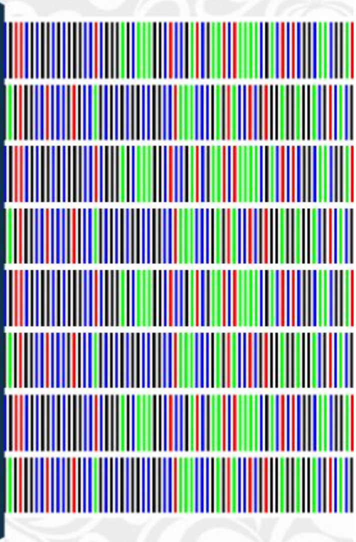
Technical Bulletin No.- 48

Raj Kumar
Pradosh Mahadani
Ravi Kishore
A. Loyanganba Meitei
D. R. Singh

DNA Barcoding of Indian Orchids



भाकृअनुप - राष्ट्रीय आर्किड्स अनुसंधान केंद्र
पक्योंग-७३७ १०६, सिक्किम, भारत

ICAR-National Research Centre for Orchids
Pakyong-737 106, Sikkim, India

Technical Bulletin No.- 48

DNA Barcoding of Indian Orchids

Raj Kumar
Pradosh Mahadani
Ravi Kishore
A. Loyanganba Meitei
D. R. Singh

भा.कृ.अनु.प.– राष्ट्रीय आर्किड्स अनुसंधान केन्द्र
पाक्योंग, सिक्किम– 737 106

ICAR-National Research Centre for Orchids
Pakyong, Sikkim-737106

DNA barcoding of Indian Orchids

October, 2016

Compiled and Edited by : Pradosh Mahadani, Raj Kumar, Ravi Kishore,
A. Loyanganba Meitei, D. R. Singh

Designed and Printed by : Astral International (P) Ltd., New Delhi
mail: info@astralint.com

Preface

The use of a universally accepted short DNA sequence for identification of species has been proposed for application across all forms of life. Such a "DNA barcode," a term first coined less than a decade before the publication of the present technical bulletin, in its simplest definition is one or more short gene sequences (<700 base pairs) taken from a standardized portion of the genome that are used to identify species through reference to DNA sequence libraries or databases. We recognized that DNA barcoding is much more than the sequencing of one or two genes from an organism. The endeavor has come to encompass many elements, from campaigns that provide a deterministic framework for how to build specimen libraries, to the bioinformatic systems needed to track the many samples and sequences.

The present technical publication '**DNA barcoding of Indian Orchids**' under the project "**National Mission on Himalayan Studies**" covers the wide aspect of molecular identification and conservation of orchids.

This technical bulletin should be of benefit and interest to all orchidologist and technicians interested in the relevance and application of molecular biology and DNA sequencing to identification, taxonomy, evolution, and ecology.

Authors

Contents

1

Introduction to DNA barcoding

The word ‘Barcode’ is a terminology which when we breaks into ‘bar’ and ‘code’ explains it as, there are some ‘bars’ which codes for something. It is mainly a technique which is adopted to give a unique identity to any product in a supermarket. The technological aspects of barcode involve an optical machine-readable representation of data, which gives information about the object to which it attaches. Originally barcodes represents data by varying the widths and spacing of parallel lines. It is treated as Universal Product Code and have become a ubiquitous element of modern civilization, as evidenced by their enthusiastic adoption by stores around the world; almost every item other than fresh produce from a grocery store, department store, and mass merchandiser has a barcode on it. The major benefit on adopting the technique involves quick and easy access to any product from a mass of commodities and thus the entire world has a common platform of product investigation.

The Earth’s biota has a wide range of diversity and thus represents a situation similar as mighty collection of products in a supermarket. The major entity which hurdles inventorying of the biodiversity is the entity ‘species’ as species identification and classification have traditionally been the specialist domain of taxonomists. Indeed, today’s society has to resolve many crucial biological issues, among which are the need to maintain biodiversity, to ensure bio-security, to protect species and to avoid pandemics. The achievement of such goals and the success of subsequent action programs require efficient global networks and rely on our capacity to identify any described species. A major solution regarding characterization of the mighty Earth's Biota may be thought in persistence to characterization of unique product using the Universal Product Code or barcode and the concept being termed as DNA barcoding.

DNA barcoding

In the post-genomic era, molecular biologists introduced a concept of DNA-Barcoding as a global standard for biological identification. It relies on the use of a standardized DNA region as a tag for rapid and accurate identification of the species of biological origin. In 1993, the term 'DNA Barcode was first used in scientific community but did not receive much attention. In 2003, however the golden age of barcoding began when a paper was published saying "*We are convinced that the sole prospect for a sustainable identification capability lies in the construction of system that employ DNA sequences as taxon barcodes*." (Hebert *et al*; 2003). Hebert successfully identified ~648bp of mitochondrial *Cytochrome C oxidase* gene as a potential barcode for animals and can serve as the core of a global identification system for animals. A remarkably short DNA sequence can contain more than enough information to resolve 10 or even 100 million species. For example, a 600-nucleotide segment of a protein-coding gene contains 200 nucleotides that are in the third position within a codon. At these sites, substitutions are (usually) selectively neutral and mutations accumulate through random drift. Even if a group of organisms was completely biased to either adenosine or thymine (or alternatively, to either guanidine or cytosine) at third nucleotide positions, there would still be 2^{200} or 10^{60} possible sequences based on third-position nucleotides alone. DNA sequence analysis of a uniform target gene to enable species identification has been termed DNA barcoding, by analogy with the Uniform Product Code barcodes on manufactured goods. It is based on a relatively simple concept of nucleotide difference between sequences, whereby, a gap of difference is always found functional between individuals of different species in comparison to individuals of same species. The above gap is referred as 'barcode gap' and is considered as threshold value for species discrimination. DNA barcoding is considered among those concepts where the modern knowledge of biological science are applied for understanding biodiversity and is a search towards long demanding common universal way to define a species. The technique has attracted attention from taxonomists, ecologists, conservation biologists, agriculturists, plant-quarantine officers and others, and the number of studies using the DNA barcode has been rapidly increasing. In 2004, the now well established Consortium for the Barcode of Life, an international initiative started supporting the development of DNA Barcoding, aims to promote both global standards and co-ordinate research in DNA Barcoding. It aims at establishing a public library of sequences and promotes development of portable devices for barcoding. The Rockfeller University in collaboration with two other more Organisations in 2004 have put up the various reasons for 'Barcode of Life'. The DNA barcoding is rapidly evolving but it is yet to provide full agreement on which region(s) of DNA should be universally used for plants. A large number of molecular techniques have been used to authenticate plants based on species-specific variations in the sequences of various chloroplast and nuclear DNA regions. Several studies tested the efficiency of seven leading candidate plastid DNA regions (*atpF–atpH*, *psbK–psbI*, *trnH–psbA spacers* and *matK, rbcL, rpoB, rpo*C1 genes) and the Plant Working Group of the Consortium for the Barcode of Life (CBOL) recommended the two-marker combination *rbcL/ matK* as the standard DNA barcode for plants to be supplemented with additional markers as required (Hollingsworth *et al.,* 2009). China Plant BOL Group (2011)

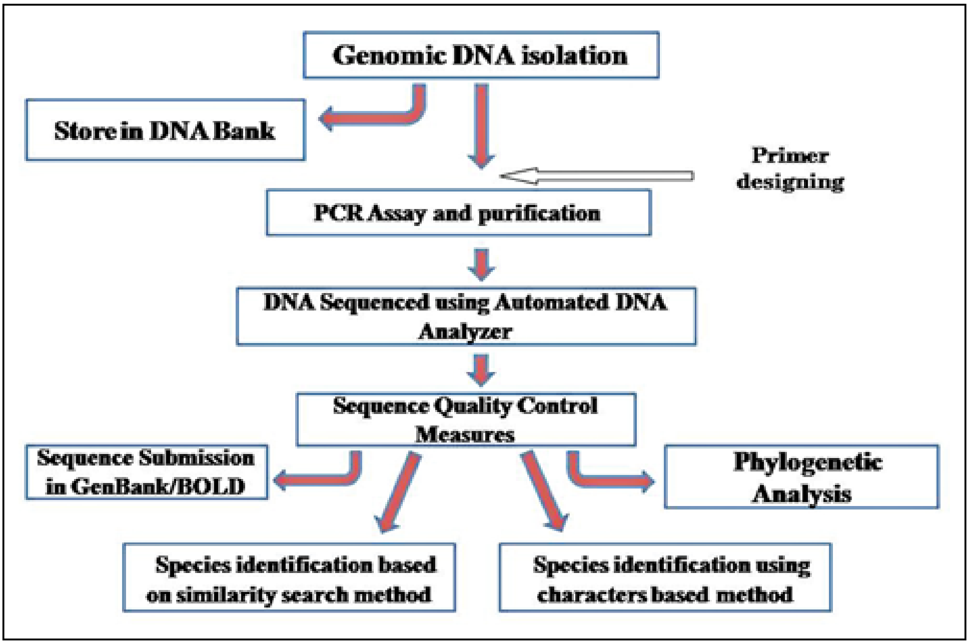

Genomic DNA isolation
Store in DNA Bank
Primer designing
PCR Assay and purification
DNA Sequenced using Automated DNA Analyzer
Sequence Quality Control Measures
Sequence Submission in GenBank/BOLD
Phylogenetic Analysis
Species identification based on similarity search method
Species identification using characters based method

Fig. 1.1: Flow chart of DNA Barcoding

proposed to include ITS/ITS2 as standard DNA barcode for seed plants. In fact, molecular taxonomists now envision cataloging all living species on earth using DNA barcoding. The generation of molecular "barcodes" of orchids species and deposition of sequence data in publicly accessible databases will be worth by the concerted effort of the Orchid research community and contribute to the ongoing effort of defining barcodes for every (plant) species on earth.

Chloroplast Genome

Chloroplast genome sequences are of broad significance in plant biology, due to frequent use in molecular phylogenetics, comparative genomics, population genetics, and genetic modification studies. The plastome is a circular molecule

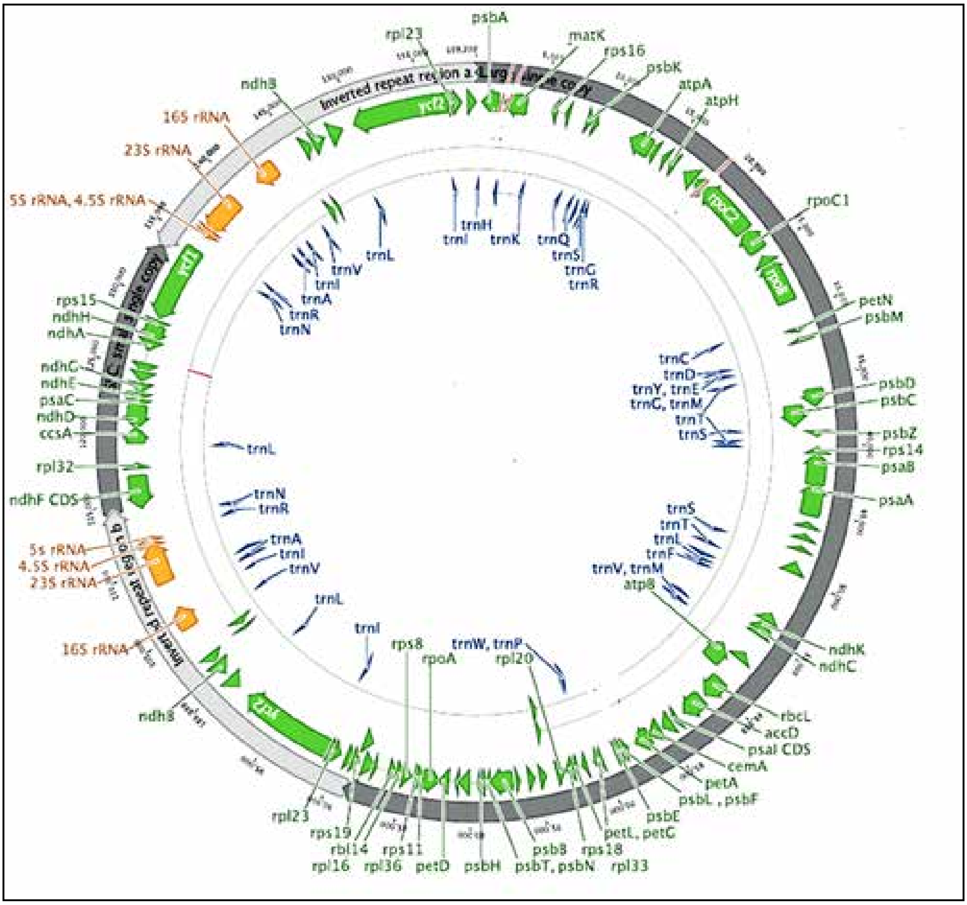


Fig. 1.2: Schematic diagram of plant Chloroplast Genome

that, in most taxa, varies between 108 and 218 kb in size and generally displays a high level of conservation between land plant species. The typical plastome is composed of two identical inverted repeats (IR, 20–76 kb) that separate a large single-copy (LSC, 60–90 kb) and a small single-copy (SSC, 7–27 kb) region. Among the land plant, cp-genome usually contains 100-120 genes. Most protein coding gene involved in photosynthesis or gene expression. Advantages of plastid DNA include I. it is monomorphic i.e separation of alleles is not required, ii. high copy number.

Maturase K (matK)

Maturase K of chloroplast is immerged as a most conserved gene in plant kingdom and functioned as Group II intron splicing. About 1500 bp long, *matK* gene is found within the intron of *trnK* of chloroplast DNA and encodes maturase like protein. The gene contains high substitution rates within the species and is emerging as potential candidate to study plant systematics and evolution. A homology search for this gene indicates that the 102 amino acids at the carboxyl terminus are structurally related to some regions of maturase-like polypeptide and this might be involved in splicing of group II introns. It is another emerging gene with potential contribution to plant molecular systematics and evolution. The *matK*-trnH gene complex is commonly used for plant evolution studies and addresses the solution for various taxonomic levels. The *matK* gene has ideal size, high rate of substitution, large proportion of variation at nucleic acid level at first and second codon position, low transition/transversion ratio and the presence of mutationally conserved sectors. These features of *matK* gene are exploited to resolve family and species level relationships. The second half (5' end) of the *matK* exon is easy to amplify and align; we propose that *matK* is used as a preferred universal DNA barcode for flowering plants.

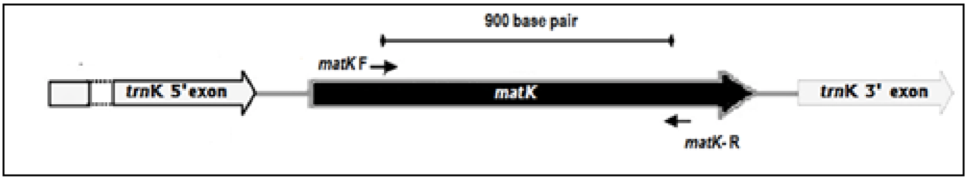


Fig. 1.3: Schematic diagram of maturase K gene of Cp-DNA

rbcL

The protein-encoding plastid gene *rbcL* has been proposed as a potential plant barcode by several sets of researchers (Chase *et al*; 2005) usually in conjunction with one or more other markers. One benefit of this region is that a large amount of information (more then 10,000 *rbcL* sequences) are already available in GenBank. Furthermore, studies by Newmaster *et al*; 2006 revealed that there was a fair degree of success in discriminating species using *rbcL* sequences used which are at least ~1300bp long. An ideal DNA barcoding region should be short enough to be amplified so there was a strive for developing primer sets for short sequence.

