STUDIES OF THE MOLECULAR EVOLUTION OF COI

STUDIES OF THE MOLECULAR EVOLUTION OF COI

by
MELANIE LOU, B.Sc. (Hons.)

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science

McMaster University

MASTER OF SCIENCE (2007)
(Biology)

McMaster University
Hamilton, Ontario

TITLE: Studies of the molecular evolution of COI

AUTHOR: Melanie Lou, B.Sc. Hons. (York University)

SUPERVISOR: Dr. G. Brian Golding

NUMBER OF PAGES: [x], 62

PREFACE

Each chapter of this thesis has been written as a separate manuscript. Programming, data collection, analysis and manuscript preparation for each chapter was primarily an individual effort, with contributions in data preparation and editing from G. Brian Golding. Chapter two was recently accepted for publication and is now in press.

ABSTRACT

There has been an increasing value in the ability to describe the world's diversity for the purpose of enhancing research and conservatory efforts. Characterizing the level of heterogeneity of particular molecular markers and verifying its suitability as an identifier of new specimens provides a way of quantifying biodiversity. One such molecular marker is the mitochondrial cytochrome c oxidase subunit I (COI). An analysis of the evolutionary rates among and within taxonomic groupings of 13,641 insect COI sequences revealed that the evolutionary rate of some species increased or decreased, sometimes by an order of a magnitude. Furthermore, the increased evolutionary rates of two species, from the Lepidopteran and Orthopteran orders, may be explained by the influence of positive selection but further analyses would be required to rule out other explanations. Overall, we deem that the rate of substitution generates enough change for COI to work sufficiently as a barcode marker in insects. As COI is not suitable for specimen identification in plants, it would be useful to be able to quickly determine if there is enough variation in COI or other molecular markers for specimen identification. In response, a visualization tool, `Fingerprint`, was developed to graphically depict 11 different types of sequence diversity. An application of the tool to Lepidopteran COI data verified the genetic diversity in insect COI and the tool's ability to sensitively detect different types of heterogeneity.

ACKNOWLEDGEMENTS

No journey is a self-sufficient one. Though, for those who know me, asking for help is not exactly my strongest trait. So to those I have asked for help from, I do so because I hold you in high opinion, and I trust that you would not lead me astray or let me down.

I am grateful to many people for help, both direct and indirect, in the journey to the production of this thesis.

To my parents, Young-Chang and Laura Lou, and my sister, Angela Lou, for their continual support throughout all points of my life.

I am very grateful to Dr. G. 'Brian' Golding for taking me on as a student in his lab and introducing me to the wonderful, often tedious, world of research! This thesis could never have been completed without his continual guidance - especially in trying to keep me on schedule! Thank you very much for putting up with me and my slow ways. As a mentor, I know he is always there to listen and give advice and for that I am thankful. I believe him when he says he worries about me as I have become one of his academic children. And, like any good supervisor, he reminds us, occasionally, to have fun!

I'd like to raise my hat to (if I ever wore one) the members of the Golding lab, both past and present, for being there to bug, have light-hearted discussions and laugh with after being at the computer for too long.

Weilong has been, and continues to be, a great workmate ever since I started my studies at McMaster. I would never have completed my Bio 720 project if it was not for his expertise in everything related to generating and displaying phylogenetic trees. I marvel at the fact that he always has a simple, yet effective, analogy up his sleeve when explaining concepts to me.

Life at McMaster is greatly enhanced by the support, guidance, and, most importantly, friendship of the coffee-break girls to whom I owe thanks for wonderful memories and for helping me keep my sanity: Maria 'Mia' Abou Chakra, Abha Ahuja, Dr. Melanie 'Senior' Huntley, Danya Konrad, Andrea Morash, Laura Smallbone, and Allyson Maclean.

I should like to acknowledge Mike 'Bacon' Bui for his companionship prior and during the early part of my time at McMaster. Without his technical expertise, this grasshopper would never have learned the basic properties of networking.

To Lance 'BB' Ferris for his continual support predominantly through numerous emails (and occasionally by phone). If not for his supportive words and listening ear, I may have been lost!

Thanks to my committee members for the generosity of their time, Dr. Jonathon Stone, and Dr. Rama Singh.

Thanks to Pat Hayward for handling all the tedious administrative details.

Lastly, I must thank the Barcode of Life Project for providing me with financial support these past two years.

Contents

List of Figures

List of Tables

Part I

INTRODUCTION

Sequence data has been useful in divulging divergence times, generating phylogenies, and identifying the evolutionary processes affecting it. Mitochondrial genes are commonly used as sequence markers due to their fast rate of evolution and lack of recombination.

Of particular interest is the mitochondrial cytochrome c oxidase subunit I (COI) gene. It is a core subunit of the cytochrome c oxidase (COX), which is a complex metalloprotein consisting of 13 subunits (Capaldi, 1990). It is the last enzyme in the electron transfer chain (ETC), responsible for transferring electrons to molecular oxygen and generating the electrochemical gradient that produces ATP (Ludwig *et al.*, 2001). As a core component of COX, COI is indispensible as its residues are involved in all activities intrinsic to COX: electron transfer, enzyme function, proton transfer, channels for O_2 transfer to reduction site, and for removing H_2O (Tsukihara *et al.*, 1996).

Consequently, it is expected that the sequence data coding for this gene has remained relatively unchanged over time and across species. However, studies have shown that COI has been successful in resolving sequence diversity across a broad range of taxa: fungi (Seifert *et al.*, 2007), gastropods (Remigio and Hebert, 2003), amphipod crustaceans (Witt, Threloff and Hebert, 2006), bats (Clare *et al.*, 2007), birds (Hebert *et al.*, 2004), fishes (Ward *et al.*, 2005), and Lepidoptera (Hebert *et al.*, 2004; Hajibabaei *et al.*, 2006). In addition, this gene has become the standard sequence in the Barcode of Life initiative which aims to identify unknown specimens (Hebert *et al.*, 2003). Thus, a complete understanding of how this molecule changes is therefore of critical importance.

The COI gene is one of the most densely sequenced genes in the world. This is especially true with regard to insects. Currently, there are a total of over 13, 600 sequences known for this gene from the insects. This provides one of the richest data sets in the world to study the molecular evolution of a gene. Though commonly viewed as pests, insects are ecologically and economically important. Consequently, more resources have been directed towards research and conservatory efforts concerning this invertebrate group (Laffin, Langor and Sperling, 2004; Alvarez *et al.*, 2005; Grimaldi and Engel, 2005; Kourti, 2006). However, success of these projects depends on being able to quantify the diversity of undescribed and described insects. The first chapter is an analysis of 13,641 insect COI sequences from 16 orders. The rates of evolution among and within each taxonomic classification are examined to determine if COI generates enough polymorphism to warrant species identification in insects.

Unlike animals, COI is a poor marker for species identification in plants (Kress *et al.*, 2005). Thus verifying that COI supports enough genetic heterogeneity to operate as a barcode marker is of importance and can be achieved graphically. The second chapter describes the creation, functionality and applicablity of an online web tool, `Fingerprint`, capable of graphically depicting different types of variation in all kinds of sequence data. Futhermore, as other molecular markers have been used as standards for insect phyloge-

netics: 16S, 18S, and elongation-1 α (Caterino, Cho and Sperling, 2000), `Fingerprint` may be used to identify other sequence regions capable of specimen identification by illustrating that the candidate sequence(s) in question generates heterogeneity.

Chapter 1

Molecular evolution of cytochrome c oxidase subunit I in Class Insecta

1.1 Abstract

A fragment of the mitochondrial cytochrome c oxidase subunit I (COI) gene sequence has been chosen as the key component of the Barcode of Life initiative for animal species. The aim of the Barcode of Life initiative is to create a library of sequences and taxa information to permit the identification of every species. The number of empirical studies to characterize the molecular evolution of COI in insects is limited, despite the large amount of sequence data that is available and the plentitude of extant insects from which to generate data; thus, the full value of COI as a discriminator in insects remains unknown. It is the purpose of this chapter to conduct an analysis of the level of variation among and within different levels of taxonomic hierarchy in COI in the Class Insecta. We analyzed 13,361 insect sequences from 16 orders and discovered that some species showed an increase or decrease, in some cases by an order of magnitude, in its molecular rate of change. To detect the occurence of positive selection, two likelihood ratio tests (LRTs) were applied. Each test statistically checks to see if the more complex model, which permits a portion of the sites to undergo positive selection, fits the dataset significantly better than the simpler model, which does not permit positively selected sites. Two species, *Bombyx mori* (Insecta: Lepidoptera) and *Melanoplus dawsoni* (Insecta: Orthoptera), generated sigificant LRT results, which suggest that they are under the influence of positive selection. *Bombyx mori* is significant for both LRTs at less than the 0.1% level, more specifically, at relatively very low p-values of 10^{-10} for both tests. Similarly, *Melanoplus dawsoni* generated the same LRT results, with relatively very low p-values of 10^{-12} and 10^{-13} for each test respectively. According to published literature, both species have the potential for adaptive

changes but further analyses are required to rule out other plausible explanations. Overall, the rate of substitution, particularly at the species level, generates enough change for COI to work sufficiently as a barcode marker in insects.

1.2 Introduction

Insects make up a significant fraction of the total biodiversity on this planet. Estimates for the total number of extant species, described and undescribed, vary between six to ten million (Chapman, 2005). As is evident by their numbers, insects are as important, if not more important, as are other animals, as they play an important role in the functioning of the terrestrial ecosystem.

With regard to their place in human society, most insects are regarded as pests, as some are parasitic, can transmit diseases, cause damage to structures, or cause damage to agricultral goods (Alvarez *et al.*, 2005; Castro, Austin and Dowton, 2002; Laffin, Langor and Sperling, 2004; Yan, Chadee and Severson, 1998).

Though commonly viewed as pests, many insects are beneficial to the environment and to humans. Some insects are involved in the pollination of flowering plants. Mutalistic wasps pollinate figs; in return, the seeds that are produced from the pollination provide nourishment for developing offspring hatched from eggs laid within the flower (Machado *et al.*, 2001; Weiblen, 2001). Without insects, we would not be able to enjoy the honey and wax produced by bees; nor the silk produced by silkworms, which not only has played a significant role in developing trading and communication in human history but is being engineered to produce useful protein products other than silk (Grimaldi and Engel, 2005). Medically, fly larvae (maggots) have been used to hamper the development of infection by consuming the dead flesh surrounding an injury. Scavengers, such as beetles, help recycle dead organic material. And, as most insects are insectivores, they are employed as useful biological control agents of insects humans deem to be pests.

Thus, understanding the number and variety of living insect species is environmentally and economically beneficial to our society. Additionally, much of what we know about insects does not include information about their habitats and ecology. Without this vital information, conservation efforts are limited. One way to differentiate insect species is to consider morphological differences, including general size and shape, specific heribvore morphology in relation to food (Bernays, Jarzzembowski and Malcolm, 1991), and differentiated genitalia as a result of sexual selection (Chapco, 2002). However, there are caveats to using only morphological taxonomy for insect identification. There are few scientists working on insect morphological taxonomy and the sheer number of extant insects makes it difficult, if not impossible, for these taxonomists to document all the unique character-

istics that defines each species. There is also evidence to suggest the existance of cryptic species (Hebert *et al.*, 2004), thus making the task even more difficult.

For these reasons, investigators have turned to sequence data, as published literature have shown its usefulness in molecular systematics. With regard to insects, several genes have been used as standards for insect phylogenetics: cytochrome c oxidase I (COI), 16S, 18S, and elongation factor-1 α (Caterino, Cho and Sperling, 2000).

Recently, much attention has been paid to the use of a 648-bp region near the 5′ end of mitochondrial COI, as it is the standard gene of a DNA-based system for specimen identification (Hebert *et al.*, 2003). Earlier studies have shown it has been successful in resolving sequence diversity in fungi (Seifert *et al.*, 2007), gastropods, (Remigio and Hebert, 2003), amphipod crustaceans (Witt, Threloff and Hebert, 2006), bats (Clare *et al.*, 2007), birds (Hebert *et al.*, 2004), fishes (Ward *et al.*, 2005), and Lepidoptera (Hebert *et al.*, 2004; Hajibabaei *et al.*, 2006).

