Mitochondrial gene organisation

Mitochondrial markers

Even though mitochondrial DNA (mtDNA) represents a tiny fraction of organism all genome size, they are often chosen for evolutionary studies. mtDNA is relatively easy to amplify because it appears in multiple copies in cell, the gene content is strongly conserved across animals with very few duplications, intronless and very short intergenic regions (Simon et al., 1994).

Mitochondrial tRNAgenes

Nucleotide sequences and the rearrangements of mitochondrial DNA gene order may be used as a tool for describing the pattern of evolutionary variations in arthropod taxa. The positions of tRNA mitochondrial genes have been used to study the phylogenetic relationships of the major arthropod groups (Roehrdanz *e t a l* ., 2002; Boore *e t a l* ., 1995). Mitochondrial tRNA genes are less constrained than their highly conserved nuclear counterparts, evolving at a higher rate. However, mitochondrial tRNA genes evolve more slowly than mitochondrial protein coding genes, suggesting that they are under structural and functional constraints (Simon *e t a l* ., 1994). mtDNA tRNA genes can fold into four-armed (clover-leaf) secondary structures, displaying a size variation in both the sequence of the loops of the dihydrouridine (DHU) and T-loop (TψC) (Wolstenholme, 1992).

Rearrangements of tRNA genes have been reported in insect orders which are useful in gene evolution studies. Dowton *e t al* . (2009) evaluated the hymenopteran mitochondrial genome to investigate the mode of mitochondrial gene order evolution and its accelerated rate of gene rearrangement. Nine mitochondrial genome arrangements from the

Hymenoptera were assessed; six of these mitochondrial genomes from the Aculeata (Apoidea, Chrysidoide, and Vespoide), together with two additional non-apocritans (*C e p hu s* and *O r u s s u s*) and three additional apocritans (*S c h l et t e r e r i u s* , *E n i c o s p il u s* , and *V e n t u r i a*). Rearrangements of the tRNA genes were accounted for with tRNAMet arrangement being the most common followed by tRNAIle changing position seven times, and tRNAGln changing position six times. Genes located close to the noncoding region (the IQM cluster) were found to be the most versatile and gene rearrangements within the mitochondrial genome were in all likelihood selectively neutral (Dowton *e t a l* ., 2009).

Protein coding genes

The protein coding genes are the most frequently sequenced mitochondrial genes for evolutionary studies and phylogenetic analysis. Protein coding genes commonly analyzed include; COI, COII, COIII, atp6, atp8, nad1, nad2, nad3, nad4, nad5, nad4L, nad6 and cob. Of these COI has been widely sequenced, yet the specific region chosen has varied from study to study (Caterino *e t a l* ., 2000, Schroeder *e t a l* ., 2003). COI and COII have been sequenced over a wide variety of taxa with homologous sequences available for nearly all orders such as Lepidoptera, Orthoptera, Hemiptera, Hymenoptera, Diptera, Coleoptera, Neuroptera and Siphonaptera (Caterino *e t a l* ., 2000).

Clary and Wolstenholme (1985) identified first subunit of the mtDNA CO gene of *Drosophila yakuba* and has been recognized as an area of interest for ‘ DNA barcoding’ (Nelson *et al* ., 2007). This region has a rate of molecular evolution that is about three times greater than that of 12S or 16S

rDNA, its third-position nucleotides showing a high frequency of base substitutions. The utilization of DNA sequences as genetic ‘ barcodes’ might conceivably be utilized as a bio identification system for all animals and has proven to be a useful identification tool for vertebrates such as birds (Hebert *et al* ., 2004), fish (Ward *et al* ., 2005), and invertebrates as in neotropical skipper butterfly *Astraptes fulgerator* (Hebert *et al* ., 2004), tropical parasitoid Dipterans: Tachinidae (Smith *et al* ., 2007) and in scuttle flies, Dipteria : Phoridae (Boehme *et al* ., 2010). Wallman *et al* . (2005) utilized a blend of mitochondrial genes to improve the genetic identification of carrion-breeding blowflies. The gene sequences of COI and COII were combined with gene sequences of ND4 and ND4L in an effort to clarify evolutionary relationships and correctly identify the species.