ITS

Internal transcribed spacer (ITS) of the nuclear DNA is one of the most popular loci in systematic and phylogenetic studies. About 400-800bp of ITS, makes it easier for sequencing and provide sufficient discrimination power among the species. Regions such as the internal transcribed regions of nuclear ribosomal DNA (ITS), although often highly variable in angiosperms as the generic and species level are not a practical option in several groups owing to peculiarities in their evolution and the fact that divergent copies are often present within single individuals. Single or low-copy nuclear regions are technically difficult to sequence and hence often not recoverable from degraded DNA. Therefore it is aimed to asses a large no. of plastid regions both coding and non-coding for their potential as a land plant barcode, both taking in to account the above attributes and striving to overcome some of the limitations inherent in regions such as ITS.

Application

The barcode of life provides an additional master key to knowledge about a species. Compiling a public library of sequences linked to named specimens, plus faster and cheaper sequencing, will make this new barcode key increasingly practical and useful.

Works with fragments: Barcoding can identify a species from bits and pieces. When established, barcoding will quickly identify undesirable animal or plant material in processed foodstuffs and detect commercial products derived from regulated species. Barcoding will help reconstruct food cycles by identifying fragments in stomachs and assist plant science by identifying roots sampled from soil layers. For example, the identification of organisms contained in stomach extracts allows the elucidation of wild animal diets, especially when behavioural studies are not feasible. DNA barcoding could also become an efficient tool to clarify host parasite and symbiotic relationships and in turn give new insights on host spectra,

Works for all stages of life: Barcoding can identify a species in its many forms, from seed, through seedlings, to adults and flowers.

Unmasks look-alikes: Barcoding can distinguish among species that look alike, uncovering dangerous organisms masquerading as harmless ones and enabling a more accurate view of biodiversity. Indeed, there is evidence from additional DNA barcoding studies revealing cryptic speciation in very distinct scenarios, among them the case of sympatric speciation in the skipper butterfly *Astraptes fulgerator*. Comprehensive analysis of DNA barcodes from adults and caterpillars of *A. fulgerator* inhabiting Costa Rica's tropical rain forest exposed a formerly described single species comprised in fact a total of 10 species, each with distinct ecological characteristics.

Reduces ambiguity: Written as a sequence of four discrete nucleotides - CATG – along a uniform locality on genomes, a barcode of life provides a digital identifying feature, supplementing the more analog gradations of words, shapes and colors. A library of digital barcodes will provide an unambiguous reference that will facilitate identifying species invading and retreating across the globe and through centuries. DNA barcode technology has already sparked US Congressional hearings by exposing widespread "fish fraud", mislabelling cheap fish as more desirable and expensive species like tuna or snapper. Until now, border inspection to keep agricultural pests, disease carrying insects and invasive species from entering a country has been a hit-and-miss effort. Barcoding offers a tool to get same-day answers for accepting or rejecting imports, an issue of acute economic importance for every country.

Makes expertise go further: The bewildering diversity of about 2 million species already known confines even an expert to morphological identification of only a small part of the plant and animal kingdoms. Foreseeing millions more species to go, scientists can equip themselves with barcoding to speed identification of known organisms and facilitate rapid recognition of new species.

Democratizes access: A standardized library of barcodes will empower many more people to call by name the species around them. It will make possible identification of species whether abundant or rare, native or invasive, engendering appreciation of biodiversity locally and globally.

Opens the way for an electronic handheld field guide, the Life Barcoder: Barcoding links biological identification to advancing frontiers in DNA sequencing, miniaturization in electronics, and computerized information storage. Integrating those links will lead to portable desktop devices and ultimately to hand-held barcoders. Imagine the promise of a schoolchild with a barcoder in hand learning to read wild biodiversity, the power granted to a field ecologist surveying with a barcoder and global positioning system, or the security imparted by a port inspector with a barcoder linked to a central computer!

Demonstrates value of collections: Compiling the library of barcodes begins with the multimillions of specimens in museums, herbaria, zoos and gardens, and other biological repositories. The spotlight that barcoding shines on these institutions and their collections will strengthen their ongoing efforts to preserve Earth's biodiversity.

Case study of DNA barcoding in Orchids

Lahaye et al., (2008) undertook intensive field collection (more than 1600 samples) in two biodiversity hot spot (Mesoamerica and southern Africa). They compared eight potential barcodes (*matK, trnh-psbA, ycf5, rbcL, rpoB, ndhJ, accD* and *rpoc1*) in more than 1000 species Mesoamerican orchids. Based on barcode gap, easy amplification, and alignment, they identified a portion of the plastid *matK* gene as a universal DNA barcode for Orchids. Recently, *Ycf1* (hypothetical chloroplast open reading frame) region of chloroplast DNA emerged as a potential candidate for DNA barcode loci for plants.

Identification of *Dendrobium*

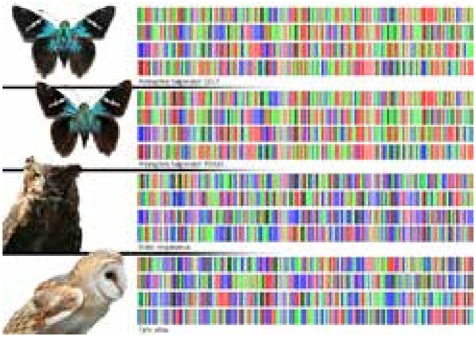
Dendrobium, the second largest genus of Orchidaceae family is equally important for both ornamental and medicinal value. Singh et al., (2012) compared seven loci of plant DNA barcoding among multiple accessions of 36 Indina *Dendrobium* species. The *trnH-psbA* spacer showed problematic in sequence quality and *ITS* provided 100% species identification. Another locus *matK* resolved 80.56% of 36 species. They recommended combination of *matK*, *rpoB* and *rpoC1* to resolve the maximum number of species.

Identification of *Paphiopedilum*

The species of *Paphiopedilum* are commonly referred to as 'Lady's or Venus's Slipper' orchids. *Paphiopedilum* is native to south-east Asia, northern India, southern China, Myanmar, Thailand and New Guinea, with 80 species distributed worldwide (Chung et al. 2006). Of the nine species of Paphiopedilum occurring in India. Parveen *et al.,* (2012) tested the efficiency of RNA polymerase-b subunit (*rpoB*), RNA polymerase-b' subunit (*rpoC1*), Rubisco large subunit (*rbcL*) and maturase K (*matK*) from the chloroplast genome and nuclear ribosomal internal transcribed spacer (nrITS) from the nuclear genome with in the Indian species of *Paphiopedilum*. This study unequivocally demonstrated that the DNA barcoding with *matK* was the signature sequence for the identification of closely related endangered species of Indian *Paphiopedilums* and also in elucidatec the parentage of their inter-specific hybrids. However, Gou *et al.,* (2016) recommended the combination of *matK* + *atpF-atpH* + ITS as a barcode for Venus slippers. They evaluate the efficiency of nine loci with large data set of 87 species.

1 25002 74135 0

Universal Product Barcode



DNA Barcode

2

Indian Orchid Biodiversity

Orchids, one of the most fascinating creations of the nature are one of the most widely distributed groups of flowering plants on the earth. They are cosmopolitan in distribution and known to occur in a wide range of climatic conditions from the alpine regions to humid climate, except the icy continent of Antarctica to dry sandy African and Australian deserts. They are abundant in tropical regions of the South East Asian countries Like India, China, Malaysia, Laos, Myanmar, Nepal, Bhutan Japan, Philippines, Australia, Europe, South & Central America and South Africa etc. The family Orchidaceae, is one of the oldest known and world’s largest family of flowering plants comprising of over 800 genera and between 25000 – 35,000 species. R. Govaerts of the Royal Botanic Garden, Kew mentioned in a personal communication that during the project ‘World Checklist of Selected Plant Families’ a total of 27,230 accepted taxa have been enumerated for the family Orchidaceae. The family exhibits the peak of evolution amongst the Monocots.

The first scientific account of Indian orchids was provided by the then Dutch Governor of Malabar, Von Rheede (1678 – 1703) in his monumental work ‘Hortus Malabaricus’. William Roxburgh (1832), the ‘Father of Indian Botany’, provided a treatment of 57 species in his ‘Flora Indica, vol. III’. But the most significant contribution to Indian orchids was made by Sir J.D. Hooker (1888, 1890) in Flora of British India (Vol. 5 & 6), who described about 1600 species of orchids from the erstwhile British India.

In India, the family Orchidaceae is widely distributed from alpine to coastal regions and islands but their maximum diversity occurs in the Eastern Himalayan and Peninsular regions respectively. According to the reports (Misra, 2007) the family has about 186 genera, 1298 species, 5 subspecies and 28 varieties in India. After the establishment of Botanical Survey of India in 1890 at Royal Botanic Garden, Calcutta (now Acharya Jagadish Chandra Bose Indian Botanic Garden, Shibpur, Howrah) with Sir George King as its Director, large scale studies on the Indian flora including orchids were launched. These studies resulted in the publication of a large number of floristic accounts pertaining to different parts of India including several classical works dealing with Indian orchids of which 'The Orchids of Sikkim Himalayas' by King & Pantling (1898), 'The Orchids of North-Western Himalaya' by Duthie (1906) are worth to mention. Apart from these, regional floras like Bengal (Prain, 1903), Presidency of Bombay (Cook, 1908), Travancore (Rao, 1914), Nilgiri and Pulney hilltops (Fyson, 1915), Bihar and Orissa (Haines, 1924), Presidency of Madras (Fischer, 1928) and many more also contributed significantly towards the knowledge of orchids of these regions. But all further studies suddenly came to a halt for few decades during pre-independent period and there was a sudden lull in the study of Indian flora. However, after gaining independence and with the reorganization of Botanical Survey of India in 1954, the task of inventorization of the country's floral wealth was resumed. As a result several under and un-explored areas were intensively surveyed and a large number of publications dealing with State and district floras, flora of fragile ecosystems and protected areas as well as monographs and other revisionary works on different plant groups were brought out. However, a number of publications dealing with detailed or brief accounts of orchids, like of Meghalaya (Kataki, 1986); North-West Himalaya (Deva & Naithani, 1986); Nilgiri (Joseph, 1987); Arunachal Pradesh (Chowdhery, 1998); Nagaland (Hynniewata et al., 2000); Kamrup (Barua, 2001); Orissa (Misra, 2004); Kerala (Sathish Kumar &Manilal, 2004); Manipur (Sathish Kumar & Suresh Kumar, 2005); Sikkim and North-East India (Lucksom, 2007) were also brought out making significant contribution to the existing knowledge of Indian orchids.

The orchids are under cultivation since 500 B.C. in the orient for ornamental and medicinal use. They produces flowers, which are most exotic, amazingly attractive, in bewitchingly curious shapes, colours, longer self-life (2 – 3 months) made them one of the top ten 'cut flowers' in international flower market. Now-a-days they occupy a major share in the global floricultural trade with extremely high returns. Horticulturists worldwide today grow orchids not only because they are curious, but mainly due to their great demand and high price. The orchids are most commonly used for cut flowers and pot plants, except very few in the 'Jewel orchids' group that are used for their very decorative leaf patterns but in addition to ornamental value, orchids have various other commercial uses too. The Indian orchids were brought to the notice of the world by Charak, the great Indian medicine man as back as in 125 A.D, who described '*Vanada*' and several other orchids in his book - 'Charak Samhita' which provides description of present known orchids like *Flickingeria, Malaxis* and *Eulophia* species (Kutumbiah, 1962). Having tremendous horticultural and medicinal value, the family Orchidaceae has been paid adequate attention in many countries throughout the world to study their biology, evolution, taxonomy, cytology, chemistry, hybridization and cultivation etc.

Many orchids are known for their wonderful fragrance and it is believed that more than 75% orchids are fragrant species. The orchid fragrance is due to the presence of volatile aromatic oils produced in minor quantities in floral parts like sepals, petals, calluses, basal spurs to petioles. Floral scent emission shows diurnal rhythm and is controlled by internal biological clock. Some scent species emit fragrance at morning while others develop fragrance at late noon, evening or night. Orchid fragrance is a chemical messanger between the plant and its pollinator. Night pollinated flowers have peak emissions at night while the situation is reversed in day pollinated flowers. Orchids flowers have specialized scent glands called osmophores that ooze liquid scent, which evaporates on contact with the air. An orchid fragrance ranges from warm, sweet & highly diffusive notes to stinky and offensive odour. The pleasant scented orchid flowers are often compared to fragrance of other flowers like rose, hyacinth, jasmine, freesia, lily, narcissus, sweet pea or easily identified scents like lemon, chocolate, vanilla, orange, coconut, cardamom, musk, honey, mint etc. On the other hand, the flower of the bee orchid smells and looks, to the male bee, just like a female bee. The duped male bee attempts to copulate with the orchid's petals, and the insect spreads pollen between the deceptive flowers. And flowers don't just imitate bees. A few orchid species smell like female flies, and others replicate the aroma and texture of scarab beetles.

Table 2.1: State wise List of orchids available in India

S. No.	State	No. of genus	No. of species/taxa
1	Andaman & Nicobar Islands	66	143
2	Andhra Pradesh (including Telangana State)	40	83
3	Arunachal Pradesh	144	558
4	Assam	74	182
5	Himachal Pradesh	-	76
6	Jammu & Kashmir	-	46
7	Jharkhand	-	63
8	Karnataka	-	175
9	Kerala	-	186
10	Madhya Pradesh	34	89
11	Maharastra	36	122
12	Manipur	69	251
13	Meghalaya	116	459
14	Mizoram	74	249
15	Nagaland	92	396
16	Orissa	-	128
17	Sikkim	132	540
18	Tamil Nadu	-	72
19	Tripura	37	66
20	Uttarakhand	71	237
21	West Bengal	111	467

Endemic Orchids

The North East India has highest flora of monotypic orchid genera (Tandon et al. 2007). North East India is reported to harbour a large number of valuable threatened orchids also. It is to be noted that there are some orchid species which are endemic not only to this region, but also to the home states in which they are distributed like in Sikkim and Arunachal Pradesh Himalayas, the Naga and

Manipur hills, the Lusai - Mizo hills and Khasi –Jaintia hills (Nayar 1996). Arunachal Pradesh has the highest number of orchid species (558 species) followed by Sikkim (540 species) Meghalaya (479 species) Assam (182 species), Nagaland (396 species), Mizoram (249 species), Manipur (251 species) and Tripura (66 species)

Table. 2.2: Orchid species endemic to particular North-Eastern states

Sl.No.	State	Endemic Orchid Species
1	**Arunachal Pradesh**	*Biermannia arunachalensis* A.N.Rao.
		Biermannia jainiana Hedge et A.N.Rao.
		*Cheirostylis gunnar*ii A.N.Rao
		Cheirostylis sessanica A.N.Rao
		Cheirostylis tippica
		Cylidrolobus arunachalensis A.N.Rao
		Cylidrolobus hegdei (Agarwal & Chowdh.)
		Cylidrolobus lohitensis (A.N.Rao, Hari. & Hegde)
		Dendrobium arunachalense C. Deori, Sarma, Phukan and Mao.
		Dendrobium josephii
		Dendrobium kamalangensis
		Dendrobium nareshbahadurii
		Dendrobium numaldeorii
		Epipogium arunachalense
		Gastrochillus arunachalensis A.N.Rao
		Gastrochillus sessanicus A.N.Rao
		Gastrodia arunachalensis Hegde & Rao
		Herminium haridasinii A.N.Rao
		Herminium kamengensis A.N.Rao
		India arunachalensis A.N.Rao
		Oberonia arunachalensis A.N.Rao
		Oberonia kamlangensis A.N.Rao
		Oberonia katakiana A.N.Rao

Sl.No.	State	Endemic Orchid Species
		Oberonia sulcata Jos & S. Chowdh.
		Propax seidenfadenii A.N.Rao
		Rhamboda arunachalensis A.N.Rao
		Sarcoglyphis arunachalensis A.N.Rao
		Taenophyllum arunachalensis A.N.Rao and J.Lal
2	**Assam**	*Chrysoglossum assamicum* **Hook.f.**
		Cleistotoma arientinum (Rchb.f.) Garay
		Coelogyne rossiana Rchb.f.
		Dendrobium assamicum Chowdhury
		Dendrobium griffithianum Rcbh
		Dendrobium keithii
		Dendrobium miserum Rchb.f.
		Erythrorchis altissima (Bl.) Bl.????
		Pholidota undulate Lindl.
		Rhynchostylis albiflora Barua et Bora
		Zexuine debranjiana Chowdhury
3	**Manipur**	*Anoectochilus tetrapterus* Hk.f
4	**Meghalaya**	*Anoectochilus crispus*
		Corymbus purpurens
		Eria ferrugina
		Eria pusilla
		Gastrodia exilis
		Goodyera recurva
		Habenaria concinna
		Habenaria furfuracea
		Habenaria khasiana

4	**Meghalaya**	*Anoectochilus crispus*
		Liparis acuminate
		Trias pusilla
		Pantlingia serrata N.C. Deori
		Pennilabium proboscideum A.S.Rao et. J.Joseph.
5	**Mizoram**	**Bulbophyllum parryae,**
		Eria lacei
		Sterogyne lushaiensis
6	**Nagaland**	*Coelogyne hitendrae* **Das et. Jain**

3

Molecular Biology methods in DNA barcoding

Cells are the basic building units of all living things, having all the body's hereditary material, capable of making self copies. The nucleus if takes the roles of the command centre, controlling the various operations of cell (divide, mature, grow or die); more importantly the hereditary material, DNA (deoxyribonucleic acid), resides where they are packaged into thread-like structures called chromosomes, by coiling around the histones proteins which is the backbone of the structure.