Given its success in phylogenetics, mostly attributable to its rate of evolution that permits it to determine evolutionary histories at the family, genus, and species levels (Caterino, Cho and Sperling, 2000), it comes as no surprise that various studies in several insect orders, including including Collembola (springtails) (Hogg and Hebert, 2004), Ephemeroptera (mayflies) (Ball and Hebert, 2005), Hymenoptera (Madagascar ants) (Smith, Fisher and Hebert, 2005), and Diptera (parasitoid flies (Smith *et al.*, 2006, 2007), mosquitoes (Cywinska, Hunter and Hebert, 2006) and cryptic Chironomus larvae (Pfenninger *et al.*, 2007)) have shown the effectiveness of barcoding in the identification of insect specimens. Nevertheless, these studies are few and they only cover a small fraction of the species richness that exists.

There has been some debate as to the use of barcoding as a poor solution to the old-age species problem. Being able to quantify the world's biodiversity relies on the definition of the smallest unit of classification: the species. With respect to barcoding, currently, an arbitrary similarity criterion has been set at 3% divergence for insects and 2% for birds and mammals (Hebert, Ratnasingham and deWaard, 2003). Essentially, a divergence of less than 3% between an unknown insect specimen and a characterized insect in the database indicates that the unknown sample has found its closest species match and thus has been identified; a divergence above 3% would constitute a new species. However, it has been suggested that this cut-off value is unreliable because it defines an arbitrary predefined level of divergence as a species boundary without consideration of morphology, ecology and behaviour (Rubinoff, 2006).

The species problem has produced at least 26 species concepts (Mayden, 1997). The argument that one species concept is less valid than another is futile to begin with given that these definitions represent human attempts to bring order and structure to an entity that is constantly changing. Secondly, barcoding's primary aim is to identify unknown

specimens - not delimit species; furthermore, the by-product event of being unable to match an unknown sequence to reference database entries only suggests the *possibility* of a new species.

Thus, our attempt to determine whether the similarity criterion is accurate for insects depends on our knowledge of the patterns of evolution in insect COI. Unfortunately, the number of empirical studies examining the level of polymorphism across a broad range of insects is limited; consequently, it leaves us unable to assess if the similarity criterion is just or in need of a redefiniton and if COI is an appropriate marker for barcoding insects. In addition to establishing a clearer image of the current species concept and developing a greater understanding of the evolutionary dynamics of insect COI, the characterization of the evolution of COI would be informative for elucidating mechanisms or processes influencing its rate of change over time.

Armed with a better understanding of the molecular evolution of COI contributes to the success of barcoding which, in turn, generates more accurate information about the abundance and distribution of extant insect diversity. Consequently, questions concerning the influence of habitat, ecology, and selective forces on shaping existing biodiversity can be formed and answered. Furthermore, increased knowledge of COI leads to more-informed use of insect COI sequences in phylogenetic analyses, contributes to the preservation of biodiversity by identifying insect species that are or on the verge of being threatened and endangered, provides insight into resistant plant varieties, and affects the success of biological control agents against pests.

The current standards of technology have permitted researchers to collect vast amounts of insect sequence information that is easily stored and accessible. Thus, we are provided an optimal dataset from which to study the genetic variability between and within taxonomic grouping of insects.

Similar comprehensive investigations in vertebrate (Ward *et al.*, 2005; Kerr *et al.*, 2007) and smaller invertebrate groups (Hajibabaei *et al.*, 2006) have shown that COI generates enough polymorphism to be able to distinguish specimens. In fishes, it was determined, through the use of Nei-Gojobori models (Nei and Gojobori, 1986), that strong purifying selection was acting on COI (Ward *et al.*, 2005). To our knowledge, a comparative study of the molecular evolution of COI in insects has not been conducted yet. An aim of the present work is to examine the evolutionary patterns in COI across a broad phylogeny of 13,641 insect sequences from 16 orders. We are particularly interested in examining the rates of evolution among and within each taxonomic classification. Particular targets of interest would be species that evolve faster or slower relative to related taxa, and those of environmental and economical interest.

Unlike the methodology used in the vertebrate assessment of COI (Ward *et al.*, 2005), we follow the established methodology of Hebert, Ratnasingham and deWaard (2003)

whereby evolutionary rates are estimated from NJ (Neighbor-joining) trees (Saitou and Nei, 1987) based on K2P (Kimura two-parameter) distance matrices (Kimura, 1980). To infer what selective forces are operating on insect COI, estimation of the rate of nonsynonymous to synonymous changes ($\omega = d_N/d_S$) is conducted using PAML.

Table 1.1: The distribution of sequences among taxa

Order	Number of Sequences
Coleoptera	3861
Hymenoptera	2430
Lepidoptera	2297
Diptera	2252
Hemiptera	1121
Phthiraptera	709
Orthoptera	332
Odonata	151
Collembola	144
Ephemeroptera	116
Thysanoptera	94
Trichoptera	59
Isoptera	39
Psocoptera	27
Strepsiptera	7
Thysanura	2

1.3 Materials and Methods

Data collection and preparation

The sequences were collected from the National Center Biotechnology Institute's (NCBI) website between September and October 2005. Among the insects, a total of 13,641 sequences of COI were obtained. These fall among the insect orders as diagrammed in Table 1.1. Each of the 13,641 sequences are identified by an accession number and have been taxonomically classified according to: order, family, genus, and species. Taxa designations were taken from the GenBank entries.

The amino acid sequences were aligned using Muscle (Edgar, 2004). Some sequences provided incorrect alignments because some seqences did not overlap; therefore, a scaffolding sequence, spanning the length of the entire COI gene, was used to aid the protein alignment. We also ensured that nuclear-encoded pseudogenes, sequences with lots of unknown amino acids (i.e. X), short sequences, and duplicate sequences were removed from the dataset. Extensive efforts were made to manually align ambiguous alignments. The resulting aligned amino acid sequences were used to obtain the corresponding nucleotide sequence alignments using `TRANALIGN`.

The aligned nucleotide dataset was partitioned according to the taxonomic groupings: order, family, genus and species. Using programs from the PHYLIP package (Felsenstein, 1989), trees were constructed using the Neighbor-joining (NJ) algorithm (Saitou and Nei, 1987) based on Kimura two-parameter (K2P) distances (Kimura, 1980) for each group.

Preliminary data statistics

Given the re-constructed trees, the phylogenetic diversity or tree length (sum of all branch lengths; Faith, 1994), the average branch length (tree length divided by the total number of branches; the total number of branch lengths is determined by $2n$-2 where n represents the total number of taxa) and the expected heterozygosity measure per taxonomic grouping was calculated (Li and Graur, 1991). Note that the tree length represents the estimated nucleotide substitutions per site.

Identifying variation between and within taxonomic groupings

To analyze the variation within and between taxonomic groupings we looked at the rate of change at each codon partition (1^{st}, 2^{nd}, and 3^{rd}). Each file, within each taxonomic grouping, was separated into three individual files, each containing sequence data from one of the codon partitions; these files are denoted C_1, C_2 and C_3 for the 1^{st}, 2^{nd} and 3^{rd} codon partitions, respectively. For each codon partition file, 100 bootstrapped datasets were generated using SEQBOOT of the PHYLIP package (Felsenstein, 1989) and separated into individual files resulting in a total of 300 individual datasets, 100 for each of the codon partitions; they are denoted as $C_{1\{i\}}$, $C_{2\{i\}}$ and $C_{3\{i\}}$ where $i = \{1 \ldots 100\}$. Starting with the first bootstrapped sequence files from each codon position, $C_{1\{i\}}$, $C_{2\{i\}}$, and $C_{3\{i\}}$ where i is initially set to 1, the three files were concatenated to produce a file, denoted B_i, similar to what we would get if we were bootstrapping the original file. A NJ-tree, $T_{1\{i\}}$, (Saitou and Nei, 1987) based on K2P distances (Kimura, 1980) was constructed. $T_{1\{i\}}$ was used in combination with each of the first bootstrapped sequence files, $C_{1\{i\}}$, $C_{2\{i\}}$ and $C_{3\{i\}}$, to construct three new trees, $T_{2\{1\}\{i\}}$, $T_{2\{2\}\{i\}}$, and $T_{2\{3\}\{i\}}$, one for each codon position, and from these we derive tree lengths. This process was repeated for the remaining 99 bootstrappped datasets, $i = \{2 \ldots 100\}$. Once applied to all the sequence sets, median tree lengths, using T_2 trees, were determined. The topology of T_2 trees is based on T_1 with the goal of generating new branch lengths representing the change occurring at each codon position. The average tree length and standard deviation was calculated for each taxonomic group; the average tree length is based on cumulative tree length values. To permit comparisions between and within taxonomic groupings, the average tree lengths were normalized based on the rate of change found at the 3^{rd} codon position. Due to the existance

of outliers affecting the average tree length (by substantially increasing or decreasing the value), the median tree length was identified for each codon partition; the individual tree lengths summed to produce the cumulative tree length were used to determine the median tree length.

Identifying types of selective forces in action via PAML and LRT

To help explain the different molecular rates between taxa, understanding the type of selective forces in operation on the data is imperative. We are particularly interested in identifying if positive selection can explain the rate changes. To collect this information, we employed the use of the PAML package (Yang, 1997). Of the codon substitution models available, we focused on site models which allow the ω ($\omega = d_N/d_S$) ratio to vary among sites. Specifically, given a tree and a sequence file, codeml (seqtype = 1) was applied to each sequence file for each of the four major taxonomic groupings of the 16 orders of insects.

It is recommended that multiple models and tests be used in real data analysis (Anisimova, Bielawski and Yang, 2001). To detect positive selection, two likelihood ratio tests (LRTs) were conducted using 4 site models. Each test consists of a general model representing the null hypothesis where the proportion of sites undergoing positive selection is set to zero; this model is essentially a special case of the second, more-complex model in the test.

Each of the four site models used allow the ω ratio to vary among sites. The most basic of all the site models used is M0 which assumes one ω for all sites (Goldman and Yang, 1994). As of PAML version 3.14, the nearly neutral model (M1a) assumes a proportion p_0 of conserved sites with ω_0 estimated from the data under the constraint $0 < \omega < 1$, while the rest p_1 = 1 - p_0 are neutral sites with ω_1 = 1 (Yang *et al.*, 2000). The selection model (M2a) adds an additional class of sites, to M1a, with frequency p_2 = 1 - p_0 - p_1 and with ω_2 estimated from the data to detect positively selected sites (Yang *et al.*, 2000). Comparison of M1a, the null model, with M2a is a test for positively selected sites; this comparison is denoted as test 1 throughout the rest of the chapter. Similar to M1a and M2a, the more general model, M7 (beta) does not allow for positively selected sites while M8 (beta+ω) adds an extra component, p_1, to account for the possible occurrence of positively selected sites (Yang *et al.*, 2000). Comparison of M8 with M7 is denoted as test 2 throughout the rest of the chapter. The difference between the two tests is the method used for estimating the ω ratio for the proportion of conserved sites. In M1a, the ω ratio is estimated from the data under the constraint $0 < \omega < 1$. In M7, ω is still subject to the same constraint, however, it is estimated using the beta distribution where a higher level of flexibility is achieved because two parameters, p and q, are manipulated. Comparison of the two models in both

tests can detect positively selected sites using a LRT. When two models are nested, as are the ones in test 1 and 2, the LRT can be used. The LRT compares twice the log-likelihood difference with a χ^2 distribution with the degrees of freedom (df) equal to the difference in the number of parameters between the two models. In this case, the df of both tests is 2. $\omega < 1$ or $\omega = 1$ infers purfying and neutral selection, respectively. If $\omega > 1$, this is an indicator that sites are under the influence of positive selection.

From the results of the codeml runs, the likelihood values, the ω ratio, and potentially positively selected sites were extracted from the output. The extracted likelihoods were used to conduct likelihood ratio tests for each of the models.