Evidences supported the use of cytochrome oxidase subunits in molecular identification; however, some research has reported difficulties caused by a maternally transmitted bacterial infection of insects. The endosymbiotic bacterium *Wolbachia* is maternally transmitted and causes for various reproductive alterations in their hosts. When a population gets contaminated with *Wolbachia* , the bacteria will spread quickly and the mtDNA related to the initial contamination will spread through the population by indirect selection. *Wolbachia* influences mtDNA variation at the intraspecific or interspecific level. At the intraspecific level, *Wolbachi* a reduces mtDNA polymorphism limiting the suitability of *Wolbachia* -infected species data in population studies, but may pose no issue for barcoding. In closely related species *Wolbachia* can also affect mtDNA variations between species and results introgression (Whitworth *et al* ., 2007).

Mitochondrial ribosomal RNA genes (12S and 16S)

Ribosomes are large complex of both RNA and protein, the RNA component catalyses protein synthesis. The ribosomal RNA (rRNA) is present in all existing species and therefore use to establish the evolutionary relationships among them. RNA molecules fold into characterized structures that are basic for their biological functions. Throughout RNA evolution, the structure is considerably more conserved than the sequence. Structure and function of the RNA molecule controls the sequence variations between the species (Smit *et al* ., 2007).

Clary and Wolstenholme (1985) in their studies on the ribosomal RNA genes of *Drosophila yakuba* found that the 5′ regions of the small rRNA gene and of the large rRNA gene are extremely low in G+C. The ability of *D. yakuba* mt-rRNA genes to fold into secondary structures with low G+C content to that proposed for rRNAs results in sequence homologies between these regions and the corresponding regions

In the phylogenetic analysis of dipteran and forensically important flies 12S and 16S rDNA genes have been used. Spicer and Bell (2002) by using 12S and 16S rDNA separated four subclades of *Drosophila virilise* (Diptera: Drosophilidae).

16S rDNA is a molecular marker existing in mitochondria and has a high conserved region. Li *et al* . (2010) analysed the morphological characteristics and 16S rRNA sequences for the identification of Muscidae species. A 289 bp fragment of the mitochondrial 16S rRNA gene was successfully sequenced for specimens from *M. domestica* , O *. spinigera* , *O. chalcogaster* , *M.*

Hortorum and *F. canicularis* collected from 13 different locations. The interspecific variation between *O. spinigera* and *O. chalcogaster* was high and it allows discrimination between the two species and concludes that the mitochondrial 16S rRNA gene can be used to distinguish the species (Li *et al* ., 2010).

Mitochondrial DNA heteroplasmy

The state in which when there is more than one type of mtDNA genome present, is known as heteroplasmy (Tsang and Lemire, 2002). Ivanov *et al* . (1996) reported single nucleotide heteroplasmy (or point heteroplasmy) in human mtDNA from the remains of Russian Tsar Nicholas II. A+ T rich tandem duplicated sequence in control region are found in various variety of animal groups. Heteroplasmy, due to the distinction within the range of tandemly repeated sequences, has been described for many hexapod taxa, including representatives from the Orthoptera, Diptera, Coleoptera, and Hymenoptera (Zhang and Hewitt, 1997; Nardi *et al* ., 2001).

The mtDNA control region of insects vary largely in size due to variation in the size and duplicate number of repeat units and accounts for most size variation of the mitochondrial genome. Variation in the copy number of tandem repeats in the mtDNA has also been reported within individuals, resulting in length heteroplasmy of certain insect populations (Zhang and Hewitt, 1997).

As the sequencing of the mtDNA genome is expanding, more sequence heteroplasmy has been accounted for. The nematode *Caenorhabditis elegans* exhibited the deletion of 11 genes from the mtDNA genome in

particular strains. Deletion reported includes 4 protein-coding and 7 tRNA genes (Tsang and Lemire, 2002). Nardi *et al* . (2001) reported length heteroplasmy in *Tetrodontophora bielanensis* as an after effect of tandem repetition within A+T rich mtDNA control region. In the grasshoppers genus *Chitaura* the mitochondrial DNA sequencing revealed an unexpectedly high incidence of heteroplasmy at the level of base substitutions in the cytochrome oxidase I (COI) region (Walton *et al* ., 1997).