Mitochondria are complex organelles that convert energy from food into a form that the cell can use. They have their own genetic material, apart from the DNA in the nucleus having independent replicating capabilities. While, Chloroplasts are similar to mitochondria, but chloroplasts are found only in plants and some eukaryotic organisms. The material within the chloroplast is called the stroma containing one or more molecules of small circular DNA.

In modern molecular biology and genetics, the genome is the entirety of an organism's hereditary information. It is encoded either in DNA or, for many types of virus, in RNA. The genome includes both the genes and the non-coding sequences of the DNA/RNA. Some organisms have multiple copies of chromosomes. In eukaryotes such as plants, protozoa and animals, however, "genome" carries the typical connotation of only information on chromosomal DNA. In fact, mitochondria are sometimes said to have their own genome often referred to as the "mitochondrial genome" where as that of chloroplast may be referred to as the "plastome". A genome does not capture the genetic diversity or the genetic polymorphism of a species. This point explains the common usage of "genome" to refer not to the information in any particular DNA sequence, but to a whole family of sequences that share a biological context. Also, eukaryotic cells seem to have experienced a transfer of some genetic material from their chloroplast and mitochondrial genomes to their nuclear chromosomes.

Since the late 1950s and early 1960s, molecular biology evolved as the branch of biology that deals with the molecular basis of biological activity. Molecular biology chiefly concerns itself in understanding the interactions between the various systems that exist in a cell. Broadly this is the study of molecular unfolding of the processes of replication, transcription, translation, and cell function.

Much of the work in molecular biology is quantitative, and recently much work has been done at the interface of molecular biology and computer science in bioinformatics and computational biology. Increasingly many other loops of biology focus on molecules, either directly studying their interactions in their own right such as in cell biology and developmental biology, or indirectly, where the techniques of molecular biology are used to infer historical attributes of populations or species, as in fields in evolutionary biology such as population genetics and phylogenetics.

DNA barcoding is a Molecular tool for taxonomist for proper identification of any biological samples, which could be from different life stages. This technique uses a short stretch of DNA (Mitochondrial CO1 DNA for animals and Chloroplast DNA for plants) of the organisms as a genetic marker to identify the organism to a species level. For this purpose certain molecular techniques such as isolation of genomic DNA, target specific amplification of DNA, sequencing of the target DNA fragment, aligning the sequence with the database etc. are performed. For the better understanding of the entire procedure, it is very much necessary to understand the materials that constitute the genome of the organism, i.e. the nucleic acids which are mainly the DNA (Deoxyribonucleic acid) and the RNA (Ribonucleic acid). Of the nucleic acids, DNA constitutes as Genetic as well as structural part of the genome in all the organisms.

Extraction of DNA

1. Crush the young leaves tissue sample in Liquid Nitrogen thoroughly and taken in a sterile micro centrifuge tube. Add DNA Extraction Buffer (1 M Tris-Cl, pH-8, 5 M NaCl and 0.5 M EDTA pH-8) in micro centrifuge tube.
2. Immediately add 10% SDS and 2 µl of β- mercaptoethanol incubate at 65^0 C for 40 mins. The tube was inverted at every 10 mins interval to ensure adequate mixing.
3. Add 200 µl of 5 M potassium acetate (pH-9) in micro centrifuge and kept -20^o C for 20 min.
4. Centrifuge at 12000 rpm for 15 mins.
5. Remove the top Aqueous phase carefully into a new centrifuge tube.
6. Add RNase (1 µl per 10 µl concentrations) and incubate at 37^0 C for 1-1.30 hour.
7. After incubation, Add equal volume of Phenol: Chloroform: Isoamylalchol: (25:24:1) to the tube and shook the tube gently and centrifuge at 12000 rpm for 10 mins.
8. Take the upper aqueous phase into a new centrifuge tube very carefully, keeping in mind not to disturb the debris of interphase.
9. Add Equal volume of Chloroform: Isoamylalchol (24:1), shook the tube gently and centrifuge.
10. Take the supernatant into new centrifuge and add double volume of chilled ethanol (absolute) and keep in -20^0 C for 2 hour for precipitation.
11. Centrifuge at 10000 rpm for 10 min.
12. Discard the supernatant gently and retain the pellet. To it, add 1ml 70% ethanol for washing the pellet and centrifuged step. Subsequently, keep the pellet for air dry until smell of the alcohol remove.
13. Dissolve in Nuclease free water or 1X TE for long term preservation and stored in -80^0 C.

Quantification of DNA

Quantification and analysis of quality of isolated DNA is necessary to ascertain the approximate quantity of DNA obtained and the suitability of DNA sample for further analysis. This is important for many applications including digestion of DNA by restriction enzymes or PCR amplification of target DNA.

The most commonly used methods for quantifying nucleic acid in a preparation are: (i) gel electrophoresis; and (ii) spectrophotometric analysis

i) Agarose Gel Electrophoresis for DNA Quantification and Quality Analysis

This method of quantification is based on the Ethidium bromide, a fluorescent dye which intercalates between the stacked bases, staining of DNA.

- Nucleic acids (gDNA) are electrophoretically separated on a 0.7-0.8% agarose gel containing Ethidium bromide at a final concentration of 0.5 µg/ml.
- Quantity of DNA is estimated by comparing the fluorescent yield of the samples with DNA marker at varying known concentrations.
- Native DNA, which migrates as a tight band of high molecular weight (≥40 kb), presence of RNA, and degraded/sheared DNA, if any, can be visually identified on the gel.

If the sample amount is less, this method is usually preferred.

ii) Spectrophotometric Determination

Analysis of UV absorption by the nucleotides provides a simple and accurate estimation of the concentration of nucleic acids in a sample. The ratio of OD_{260}/OD_{280} should be determined to assess the purity of the sample. This method is however limited by the quantity of DNA and the purity of the preparation. Accurate analysis of the DNA preparation may be impeded by the presence of impurities in the sample or if the amount of DNA is too little.

- A ratio between 1.8-2.0 denotes that the absorption in the UV range is due to nucleic acids.

- A ratio lower than 1.8 indicates the presence of proteins and/or other UV absorbers.
- A ratio higher than 2.0 indicates that the samples may be contaminated with chloroform or phenol. In either case (<1.8 or >2.0) it is advisable to re-precipitate the DNA.

The amount of DNA can be quantified using the formula:

$$\text{DNA concentration } (\mu g/ml) = OD_{260} \times \frac{100 \text{ (dilution factor)} \times 50 \text{ mg/ml}}{1000}$$

Spectrophotomteric Conversions for Nucleic Acids

A 260 of ds DNA	**= 50 mg/ml**
A 260 of ss oligonucleotides	**= 33 mg/ml**
A 260 of ss RNA	**= 40 mg/ml**

PCR amplification of Barcode regions

Polymerase Chain Reaction, allowed the *in vitro* amplification of specific DNA from a complex DNA template in a simple enzymatic reaction. DNA polymerase uses single stranded DNA as template for the synthesis of a complementary new strand and it also requires a small section of double stranded DNA to initiate the synthesis. Thus, the starting point of DNA synthesis can be specified by supplying an oligonucleotide primer that anneals to the template at that point. Both strands can serve as templates for synthesis provided an oligonucleotide primer is supplied for each strand. For PCR, the primers are chosen to flank the regions of DNA that is to be amplified so that newly synthesized strands of DNA, starting at each primer extend beyond the position of the primer on the opposite strand. Thus, new primer sites are generated on each newly synthesized DNA strand.

Template DNA: This is the original genomic DNA material. Use 1–2 ml DNA solution (obtained from extraction) with a concentration between 20 and 100 ng/ml. Usually, PCR also works well with lower concentrations.

PCR Buffer: For optimal DNA Polymerase reaction activity, PCR buffers are used containing Tris–HCl, KCl, and, optional, $MgCl_2$. Buffers are provided by the supplier together with Taq Polymerase. It is important to use Polymerase and PCR buffer from the same manufacturer.

Primers: PCR primers are short, singlestranded DNA fragments (usually, 20–30 nucleotides). PCR requires one forward and one reverse primer to assign the favored fragment of the DNA. Primers are usually delivered in desalted mode. Resuspend primers in molecular water to astock concentration of 100 pmol/ml; prepare aliquots of working solutions with a concentration of 10 pmol/ml ready to use for PCR. There should be a surplus of primers in the reaction mix, but too much of them may lead to unspecific reactions. For standard PCR, 0.5 ml of each primer working solution (10 pmol/ml) is enough.

***Taq* polymerase :** Taq DNA Polymerase is commonly used for standard PCR. It is a thermostable enzyme of the thermophilic bacterium Thermus aquaticus and is, therefore, able to synthesize DNA at high temperatures. Usually, 0.025 U of Taq DNA Polymerase are used per ml of the PCR reaction.

Deoxynucleotide triphosphates (dNTPs): dNTPs (dATP, dTTP, dGTP, and dCTP) are the nucleotide bases added by the DNA Polymerase during synthesis of the template strand. They are available as single ingredients or as a dNTP mix. For PCR, a final concentration of 2 mM dNTPs (which means 2 mM of each type of nucleotide!) is applicable. In case of the nucleotide premix (all nucleotides in a total concentration of 10 mM), the solution has just to be diluted to the desired concentration. In case of single nucleotides, add 20 □l (100 mM stocks) of each dNTP to 920 □l molecular water.

Molecular water: Use only ultra pure and nuclease-free water for PCR. Water is used to fill the mix of ingredients up to the desired volume, which is normally 10–25 □l.

Table 3.1 : List of primer universally used in plant DNA barcode studies

Region	Primer Name	Sequences (5' to 3')
matK	*matK X F*	TAA TTT ACG ATC AAT TCA TTC
	matK 5r	GTT CTA GCA CAA GAA AGT CG
	3F_Kim matK	CGTACAGTACTTTTGTGTTTACGAG
	1R_Kim matK	ACCCAGTCCATCTGGAAATCTTGGTTC

Region	Primer Name	Sequences (5′ to 3′)
ITS	ITS 5a	CCTTATCATTTAGAGGAAGGA
	ITS 4	TCCTCCGCTTATTGATATGC
rbcL	*rbcLa-F*	ATGTCACCACAAACAGAGACTAAAGC
	rbcLa-R	GTAAAATCAAGTCCACCRCG
ycf1	*ycf1_OF1*	ATACATATCCACGTAATGGAAGA
	ycf1_OR1	TCTCTCCGAAAATCCGACTGTTGGGAAT
	ycf1_OF2	TTACATGTAAAAGTGATGGTAAA
	ycf1_OR3	TTGCGACGAAAATCCGATTGTTGTGAGT
trnH-psbA	*trnH-F*	CGC GCA TGG TGG ATT CAC AAT CC
	psbA-R	GTT ATG CAT GAA CGT AAT GCT C

Polymerase Chain Reaction (PCR)

More than 35 years ago, the introduction of recombinant DNA technology as a tool for the biological sciences revolutionized the study of life. Molecular cloning allowed the study of individual genes of living organisms; however this technique was dependent on obtaining a relatively large quantity of pure DNA. This depended on the replication of the DNA of plasmids or other vectors during cell division of microorganisms. Researchers found it extremely laborious and difficult to obtain a specific DNA in quantity from the mass of genes present in a biological sample. Recombinant DNA technology made possible the first molecular analysis and prenatal diagnosis of several human diseases. Foetal DNA obtained by amniocentesis sampling could be analyzed by restriction enzyme digestion, electrophoresis, southern transfer and hybridization to a cloned gene or oligonucleotide probes. However, southern blotting permitted only rudimentary mapping of genes in unrelated individuals (4). Polymerase Chain Reaction (PCR) is a technique sensitive enough to amplify small DNA fragments a billion-fold. The generations of amplified fragments (amplicon) by conventional PCR or by modified methods of these techniques are extensively used to address a variety of issues related to biology, medicine and forensic sciences. By PCR, not only the genetic diseases are diagnosed rapidly but ailment caused due to several other reasons such as protozoan, parasites, bacteria or even viruses may be diagnosed and monitored with much ease and accuracy The present topic covers the basic principle and some of its novel applications.

PCR, an acronym for Polymerase Chain Reaction, allowed the production of large quantities of a specific DNA from a complex DNA template in a simple enzymatic reaction. PCR is a recently developed procedure for the *in vitro* amplification of DNA. PCR has transformed the way that almost all studies requiring the manipulation of DNA fragments may be performed as a result of its simplicity and usefulness. In the 1980s, Kary Mullis and a team of researchers at Cetus Corporation at Cetus Corporation conceived of a way to start and stop a polymerase's action at specific points along a single strand of DNA. Mullis also realized that by harnessing this component of molecular reproduction technology, the target DNA could be exponentially amplified. This DNA amplification procedure was based on an in vitro rather than an in vivo process. Cell-free DNA amplification by PCR was able to simplify many of the standard procedures for cloning, analyzing, and modifying nucleic acids. Previous techniques for isolating a specific piece of DNA relied on gene cloning – a tedious and slow procedure. PCR, on the other hand Kary Mullis stated "lets you pick the piece of DNA you're interested in and have as much of it as you want". When other Cetus scientists eventually succeeded in making the polymerase chain reaction perform as desired in a reliable fashion, they had an immensely powerful technique for providing essentially unlimited quantities of the precise genetic molecular biologists and others required for their work. Since the first report in1985, more than 5000 scientific papers were published by 1992 and more than 5 million papers by 2007. Furthermore, the large number of publications of course makes it impossible to review all the important contributions to the development and application of PCR technology; however we will attempt to review here the most important developments in the practice of basic PCR.