1.4 Results

Preliminary data statistics

Preliminary statistics are given in Table 1.2. As we travel from deeper (order) to shallower (species) portions of the phylogeny, the average number of taxa per group decreases. Similarly, the average tree length decreases; exceptions include Hymenoptera, Lepidoptera, Hemiptera, Collembola, and Psocoptera. When it comes to average branch length, there doesn't seem to be any evident pattern; it is found to increase (Odonata, Collembola, Thysanoptera, Isoptera, and Strepsiptera), or decrease (Coleoptera, Hymenoptera, Lepidoptera, and Pscoptera), or show no pattern at all (Diptera, Hemiptera, Orthoptera, Phthiraptera, Ephemeroptera, and Trichoptera). Generally, average heterozygosity decreases as we travel from order to species.

Table 1.2: The distribution, average tree length, average branch length and heterozygosity of taxa among different taxonomic groupings

Order	Grouping	Count	$Avg.S^a$†	$Avg.TL^b$†	$Avg.BL^c$	$Avg.H^d$
Coleoptera	Order	1	3861	0^e	0^e	0.20
	Family	29	133.14	3.228	0.024	0.08
	Genera	449	8.60	1.015	0.044	0.03
	Species	1374	2.81	0.884	0.124	0.0086
Hymenoptera	Order	1	2430	0^e	0^e	0.26
	Family	41	59.27	3.697	0.041	0.07
	Genera	298	8.15	1.902	0.108	0.03

Continued on next page...

Table 1.2 – Continued from previous page...

Order	Grouping	Count	$Avg.S^a$†	$Avg.TL^b$†	$Avg.BL^c$	$Avg.H^d$
	Species	957	2.54	2.914	0.566	0.0061
Lepidoptera	Order	1	2297	0^e	0^e	0.13
	Family	27	85.07	2.454	0.051	0.05
	Genera	322	7.13	1.046	0.151	0.01
	Species	913	2.52	3.699	0.731	0.0023
Diptera	Order	1	2252	0^e	0^e	0.15
	Family	24	93.83	2.511	0.079	0.06
	Genera	132	17.06	2.176	0.207	0.02
	Species	485	4.64	0.973	0.121	0.0046
Hemiptera	Order	1	1121	0^e	0^e	0.25
	Family	52	21.56	2.278	0.041	0.06
	Genera	272	4.12	0.920	0.126	0.02
	Species	403	2.78	1.023	0.124	0.0093
Phthiraptera	Order	1	709	37.22	0.026	0.10
	Family	14	50.64	5.730	0.048	0.08
	Genera	111	6.39	1.198	0.051	0.05
	Species	236	3.00	0.715	0.046	0.0142
Orthoptera	Order	1	332	6.34	0.010	0.13
	Family	5	66.4	3.107	0.042	0.06
	Genera	47	7.06	0.285	0.014	0.02
	Species	124	2.68	0.054	0.005	0.0044
Odonata	Order	1	151	11.94	0.040	0.04
	Family	4	37.75	0.610	0.007	0.04
	Genera	9	16.78	0.424	0.005	0.02
	Species	43	3.51	0.089	0.003	0.0016
Collembola	Order	1	144	4.338	0.015	0.13
	Family	4	36	4.503	0.030	0.05
	Genera	9	16	4.298	0.054	0.04
	Species	26	5.54	0.109	0.005	0.0150
Ephemeroptera	Order	1	116	1.904	0.008	0.11
	Family	2	58	0.919	0.016	0.11
	Genera	4	29	0.785	0.013	0.05
	Species	17	6.82	0.276	0.002	0.0058
Thysanoptera	Order	1	94	0^a	0^a	0.13
	Family	2	47	2.307	0.028	0.09
	Genera	18	5.22	0.594	0.027	0.03
	Species	49	1.92	0.207	0.034	0.0070

Continued on next page...

Table 1.2 – Continued from previous page...

Order	Grouping	Count	$Avg.S^a$†	$Avg.TL^b$†	$Avg.BL^c$	$Avg.H^d$
Trichoptera	Order	1	59	2.663	0.023	0.18
	Family	5	11.8	0.526	0.030	0.07
	Genera	10	5.9	0.473	0.019	0.03
	Species	26	2.27	0.038	0.007	0.0056
Isoptera	Order	1	39	0^e	0^e	0^e
	Family	3	13	0^e	0^e	0^e
	Genera	4	9.75	0^e	0^e	0^e
	Species	12	3.25	0^e	0^e	0^e
Psocoptera	Order	1	27	3.14	0.060	0.13
	Family	17	1.59	0.390	0.089	0.03
	Genera	25	1.08	0^f	0^f	0.01
	Species	25	1.08	0.533	0.133	0.0067
Strepsiptera	Order	1	7	1.312	0.109	0.19
	Family	2	3.5	0.892	0.089	0.08
	Genera	4	1.75	0.261	0.065	0.05
	Species	4	1.75	0.261	0.065	0.0497
Thysanura	Order	1	2	0^f	0^f	0
	Family	2	1	0^f	0^f	0
	Genera	2	1	0^f	0^f	0
	Species	2	1	0^f	0^f	0

† For the calculation of avg. tree length and avg. branch length, the number of groups used to calculate the average does not equal the total number of groups available since trees could not be constructed for some groups.

[a] Represents the average number of species per group.

[b] Represents the average tree length per group.

[c] Represents the average branch length per group.

[d] Represents the average heterozygosity per site per group.

[e] Empty distance matrix due to the existance of no overlaps between most sequences.

[f] No trees obtained due to lack of taxa (3 minimum).

Identifying variation between and within taxonomic groupings

To quantify sequence variation, we examined the nucleotide substitution rates at the 1^{st}, 2^{nd}, and 3^{rd} positions. We expected the rates to differ between different partitions and that the rates of change may not be uniform among the different taxonomic groupings within each order.

The data in Table 1.3 reveal differences in the rate of evolution between codon partitions for each taxonomic grouping within each order. Rates of evolution are relatively consis-

tent across taxonomic groups within each order; however, some species groups show an increase or decrease, sometimes by an order of magnitude, in the molecular rate. Collembola, Odonata, Strepsiptera, Hymenoptera (only at the 1^{st} codon position), Trichoptera, Phthiraptera, Orthoptera and Diptera yield changes of an order of magnitude or more, whereas Lepidoptera, Hymenoptera (only at the 2^{nd} codon position), and Thysanoptera have slighly increased or decreased rates of change.

Table 1.3: Median tree lengths for each codon partition relative to the 3^{rd} codon position

Order	Group	1^{st} Position	2^{nd} Position	3^{rd} Position
Coleoptera	Order	0^a	0^a	0^a
	Family	0.12352	0.03046	1.000
	Genus	0.11070	0.02108	1.000
	Species	0.14671	0.02871	1.000
Hymenoptera	Order	0^a	0^a	0^a
	Family	0.15461	0.05108	1.000
	Genus	0.12188	0.03419	1.000
	Species	0.27083	0.10069	1.000
Lepidoptera	Order	0^a	0^a	0^a
	Family	0.13785	0.03281	1.000
	Genus	0.12604	0.01917	1.000
	Species	0.22377	0.04680	1.000
Diptera	Order	0^a	0^a	0^a
	Family	0.11604	0.02097	1.000
	Genus	0.14251	0.02217	1.000
	Species	0.09634	0.00581	1.000
Hemiptera	Order	0^a	0^a	0^a
	Family	0.16347	0.0331	1.000
	Genus	0.14265	0.03504	1.000
	Species	0.15639	0.06291	1.000
Phthiraptera	Order	0.13676	0.04245	1.000
	Family	0.15732	0.05508	1.000
	Genus	0.08842	0.01941	1.000
	Species	0.08622	0.00242	1.000
Orthoptera	Order	0.14490	0.03810	1.000
	Family	0.04510	0.00547	1.000
	Genus	0.14308	0.02211	1.000
	Species	0.13222	0.00804	1.000
Odonata	Order	0^a	0^a	0^a
	Family	0.12132	0.04861	1.000

Continued on next page...

Table 1.3 – Continued from previous page...

Order	Group	1^{st} Position	2^{nd} Position	3^{rd} Position
	Genus	0.15614	0.07000	1.000
	Species	0.24960	0.12737	1.000
Collembola	Order	0.13479	0.03526	1.000
	Family	0.11535	0.02654	1.000
	Genus	0.12455	0.03661	1.000
	Species	0.00087	0.00087	1.000
Ephemeroptera	Order	0.06423	0.00289	1.000
	Family	0.04429	0.00234	1.000
	Genus	0.05207	0.00338	1.000
	Species	0.08844	0.00483	1.000
Thysanoptera	Order	0[a]	0[a]	0[a]
	Family	0.16906	0.04879	1.000
	Genus	0.10779	0.01908	1.000
	Species	0.15846	0.01222	1.000
Trichoptera	Order	0.15166	0.04483	1.000
	Family	0.11463	0.01933	1.000
	Genus	0.12070	0.01406	1.000
	Species	0.00643	0.00292	1.000
Isoptera	Order	0[a]	0[a]	0[a]
	Family	0[a]	0[a]	0[a]
	Genus	0[a]	0[a]	0[a]
	Species	0[a]	0[a]	0[a]
Psocoptera	Order	0.15585	0.05367	1.000
	Family	0.10031	0.01985	1.000
	Genus	0[a]	0[a]	0[a]
	Species	0[a]	0[a]	0[a]
Strepsiptera	Order	0.22435	0.05819	1.000
	Family	0.15162	0.04104	1.000
	Genus	0.05778	0.00010	1.000
	Species	0.05778	0.00010	1.000
Thysanura	Order	0[b]	0[b]	0[b]
	Family	0[b]	0[b]	0[b]
	Genus	0[b]	0[b]	0[b]
	Species	0[b]	0[b]	0[b]

[a] Empty distance matrix due to the existance of no overlaps between most sequences.

[b] No trees obtained due to lack of taxa (3 minimum).

Identifying types of selective forces in action via PAML and LRT

A comprehensive listing of information about groups that were passed through PAML and returned with likelihoods, whether significant or not is given in Table 1.4. Thirty significant LRTs, for test 1 or test 2 or both, were found in the following orders: Hymenoptera, Lepidoptera, Diptera, Hemiptera, Phthiraptera, Orthoptera, and Trichoptera. Significant results were typically only found in the trees based on taxonomic groupings of genus and species. However within the Hemiptera and Phthiraptera, each had one significant LRT based on the tree for an entire family.

The likelihood values, the ω ratio, and potential sites of positive selection of groups that resulted in a signficant LRT are given in Table 1.5; groups processed through PAML that did not have significant LRT values are not shown.

Of the 30 significant LRTs, 10 cases had an ω ratio was above 1. *Hylaeus connectens* (Insecta: Hymenoptera) is significant for test 2 at less than the 5% level, specifically at a p-value of 10^{-2}. In the flies, (Insecta: Diptera), *Paragus tibialis* is significant for both tests at less than the 5% level, specifically at a p-value of 10^{-2}. Within the 'true bug' classification, Hemiptera, the family Cicadellidae is significant for test 2 at less than the 0.1% level, specifically at a p-value of 10^{-4}. Several cases within Lepidoptera were found to be significant. Two independent results for *Sesamia nonagrioides* (one based on the genus sequences and the other based on the species sequences) is significant for both tests at less than the 0.1% level, specifically at p-values of 10^{-4} and 10^{-5} for test 1 and 2 respectively. *Maculinea arionides* is significant for both tests at less than the 5% level, specifically at a p-value of 10^{-2}. Lastly, *Bombyx mori* is significant for both tests at less than the 0.1% level, specifically at a p-value of 10^{-10}. In the order Orthoptera, both *Melanoplus dawsoni* and *Melanoplus infantilis* are significant for both tests at less than the 0.1% level, specifically at p-values of 10^{-12} and 10^{-13} for test 1 and 2 respectively. The family Ricinidae of lice order, Phthiraptera, is significant for test 2 at less than the 5% level, specifically at a p-value of 10^{-3}.

Of the significant LRTs with an ω ratio above 1, the highest ω value is 255.9 (*Melanoplus dawsoni*: Orthoptera), followed by 49.62 (*Melanoplus infantilis*: Orthoptera), and 46.24 (*Bombyx mori*: Lepidoptera). The lowest ω value above 1 is 1.248 (*Hylaeus connectens*: Hymenoptera), followed by 1.346 (*Sesamia nonagrioides*: Lepidoptera), and 1.614 (Ricinidae: Phthiraptera). With the exception of the family Ricinidae, the rest are found at the species level.