The mtDNA control region of insects

Mitochondrial genome (mtDNA) of insects contains a single major non-coding region, which is believed to be responsible for the regulation of transcription and control of DNA replication. This major non-coding region has therefore been called the ‘ control region’. The region is rich in adenine and thymidine nucleotides and is also known as the A+T rich region. This region has been shown to contain upto 96% adenine and thymidine (in the honey bee and the fruitfly) (Zhang and Hewitt, 1997). A+T nucleotide substitution pattern within invertebrate mtDNA is due to long-term evolution, the various enzymes responsible for transcription and replication of DNA functioning optimally with a high A+T content (Clary and Wolstenholme, 1985). In phylogenetic studies difficulties arises when using the control region, as it is more A+T rich than the rest of the genome. The control region is subject to extensive length variation (Zhang and Hewitt, 1997). In crickets *Gryllus firmus* extreme length variation due to the presence or absence of tandem repeats and homopolymer runs, has been demonstrated (Rand and Harrison, 1989). Hale and Singh (1986) have also reported the extreme length variation in the control region of *Drosophila melanogaster.* This suggests that care should be

taken with alignment and significant testing is needed to distinguish between random matches and homology (Simon *et al* ., 1994).

Mitochontrial genome control region location of invertebrates shows great diversity and as a result primers designed to amplify these region and surrounding genes will work well for one taxon, but may fail for others. Comparision of the control region sequences of insects *Schistocerca gregaria* (desert locust), *Chorthippus parallelus* (grasshopper) and the fruit flies *Drosophila virilise* and *Drosophila tristis* indicated that sequence conservation patterns differed between *Drosophila* species and other insects, forming two distinct groups, Group 1 and Group 2 (Zhang and Hewitt, 1997).

The fruit flies (Zhang and Hewitt, 1997), calliphorid flies (Junquerra *et al* ., 2004; Duarte *et al* ., 2008) and muscid flies (Oliveira *et al* ., 2007) comes in the taxa with Group 1 control regions and these taxa have two distinct domains within the control region. The first domain lies adjacent to the tRNA ^{Ile} gene, which is highly conserved in *Drosophila* , and contains a putative replication origin and conserved sequence blocks (CSBs). The second domain includes the rest of the region and is variable in nucleotide sequence and length due to large repeated fragments, tandem repeats and indels (Duarte *et al* ., 2008).

In Group 2 insects as delineated by Zhang and Hewitt (1997) have a control region that cannot be separated into distinct conserved or variable domains. Insects with Group 2 taxa control regions include grasshoppers, butterflies and mosquitoes. Short conserved sequence blocks are scattered through the entire region, while tandem repetition has also been observed. The

distribution pattern of conserved structural elements demonstrates that the conserved domain in Group 2 control regions indicates their homologous relationship (Zhang and Hewitt, 1997).

DNA barcoding

DNA barcoding involves sequencing a short section of DNA from a homogenous region of the genome for species identification. Various authors have proposed distinctive purposes for DNA barcoding, but the most prevalent concept of barcoding is the production of a library of sequences that could be utilized to identify previously described taxa (Meusnier *et al* ., 2008; Rubinoff, 2006a).

The mitochondrial genome is known for having relatively well conserved regions that are excellent for primer creation. Mitochondrial DNA (mtDNA), unlike nuclear DNA, has no introns, rarely experiences recombination, and is maternally inherited in a haploid manner (Hebert *et al* ., 2003; Rubinoff, 2006b; Ballard *et al* ., 2005). Rapid rate of evolution of mitochondrial DNA can lead to homoplasy since frequent base pair changes might result in convergently similar sequences in two unrelated taxa (Rubinoff, 2006a).

Phylogenetics uses mtDNA when examining species level connections and as of late saperated taxa (Hebert *et al.,* 2003). However the brisk rate of evolution makes mtDNA so valuable for recent divergences, becomes problematic when divergences dating to the Mesozoic or earlier are examined (Mitchell, 2008). This rapid rate of evolution of mitochondrial DNA can prompt homoplasy since frequent base pair progressions might result in convergently similar sequences in two unrelated taxa (Rubinoff, 2006a).