Basic Principle of PCR

Polymerase chain reaction can best be understood following the principle of DNA replication. DNA polymerase uses single stranded DNA as template for the synthesis of a complementary new strand and it also requires a small section of double stranded DNA to initiate the synthesis. Thus, the starting point of DNA synthesis can be specified by supplying an oligonucleotide primer that anneals to the template at that point. Both strands can serve as templates for synthesis provided an oligonucleotide primer is supplied for each strand. For PCR, the primers are chosen to flank the regions of DNA that is to be amplified so that newly synthesized strands of DNA, starting at each primer extend beyond the position of the primer on the opposite strand. Thus, new primer sites are generated on each newly synthesized DNA strand. Generation of amplicons with a pair of primers is usually known as PCR or symmetrical PCR.

Reaction Conditions for PCR

Besides target DNA and primer(s), the reaction mixture requires an appropriate buffer containing Tris/HCl, $MgCl_2$, KCl, Triton-X gelatin, four deoxyribonucleotide triphosphates dATP, dCTP, dTTP and dGTP enabling the thermostable *Taq* polymerase to elongate the complementary chains concurrently on both the strands. The duplex formation between the primer and target DNA are both sequence and temperature dependent. Therefore, to achieve an optimal amplification, it is advisable, to empirically determine the molarity of each chemical ingredient, the amount of DNA polymerase, the temperature and number of cycles. Each cycle requires three different temperatures in quick successions for template (sample DNA) denaturation, primer annealing and extension of the regions between the primers. This is achieved by using automatic thermal cycler reactor. Usually, sufficient amplification of a target DNA is obtained in about 25-30 cycles. Double amount of DNA is synthesized in each cycle resulting in an exponential accumulation of the PCR product. Initially, PCR was carried out by using DNA polymerase I (klenow fragment) from *E. coli*, but this enzyme is not thermostable. Now a choice of several thermostable *Taq* polymerases useful for PCR amplification is available.

Detection of PCR amplicons

PCR amplified product is usually detected either by agarose or polyacrylamide gel electrophoresis and ethidium bromide staining or by hybridizing the amplicons with the specific DNA probes. If the non-specific amplicons have electrophoretic mobility similar to that of specific amplified product, the result would be misleading. Therefore, in gel- detection procedure alone is not always sufficient unless it is substantiated by additional analysis.

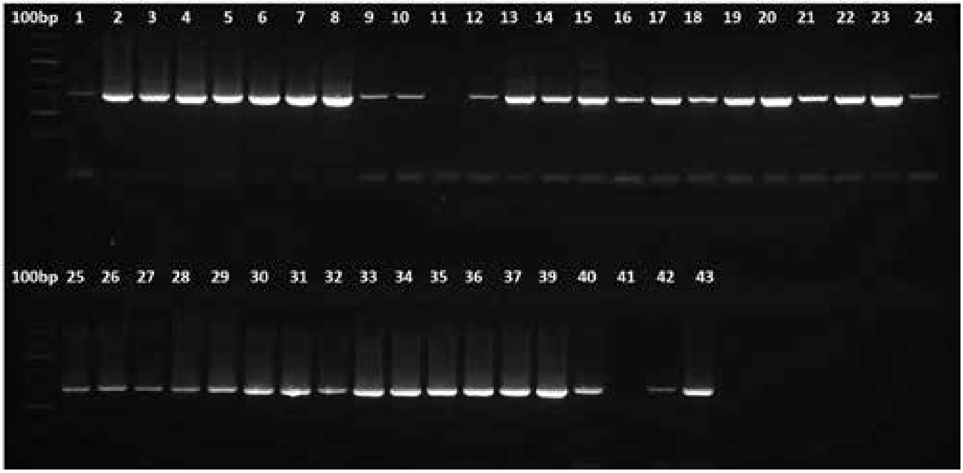

100bp 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
100bp 25 26 27 28 29 30 31 32 33 34 35 36 37 39 40 41 42 43

Fig. 3.1: Agarose gel electrophoresis of 600 bp of amplification products of ITS

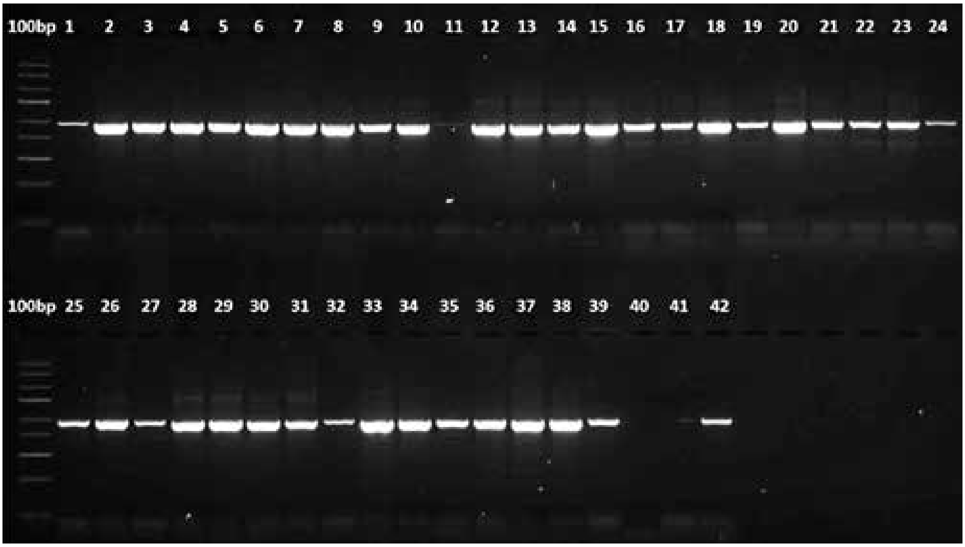

100bp 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
100bp 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42

Fig. 3.2: Agarose gel electrophoresis of 800 bp of amplification products of *matK*

Purification of PCR product and Sequencing

The PCR amplified products of expected size are extracted from the gel and purified. This DNA is then subjected to DNA sequencing.

4

DNA Sequence Quality check and Submission

DNA barcode sequences can be submitted to GenBank (the genetic sequence database at the National Center for Biotechnology Information, NCBI) using several methods. Quality checking is one of the most crucial step for the generation of DNA barcode sequences. The purpose of bidirectional sequencing is to increase the confidence of sequence quality. There are two check points; one is trimming from raw trace file and second, 3'and 5' terminals were clipped to generate consensus sequences and checking of open reading frame (ORF) for the protein coded sequences. Trace file are assemble in sequence editing software and sequence greater than 2% ambiguous bases are to be discarded, using quality value of 40 for bidirectional reads. Manual editing of raw traces and subsequent alignments of forward and reverse sequences enabled us to assign edited sequences for most species.

Important tools for sequence quality check:

BioEdit - BioEdit is a mouse-driven, easy-to-use sequence alignment editor and sequence analysis program. BioEdit is intended to supply a single program that can handle most simple sequence and alignment editing and manipulation functions that researchers are likely to do on a daily basis, as well as a few basic sequences analyses.

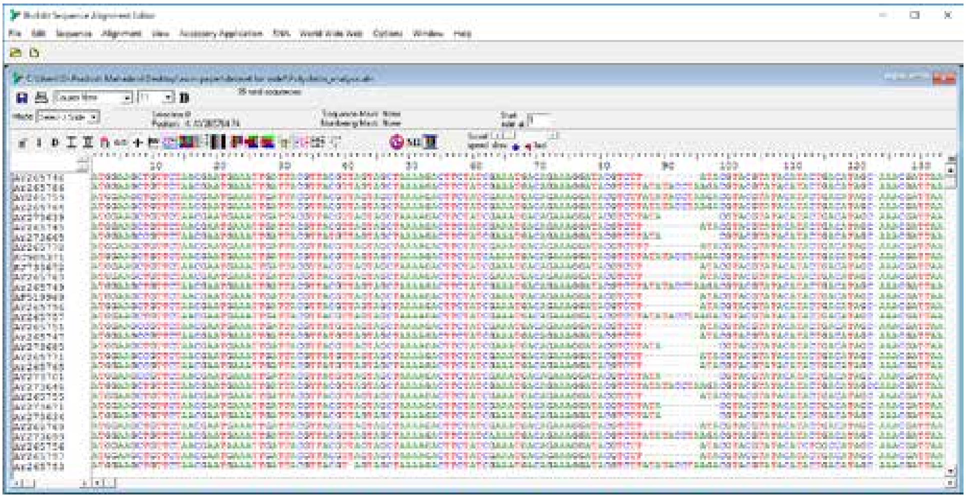
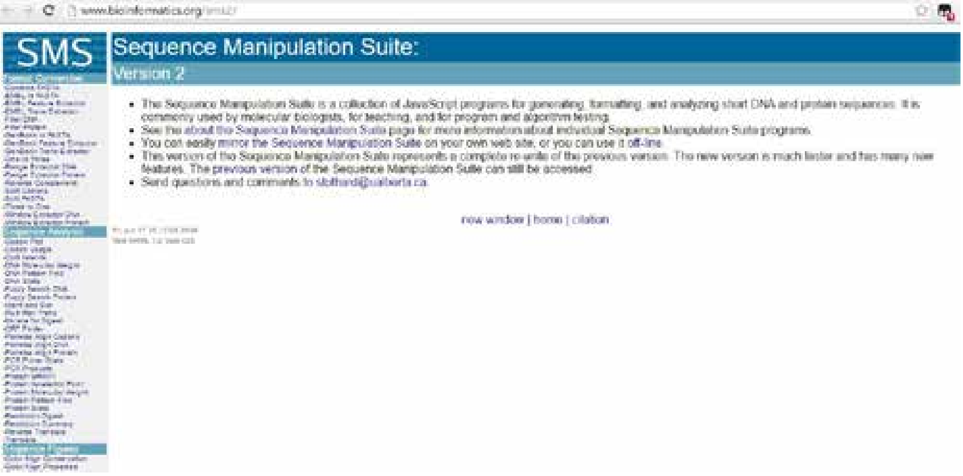

BioEdit Sequence Alignment Editor
File Edit Sequence Alignment View Accessory Application RNA World Wide Web Options Window Help

Fig. 4.1: Showing alignment though BioEdit.

Sequence Manipulations Suit - The Sequence Manipulation Suite (http://www.bioinformatics.org/sms2/) is written in *JavaScript* 1.5, which is a lightweight, cross-platform, object-oriented scripting language. *JavaScript* is now standardized by the ECMA (European Computer Manufacturers Association). The first version of the ECMA standard is documented in the ECMA-262 specification. The ECMA-262 standard is also approved by the ISO (International Organization for Standards) as ISO-16262. JavaScript 1.5 is fully compatible with ECMA-262, Edition 3. Sequences submitted to the Sequence Manipulation Suite



Sequence Manipulation Suite:

Version 2

- The Sequence Manipulation Suite is a collection of JavaScript programs for generating, formatting, and analyzing short DNA and protein sequences. It is commonly used by molecular biologists, for teaching, and for program and algorithm testing.
- See the about the Sequence Manipulation Suite page for more information about individual Sequence Manipulation Suite programs.
- You can easily mirror the Sequence Manipulation Suite on your own web site, or you can use it off-line.
- This version of the Sequence Manipulation Suite represents a complete re-write of the previous version. The new version is much faster and has many new features. The previous version of the Sequence Manipulation Suite can still be accessed.
- Send questions and comments to stothard@ualberta.ca.

new window | home | citation

Fig. 4.2: Sequence Manipulation Suit web portal.

are manipulated by the web browser, which are executed by the JavaScript. The Sequence Manipulation Suite was written by Paul Stothard (University of Alberta, Canada).

Reverse Complement - converted a DNA sequence into its reverse, complement, or reverse-complement counterpart. The entire IUPAC DNA alphabet is supported, and the case of each input sequence character are maintained.

ORF Finder - searched for open reading frames (ORFs) in the DNA sequence. The program returned the range of each ORF, along with its protein translation. ORF Finder supports the entire IUPAC alphabet and several genetic codes. Here, bacterial genetic code was selected in chloroplast *matK* sequences. ORF Finder was used to search newly sequenced DNA for potential protein encoding segments.

Pairwise Align DNA - accepted two DNA sequences and determined the optimal global alignment. Pairwise Align DNA was used to look for conserved sequence regions.

Sequence submission in Genbank:

BankIt submission:

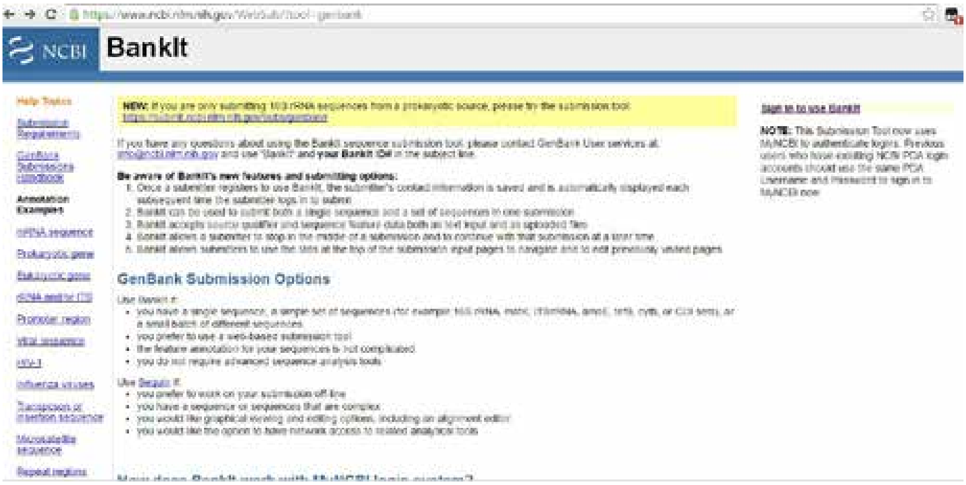
On entering the BankIt submission, user is asked about the contact person (the individual to whom the database staff may address any questions), the citations (who gets the scientific credit), the organism (the top 100 organisms are on the form; all others must be typed in), the location (nuclear *vs*. organelle), some map information, and the nucleotide sequence itself.

The data that you will need consist of 3 parts:

Sequence file: The sequences themselves, in FASTA format

Here is a example of sequence file:

>SCPM-01 [Organism=*Dendrobium jenkinsii*] internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence;



https://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank

NCBI BankIt

Help Topics

Submission Requirements

GenBank Submissions Handbook

Annotation Examples

mRNA sequence

Prokaryotic gene

Eukaryotic gene

rRNA and/or ITS

Promoter region

Viral sequence

HIV-1

Influenza viruses

Transposon or insertion sequence

Microsatellite sequence

Repeat regions

NEW: If you are only submitting 16S rRNA sequences from a prokaryotic source, please try the submission tool: https://submit.ncbi.nlm.nih.gov/subs/genbank/

If you have any questions about using the BankIt sequence submission tool, please contact GenBank User services at: info@ncbi.nlm.nih.gov and use "BankIt" and **your BankIt ID#** in the subject line.

Be aware of BankIt's new features and submitting options:

1. Once a submitter registers to use BankIt, the submitter's contact information is saved and is automatically displayed each subsequent time the submitter logs in to submit.
2. BankIt can be used to submit both a single sequence and a set of sequences in one submission.
3. BankIt accepts source qualifier and sequence feature data both as text input and as uploaded files.
4. BankIt allows a submitter to stop in the middle of a submission and to continue with that submission at a later time.
5. BankIt allows submitters to use the tabs at the top of the submission input pages to navigate and to edit previously visited pages.

GenBank Submission Options

Use BankIt if:

- you have a single sequence, a single set of sequences (for example 16S rRNA, matK, ITS/rbcL, amoA, trnL, cytb, or COI sets), or a small batch of different sequences
- you prefer to use a web-based submission tool
- the feature annotation for your sequences is not complicated
- you do not require advanced sequence analysis tools

Use Sequin if:

- you prefer to work on your submission off-line
- you have a sequence or sequences that are complex
- you would like graphical viewing and editing options, including an alignment editor
- you would like the option to have network access to related analysis tools

How does BankIt work with MyNCBI login system?