Table 1.4: Groups with significant LRT tests for positive selection

Order	Grouping	NS^a	NG^b	G^c	NS/G^d	$Test^e$	P-value
Coleoptera	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	166	11	n/a	n/a	n/a	n/a
	Genus	1009	43	*Amblystomus*	3	2	1.45×10^{-2}
	Species	2209	197	*A. obtectus*	49	1	3.07×10^{-3}
						2	8.35×10^{-4}
				P. strobi	37	2	2.58×10^{-2}
Hymenoptera	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	487	18	n/a	n/a	n/a	n/a
	Genus	914	70	*Andrena*	106	1	2.15×10^{-3}
				Kradibia	6	2	3.39×10^{-2}
				Peristenus	9	1	1.62×10^{-3}
						2	1.79×10^{-4}
	Species	1277	117	*L. acervorum*	34	2	5.00×10^{-3}
				H. connectens	3	2	1.68×10^{-2}
				C. obscurior	7	2	4.66×10^{-2}
				C. sp. C SPQ-2003	5	2	1.52×10^{-2}
Lepidoptera	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	145	9	n/a	n/a	n/a	n/a
	Genus	1039	62	*Charissa*	4	1	3.99×10^{-3}
						2	3.51×10^{-4}
				S. nonagrioides	4	1	2.72×10^{-4}
						2	9.05×10^{-5}
	Species	1034	117	*P. phoebus*	75	1	8.44×10^{-6}
						2	3.78×10^{-7}
				M. arionides	4	1	4.78×10^{-2}
						2	1.62×10^{-2}
				B. mori	14	1	4.82×10^{-10}
						2	4.73×10^{-10}
				S. nonagrioides	4	1	1.36×10^{-4}
						2	1.37×10^{-4}
				P. phaon	4	2	1.37×10^{-4}
Diptera	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	357	7	n/a	n/a	n/a	n/a
	Genus	845	26	n/a	n/a	n/a	n/a

Continued on next page...

Table 1.4 – Continued from previous page...

Order	Grouping	NS^a	NG^b	G^c	NS/G^d	$Test^e$	P-value
	Species	1387	97	*A. aquasalis*	15	1	2.14×10^{-3}
						2	4.41×10^{-4}
				A. culicifacies	13	1	2.80×10^{-2}
						2	1.82×10^{-3}
				B. tau	3	1	1.64×10^{-2}
						2	1.06×10^{-2}
				P. tibialis	4	1	3.67×10^{-2}
						2	3.43×10^{-2}
Hemiptera	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	198	15	Cicadellidae	3	2	1.51×10^{-4}
	Genus	437	42	*Halobates*	19	2	2.24×10^{-2}
	Species	349	39	*G. pallescens*	3	2	4.87×10^{-2}
Phthiraptera	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	199	6	Pediculidae	168	2	2.64×10^{-2}
				Ricinidae	5	2	8.28×10^{-3}
	Genus	328	21	n/a	n/a	n/a	n/a
	Species	444	34	n/a	n/a	n/a	n/a
Orthoptera	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	57	3	n/a	n/a	n/a	n/a
	Genus	295	14	n/a	n/a	n/a	n/a
	Species	230	31	*M. dawsoni*	4	1	9.29×10^{-13}
						2	1.21×10^{-12}
				M. infantilis	6	1	2.32×10^{-12}
						2	1.69×10^{-12}
				M. missouli		1	5.24×10^{-3}
						2	5.51×10^{-3}
				M. triangularis	8	1	1.71×10^{-2}
						2	2.14×10^{-3}
Odonata	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	150	30	n/a	n/a	n/a	n/a
	Genus	144	3	n/a	n/a	n/a	n/a
	Species	99	5	n/a	n/a	n/a	n/a
Collembola	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	63	2	n/a	n/a	n/a	n/a
	Genus	21	2	n/a	n/a	n/a	n/a

Continued on next page...

Table 1.4 – Continued from previous page...

Order	Grouping	NS^a	NG^b	G^c	NS/G^d	$Test^e$	P-value
	Species	126	11	n/a	n/a	n/a	n/a
Ephemeroptera	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	116	2	n/a	n/a	n/a	n/a
	Genus	114	2	n/a	n/a	n/a	n/a
	Species	102	3	n/a	n/a	n/a	n/a
Thysanoptera	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	21	1	n/a	n/a	n/a	n/a
	Genus	30	5	n/a	n/a	n/a	n/a
	Species	35	9	n/a	n/a	n/a	n/a
Trichoptera	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	58	4	n/a	n/a	n/a	n/a
	Genus	48	3	n/a	n/a	n/a	n/a
	Species	32	8	*N. yamagataensis*	3	1	9.90×10^{-3}
						2	1.53×10^{-3}
Isoptera	Order	39	1	n/a	n/a	n/a	n/a
	Family	39	3	n/a	n/a	n/a	n/a
	Genus	38	3	n/a	n/a	n/a	n/a
	Species	31	4	n/a	n/a	n/a	n/a
Psocoptera	Order	27	1	n/a	n/a	n/a	n/a
	Family	13	4	n/a	n/a	n/a	n/a
	Genus	n/a	n/a	n/a	n/a	n/a	n/a
	Species	3	1	n/a	n/a	n/a	n/a
Strepsiptera	Order	7	1	n/a	n/a	n/a	n/a
	Family	6	1	n/a	n/a	n/a	n/a
	Genus	3	1	n/a	n/a	n/a	n/a
	Species	3	1	n/a	n/a	n/a	n/a
Thysanura	Order	n/a	n/a	n/a	n/a	n/a	n/a
	Family	n/a	n/a	n/a	n/a	n/a	n/a
	Genus	n/a	n/a	n/a	n/a	n/a	n/a
	Species	n/a	n/a	n/a	n/a	n/a	n/a

[a] Represents the total number of sequences that was passed through PAML and returned with quantifiable results.
[b] Represents the total number of groups, in the specified taxonomic grouping, that was passed through PAML and returned with quantifiable results.
[c] The name of a group with a signficant LRT.
[d] Represents the number of sequences in the significant LRT group.
[e] Test 1: M1a-M2a; Test 2: M7-M8.

Table 1.5: Likelihood values for groups with significant LRT

Order	Group	Group Name	Model code	likelihood	ω (d_n/d_s)	Sites
Coleoptera	Genus	*Amblystomus*	M0	-1326.69	0.021	
			M1	-1321.72	0.026	
			M2	-1320.23	0.226	495, 515
			M7	-1324.82	0.022	
			M8	-1320.58	0.392	515
	Species	*A. obtectus*	M0	-722.59	0.064	
			M1	-713.59	0.042	
			M2	-707.81	0.173	87
			M7	-714.90	0.100	
			M8	-707.81	0.173	87
		P. strobi	M0	-1464.28	0.037	
			M1	-1451.44	0.027	
			M2	-1450.18	0.035	
			M7	-1453.84	0.043	
			M8	-1450.18	0.034	217
Hymenoptera	Genus	*Andrena*	M0	-227.49	0.114	
			M1	-234.16	0.342	
			M2	-228.02	0.104	
			M7	-227.55	0.101	
			M8	-226.30	0.105	
		Kradibia	M0	-2329.01	0.013	
			M1	-2306.76	0.030	
			M2	-2306.76	0.030	
			M7	-2274.30	0.014	
			M8	-2270.91	0.020	

Continued on next page...

Table 1.5 – Continued from previous page...

Order	Group	Group Name	Model code	likelihood	ω (d_n/d_s)	Sites
		Peristenus	M0	-1955.63	0.024	
			M1	-1913.41	0.056	
			M2	-1906.98	0.541	1, 2, 3, 4, 9
			M7	-1901.14	0.037	
			M8	-1892.51	0.434	1, 2, 3, 4, 9
	Species	*L. acervorum*	M0	-1390.55	0.045	
			M1	-1382.28	0.040	
			M2	-1379.75	0.073	251
			M7	-1385.06	0.050	
			M8	-1379.77	0.073	251
		Hylaeus connectens	M0	-1396.67	0.013	
			M1	-1391.64	0.015	
			M2	-1390.32	1.248	207
			M7	-1394.64	0.013	
			M8	-1390.55	2.895	207
		C. obscurior	M0	-1112.17	0.012	
			M1	-1108.23	0.017	
			M2	-1106.99	0.310	213
			M7	-1109.74	0.014	
			M8	-1106.67	0.036	213, 217
		C. sp. C SPQ-2003	M0	-921.67	0.046	
			M1	-914.30	0.043	
			M2	-911.58	0.113	44
			M7	-915.77	0.075	
			M8	-911.58	0.113	44
Lepidoptera	Genus	*Charissa*	M0	-2113.78	0.003	

Continued on next page...

Table 1.5 – Continued from previous page...

Order	Group	Group Name	Model code	likelihood	ω (d_n/d_s)	Sites
			M1	-2112.80	0.008	
			M2	-2107.28	0.376	92
			M7	-2111.64	0.003	
			M8	-2103.68	0.973	92, 94, 95
		Sesamia nonagrioides	M0	-1656.33	1.294	
			M1	-1654.62	0.659	
			M2	-1646.41	0.409	
			M7	-1642.74	0.100	
			M8	-1633.43	1.356	151, 153, 154, 340
	Species	*P. phoebus*	M0	-1705.14	0.053	
			M1	-1684.69	0.045	
			M2	-1673.01	0.154	261
			M7	-1688.92	0.066	
			M8	-1674.13	0.157	261
		Maculinea arionides	M0	-1293.33	0.047	
			M1	-1287.89	0.039	
			M2	-1284.85	0.663	284
			M7	-1289.00	0.051	
			M8	-1284.87	3.226	284
		Bombyx mori	M0	-576.75	0.305	
			M1	-569.39	0.139	
			M2	-547.94	46.06	97, 98, 99, 100, 101, 122
			M7	-569.39	0.100	
			M8	-547.92	46.24	97, 98, 99, 100, 101, 122

Continued on next page...

Table 1.5 – Continued from previous page...

Order	Group	Group Name	Model code	likelihood	ω (d_n/d_s)	Sites
		Sesamia nonagrioides	M0	-1646.41	0.410	
			M1	-1642.33	0.172	
			M2	-1633.43	1.346	151, 153, 154, 340
			M7	-1642.33	0.200	
			M8	-1633.43	1.354	151, 153, 154, 340
		P. phaon	M0	-1060.32	0.011	
			M1	-1047.88	0.011	
			M2	-1046.10	0.033	1
			M7	-1053.16	0.013	
			M8	-1044.27	0.028	1
Diptera	Species	*A. aquasalis*	M0	-944.36	0.030	
			M1	-930.49	0.021	
			M2	-924.34	0.142	31
			M7	-934.70	0.040	
			M8	-926.97	0.052	31, 160
		A. culicifacies	M0	-824.64	0.022	
			M1	-811.89	0.023	
			M2	-808.31	0.133	2
			M7	-815.67	0.031	
			M8	-809.36	0.071	2, 61
		B. tau	M0	-1053.18	0.036	
			M1	-1034.26	0.065	
			M2	-1030.14	0.329	31, 32, 33, 34, 36
			M7	-1034.56	0.100	
			M8	-1030.01	0.334	13, 24, 31, 32, 33, 34, 36

Continued on next page...

Table 1.5 – Continued from previous page...

Order	Group	Group Name	Model code	likelihood	ω (d_n/d_s)	Sites
		Paragus tibialis	M0	-1358.58	999.0	
			M1	-1359.22	0.666	
			M2	-1355.92	36.17	138, 254, 277, 324, 329, 341
			M7	-1359.29	1.000	
			M8	-1355.92	37.45	138, 254, 277, 324, 329, 341
Hemiptera	Family	Cicadellidae	M0	-2020.44	0.005	
			M1	-1993.66	0.023	
			M2	-1993.26	0.675	30, 91, 313, 314, 316
			M7	-2000.44	0.007	
			M8	-1991.64	3.551	30, 91, 313, 314, 316
	Genus	*Halobates*	M0	-3637.57	0.002	
			M1	-3647.63	0.005	
			M2	-3647.63	0.005	
			M7	-3631.92	0.002	
			M8	-3628.12	0.002	
	Species	*G. pallescens*	M0	-2191.10	0.009	
			M1	-2187.25	0.014	
			M2	-2186.07	0.430	382
			M7	-2189.01	0.010	
			M8	-2185.99	0.610	382
Phthiraptera	Family	Pediculidae	M0	-1167.13	0.009	
			M1	-1160.32	0.014	

Continued on next page...

