|---|
| Jukes and Cantor (1969)<br>(Rzhetsky et al., 1993) | Due to equal base frequencies, substitutions have same probability. |
| Kimura Model<br>(Rzhetsky et al., 1993) | Transition and Transversion rate is the same with same base frequency. |
| Hasegawa (1985)<br>(Hasegawa et al., 1985) | Transition and Transversion rate is the same with same base frequency. |
| Tamura 3 parameter<br>(Tamura., et al 1992) | Difference in Transition and Transversion rates with G+C content. |

This leads to incorporation of a procedure to identify a model which can best represent mutational processes (Sullivan & Joyce 2005) such as the use of Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) (Kass & Wasserman, 1994). A lower AIC and BIC means a model is considered to be closer to the truth.

## 2.3 Methods of Species Discrimination

Various methods of species discrimination are available. There are similarity based methods such as Basic Local Alignment, BLAST as well as distance based methods such as using the barcode gap or tree topology.

## 2.4 Barcode gap

Evolutionary divergence amongst the sequence can be measured via estimating the ratio of different nucleotide *p, n* being the complete number of nucleotides and $n_d$ represents different number of nucleotides for the pair.

$$\hat{p} = \frac{n_d}{n} \qquad (1)$$

The equation above measures the evolutionary divergence between the two aligned sequences. The p-distance is part of distance method (Hebert et al. 2003) on which the barcode gap is based, This method theorizes that species which are distantly related will be represented by a larger barcoding gap (Gordh et al., 2014), in affect the intraspecific divergence is smaller than interspecific divergence (Dasmahapatra et al., 2006;Yassin et al., 2010 ). Identification may be based on using a threshold distance gap of 2–3% (Hebert et al., 2003). It may use average intraspecific distances (Hebert et al. 2004; Zhou et al. 2009) or resolved if its minimum inter-specific distance is greater than maximum intra-specific distance. The population size has a significant influence on the intraspecific variation hence larger population will have species with larger intraspecific variation (Nichols et al., 2001). Another factor is mutation rate, which shows