Sign in to use BankIt

NOTE: This Submission Tool now uses MyNCBI to authenticate logins. Previous users who have existing NCBI PDA login accounts should use the same PDA Username and Password to sign in to MyNCBI now.

Fig. 4.3: web portal of Bankit in NCBI

AACGAGCGATTTAGAGAACCTGTTAAAATAATCGGTGGCTGTTGTTACCGTGATAAATTCCATCCAAGTCGTTGCCTCATGTCCTCTTGGGGGCTGGATGCGATGAAGGATGGATGAACACTCAAACCGGCGCAGCATCGCGCCAAGTCAAAATATTGAAAGACAAGCCCTTAAAAGGGTTTGGTGGC

>SCPM-02 [Organism=*Dendrobium loddigesii*] internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence;

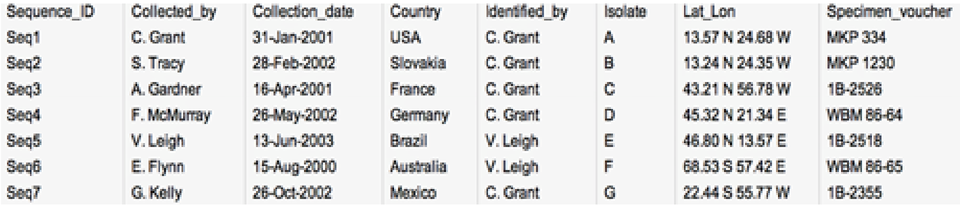
TTGTCGAGACTGAAATATATTGAGCGATCTTGAGAACCCGTCAAAATAAGCGATGGCTATAGTTGTCGAGATAAAATTCATCCCAGTCGTCATGTCATCCTCTTTTGCGGGGTTGGGGACATGATGAAGGATGGATGAACCCACAAATCGGCGCAGCATCGCGCCAAGGAAATAATGAAATACGAGCCCTAAAATGGGTTTTATGAAATGGGGTGTTGTTGCATTCCTTATGATTGACAT

>SCPM-03 [Organism=*Dendrobium amoenum*] internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence;

TCGAGACTGAAAAACGAGCGATTTTGAGAACCCGTAAAAATAAGCGGCGGCTCTTGCTGCTGAGACAAAATCCAGCCTGGTCATCGCCTCTTCCCCTTTCCGGGGTGGGGACGTGATCAAGGATGGATGAACCCTCAAATCGGCGCAGCCTTG

- Source modifier: Information about the sequence and the specimen from which it is derived.

Here is a sample source modifier table:



Sequence_ID	Collected_by	Collection_date	Country	Identified_by	Isolate	Lat_Lon	Specimen_voucher
Seq1	C. Grant	31-Jan-2001	USA	C. Grant	A	13.57 N 24.68 W	MKP 334
Seq2	S. Tracy	28-Feb-2002	Slovakia	C. Grant	B	13.24 N 24.35 W	MKP 1230
Seq3	A. Gardner	16-Apr-2001	France	C. Grant	C	43.21 N 56.78 W	1B-2526
Seq4	F. McMurray	26-May-2002	Germany	C. Grant	D	45.32 N 21.34 E	WBM 86-64
Seq5	V. Leigh	13-Jun-2003	Brazil	V. Leigh	E	46.80 N 13.57 E	1B-2518
Seq6	E. Flynn	15-Aug-2000	Australia	V. Leigh	F	68.53 S 57.42 E	WBM 86-65
Seq7	G. Kelly	26-Oct-2002	Mexico	C. Grant	G	22.44 S 55.77 W	1B-2355

- Feature table: Annotation of particular DNA sequence region.

Example of preparation of feature for a protein Coding Sequence

```
>Feature SGPM-MP1

<1  >669    gene

                    gene    matK

<1  669     CDS

                    productmaturase K

                    codon_start     1
```

BOLD SUBMISSION:

BOLD provides users with several pathways for direct submission of their data, which can include specimen collaterals, sequences, trace files and images. Once data have been injected, there are several tools to aid their review, including a search engine that enables the retrieval of records based on multiple criteria. Specimen records in each project can also be sorted by a variety of factors including taxonomy, sequence length and specimen record number. Finally, because specimen information may need updating, an edit function is accessible from this page.

However, the specimen record will not gain formal barcode status until seven data elements are in place:

1. **Species name:**
2. **Voucher data:** catalogue number and institution storing
3. **Collection record:** collector, collection date and location with GPS coordinates.
4. **Identifier of the specimen:**
5. **Barcode sequence**: COI sequence of at least 500 bp from the 5'-end.
6. **Polymerase chain reaction (PCR) primers:** used to generate the amplicon.
7. **Sequence Trace files:**

Specimen data can be uploaded to bold using either online forms (for small numbers of specimen records) or through standardized spreadsheets. Trace files, specimen images and sequence records are also uploaded directly, allowing users to have immediate access to their submitted data. Although bold currently only supports trace files from ABI sequencers, other trace formats will be added as the need arises. Data that reside in bold can be readily exported for use in other analytical packages. The simplest forms of data export, downloads from single projects, are available directly from the project management console. The sequence-export function generates a FASTA file for all sequence records in a project, each labeled with a species name and specimen identifiers (*i.e.* voucher, sequence ID). Another function, the data workbook, generates an Excel spreadsheet that includes all collateral details (taxonomy, collection dates, *etc.*) for each specimen in a project, while the label-maker function generates labels in varied formats (pinned insects to vertebrate skins) to identify specimens which have been barcoded. Finally, the trace-file function provides access to the raw sequence traces that underpin the sequence records.

5

Bioinformatics Analysis

Bioinformatics is a biological science, the science of using information to understand biology. In a literal sense Bioinformatics is the integration of life sciences and information science. A common definition is the 'Science of organizing and analyzing increasingly complex biological data resulting from modern molecular and biochemical techniques'. But a classical definition describes it as 'the mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information'. Bioinformatics is conceptualizing biology in terms of molecules and applying 'informatics techniques' to understand and to organize the information associated with these molecules, on a large scale. In short, information science has been applied to manage the information science has been applied to manage the information generated in molecular biology to produce the field called ***Bioinformatics.***

Towards practical uses of DNA barcoding approach for assigning sequences to species, various methods have been proposed. The most common bioinformatics method for species identification are similarity search, phylogenetic analysis, genetic distance calculation, character based method. For this purpose, DNA barcoding researchers utilizes DNA databases (GenBank, BOLD), similarity search tools (BLAST), tree construction methods (NJ), genetic distance calculation or barcode gap or intra and inter specific (K2P), character based methods (Multiple sequence alignment).

Database

A database is a collection of information stored in a computer in a systematic way, such that a computer program can access it easily. Databases that are available via the web also became an indisable tool for biological research. The store data need to be accessed in a meaningful way, and often contents of several databanks or databases have to accessed simultaneously and correlated with each other. Bioinformatics databases are publicly available and are designed, developed and maintained by different organization located across countries in the world. The Barcode of Life Data System (BOLD) has evolved into primary resources for the DNA barocding community as well as NCBI or its sister genomic repositories, DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory (EMBL).

The Barcode Of Life Data System (BOLD):

The Barcode Of Life Data System - www.barcodinglife.org is an informatics work beach aiding the acquisition, storage, analysis and publication of Dan barcode records. By assembling molecular and morphological and distribution data, it's bridge a traditional bioinformatics chasm (Ratnasingham and Hebert 2007). BOLD is freely available to any researcher with interest in DNA barocding.

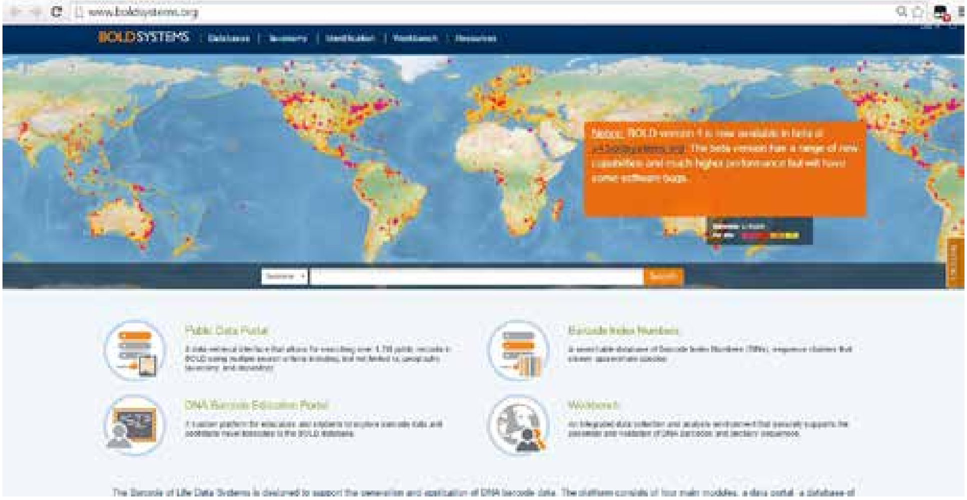
BOLD database resources are four different parts.

Public data portal: contain the entire DNA barcode sequences on BOLD. This database can be used to access and download the associated specimen data and sequences.

BIN database: Barcode Index Numbers (BINs) are an interim taxonomic system for animals. Barcodes are clustered algorithmically, generating a web page for each cluster which is deposited in this database.

Primer database: A searchable database of barcode primers, which includes primer statistics.

Publication database: A searchable, community maintained database of barcode papers linked to published datasets. Search by title, abstract or author keywords. Any registered researcher can start new barcode project for improving communication and preventing data loss or duplication. Projects are not subjected to any centralized review: data quality is ultimately depend upon the project participants.


www.boldsystems.org
BOLDSYSTEMS
Databases | Taxonomy | Identification | Workbench | Resources
Search
Public Data Portal
Barcode Index Numbers
DNA Barcode Education Portal
Workbench

Notice: BOLD version 4 is now available in beta at v4.boldsystems.org. The beta version has a range of new capabilities and much higher performance but will have some software bugs.

Public Data Portal: A data retrieval interface that allows for searching over 1.7M public records in BOLD using multiple search criteria including, but not limited to, geography, taxonomy, and depository.

Barcode Index Numbers: A searchable database of Barcode Index Numbers (BINs), sequence clusters that closely approximate species.

DNA Barcode Education Portal: A custom platform for educators and students to explore barcode data and contribute novel barcodes to the BOLD database.

Workbench: An integrated data collection and analysis environment that securely supports the assembly and validation of DNA barcodes and ancillary sequences.

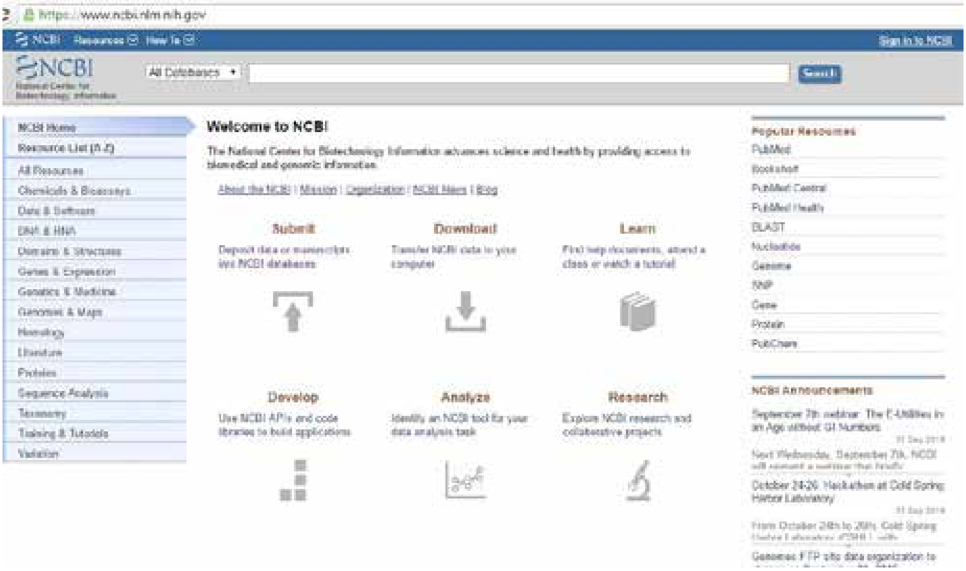
The Barcode of Life Data Systems is designed to support the generation and application of DNA barcode data. The platform consists of four main modules, a data portal, a database of

Fig. 5.1: Barcode of life webpage portal

In BOLD, algorithms for recognition and identification align barcodes from a new specimen to the existing multiple alignment to calculate "distances" from the new specimen to database specimens. The new specimen can have one or several "nearest neighbors" in the database, because the new specimen might be at the nearest distance to several database specimens. Based on the nearest neighbors in the database, the algorithms then yield: *(1)* a list of possible species to which the specimen belongs; and *(2)* the nearest-neighbor distance, which indicates how probable it is that the specimen belongs to one of the species. The distance has the smallest value of 0, achieved for identical sequences (and possibly others); and its increase suggests a decreasing probability of correct identification. Most algorithms have similar output as nearest-neighbor algorithms, substituting only another type of number, e.g., a probability for a distance. No algorithm seems to improve noticeably on the identifications from a nearest neighbor identification algorithm (Austerlitz, 2007). To recognize a known species, BOLD applies a threshold to the nearest-neighbor distance. If the distance is less than the threshold, the specimen is recognized as a known species. Whereas a probabilistic algorithm can automatically adjust parameters as species and sequences enter a barcode database, a nearest-neighbor recognition algorithm must make ad hoc adjustments to its species specific thresholds. Alignments between intergenic barcode sequences generally produce few sites displaying point mutations. At present, evolutionary distances examine only point mutations: they discard the insertions and deletions that occur in intergenic barcode alignments. Thus, in theory, nearest-neighbor algorithms for intergenic barcodes should probably use alignment distances (which account for gaps), and not evolutionary distances (which do not). Preliminary results indicate that even with evolutionary distances, the identification algorithms perform well with intergenic barcodes, however, because of their enhanced variability.

Genbank

GenBank is a comprehensive database of publicly available DNA sequences for more than 300,000 named organisms, obtained through submissions from individual laboratories and batch submissions from large-scale sequencing projects. GenBank is maintained and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), at the US National Institutes of Health (NIH) in Bethesda, MD. NCBI builds GenBank from several sources including the submission of sequence data from authors and from the bulk submission of expressed sequence tag



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

NCBI Announcements

September 7th webinar: The E-Utilities in an Age without GI Numbers
[illegible]
Next Wednesday, September 7th, NCBI will present a webinar that briefly

October 24-26: Hackathon at Cold Spring Harbor Laboratory
[illegible]
From October 24th to 26th, Cold Spring Harbor Laboratory (CSHL), with

Genomes FTP site data organization to

(EST), genome survey sequence (GSS), whole genome shotgun (WGS) and other high-throughput data from sequencing centers. The U.S. Office of Patents and Trademarks also contributes sequences from issued patents.