Table 1.5 – Continued from previous page...

Order	Group	Group Name	Model code	likelihood	ω (d_n/d_s)	Sites
			M2	-1160.10	0.019	78
			M7	-1163.93	0.011	
			M8	-1160.30	0.019	78
		Ricinidae	M0	-806.06	0.013	
			M1	-796.36	0.023	
			M2	-795.92	0.595	84
			M7	-798.83	0.014	
			M8	-794.04	1.614	84
Orthoptera	Species	*Melanoplus dawsoni*	M0	-699.87	0.582	
			M1	-679.60	0.322	
			M2	-651.90	255.6	3, 5, 16, 20, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119
			M7	-679.33	0.300	
			M8	-651.89	255.9	3, 5, 16, 20, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119
		Melanoplus infantilis	M0	-652.54	0.819	

Continued on next page...

Table 1.5 – Continued from previous page...

Order	Group	Group Name	Model code	likelihood	ω (d_n/d_s)	Sites
			M1	-644.10	0.161	
			M2	-617.31	49.62	1, 2, 3, 4, 5, 6, 7
			M7	-644.41	0.100	
			M8	-617.31	47.48	1, 2, 3, 4, 5, 6, 7
		M. missouli	M0	-1412.08	0.170	
			M1	-1406.79	0.110	
			M2	-1401.54	0.356	185, 187
			M7	-1406.74	0.100	
			M8	-1401.54	0.357	185, 187
		M. triangularis	M0	-1004.21	0.019	
			M1	-1000.25	0.017	
			M2	-996.18	0.169	213
			M7	-1001.81	0.021	
			M8	-995.67	0.158	213
Trichoptera	Species	*N. yamagataensis*	M0	-723.47	0.027	
			M1	-719.19	0.036	
			M2	-714.58	0.667	121
			M7	-720.59	0.042	
			M8	-714.10	0.717	50, 121, 137

Comparison of median rates of change

A comparison of the median rate of evolution of groups with significant LRTs and ω ratios greater than 1 against the median rate of change for the parent grouping in which the significant group is found is given in Table 1.6. *Hylaeus connectens* is evolving a magntitude slower than the rest of the Hymenoptera species. *Paragus tibialis* is evolving 2-fold faster than the rest of the Dipteran species. In comparison to the median rate of change for the Hemiptera family, Cicadellidae is evolving a magnitute slower at the 1^{st} position but the rate remains similar for the 2^{nd} position. Of the signficant cases for Lepidoptera, results for *Sesamia nonagrioides* indicates that the group is evolving approximately 4 times faster than other genera and slightly faster at the 1^{st} position and similarly at the 2^{nd} among species. *Maculinea arionides* is evolving slightly faster at the first and an order of magnitude slower at the 2^{nd} and *Bombyx mori* is evolving at rates that are an order of magnitude faster than other species. Within the Orthopteran sequences, *Melanoplus dawsoni* is changing at a rate that is 5 times faster at the 1^{st} and 2-fold faster at the 2^{nd} positions whereas *Melanoplus infantilis* is evolving at a very slow rate versus other species. The Ricinidae family of Phthiraptera evolves at a slightly faster rate only at the 1^{st} position.

Table 1.6: Normalized median tree lengths for each data partition relative to the 3^{rd} codon position for LRT significant groups

Order	Group	1^{st} Position	2^{nd} Position	3^{rd} Position
Coleoptera	Genus	0.111	0.021	1.000
	Amblystomus	0.125	0.062	1.000
	Species	0.147	0.029	1.000
	A. obtectus	0.317	0.291	1.000
	P. strobi	0.036	0.020	1.000
Hymenoptera	Genus	0.122	0.034	1.000
	Andrena	0.216	0.226	1.000
	Kradibia	0.184	0.121	1.000
	Peristenus	0.221	0.144	1.000
	Species	0.271	0.101	1.000
	L. acervorum	0.141	0.026	1.000
	Hylaeus connectens	0.076	0.021	1.000
	C. obscurior	0.130	0.052	1.000
	C. sp. C SPQ-2003	0.163	0.046	1.000
Lepidoptera	Genus	0.126	0.019	1.000

Continued on next page...

Table 1.6 – Continued from previous page...

Order	Group	1^{st} Position	2^{nd} Position	3^{rd} Position
	Charissa	0.051	0.001	1.000
	Sesamia nonagrioides	0.531	0.031	1.000
	Species	0.224	0.047	1.000
	P. phoebus	0.214	0.095	1.000
	Maculinea arionides	0.341	0.007	1.000
	Bombyx mori	1.554	0.831	1.000
	Sesamia nonagrioides	0.531	0.031	1.000
	P. phaon	0.105	0.043	1.000
Diptera	Species	0.096	0.006	1.000
	A. aquasalis	0.059	0.008	1.000
	A. culicifacies	0.126	0.020	1.000
	B. tau	0	0	0
	Paragus tibialis	6.000	6.000	1.000
Hemiptera	Family	0.163	0.033	1.000
	Cicadellidae	0.086	0.020	1.000
	Genus	0.143	0.035	1.000
	Halobates	0.074	0.007	1.000
	Species	0.156	0.063	1.000
	G. pallescens	0.131	0.029	1.000
Phthiraptera	Family	0.157	0.055	1.000
	Pediculidae	0.133	0.074	1.000
	Ricinidae	0.216	0.057	1.000
Orthoptera	Species	0.132	0.008	1.000
	Melanoplus dawsoni	0.705	0.573	1.000
	Melanoplus infantilis[a]	0	0	0
	M. missouli	0.225	0.527	1.000
	M. triangularis	0.089	0.011	1.000
Trichoptera	Species	0.006	0.003	1.000
	N. yamagataensis	0.071	0.001	1.000

[a] Median tree lengths could not be generated from all 100 boostrapped trees.

1.5 Discussion

1.5.1 Identifying variation between and within taxonomic groupings

As expected, variation seen at the three data partitions, 1^{st}, 2^{nd}, and 3^{rd} codon positions, generally reflect rates typically found at each codon position. The rates at the 3^{rd} position are the highest of the three data partitions, and, as changes at the 3^{rd} position are generally synonymous, it accurately reflects its status as the site predominantly responsible for generating heterogeneity. The fastest rates following those at the 3^{rd} are generally found at the 1^{st} codon position. As changes at this position can result in both synonymous and nonsynonymous substitutions, more constraint is applied. Lastly, the slowest molecular rates are found at the 2^{nd} position, also fitting, as changes at this position result in nonsynonymous changes.

The data in Table 1.3 reveal that there are differences in the rates of change among different taxonomic groupings. Changes in the substitution rate may be a result or composite of a number of factors. With respect to elevated rates, such a result may be explained by the sequences themselves; perhaps, for some species, the sequences used are part of the same sequence, but each portion specifies a different part of the complete sequence. To remedy this lack of overlap between these sequences, an increase in tree length might result. Given the size of our dataset, this is certainly a possibility, as it is difficult to account for each sequence, though much effort has been invested in generating a robust dataset. There is also the factor of population size; small effective sizes are said to experience faster rates of evolution because of drift (Castro, Austin and Dowton, 2002); however, according to Bazin, Glemin and Galtier (2006), "population size does not influence mitochondrial genetic diversity in animals". Regarding data collection, as our sequences were obtained from NCBI's website, we cannot guarantee unbiased sampling because it is influenced by the particular aims of phylogeographical studies. Additionally, factors affecting the mutation rate can cause increased or decreased rates in mtDNA evolution. Specific differences between the ecology or physiology of disparate groups may account for observed differences in the molecular rates. For instance, the efficiency of DNA repair may differ among organisms, thus affecting the mutation rate (Castro, Austin and Dowton, 2002; Li and Graur, 1991). A shorter generation time, marked by a greater number of germline cell divisions, and increased rates of DNA replication and nucleotide replacement may lead to higher mutation rates (Castro, Austin and Dowton, 2002; Li and Graur, 1991).

1.5.2 Identifying types of selective forces in action via PAML and LRT and comparison of median rates of change

Some of the likelihoods generated by PAML produced significant results when used in a LRT. More precisely, thirty significant LRTs, in test 1 or test 2 or both, were generated in nearly half of the insect orders. Of the thirty, ten groups had ω ratios greater than 1, which suggest that some sites are undergoing positive selection. The drawbacks associated with using LRTs should be kept in mind when considering and drawing inferences from the results. For instance, both very similar sequences and ones that are too far diverged carry little information and can lead to reduced LRT power (Anisimova, Bielawski and Yang, 2001). Very divergent sequences may have been subject to multiple substitutions, which can effectively mask any useful information (Anisimova, Bielawski and Yang, 2001). However, given the ubiquitous and vital nature of COI, there is little worry about the sequences being too far removed from one another and it seems there are enough sufficient changes at the species level (Table 1.3). The χ^2 distribution can be negatively affected by insufficient sample sizes (Anisimova, Bielawski and Yang, 2001); minimally, 4 or 5 sequences might be enough if the sequence divergence is optimal (Anisimova, Bielawski and Yang, 2001; Yang and Bielawski, 2000; Yang, 2001, 2002). Of the 10 groups with a significant LRT and an omega ratio greater than 1, the minimum number of species per group was 3 species for 4/10 tests. The maximum was 14 species for 1/10 tests. Intermediate groups range from 4-6 species. As seen in the data, it is possible for a group to have a significant LRT but an ω ratio less than 1. In fact, 20 groups experienced this result. If applied correctly, a LRT of positive selection does not generally lead to an excess of false positives (Anisimova, Bielawski and Yang, 2001). That there are many significant LRT tests without the proper support of an ω ratio > 1 suggests some of the results may be attributable to the phenomenon known as multiple test significance; this event was not accounted for. The multiple test significance describes the situation where the chance of generating false positives is more likely as the number of tests conducted increases (Bland and Altman, 1995). According to Anisimova, Bielawski and Yang (2001), short sequences reduce the power of the LRT, almost to 0% in detecting adaptive evolution. However, the distribution of the LRT fits the χ^2 distribution well enough that relatively short sequences of 50 codons should do (Anisimova, Bielawski and Yang, 2001). In this chapter, sequences carry a length of approximately 611 codons; though short sequences may be padded with gaps to satisfy alignment requirements. Alternatively, LRT significance with an ω less than 1 could result from a relaxation of purifying selection or low functional constraint unless the Bayes empirical Bayes (BEB) shows a relatively high posterior probabiliy of being under selection (Anisimova, Bielawski and Yang, 2002). Accounting for sampling errors, BEB uses the maximum likelihood estimates of parameters (such as site proportions and ω ratios) to calculate posterior probabilities for site classes (sites undergoing purifying, neutral, or positive selection; Yang and Bielawski, 2000; Yang, 2001, 2002).

Despite the requirements of using the LRT properly, there are also some inherent measures to ensure conservative implementation of the test. For instance, defaults of certain models help avoid false positives. In M1a, fixing ω_1 to 1, for netural sites, helps avoid misclassifying sites under weak purifying selection into the site class of positive selection by explicitly making them neutral (Yang and Bielawski, 2000; Yang, 2001, 2002). In both of the tests, the more-complex model has 2 extra parameters than does the more-general model; 2 degrees of freedom is very conservative (Anisimova, Bielawski and Yang, 2001; Yang and Bielawski, 2000; Yang, 2001, 2002).