There are two ways to search GenBank: a text-based query can be submitted through the Entrez system at www.ncbi.nlm.nih.gov/Entrez/ or a sequence query can be submitted through the BLAST family of programs (see http://www.ncbl.nlm.nih.gov/BLAST/). To search GenBank through the Entrez system you would select the Nucleotides database from the menu. The Entrez Nucleotides Database is a collection of sequences from several sources, including GenBank, RefSeq, and the Protein Databank, so you don't actually search GenBank exclusively. Searches of the Entrez Nucleotides database query the text and numeric fields in the record, such as the accession number, definition, keyword, gene name, and organism fields to name just a few. Nucleotide sequence records in the Nucleotides database are linked to the PubMed citation of the article in which the sequences were published. Protein sequence records are linked to the nucleotide sequence from which the protein was translated. If you have obtained a record through a text-based Entrez Nucleotides Database search you can read the nucleotide sequence in the record. However, most researchers wish to submit a nucleotide sequence of interest to find the sequences that are most similar to theirs. This is done using the BLAST (Basic Local Alignment Search tool) programs. You select the BLAST program you wish to use depending upon the type of comparison you are doing (nucleotide to nucleotide, or nucleotide to protein sequence, etc.) and then you select the database to run the query in (any of several nucleotide or protein databases).

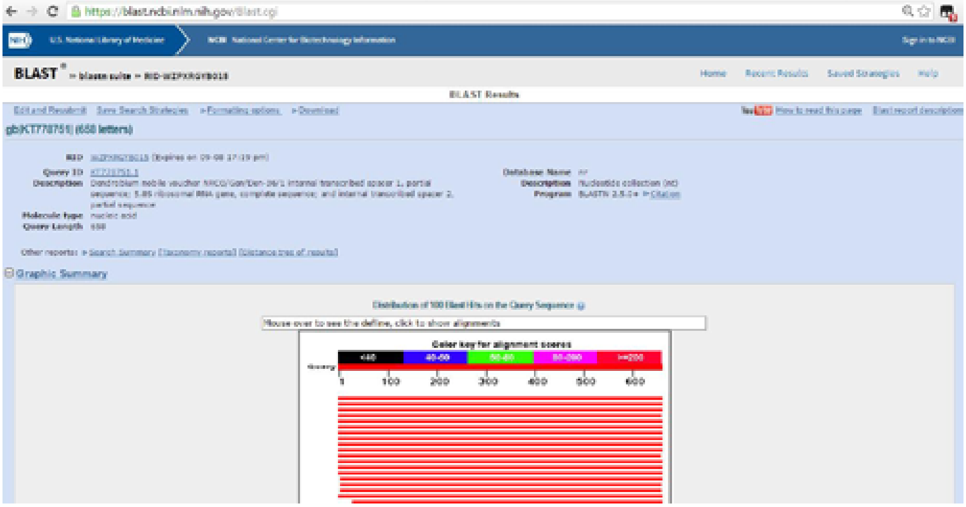
Similarity search

Database sequence similarity searching is an important methodology in DNA barocding. Database searches reveal biologically significant sequence relationships and suggest future investigation strategies. Sequence alignment provide a powerful tool to compare novel sequences with previously characterized gene in a database of potentially unrelated sequences.Sequence alignment is the most common way of comparing DNA,RNA or amino acids. Sequence alignmentis the procedure of comparing two (pariwise) or more (multiple) sequences by searching for a series of character patterns that are same oder in the alignments correspond to mutation and gap correspond to insertion and deletions. Gaps are also introduced to more similar characters between the sequence involve.

BLAST: Bioinformaticians have developed so called 'heuristic' algorithms, which allow searching a database in considerably less time. The most popular one is Basic Local Alignment Search Tool (BLAST) Percent similarity of the resulting DNA or protein sequences was analyzed through BLAST (Altschul *et al.,* 1997; http://www.ncbi.nlm.nih.gov/blast/), a choice is offered between the different BLAST programs through different hyperlinks (nucleotide blast, protein blast, blastx, tblastn, tblastx etc).

As query for barcode sequences in nucleotide BLAST (BLASTN), were used in FASTA format or Accession No. We specified the database because our sequences were from mitrochondria or chloroplast DNA. So,we choice the"others" database. We also selected megablast under the program selection header, which optimized the search for highly similar sequences and clicked on the Blast button to initiate the search. The output of the megablast search contain a table with sequencing producing significant alignments. We also used specialized BLAST in bl2seq for the alignment two (or more) sequence and primer blast to make specific primer.

BLAST uses statistical method to evaluate hits for their significance. BLAST program identifies sequences having share common words of a pre-set size


https://blast.ncbi.nlm.nih.gov/Blast.cgi
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST ® » blastn suite » RID-WZPXRGYB013
Home
Recent Results
Saved Strategies
Help
BLAST Results
Edit and Resubmit
Save Search Strategies
▸Formatting options
▸Download
How to read this page
Blast report description
gb|KT778751| (658 letters)
Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results]
Graphic Summary
Distribution of 100 Blast Hits on the Query Sequence
Mouse over to see the define, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
100
200
300
400
500
600

RID WZPXRGYB013 (Expires on 09-08 17:19 pm)
Query ID KT778751.1
Description Dendrobium nobile voucher NRCO/Gan/Den-26/1 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence
Molecule type nucleic acid
Query Length 658
Database Name nr
Description Nucleotide collection (nt)
Program BLASTN 2.5.0+ ▸Citation

Figure : Showing the BLAST result on http://blast.ncbi.nlm.nih.gov/Blast.cgi

(K-tuple) in the database sequences and these matching words are extended only if they score higher than pre defined threshold. The default threshold for the E( ) value is 10 and default word size is 11. The E- value is the probability that the query match is due to randomness. The lower the E-value to the more significant the match. The score (inbits) is a value attributed to the alignment but is independent of the scoring matrix used, while E-values of 10-3 and below are often considered indicative of statistically significant results.

Multiple Sequence Alignment

A multiple sequence alignment (MSA) arranges a set of sequences in a scheme where positions believed to be homologous are written a common column. The gap represent a deletion, an insertion in the sequences that do not have a gap, or a combination of insertion and deletions.MSA gives biologist the ability to extract biological important but perhaps widely dispersed sequence similarity that can give biologist hints about the evolutionary history of certain sequences. The MSA , homology search algorithm is some time called 'many- against- eachother' search because the input is a small defined set of sequences which are compared only against each other, not against an entire database. There are several approaches, one of the most popular being the progressive alignment strategy used by the clustal family of programs.

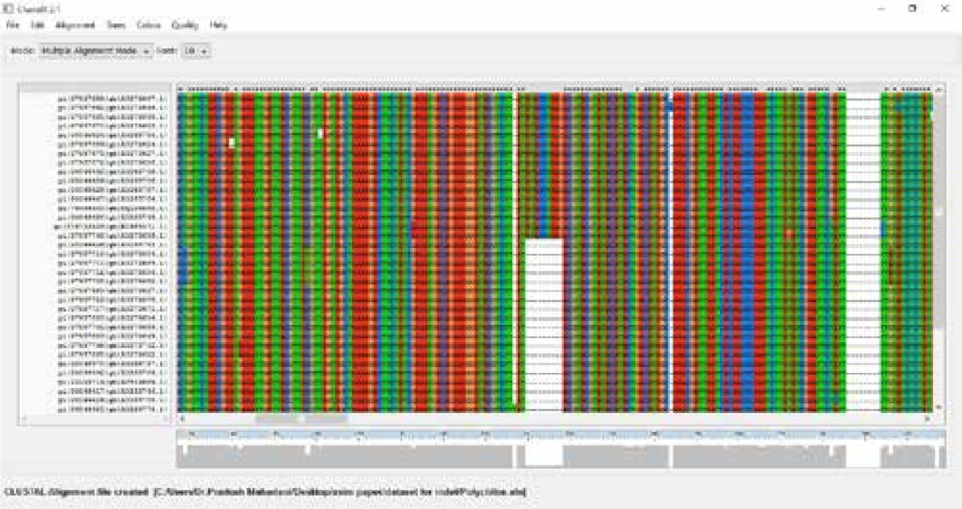
AF2587	CTGAAATTCTACTTAAACTATTCCTTGATTTCTTCCCCTAAACGACAACAATTCACCCTC
AF2592	CTGAAATTCTACTTAAACTATTCCTTGATTTCTTCCCCTAGACGACAACAATTCACCCTC
AF2591	CTGAAATTCTACTTAAACTATTCCTTGATTTCTTCCCCTAAACGACAACAATTCACCCTC
AF2594	CTGAAATTCTACTTAAACTATTCCTTGATTTCTTCCCCTAAACGACAACAACTCACCCTC
AF2593	CTGAAATTCTACTTAAACTATTCCTTGATTTCTTCCCCTAAACGACAACAATTCACCCTC

```
AF2589      CTGAAATTCTACTTAAACTATTCCTTGATTTCTTCCCCTAAACG
            ACAACAATTCACCCTC

AF2588      CTGAAATTCTACTTAAACTATTCCTTGATTTCTTCCCCTAAACG
            ACAACAATTCACCCTC

            ******************************************************************************
```

Fig. 5.3: Example of Multiple sequence alignment

CLUSLAT: The most commonly used software for progressive alignment is CLUSTALW (Thompson et al., 1994) and CLUSTALX (Thompson et al., 1997).This programs are identical to each other in term of alignment method but offer either a simple text interface (ClustalW) suitable for high- throughput tasks or a graphical interface (ClustalX). Clustal X and Clustal W will take a set of input sequences and carry out the entire progressive alignment procedure automatically. The sequences are aligned in pairs in order to generate a distance matrix that can be used to generate a distance matrix that can be used to make a simple initial tree of the sequence. Finally, the multiple sequence alignment is carried out using the progressive approach.

ClustalW and ClustalX are both freely available and can be downloaded from the EMBL/EBI file server (*ftp://ftp.ebi.ac.uk/pub/software/*) or from ICGEB in Strasbourg, France (*ftp://ftp-igbmc.u strasbg.fr/pub/*ClustalW/ and *ftp://ftp-igbmc.u-strasbg.fr/pub/*ClustalX/).In each case, ClustalX (X stands for X windows) provides a graphical user interfacewith colorful display of alignments. Open ClustalX and open the sequence file using File → Load Sequences. The graphical display allows the user to slide over the unaligned protein sequences. Select Do complete Alignment from the Alignment menu. ClustalX performs the progressive alignment (progress can be followed up in the lower left corner), and creates an output guide tree file and an output alignment file in the default Clustal format. It is, however, possible to choose a different format in the Output Format Options from the Alignment menu. ClustalX also allows the user to change the alignment parameters (from Alignment Parameters in the Alignment menu). If an alignment shows, for example, too many large gaps, the user can try to increase the gap-opening penalty and redo the alignment. ClustalX indicates the degree of conservation at the bottom of the aligned sequences, which can be used to evaluate a given alignment.



CLUSTAL Alignment file created [illegible]

Fig. 5.4: Result showing alignment though CLUSTAL

Phylogenetic Analysis Based On Dna Barcode Sequences

Phylogenetics is the science of estimating the evolutionary past, in the case of molecular phylogeny, based on the comparison of DNA or protein sequences. DNA barcodes are used both to identify species and to draw attention to overlooked and new species, they can help identify candidate exemplar taxa for a comprehensive phylogenetic study. Barcode of Life projects create a perfect taxonomic sampling environment for conducting phylogenetic studies on different branches of the Tree of Life. Barcode sequence data can also provide a shared genomic cornerstone for the variable repertoire of genes that can be used to build the phylogenetic tree. It can be used as a link between the deeper branches of the tree to its shallow, species-level branches. barcode sequences have been analyzed mainly by using phylogenetic tree reconstruction methods such as NJ, these barcode-based trees should not be interpreted as phylogenetic trees.

Basic Terminology for phylogentic:

Phylogenetic tree:

Phylogenetic tree is a graphical representation of the evolutionary relationship among three or more gene or organism.

Phylogram:

A phylogram depicts the phylogenetic relationships between a group of taxa with branch lengths representing evolutionary distances, which were inferred using a phylogenetic approach.

Cladogram:

The ancestor–descendant relationship between species (or groups of organisms) can be represented using a ***cladogram***, which is not necessarily based on phylogenetic analysis.

Rooted tree:

A single node is designated as a common ancestor and unique path leads to evolutionary time to any other node. No of possible rooted tree can be determined with following equation:

Rooted $(N_R) = (2n-3)!/2^{n-2}(n-2)!$ N=No of data set

Unrooted tree :

An ***unrooted*** tree only positions the individual taxa relative to each otherwithout indicating the direction of the evolutionary process. In an unrooted tree, there is no indication of which node represents the ancestor. No of possible unrooted tree can be determined with following equation:

Rooted $(N_u) = (2n-3)!/2^{n-3}(n-3)!$ N=No of data set

Outgroup A taxon that is used to root a phylogenetic tree and thus providing directionality to the evolutionary history. An outgroup taxon is not considered to be part of the group in question (*the ingroup*), but preferably, it is closely related to that group. In a cladistic analysis, an outgroup is used to help resolve the polarity of characters, which refers to their state being original or derived.

Midpoint rooting The midpoint rooting method places the root of a tree at the midpoint of the longest distance between two taxa in a tree.

Transition/transversion ratio In nucleotide ***substitution models***, the ratio between ***transition*** changes (purine to purine or pyrimidine to pyrimidine) and ***transversion*** changes (purine to pyrimidine or pyrimidine to purine). Since there are twice as many possible transversions and transitions, a ratio of 0.5 indicates that all base changes are equally likely.

Genetic distance: In evolutionary biology, genetic distance is a measure of the evolutionary divergence or dissimilarity between the genetic material of different species or individual of the same species. Genetic distances estimated from nucleotide sequences are generally based on pairwise difference between two data set.

Paraphyly/paraphyletic: In phylogenetics, a group of taxa is paraphyletic or represents a paraphyly if the group does not include all descendants from its inferred common ancestor.

Branch length: In sequence analysis, the number of sequence changes along a particular branch of a phylogenetic tree.

Bootstrap value*:* Bootstrap analysis is a widely used sampling technique for estimating the statistical error insituations in which the underlying ***sampling distribution*** is either unknown or difficult to derive analytically (Efron&Gong, 1983).The bootstrap method offers a useful way to approximate the underlying distribution by resampling from the original data set. Felsenstein (1985) first applied this technique to the estimation of confidence intervals for phylogenies inferred from sequence data. First, the sequence data are bootstrapped, which means that a new alignment is obtained from the original by randomly choosing columns from it with replacements. Each column in the alignment can be selected more than once or not at all until a new set of sequences, a *bootstrap replicate*, the same length as the original one has been constructed. Therefore, in this resampling process, some characters will not be included at all in a given bootstrap replicate and others will be included once, twice, or more. Second, for each reproduced (*i.e.* artificial) data set, a tree is constructed, and the proportion of each clade among all the bootstrap replicates is computed. This proportion is taken as the statistical confidence supporting the monophyly of the subset.

METHODS OF TREE CONSTRUCTION:

The methods for calculating phylogenetic trees fall into two general categories. These are distance-matrix methods, also known as clustering or algorithmic methods (e.g. UPGMA, neighbour-joining, Fitch Margoliash), and discrete data methods, also known as tree searching methods (e.g. parsimony, maximum likelihood, Bayesian methods). Distance is relatively simple and straightforward – a single statistic, the distance (roughly, the percent sequence difference), is calculated for all pairwise combinations of OTUs, and then the distances are assembled into a tree. Discrete data methods examine each column of the alignment separately and look for the tree that best accommodates all of this information. Unsurprisingly, distance methods are much faster than discrete data methods. However, a distance analysis yields little information other than the tree. Discrete data analyses, however, are information rich; there is an hypothesis for every column in the alignment, so you can trace the evolution at specific sites in the molecule. Barcode sequences have been analyzed mainly by using phylogenetic tree reconstruction methods NJ with K2P model, these barcode-based trees should not be interpreted as phylogenetic trees.

Neighbor-joining method:

Neighbor-joining (***NJ***) A heuristic method for estimating the ***minimum evolution*** tree originally developed by Saitou and Nei (1987) and modified by Studier and Keppler (1988). NJ is conceptually related to clustering, but does not require the data to be ***ultrametric***. The principle of NJ is to find pairs of operational taxonomic units (OTUs) that minimize the total branch length at each stage of clustering of OTUs starting with a star-like tree. The neighbor-joining method is therefore a special case of the ***star decomposition*** method. The ***Neighbor-joining*** (***NJ***) method constructs a tree by sequentially finding pairs of neighbors, which are the pairs of OTUs connected by a single interior node. This algorithm does not attempt to cluster the most closely related OTUs, but rather minimizes the length of all internal branches and thus the length of the entire tree. The NJ algorithm starts by assuming a star-like tree that has no internal branches. In the first step, it introduces the first internal branch and calculates the length of the resulting tree. The algorithm sequentially connects every possible OTU pair and finally joins the OTU pair that yields the shortest tree. The length of a branch joining a pair of neighbors, X and Y to their adjacent node is based on the average distance between all OTUs and X for the branch to X, and all OTUs and Y for the branch to Y, subtracting the average distances of all remaining OTU pairs. This process is then repeated, always joining two OTUs (neighbors) by introducing the shortest possible internal branch. The *Fitch–Margoliash* method is a distance-matrix method that evaluates all possible trees to find the tree that minimizes the differences between the pairwise genetic distances and the distance represented by the sum of branch lengths for each pair of taxa in the tree. NJ has the advantage of being very fast, which allows the construction of large trees including hundreds of sequences; this significant difference in speed of execution compared to other distance methods has undoubtedly accounted for the popularity of the method. Distance methods are implemented in many different software packages, including Phylip, Mega, Treecon , Paup* , Dambe , and many more.

MEGA 5:

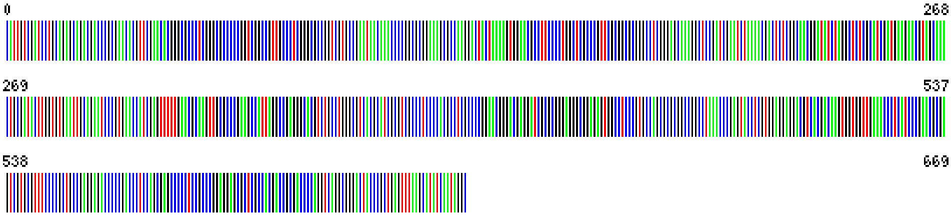
Here, We were perform tree inference using the Mega 5 program and the input file should .meg format. The file contains aligned DNA sequences in mega format. Using Mega 5 it is possible to estimate a NJ tree and perform the bootstrap test in an automated fashion. The program will display the tree in a new window and superimpose bootstrap support values along each branch of the tree. To estimate a NJ tree using Kimura-2P (K2P) corrected distances and perform bootstrap analysis on 1000 replicates, open the align file in Mega4 and select the submenu Bootstrap Test of Phylogeny > Neighbor-Joining from the Phylogeny menu in the Mega4 main window. The Analysis Preferences window will appear. Click on the green square to the right of the Gaps/missing data row and select pairwise deletion (specifying that for each pair of sequences only gaps in the two sequences being compared should be ignored). Similarly, select in the Model row Nucleotide >Kimura-2-parameter. To set the number of bootstrap replicates, click on the Test of Phylogeny tab on the top of the window and enter 1000 in the Replications cell. Select again the Option Summary tab and click on the compute button at the bottom of the window. After a few seconds (or a few minutes, depending on the speed of your computer processor) the NJ tree with bootstrap values will appear in the Tree Explorer window. By default, the tree is midpoint rooted. If the location of the root needs to be placed on any other branch of the tree, this can be done by selecting the top button on the left side of the window (the button is indicated by an icon representing a phylogenetic tree with a green triangle on its left), placing the mouse on the branch chosen as the new root and clicking on it: a re-rooted tree will be displayed in the same window. To go back to the midpoint-rooted tree, simply select Root on Midpoint from the View menu.

Vanda bicolor voucher NRCO/Gen/Van-22/D1 internal transcribed spacer 1, partial sequence

LOCUS KX298642 568 bp DNA linear PLN 12-JUL-2016
DEFINITION Vanda bicolor voucher NRCO/Gen/Van-22/D1 internal transcribed
spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete
sequence; and internal transcribed spacer 2, partial sequence.
ACCESSION KX298642
VERSION KX298642.1 GI:1042762808
ORGANISM ***Vanda bicolor***
Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
Tracheophyta;Spermatophyta; Magnoliophyta; Liliopsida; Asparagales;
Orchidaceae;Epidendroideae; Vandeae; Aeridinae; Vanda.
REFERENCE 1 (bases 1 to 568)
AUTHORS Chakrabarti,S., Mahadani,P., Jain,S.K. and Sevanthi,A.M.
TITLE DNA Barcoding of native Vanda orchids of India
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 568)
AUTHORS Chakrabarti,S., Mahadani,P., Jain,S.K. and Sevanthi,A.M.
TITLE Direct Submission
JOURNAL Submitted (16-MAY-2016) Genetics, **ICAR-National Research Centre for Orchids, Dickling Road, Pakyong, Gangtok, Sikkim 737106, India**
Location/Qualifiers source 1..568
/organism="Vanda bicolor"
/mol_type="genomic DNA"
/isolate="A"
/specimen_voucher="NRCO/Gen/Van-22/D1"

/db_xref="taxon:1712275"
/country="India"
/lat_lon="27.546 N 93.816 E"
/collection_date="23-Apr-2013"
/collected_by="S K Jain"
/identified_by="S K Jain" misc_RNA <1..>568
/note="contains internal transcribed spacer 1, 5.8S ribosomal RNA, and internal transcribed spacer 2"ORIGIN

1 tggccccccc tgtctggagg gggccgcgat gagggacggc tgaaacccca aaccggcgca
61 gactggcgcc aaggtaacta tcgaaaggca cgagcccgac atcgggtcct cgtggggcgg
121 agcggtgttg cgcaccgcac gtattgacac gactctcgac aatggatatc tcggctctcg
181 catcgatgaa gagcgcagcg aaatgcgata cgtggtgcga attgcagaat cccgcgaacc
241 atcgagtctt tgaacgcaag ttgcgcccga ggccaatcgg tcgagggcac gtccgcctgg
301 gcgtcagacg ttgcgtcgct ccgtgccaag tccacgccgc ctcaccgtag tgggggtgcc
361 gggcgaggct cggatgtgca gggtggcccg tcgtgcccat cggtgcggcg ggctgaagag
421 cgggttgtca tctcataggc cacgaacaac gaggggtgga tgaaagctgc cgcgggcaag
481 gcccgcgttg tctcgtgccg gcccgagaga agattgcacc cttcgtgcga tcccatccca
541 tgcgccgccc ccgcgcggcg gctgaacg



//

Cymbidium eburneum voucher NRCO/Gen/Cym-4/1 maturase K (matK) gene, partial cds; chloroplast

LOCUS KX298605 726 bp DNA linear PLN 12-JUL-2016

DEFINITION *Cymbidium eburneum* voucher NRCO/Gen/Cym-4/1 maturase K (matK) gene,partial cds; chloroplast.

ACCESSION KX298605

VERSION KX298605.1 GI:1042762736

SOURCE chloroplast *Cymbidium eburneum*

ORGANISM ***Cymbidium eburneum***

Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;Spermatophyta; Magnoliophyta; Liliopsida; Asparagales; Orchidaceae;Epidendroideae; Cymbidieae; Cymbidiinae; Cymbidium.

REFERENCE 1 (bases 1 to 726)

AUTHORS Chakrabarti,S.,Mahadani,P., Sevanthi,A.M. and Singh,D.R.

TITLE Authentication of Cymbidium (Orchidaceae) species through DNA barcoding from Northeast India

AUTHORS Chakrabarti,S.,Mahadani,P., Sevanthi,A.M. and Singh,D.R.

TITLE Direct Submission

JOURNAL Submitted (16-MAY-2016) Genetics, ICAR-National Research Centre for Orchids, Dickling Road, Pakyong, Gangtok, Sikkim 737106, India Location/Qualifiers

source 1..726

/organism="*Cymbidium eburneum*"
/organelle="plastid:chloroplast"
/mol_type="genomic DNA"
/isolate="A"
/specimen_voucher="NRCO/Gen/Cym-4/1"
/db_xref="taxon:160528"
/country="India"
/lat_lon="27.34 N 88.19 E"

```
                /collection_date=”26-Mar-2015”
                /collected_by=”S Chakrabarti”
                /identified_by=”S Chakrabarti”
gene            <1..>726
                /gene=”matK”
CDS             <1..>726
                /gene=”matK”
                /codon_start=3
                /transl_table=11
                /product=”maturase K”
                /protein_id=”ANQ45597.1”
                /db_xref=”GI:10427627
translation=”LFFHEYHNLNSLITSNKSIYVFSKRKKRLFWFLHNSYVYEFEYL
FLFLRKKSSYLRSISSGVFIERTLFFGKIEYLMVVCCNSFQRILWFLKDTFIHYVRYK
GKAILASKGTLILMKKWKFHLVNFWQSYFHFWFQPYRIHIKQLPNYSFSFLGYFSSVL
KNPLVVRNQMLENSFIINTLTNKLDTIAPVISLIGSLSKAQFCSVLGNPISKPIWTDL
                SDSDIIDRFCRICRNLCHYHSGS”
ORIGIN
      1 gattgttttt ccacgaatat cataatttga
atagtctgat tacttcaaat aaatctattt
     61 acgtcttttc aaaaagaaag aaaagattat
tttggttcct acataattct tatgtatatg
    121 aatttgaata tctattccta tttcttcgta
aaaagtcttc ttatttacga tcaatatctt
    181 ctggagtctt tattgagcga acacttttct
ttggaaaaat agaatatctt atggtcgtgt
    241 gttgtaattc ttttcagagg atcctatggt tcctcaaaga tactttcata cattatgttc
    301 gatataaagg aaaagcgatt ctggcttcaa aaggaactct tattctgatg aagaaatgga
    361 aatttcatct tgtgaatttt tggcaatctt attttcactt ttggtttcaa ccttatagga
    421 tccatataaa gcaattaccc aactattcct tctcttttct ggggtatttt tcaagtgtac
    481 taaaaaatcc tttggtagta agaaatcaaa tgctagagaa ttcatttata ataaatactc
    541 tgactaataa attagatacc atagccccag ttatttctct tattggatca ttgtcgaaag
    601 ctcaattttg tagtgtattg ggtaatccta taagtaaacc gatctggacc gatttatcgg
    661 attctgatat tattgatcga ttttgtcgga tatgtagaaa tctttgtcat tatcacagtg
    721 gatcct
//
```

Dendrobium nobile voucher NRCO/Gen/Den-36/1 internal transcribed spacer

LOCUS KT778751 658 bp DNA linear PLN 01-MAR-2016

DEFINITION *Dendrobium nobile* voucher NRCO/Gen/Den-36/1 internal transcribed

spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete

sequence; and internal transcribed spacer 2, partial sequence.

ACCESSION KT778751

SOURCE *Dendrobium nobile*

ORGANISM *Dendrobium nobile*

Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;Spermatophyta; Magnoliophyta; Liliopsida; Asparagales; Orchidaceae;Epidendroideae; Malaxideae; Dendrobiinae; Dendrobium.

REFERENCE 1 (bases 1 to 658)

AUTHORS Chakrabarti,S., Mahadani,P., Jain,S.K. and Sevanthi,A.M.

TITLE Authentication of medicinal Dendrobium (Orchidaceae) species

through DNA barcoding from Northeast India

JOURNAL Unpublished

REFERENCE 2 (bases 1 to 658)

AUTHORS Chakrabarti,S., Mahadani,P., Jain,S.K. and Sevanthi,A.M.

TITLE Direct Submission

JOURNAL Submitted (14-SEP-2015) Genetics, **ICAR-National Research Centre for Orchids, Dickling Road, Pakyong, Gangtok, Sikkim 737106, India**

Location/Qualifiers source 1..658

/organism=”*Dendrobium nobile*”

/mol_type=”genomic DNA”

/specimen_voucher=”NRCO/Gen/Den-36/1”

/db_xref=”taxon:94219”

/country=”India”

/lat_lon=”27.13 N 88.45 E”

/collection_date=”12-May-2013”

/collected_by=”S K Jain” misc_RNA <1..>658

/note=”contains internal transcribed spacer 1, 5.8S

ribosomal RNA, internal transcribed spacer 2”ORIGIN

1 taggtgaacc tgcggaagga tcattgtcga gactgaaaca caatgagcga ttttgtgaac

61 ctgtaaaaat aagcggtgcc tgtagtgctg cgataaaatc cactcgagtc atcgcctcat

121 cccctctttg ggttggggac gtgatgaagg atggatgaac cctcaaatcg gcgcagcgta

181 gcgccaaggg aatcttgaaa cacaacccca taaatgggtt ttgtgggatg gggtgctgtc

241 gcaccccata ttgattgaca cgactctcgg caatggatat ctcggctctc gcatcgatga

301 agagcgcagc gaaatgcgat atgtggtgcg aattgcagaa tcccgcgaac catcgagtct

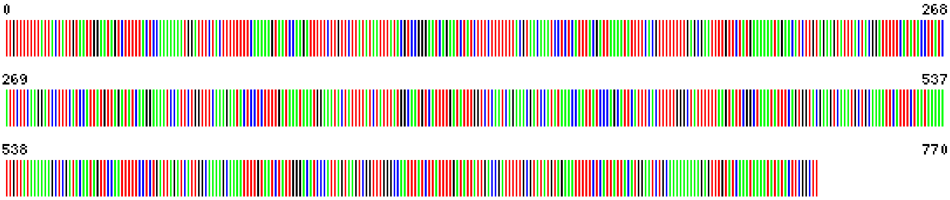
361 ttgaacgcaa gttgcgcctg aggccaatcg gttgagggca cgtccgcctg ggcgtcaagc

421 attttatcgc tccgtgccta gtctcccatc catggatgtg ttgccaaggc tcggatgtgc

481 acggtggctc gtcgtgccca ttggtgcggc gggctgaagg gcgggtcatc ttctcgttgg

541 ttgccaacaa taaggggtgg attaaataag gcctatgcta ttgtgtcaag cgcgcctgag

601 agatggtcat actttttagg tgatcccaat tcatgcgtcg atccatggat ggcgtatc



//

Satyrium nepalense voucher NRCO/Gen/Sat-1/1 maturase K (matK) gene, partial cds; chloroplast

LOCUS KX298579 724 bp DNA linear PLN 12-JUL-2016

DEFINITION *Satyrium nepalense* voucher NRCO/Gen/Sat-1/1 maturase K (matK) gene,partial cds; chloroplast.

ACCESSION KX298579

VERSION KX298579.1 GI:1042762684

KEYWORDS .

SOURCE chloroplast *Satyrium nepalense*

ORGANISM Satyrium nepalense

Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;Spermatophyta; Magnoliophyta; Liliopsida;

Asparagales;

Orchidaceae;Orchidoideae; Orchideae; Orchidinae; Satyrium.

REFERENCE 1 (bases 1 to 724)

AUTHORS Mahadani,P.,Chakrabarti,S.,Chhetri,R.,Singh,D.R.andSevanthi,A.M.

TITLE DNA barcoding of medicinal plant of Orchidaceae from Northeast India

AUTHORS Mahadani,P.,Chakrabarti,S.,Chhetri,R.,Singh,D.R. and Sevanthi,A.M.