With regard to the results of the LRTs, it is expected that estimates of branch lengths, κ, and ω should be relatively consistant among different models (Anisimova, Bielawski and Yang, 2001; Yang and Bielawski, 2000; Yang, 2001, 2002). Though we do not actually use the output generated by M0 in the LRTs, its output provides a scaffold for comparison with the output of the other models. Estimates for a couple of ω ratios in M0 were unexpected. For instance, in the species *Paragus tibialis*, an ω ratio of 999.0 was generated. Furthermore, some of the ω ratios generated in M2 or M8 or both were very large (Table 1.5). It is possible that such estimates of ω are given by the algorithm when $d_S = 0$ (Anisimova, Bielawski and Yang, 2001; Yang and Bielawski, 2000; Yang, 2001, 2002). Regardless, so long as d_N and d_S are specified in the output file, the results of an LRT, even a significant one, are still valid. Beyond the workings of the LRT, problems inherent in the data may explain the inflated ω values. Perhaps the persistance of an unresolved alignment error, or mistakes in sequencing, or both is to blame. Is it equally possible that a taxonomist may have incorrectly classified specimens as the same species, when, in fact, they are not.

Most of the groups significantly found to be under the influence of adaptive evolution were found at the genus and species level. This is expected as groups closer to the leaves of the tree are less likely to have multiple substitutions obscuring the changes; the most likely explanation for the increased rate of molecular evolution close to the tips of the phylogeny is an increased number of speciation events (Webster, Payne and Pagel, 2003).

All of the results have p-values lower than 10^{-2} but higher than 10^{-13}. The lower the p-value, the lower the chance of randomly achieving the result. Merely having a significant LRT test and proper ω ratio is not enough evidence to claim the occurrence of positive selection. Largely different molecular rates between a group that is putatively under positive selection and the taxonomic grouping in which they are situated may further attest the influence of adaptive selection or suggest otherwise.

Even though deemed significant, the likelihood of observing positive selection in *Hylaeus connectens* (Insecta: Hymenoptera), Cicadellidae (Insecta: Hemiptera), Ricinidae (Insecta: Phthiraptera) is low, as significance was observed only for test 2 with the support of relatively low p-values (Tables 1.4, 1.6). Significance based only on test 2 is more unreliable than if based only on test 1 because the M1a-M2a (test 1) is more robust than

M7-M8 (test 2) because the latter is prone to false positives (Yang and Bielawski, 2000; Yang, 2001, 2002).

There were several cases where the significance of the answer did not agree with the logic proposed by the rates. For instance, support for *Paragus tibialis* (Insecta: Diptera) was relatively low. Within Lepidoptera, the molecular rates of *Maculinea arionides*, at the 1^{st} and 2^{nd} positions, were evolving in opposite directions and *Melanoplus infantilis* was found to be evolving incredibly slow (Tables 1.4, 1.6).

Despite these inconsistencies, two species from Lepidoptera, namely *Sesamia nonagrioides* and *Bombyx mori*, and an Orthopteran species, *Melanoplus dawsoni*, showed higher substitution rates did than other related species that are reinforced by extremely low p-values, especially in the latter two cases (Tables 1.4, 1.6).

In summary, comparisons of the collected sequences suggest that COI largely remains evolutionary conserved, which implies that it evolves under the influence of purifying selection. Given its indispensable role as the terminal enzyme of the ATP-generating pathway, its conservation is likely a result of functional constraints.

Despite COI's inherent tendency to retain its original encoding, certain species are evolving at faster molecular rates. According to our data, the heterogeneity generated at the species level may simply be caused by the geographical isolation of populations, which are then subjected to different external forces that result in the natural occurrence of reproductive isolation and speciation (Coyne and Orr, 2004; Webster, Payne and Pagel, 2003). Microorganisms, such as bacteria, have the ability to cause reproductive isolation. One such organism is the bacterium *Wolbachia*, which has been shown to cause hybrid inviability across five orders of insects: Diptera, Coleoptera, Hymenoptera, Lepidoptera, and Hemiptera (Stevens and Wade, 1990); some affected species include tephritid fruit flies (Jamnongluk, Baimai and Kittayapong, 2003), leaf beetles (Keller *et al.*, 2004), parasitic wasps (Perrot-Minnot, Guo and Werren, 1996), and fire ants (Shoemaker *et al.*, 2000). We've already alluded to possible inherent biological mechanisms that are responsible for differences between species including but not limited to DNA repair efficiency, generation time, and metabolic rates. Biased sampling of the data must also be considered.

Alternatively, the overall reduced COI diversity in conjunction with increased evolutionary rates may also result from recurrent selective sweeps (Bazin, Glemin and Galtier, 2006). This can readily be applied to COI given its general sequence conservation and, according to our results, increased rate of evolution in several species. Though unrelated to COI, yellow fever mosquitoes, *Aedes aegypti* (Insecta: Diptera), have undergone a hitchhiking effect whereby genetic variation is reduced at the LF90 locus under intense selection imposed by OP (organophosphates) insecticides (Yan, Chadee and Severson, 1998). As a result of the selection pressure, mosquitoes will likely develop a resistance to the insecticide. This development increases the fitness of the mosquitoes which, in turn, helps the

mosquitoes successfully adapt to environmental conditions contaminated with insecticide (Yan, Chadee and Severson, 1998). Our data suggests that species from the Lepidopteran and Orthopteran orders may support this theory. If so, the question then becomes what is being selected for?

Since the engulfment of proteobacteria by eukaryotic cells (Li and Graur, 1991), there has been an ongoing evolutionary dynamic between mitochondria and their nuclear neighbors. For example, gene redundancy, owing to cyto-nuclear coexistance, has resulted in the transfer of mitochondrial sequences to the nucleus (referred to as Numts) (Zhang and M., 1996). The interaction between mitochondria and nucleus, within and among populations, is said to be important in maintaining polymorphism in the mitochondrial genome (Dowling, Abiega and Arnqvist, 2007). In addition to finding that marine copepod COI displays divergences above 20% for nucleotide subsitutions and up to 15% for amino acid divergence, they have found direct evidence for functional coadaptation between nuclear-encoded soluble cytochrome c (CYC) and COI (Edmands and Burton, 1999). In anthropoid primates, the interaction sites where the fastest evolving part of COX VIII-L contacts mtDNA-encoded subunit COX I provide evidence for structurally mediated nuclear-mitochondrial coevolution (Goldberg *et al.*, 2003). In *Drosophila*, through backcrossing, when the cytoplasmic genome is introduced into a foreign nuclear background, due to maternal inheritance of mtDNA, COX functionality is disrupted thus indicating the importance of mtDNA/nuclear coadaptation (Sackton, Haney and Rand, 2003). And, according to Rand, Haney and Fry (2004), amino acid substitution rates at mitochondrial COX-nuclear contact sites are different from non-contact sites. Essentially, all this evidence implies that fast-evolving sites of the mtDNA can adaptively evolve with respect to external structures they interact with, such as the nucleus, as the interaction likely provides structural stability important to enzyme function.

Futhermore, the range of mitochondrial diversity is as much a product of its interaction with the nucleus as it is of the local environment in which it exists. Similar to nuclear genes, mitochondrial variation may represent local adaptation via natural selection to local climatic conditions such as temperature (Dowling, Abiega and Arnqvist, 2007) and environmental oxygen levels (Rand, Haney and Fry, 2004).

Insect genetic diversity may also result from the development of morphological characteristics particular to different host plants they use as food. This is apparent in fruit flies of the genus *Bactrocera* (Jamnongluk, Baimai and Kittayapong, 2003) and bruchid beetles (Alvarez *et al.*, 2005; Fricke and Arnqvist, 2007).

Both *Bombyx mori* (Insecta: Lepidoptera) and *Melanoplus dawsoni* (Insecta: Orthoptera) possess high rates of evolution that may to be under the influence of adaptive selection. Long-term adaptation ability of an organism depends on the level of polymorphism. In Table 1.2, examination of the average tree length within Lepidoptera shows the

accumulation of many changes at the species level in comparison to organisms grouped at the family level (3.699 vs. 2.454). Unlike Lepidoptera, Orthoptera shows the opposite pattern (0.054 vs. 3.107). The general pattern is the same for average branch length (Table 1.2). However, average heterozygosity is seen to decrease in both orders. Furthermore, while Lepidopteran species tend to show higher median rates of evolution than do other taxonomic levels, Orthopteran species tend to be slower (Table 1.3). Given these patterns of polymorphism at the species level, both *Bombyx mori* and *Melanoplus dawsoni* have median evolutionary rates greater than the species norm (Table 1.6). It seems that the level of genetic variation occurring within either species may allow for adaptive changes. Though, further information regarding the life history of these organisms is required to determine the potential of such a force.

A search in published literature revealed that COI, along with other molecular markers, found that five large clades within the genus *Melanoplus*, in which *Melanoplus dawsoni* and *Melanoplus infantilis* are found, were phylogenetically questionable, as the branching order among these clades were indeterminate (Chapco, 2002). *Melanoplus infantilis* was found to be the least diverse of the species, thus confirming that the slow rate of change seen for this group is accurate. According to Chapco (2002), phylogenetic ambiguity is attributed to lack of lineage sorting for closely related species and the rapid, sequential burst(s) of evolution likely guided by local climatic conditions and sexual selection via male genetic traits. From this, it is possible that the sequence diversity, at the species level within Orthoptera, promotes adaptation to variable environments. Other Orthopteran species have shown an adaptation for an efficient metabolism during activity. Reinhold (1999) states that energetically demanding activities increase the mass-independent resting metabolic rate (RMR) in comparison to related species that spend less energy. Thus, the efficiency of metabolism during activities that require more energy should be under intense selection. As mitochondrial genes have an essential role in metabolism (Dowling, Abiega and Arnqvist, 2007), COI may be one of the prime targets for selection to act upon. There are several possible mechanisms upon which selection may act that could explain a resulting increase in RMR. Of particular interest is the adaptation for physiological mechanisms affecting protein activity, proton leakage, and oxygen availability; workings intricately related to COI functionality. Some Orthopteran species display acoustic advertisement signalling, which is energetically demanding behaviour (Prestwich, 1994). Two Orthopteran species, crickets and katydids, were shown to have exceptionally high pulse rates (Prestwich, 1994; Reinhold, 1999), thus suggesting a greater opportunity for the adaptive development of processes in which COI operates.

With regard to Lepidoptera, silkworms (*Bombyx mori*) have showed increased levels of COI transcript prior to termination of diapause (Hwang *et al.*, 2005). The increased consumption of oxygen, to aid tissue development, relies on COI expression and functionality; thus, COI must function efficiently, and, again, an enhanced opportunity for selection

may exist. In addition, the wings of butterflies and moths have long been morphologically diverse and most are a product of adaptive selection in response to predators in their surrounding environment and sexual selection (Beldade and Brakefield, 2002).

Additional examples for molecular diversity and the processes from which it stems are not limited to species found to be significant in this study. Within Diptera, *Drosophilia simulans* males have at least 2 distinct mtDNA haplotypes that exhibit differences in mitochondrial respiration and electron transport (Katewa and Ballard, 2007). With regard to CO1, (cytochrome c oxidase) complex IV of the electron transport chain (ETC) showed higher activity in one of the haplotypes. Thus, it seems that certain haplotypes are more metabolically efficient than are others, and it is these differences in metabolism efficiency that provides a platform on which selection may act.

Weevils of the *Pissodes strobi* complex (Insecta: Coleoptera) exhibit intergenic and interspecific differences of 12.8% and 6.0%, respectively, and the distribution of differences are not uniform across all domains of COI (Langor and Sperling, 1997). High polymorphism levels suggest this species complex evolves at faster rates. Within the *P. strobi* complex, the white pine weevil pest, *Pissodes strobi*, has no clear ancestral haplotype of the many haplotypes that it is described by, and its nucleotide diversity is the outcome of restricted gene flow or geographically separated populations or both (Laffin, Langor and Sperling, 2004).

Pollinating obligate mutualists, particularly the fig wasps of the genus *Kradibia* (Insecta: Hymenoptera), have large intergenic and interspecific sequence divergences and experiences a faster rate of evolution (Machado *et al.*, 2001; Weiblen, 2001). The increased rate of change may explain their ability to morphologically adapt to host figs (Weiblen, 2001). Branching pattern of these wasps is not well-resolved and thus further supports the notion that related figs are not necessarily pollinated by related wasps if they can readily adapt to the host fig that is available in the vicinity (Machado *et al.*, 2001).