TITLE Direct Submission

JOURNAL Submitted (16-MAY-2016) Genetics, ICAR-National Research Centre for Orchids, Dickling Road, Pakyong, Gangtok, Sikkim 737106, India

Location/Qualifiers

source 1..724

/organism="*Satyrium nepalense*"

/organelle="plastid:chloroplast"

/mol_type="genomic DNA"

/isolate="A"

/specimen_voucher="NRCO/Gen/Sat-1/1"

/db_xref="taxon:62865"

/country="India"

/collection_date="20-Apr-2013"

/collected_by="S Chakrabarti"

/identified_by="S Chakrabarti"

gene <1..>724

/gene="matK"

CDS <1..>724

/gene="matK"

/codon_start=3

/transl_table=11

/product="maturase K"

/protein_id="ANQ45571.1"

/db_xref="GI:1042762685"

/translation="RLVFHEYQNLNSLITSKKDIYVFSKRNKRFFWFLHNSYVYECEY
IFLFLRKQSSYLRSTSFEVFLERTHFYVKIEYFILVYCNSFQRIIWFLKDPFIHYVRY
QGKAIMASKGTLILMKKWKFHLVHFWQFYFHFWSQPYRIHIKELPNYSFSFLGYFLSV
LKKTLVVRNQMLENSFFINILTKKLDTIAPVISLIRALSKAQFCTVLGHPISKPIWTD
LSDSDILDRFCRICKNLCRYHSG"

```
  1 tgcgattggt tttccacgaa tatcaaaatt taaatagtct cattacttca aagaaagaca
 61 tttacgtctt ttcaaaaaga aataaaagat ttttttggtt cttacataat tcttatgtat
121 acgaatgtga atatatattc ctgtttcttc gcaaacagtc ttcttattta cgatcaacat
181 cttttgaagt ctttcttgaa cgaacacatt tttatgtaaa aatagaatat tttatattag
241 tttattgtaa ttcttttcag aggattatat ggttcctcaa agatcctttc atacattatg
301 ttcgatatca aggaaaagca attatggctt caaagggaac tctaattctg atgaagaaat
361 ggaaatttca tcttgttcat ttttggcaat tttattttca cttttggtct caaccttata
421 ggatccatat aaaggaatta cccaactatt ccttctcttt tttggggtat tttttaagtg
481 tactaaaaaa gactttggta gtaagaaatc aaatgctgga gaattctttt ttcataaata
541 ttctgactaa gaaattagat accatagccc cagttatttc tcttattaga gcattgtcaa
601 aagctcaatt ttgtactgta ttgggccatc ccattagtaa accgatatgg actgatttat
661 cggattctga tattcttgat cgattttgtc ggatatgtaa aaatctttgt cgttatcaca
721 gtgg
//
```

Rhynchostylis retusa voucher NRCO/Gen/Rhy-1/1 maturase K (matK) gene, partial cds; chloroplast

LOCUS KX298578 545 bp DNA linear PLN 12-JUL-2016

DEFINITION *Rhynchostylis retusa* voucher NRCO/Gen/Rhy-1/1 maturase K (matK)
gene, partial cds; chloroplast.

ACCESSION KX298578

VERSION KX298578.1 GI:1042762682

SOURCE chloroplast *Rhynchostylis retusa*

ORGANISM Rhynchostylis retusa
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
Spermatophyta; Magnoliophyta; Liliopsida; Asparagales; Orchidaceae;
Epidendroideae; Vandeae; Aeridinae; Rhynchostylis.

```
REFERENCE   1  (bases 1 to 545)
AUTHORS Mahadani,P.,Chakrabarti,S.,Chhetri,R.,Singh,D.R.and  Sevanthi, A.M.
TITLE  DNA barcoding of medicinal plant of Orchidaceae from Northeast
       India
 JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 545)
 AUTHORS   Mahadani,P., Chakrabarti,S., Chhetri,R., Singh,D.R. and
       Sevanthi,A.M.
 TITLE     Direct Submission
 JOURNAL   Submitted (16-MAY-2016) Genetics, ICAR-National Research Centre
for
       Orchids, Dickling Road, Pakyong, Gangtok, Sikkim 737106, India

 Location/Qualifiers    source        1..545
             /organism="Rhynchostylis retusa"
             /organelle="plastid:chloroplast"
             /mol_type="genomic DNA"
             /isolate="A"
             /specimen_voucher="NRCO/Gen/Rhy-1/1"
             /db_xref="taxon:257352"
             /country="India"
             /lat_lon="26.86 N 88.73 E"
             /collection_date="03-Aug-2013"
             /collected_by="S Chakrabarti"
             /identified_by="S Chakrabarti"
gene         <1..>545
             /gene="matK"
CDS          <1..>545
             /gene="matK"
             /codon_start=1
             /transl_table=11
             /product="maturase K"
             /protein_id="ANQ45570.1"
             /db_xref="GI:1042762683"
              /translation="YIFLFLRKQSSYLRSISSGVFLERTHFYGKIEYLRVVSCNS
```

FQR
ILWFLKDIFIHYVRYQGKAILASKGTLILMNKWKFHFVNFWQSYFHFW
FQPYRIHIKQ
LPNYSFSFLGYFSSVLKNPLVVRNQMLENSFLINTLTKKLDTIAPVIFLIG
SLSKAQF
CTVLGHPISKPIWTNLSDSDIL"ORIGIN

1 tatatattcc tttttcttcg taaacagtct tcttatttac gatcaatatc ttctggagtc
61 tttcttgagc gaacacattt ttatggaaaa atagaatatc ttagagtcgt gtcttgtaat
121 tcttttcaga ggatcttatg gttcctcaaa gatattttca tacattatgt tcgatatcaa
181 ggaaaagcga ttctggcttc aaaaggaact cttattctga tgaataaatg gaaatttcat
241 tttgtgaatt tttggcaatc ttattttcac ttttggtttc aaccttatag gatccatata
301 aagcaattac ccaattattc cttctctttt ctgggatatt tttcaagtgt actaaaaaac
361 cctttggtag taagaaatca aatgctagag aattcatttc taataaatac tctgactaag
421 aaattagata ccatagctcc cgttattttt cttattggat cattgtcgaa agctcaattt
481 tgtactgtat tgggtcatcc tattagtaaa ccgatctgga ccaatttatc ggattctgat
541 attct
//

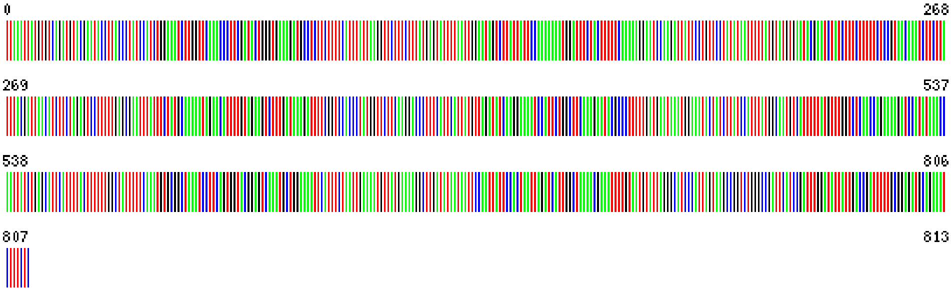
Arundina graminifolia voucher NRCO/Gen/Aru-1/1 maturase K (matK) gene, partial cds; chloroplast

LOCUS KX298566 708 bp DNA linear PLN 12-JUL-2016
DEFINITION *Arundina graminifolia* voucher NRCO/Gen/Aru-1/1 maturase K (matK)
gene, partial cds; chloroplast.
ACCESSION KX298566
VERSION KX298566.1 GI:1042762658
SOURCE chloroplast *Arundina graminifolia*
ORGANISM *Arundina graminifolia*
Eukaryota; Viridiplantae; Streptophyta; Embryophyta;

Tracheophyta;Spermatophyta; Magnoliophyta; Liliopsida; Asparagales;
Orchidaceae;Epidendroideae; Arethuseae; Arethusinae; Arundina.
REFERENCE 1 (bases 1 to 708)
AUTHORS Mahadani,P., Chakrabarti,S., Chhetri,R., Singh,D.R. and
Sevanthi,A.M.
TITLE DNA barcoding of medicinal plant of Orchidaceae from Northeast India
AUTHORS Mahadani,P., Chakrabarti,S., Chhetri,R., Singh,D.R. and
Sevanthi,A.M.
TITLE Direct Submission
JOURNAL Submitted (16-MAY-2016) Genetics, ICAR-National Research Centre
for
Orchids, Dickling Road, Pakyong, Gangtok, Sikkim 737106,
IndiaCOMMENT Location/Qualifiers source 1..708
/organism="*Arundina graminifolia*"
/organelle="plastid:chloroplast"
/mol_type="genomic DNA"
/isolate="A"
/specimen_voucher="NRCO/Gen/Aru-1/1"
/db_xref="taxon:78703"
/country="India"
/collected_by="S Chakrabarti"
/identified_by="S Chakrabarti"
gene <1..>708
/gene="matK"
CDS <1..>708
/gene="matK"
/codon_start=1
/transl_table=11
/product="maturase K"
/protein_id="ANQ45558.1"
/db_xref="GI:1042762659"
/translation="EYHNLNSLITSNKSIYVFSKRTKRFFWFLHNSYVYECEYI
FLFL
RKQSSYLRSISSGVFLERTHFYGKIEYLIVVCCNSFQRILWFLKDTFIHYV
RYQGKTI

LVSKGTLILIKKWKFHLVNFWQSYFHFWFQPYRIHIKQLPNYSFSFLGY
FSSVLKNNL
VIRNQMLENSFLINTLTKKLDTIAPVTSIIGSLSKAQFCTVLGHPISKPIW
TDLSDSD
IFDRFCRICRNLCRYHSG”ORIGIN

1 gaatatcata atttgaatag tctcattact
tcaaataaat ccatttacgt cttttcaaaa
61 agaaccaaaa gattcttttg gttcctacat
aattcttatg tatatgaatg cgaatatata
121 ttcctgtttc ttcgtaaaca gtcttcttat
ttacgatcaa tatcttctgg agtctttctt
181 gagcgaacac atttctatgg aaaaatagaa
tatcttatag tcgtgtgttg taattctttt
241 cagaggatcc tatggttcct caaagatact
ttcatacatt atgttcgata tcaaggaaaa
301 acaattctgg tttcaaaagg aactcttatt
ctgattaaga aatggaaatt tcatcttgtg
361 aatttttggc aatcttattt tcacttttgg tttcaacctt ataggattca tataaagcaa
421 ttacccaact attccttctc ttttctgggg tatttttcaa gtgtactaaa aaataatttg
481 gtaataagaa atcaaatgct agagaattca tttctaataa atactctgac taagaaatta
541 gataccatag ccccagttac ttctattatt ggatcattgt cgaaagctca attttgtact
601 gtattgggtc atcctattag taaaccgatc tggaccgatt tatcggattc tgatattttt
661 gatcgatttt gtcggatatg tagaaatctt tgtcgttatc acagcgga



//

References

China Plant BOL Group, Li, D., Gao, L., Li, H., Wang, H., Ge, X., Liu, J., Chen, Z., (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proc Natl Acad Sci USA* 108, 19641–19646.

Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi, S., Cheng, T., Guo, J., Zhou, S., (2015). ycf1, the most promising plastid DNA barcode of land plants. *Scientific Reports* 5, 8348.

Lahaye, R., van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T.G., Savolainen, V., (2008). DNA barcoding the floras of biodiversity hotspots. Proc Natl Acad Sci U S A 105, 2923-2928.

Lucksom, S.Z., (2007). The Orchids of Sikkim and North East Himalaya. Lucksom, S.Z, Gangtok.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25 (17):3389-3402.

Asahina H, Shinozaki J, Masuda K, Morimitsu Y, Satake M (2010). Identification of medicinal Dendrobium speciesby phylogenetic analyses using *matK* and *rbcL* sequences. *Journal of Natural Medicines* 64 (2):133-138.

Barthet MM, Hilu KW (2008) Evaluating evolutionary constraint on the rapidly evolving gene *matK* using protein composition. *Journal of Molecular Evolution* 66 (2):85-97.

Chase MW, Cowan RS, Hollingsworth PM, Van den Berg C, Madrinan S, Petersen G, Seberg O, Rgsensen T,Cameron KM, Carine M, Pedersen N, Hedderson TAJ, Conrad F, Salazar GA, Richardson JE, Hollingsworth ML, Barraclough TG, Kelly L, Wilkinson M (2007). A proposal for a standardised protocol to barcode all land plants. *Taxon* 56:295-299.

Ghosh SK, Bhattacharjee MJ, Mahadani P, Laskar BA (2012). Fundamentals of DNA barcoding. In: Ghosh SK (ed.) A text book of DNA barcoding. Book Space, Kolkata.

Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98.

Hao DC, Chen SL, Xiao PG (2010). Sequence characteristics and divergent evolution of the chloroplast *psbA-trnH* noncoding region in gymnosperms. *Journal of Applied Genetics* 51 (3):259-273.

Hebert PD, Cywinska A, Ball SL, deWaard JR (2003a). Biological identifications through DNA barcodes. Proceedings Biological sciences - The Royal Society 270 (1512):313-321.

Hebert PD, Ratnasingham S, deWaard JR (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings Biological sciences / The Royal Society 270 Suppl 1:S96-99.

Hebert PD, Stoeckle MY, Zemlak TS, Francis CM (2004a). Identification of Birds through DNA Barcodes. *Plos Biology* 2 (10): e312.

Hilu KW, H. L (1997). The Matk gene: sequence variation and application in plant systematics. *American Journal of Botany* 84:830-839.

Hogg ID, Hebert PDN (2004). Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. *Canadian Journal of Zoology* 82:749–754.

Hollingsworth PM, Graham SW, Little DP (2011). Choosing and using a plant DNA barcode. *PloS one* 6 (5):e19254.

Hollingsworth MP, Forrest LL, Spouge LJ, Hajibabaei M, Ratnasingham S, van der Bank M, Chase WM, Cowan SR, et al(2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America* 106 (31):12794-12797.

Kress WJ, Erickson DL (2012). DNA barcodes: methods and protocols. *Methods in Molecular Biology* 858:3-8.

Kress WJ, Erickson DL, Jones FA, Swenson NG, Perez R, Sanjur O, Bermingham E (2009). Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences of the United States of America* 106 (44):18621-18626.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* 102 (23):8369-8374.

Mahadani P, Devi KM, Das MM, Chakraborty M, Rahman F, Hansa J, Ghosh SK (2012). Bioinformatics in DNA barcoding. In: Ghosh SK (ed.) A text book on DNA barcoding. Book Space Kolkata, pp 105-136.

Mahadani P, Ghosh SK (2013a). DNA Barcoding: A tool for species identification from herbal juices. *DNA Barcodes* 1:35-38.

Mahadani P, Sharma GD, Ghosh SK (2013b). Identification of Ethnomedicinal plants (Rauvolfioideae: Apocynaceae) through DNA Barocding from Northeast India. Pharmacognosy Magazine (In Press).

Ratnasingham S, Hebert PD (2007). bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes* 7 (3):355-364.

Saitou N, Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4 (4):406-425.

Sambrook JF, Russell DW (2001). Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring.

Singh HK, Parveen I, Raghuvanshi S, Babbar SB (2012). The loci recommended as universal barcodes for plants on the basis of floristic studies may not work with congeneric species as exemplified by DNA barcoding of *Dendrobium* species. *BMC Research Notes* 5:42.

Stoeckle MY, Gamble CC, Kirpekar R, Young G, Ahmed S, Little DP (2011). Commercial teas highlight plant DNA barcode identification successes and obstacles. Scientific Reports 1:42.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface:flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* 25 (24):4876-4882.

Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22 (22):4673-4680.

भाकृअनुप - राष्ट्रीय आर्किड्स अनुसंधान केंद्र
पक्योंग-७३७ १०६, सिक्किम, भारत

ICAR-National Research Centre for Orchids
Pakyong-737 106, Sikkim, India