Within the same genus as *Melanoplus infantilis*, *M. missouli* (Insecta: Orthoptera) has a paraphyletic geneology, which suggests that it is a species of recent origin. A group of organisms is paraphyletic if the group contains its most recent common ancestor but does not contain all the descendents of that ancestor. The grasshoppers' location is relatively isolated and has limited contact with other taxa. Though species in this genus are morphologically similar, sequence diversity is possibly reflected in the shape of male genitalia, which differs among these species; morphological differences are likely directed by sexual selection and reproductive isolation (Knowles, 2000).

Most evidence and studies suggest that the molecular variation is largely due to restricted gene flow from geographically disparate populations, biological differences, infection by *Wolbachia*, and biased sampling. However, published literature reveals that mtDNA can be equally affected by cyto-nuclear interactions and that some species, such as

pests and pollinators, must 'actively' adapt to their local surroundings to survive. And despite being largely conserved, empirical studies show that differential fitness among insect mtDNA haplotypes suggest a role for selection in maintaining mtDNA polymorphism. Insect evolution is also driven by the metabolic demands associated with activities requiring lots of energy including acoustic communication in Orthoptera (Prestwich, 1994; Reinhold, 1999), and termination of diapause in silkworms (Hwang *et al.*, 2005); these activities may suggest adaptation of metabolism during activity. There are also cases where sexual selection may commonly accelerlate adaptation under directional natural selection; however, sexual selection may also tend to depress population fitness under stabilizing natural selection (Fricke and Arnqvist, 2007). Sexual selection has been documented in grasshoppers (Insecta: Orthoptera; Chapco, 2002) and silkworms (Beldade and Brakefield, 2002) and many other insect species.

To make a connection between pattern and process, we must consider the evolutionary processes and demographic factors contributing to genetic variation, including but not limited to genetic drift, selection, mutation, and population history. Additionally, it is also important to ensure that assumptions regarding population size and structure, sampling, and sampling size have not been violated. Our data provides strong evidence that two species have the potential to have undergone adaptive evolution; however, further rigorous analyses are required to rule out alternative explanations. The increased molecular rates of change, at the species level, generates enough genetic divergence for COI to work effectively in barcoding insects.

1.6 Future work

A robustness analysis of the results retrieved from PAML should be conducted. One recommendation would be to increase sample size, especially for those groups proposing the workings of adaptive selection. The larger the sample size, the larger the proportion of changes from which to make an inference as to the whether there are more nonsynonymous changes relative to synonymous changes (Anisimova, Bielawski and Yang, 2001, 2002; Yang and Bielawski, 2000; Yang, 2001, 2002). Another option would be to consider changing the parameter CodonFreq to see the effect of codon usage and whether the results remain consistent (Yang and Bielawski, 2000; Yang, 2001, 2002).

Another interesting task would be the determination of structural domains in insect COI. By doing so, we might gather more information to explain how and why rate variation occurs by examining the rate changes occurring in different structural domains of COI. In some species, the level of polymorphism differed across structural domains of COI; this phenomenon occurs in the *Pissodes strobi* species complex (Langor and Sperling, 1997), ground beetles (Martinez-Navarro, Galian and Serrano, 2005), and a pest of maize, the corn

stalk borer (Kourti, 2006).

In bovine heart cytochrome c oxidase, the COI subunit consists of 12 transmembrane and 2 extra-membrane alpha-helices (Tsukihara *et al.*, 1996). Structural information is also given in the *Bos taurus* GenBank entry (accession: P00396) though its accuracy is uncertain. Given information about the structural boundaries of COI in bovine, it is assumed that bovine and insect COI are similar enough such that we can project the structural domains of the former onto the latter. One means to do so is to use WURST, a server that takes a protein sequence and performs sequence to structure alignments (Torda, Procter and Huber, 2004). BLOCKS identifies putative protein domains, in a given unknown sequence, based on comparisons against highly conserved regions of proteins from a protein database. Alternatively, we can conduct a manual profile alignment of the bovine COI sequence from the GenBank entry alongside a pre-aligned insect dataset of sequences across each order; this profile alignment would be accomplished through the use of MUSCLE (Edgar, 2004). A combination of the three methods would be the best approach to increase the accuracy and robustness of the results.

Once insect sites have been partitioned into different structural classes, the structural location of putative sites of positive selection is easily determined. Results may explain why these sites undergo rapid change in comparison to adjacent sites. In general, different structural domains are expected to be under different selective pressures. The *Pissodes* species complex showed regions of high and low variability in particular functional domains of COI (Langor and Sperling, 1997). This pattern was also reported in other orders, namely Diptera, Hymenoptera and Orthoptera, thus suggesting that the evolutionary rates have been consistent throughout most of insect history (Langor and Sperling, 1997). One way to illustrate this is to adopt an approach similar to previous analyses, whereby we determine the rates of change at each data partition for each structural domain. Another plan employs the use of PAML to implement maximum likelihood (ML) fixed-site models that assign and estimate different ω parameters for different structural partitions (Yang and Swanson, 2002). Results, from either procedure, may further corroborate rates and ω ratios seen for sites of adaptive evolution and may reveal new sites or domains of interest. Such research may identify sites integral for protein function.

Alternatively, with emphasis on codon and amino acid data, one might classify sites into different classes of variability (invariant, variant, or neutral) and determine if there is a correspondence between its variability class and its structural location. If so, putative sites of positive selection, identified by our previous analyses, analyses can be mapped accordingly.

Employing one of the three approaches proposed to determine structural partitions, a putative mapping of the selected sites (Table 1.5), within a 10 amino acid sliding window, on supported COI domains (Table 1.7) inferred by the online webserver tool, BLOCKS,

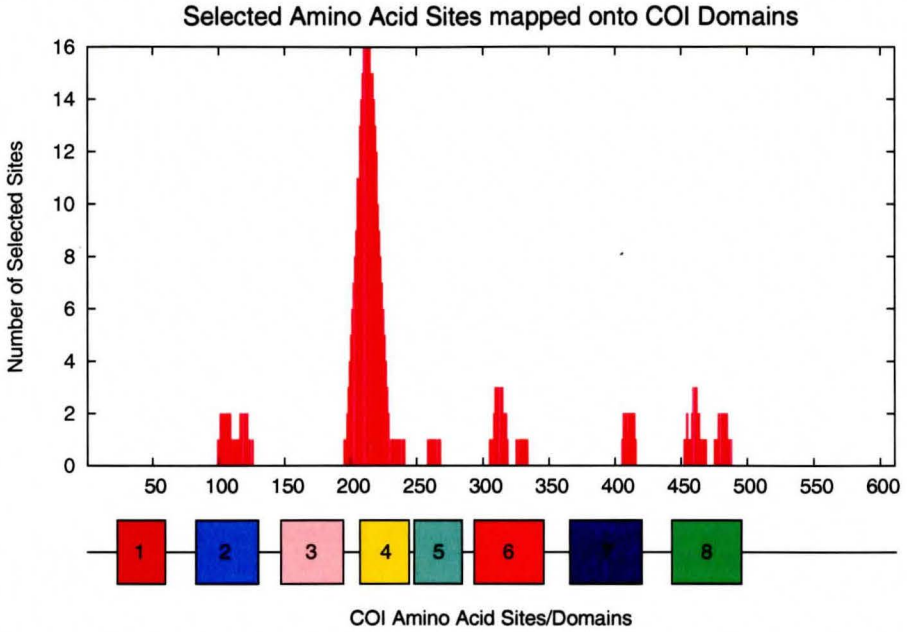

Figure 1.1: Visual distribution of selected sites along COI Blocks

(Henikoff and Henikoff, 1994) is given in Figure 1.6. BLOCKS detects and verifies protein sequence homology by comparing a protein or DNA sequence to the current database of protein 'blocks'; a match between short mutiply aligned ungapped segments with high conserved regions of protein are designated as 'blocks' or protein domains (Henikoff and Henikoff, 1994).

As seen in figure 1.6 and table 1.7, selected sites are distributed throughout COI with the highest frequency of sites concentrated in block 4. Based on a few annotated insect COI sequences and the fully characterized crystal structure of bovine heart COX (Tsukihara *et al.*, 1996), sites in block 4 are generally found on α-helices, which are known to have high substitution rates (Goldman, Thorne and Jones, 1998; Bustamante, Townsend and Hartl, 2000) and they may be responsible for the proper catalytic function of COX through structural maintenance. A residue within COI, histidine, associates with and holds in place the redox centers, such is the Cu_B found in domain 3, instrinsic to COX (Tsukihara *et al.*, 1995, 1996). Generally, selected sites would likely represent compensatory changes, with respect to cyto-nuclear interactions and structural stability for COX functionality.

A combination of structural-delimiting approaches in conjunction with information pertaining to amino acid characteristics, such as charge, hydrophobicity, and solvent accessibility, will help elucidate why and how certain sites are selected for in insect COI.

Table 1.7: Distribution of selected sites among supported COI Blocks. Values plotted based on a 10 amino acid sliding window

B[a]	Location (aa)	Block E-value	Insect[b]	Bovine [c]	Sites[d]
B1	21-61	6.4 x 10^{-34}	Fe (heme a)	I, II, Fe (heme a)	
B2	80-129	1.9 x 10^{-51}		II, III	**
B3	148-197	3 x 10^{-22}		IV, V	
B4	204-244	6.1 x 10^{-36}	Cu_B, His-Tyr	V, VI, Cu_B, O_2, His-Tyr	***
B5	246-283	2.3 x 10^{-16}	Cu_B	VI, VII	*
B6	293-345	2.1 x 10^{-19}		VIII, IX	**
B7	368-420	5.2 x 10^{-47}	Fe (heme a, a_3)	X, XI, Mg, Fe (heme a_3)	**
B8	445-490	5.4 x 10^{-5}		XII	**

[a]Block

[b]Based on select entries used to make blocks.

[c]Based on the *Bos taurus* entry used to make blocks.

[d]Relative frequency of selected sites. The number of asterisks (*) roughly correlate to the frequency of selected sites seen in Figure 1.6.

1.7 Acknowledgements

This work was supported through funding to the Canadian Barcode of Life Network from Genome Canada through the Ontario Genomics Institute, NSERC, and other sponsors listed at www.BOLNET.ca.

Chapter 2

`Fingerprint`: Visual depiction of variation in multiple sequence alignments

Lou, M., and Golding, G.B. (2007) *Molecular Ecology Notes* in press.

2.1 Abstract

There are a lack of programs available that focus on providing an overview of an aligned set of sequences such that the comparison of homologous sites becomes comprehensible and intuitive. Being able to identify similarities, differences, and patterns within a multiple sequence alignment is biologically valuable because it permits visualization of the distribution of a particular feature and inferences about the structure, function, and evolution of the sequences in question. We have, therefore, created a web tool, Fingerprint, which combines the characteristics of existing programs that represent identity, variability, charge, hydrophobicity, solvent accessibility, and structure along with new visualizations based on composition, heterogeneity, heterozygosity, d_N/d_S, and nucleotide diversity. Fingerprint is easy to use and globally accessible through any computer using any major browser. Fingerprint is available at http://evol.mcmaster.ca/fingerprint/.

2.2 Introduction

The mitochondrial gene, cytochrome c oxidase subunit I (COI), is the terminal enzyme in the electron transfer chain that transfers electrons to molecular oxygen without forming reactive oxygen species (Ludwig *et al.*, 2001). It helps form the electrochemical gradient across the inner mitochondrial membrane by pumping positively charged particles across it (Ludwig *et al.*, 2001). COI is a vital player in generating energy and is found broadly across many taxonomic categories. The Barcode of Life Initiative has employed COI as the standard gene because it is able to discriminate between many closely related animal species (Hebert, Ratnasingham and deWaard, 2003) and there is evidence to suggest that it also works well in algae (Saunders, 2005), arthropods (Smith, Fisher and Hebert, 2005), fish (Ward *et al.*, 2005) and some plants (Kress *et al.*, 2005). Identifying sequence changes in homologous sites provides insights about the structure, functional genomics, and evolution of a protein. Although some tools are currently available through the Barcode of Life Database (BOLD) for COI analysis, it is a continuing goal of the project to develop tools that can analyze and display data effectively.

There are various graphical multiple alignment editors, such as `ClustalX` (Thompson *et al.*, 1997), `Seaview` (Galtier, Gouy and Gautier, 1996), and `Jalview` (Clamp *et al.*, 2004), that display an alignment in its entirety. The problem is that it becomes difficult to summarize the characteristics or diversity of a site relative to other sites within a multiple sequence alignment. To qualitatively analyze up to 1000 sequences or more at lengths of over 1000 residues is very tedious, time consuming and difficult. To aid in such a task, there are a variety of multiple alignment shading programs available: `Alscript` (Barton, 1993), `ESPript` (Gouet *et al.*, 1999), `BoxShade`, `AMAS`, `WebLogo` (http://weblogo.berkeley.edu/) (Crooks *et al.*, 2004), `Sequence Similarity Presenter` (Fröhlich, 1994) and TEXShade (Beitz, 2000). Unfortunately, most of these programs require download and installation of software, support complicated documentation, impose a fee or limit the number of sequences allowed in the input file. Furthermore, most of the programs focus on providing sequence-by-sequence representations and not alignment overviews. With the continued advancement in technology, increasing amounts of sequence data are becoming readily available, which spurs the need for more visualization software.

In this paper, we introduce Fingerprint, a web server application that produces diagrams called fingerprints. A fingerprint is a horizontal bar made up of coloured or grey-scale vertical lines representing an overview of a desired feature in a sequence or in a set of aligned sequences. The concept of the alignment fingerprint was first introduced by Fröhlich (1994) in his `Sequence Similarity Presenter` and was subsequently adopted and updated by Beitz (2000) in his TEX-based alignment shading package. Though these programs do produce fingerprints, only one feature is available for representation or

the user is required to learn how to format documents in LaTeX to use the shading package, respectively. With new developments on five features, our tool provides options for a total of eleven distinct types of fingerprints, each depicting a different feature or 'flavour' of variation, and requires little to no overhead in learning how to use the program. The fingerprint concept has been incorporated into an online web interface, thus making it globally accessible via any major web browser. By default, information regarding the number of sequences and the average branch length of the aligned sequence set is given to provide a crude estimate of the significance of the fingerprint and a confidence level in the data presented.

Though the development of this tool was geared towards identifying diversity in COI barcodes, Fingerprint can be applied to a wide variety of datasets from any sequence data. Overall, Fingerprint is an effective tool to quickly and intuitively view the similarities, differences, and patterns in a multiple sequence alignment. The human eye can quickly assimilate these patterns making data exploration much easier.

2.3 System and Methods

Fingerprint was written using PHP, Perl, PostScript and the PHYLIP (Felsenstein, 1989) suite of programs. It was tested with Internet Explorer (IE), Konqueror, and Mozilla.

2.3.1 Algorithm and Implementation

Fingerprint is available online freely; no registration or download is necessary. As input, the user can choose to upload a single file or multiple files containing a sequence or a set of aligned sequences in FASTA format. Depending on the preferences of the user, the output can be placed in a single PDF file or multiple PDF files, which can be viewed in Acrobat Reader (free downloadable software) or any other PDF viewer.

The tool is currently capable of producing eleven different types of fingerprints, each depicting a particular feature or 'flavour' of variation. The fingerprint is a consensus overview of the desired feature within the aligned sequences.

2.3.2 Composition and Heterogeneity

In a *composition* fingerprint, each residue is represented by its own colour. This fingerprint depicts the unique composition of elements encoded by a sequence or a unique consensus

of a set of sequences which can be used to differentiate species based on the colouring and pattern of the residues (Figure 2.1A).

With regard to a consensus composition fingerprint, there is a loss of information since the tool represents the residues with the highest frequency of occurrence. To prevent this loss of information an alternate presentation is encoded. Each possible residue at a given site corresponds to a distinct coloured percentage of the vertical line drawn to represent a site. The heterogeneous composition of an alignment is viewed using a *heterogeneity* fingerprint. For example, invariable sites (represented by only 1 residue) are represented by one colour that extends for the entire length of the vertical line representing that particular site; the colour is determined by the residue. If, at a particular site, one residue occurs with a frequency of 0.25 and the second occurs with a frequency of 0.75, then the former colour will represent 25% of the height of the drawn line, and the latter will represent the remaining 75% (Figure 2.1B).

2.3.3 Identity, Variability, Heterozygosity, and Nucleotide Diversity

The diversity at sites possessing more than one residue is quantified and graphically depicted in different types of fingerprints. An *identity* fingerprint differentiates between invariant (identical residues) and variable (more than one residue possible) sites (Figure 2.1C). More information about the variable sites is obtained in a *variability* fingerprint. The variability of a site is quantified by considering the number of possible residues occurring at a site and is coloured accordingly. Thus, sites with the highest variability, are coloured black; in contrast, invariant sites are coloured white (Figure 2.1D1). Depending on user preference, the opposite colour scheme can be selected as a preference (Figure 2.1D2). Sites existing between these two extremes are shaded/coloured accordingly. Measures of diversity are calculated and graphically depicted in a *heterozygosity* fingerprint; this calculates the expected heterozygosity measure according to the equation

$$1 - \sum_{i=1}^{m} x_i^2$$

(Li and Graur, 1991) where x_i is the frequency of the i^{th} residue at a particular site. The value can also be interpreted as the probability that two residues chosen at random are different from each other. Highly variable sites possess high heterozygosity measures; those with the highest heterozygosity measures are coloured black. In contrast, invariant or invariable sites (one to a few residues) possess low heterozygosity measures; these sites are coloured white or close to it (Figure 2.1E). The heterozygosity measure, however, may not be accurate for nucleotide sequences due to the more extensive variation at the DNA level over large sequence lengths (Li and Graur, 1991). For nucleotide sequence data, the

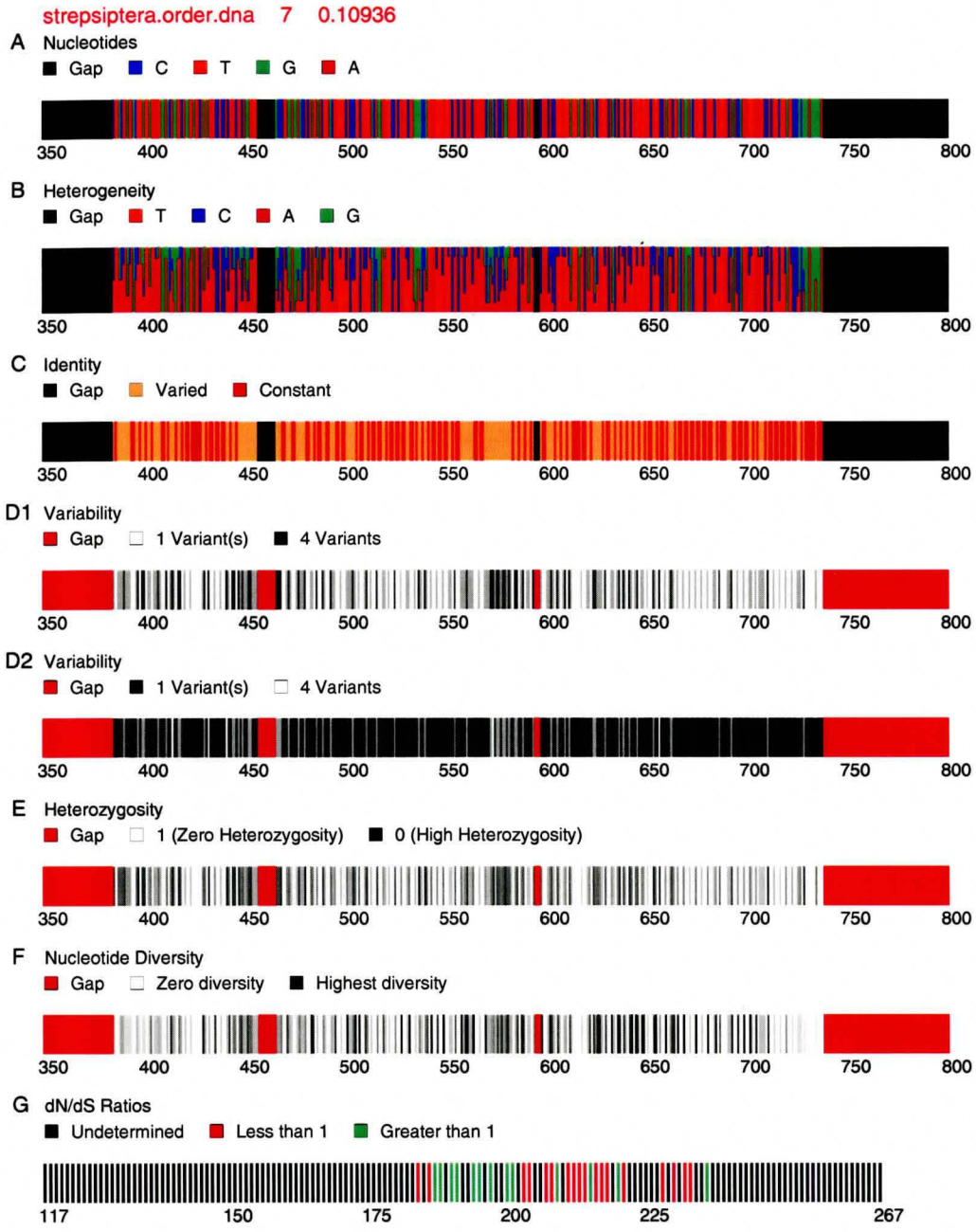

Figure 2.1: Nucleotide fingerprints based on the cytochrome c oxidase I (COI) gene from the order Strepsiptera (twisted-wing parasites). A. Composition, B. Heterogeneity, C. Identity, D1. Variability (Black), D2. Variability (White), E. Heterozygosity, F. Nucleotide Diversity, G. d_N/d_S Ratio. All fingerprints were constructed using the same input file

nucleotide diversity measure is calculated for each site using the equation:

$$\sum_{ij} x_i x_j \pi_{ij}$$

(Li and Graur, 1991) where x_i and x_j are the frequencies of the i^{th} and j^{th} residues at a particular site, respectively, and π_{ij} is either 1 or 0 if there is or is not a difference, between the i^{th} and j^{th} residues, respectively. Like the heterozygosity fingerprint, the *nucleotide diversity* fingerprint lies on a gray scale where sites possessing high nucleotide diversity measures are coloured black or close to it, and sites with low nucleotide diversity measures are coloured white or close to it (Figure 2.1F).

2.3.4 d_N/d_S Ratio

To gain some insight about the type of selective forces in operation, Fingerprint calculates the d_N/d_S ratio for each codon (triplet of nucleotides) within a sequence or set of sequences. The d_N/d_S fingerprint maps possible sites of purifying, neutral, and adaptive evolution (Yang, 1997, Figure 2.1G). Note that this algorithm is computationally extensive and may take time to complete. Also beware that this algorithm makes use of a simple NJ tree that could be easily be improved; hence, these results should be used only in a data exploration framework.

2.3.5 Charge, Hydrophobicity, Solvent Accessibility, Structure

The definitions for residue groupings, charge, structure, hydrophobicity, and solvent accessibility, were taken from Beitz (2000). The *charge* fingerprint identifies sites that are charged and uncharged (Figure 2.2A). The user may choose to differentiate the charged sites as either acidic or basic (Figure 2.2B). A *hydrophobicity* fingerprint categorizes sites as being acidic, basic, hydrophobic, or hydrophilic (Figure 2.2C). In a *solvent accessibility* fingerprint, each residue is categorized according to experimentally determined solvent accessibilities based on the position that such a residue is usually found in a folded protein (Figure 2.2D). A *structure* fingerprint identifies sites that are usually localized in the core (internal), on the surface (external) or neither of a globular protein (Figure 2.2E). Similar to the *composition* fingerprint, these features work best for a data set consisting of one sequence. Given a multiple sequence data set, the residue with the highest frequency of occurrence is used to represent that site.

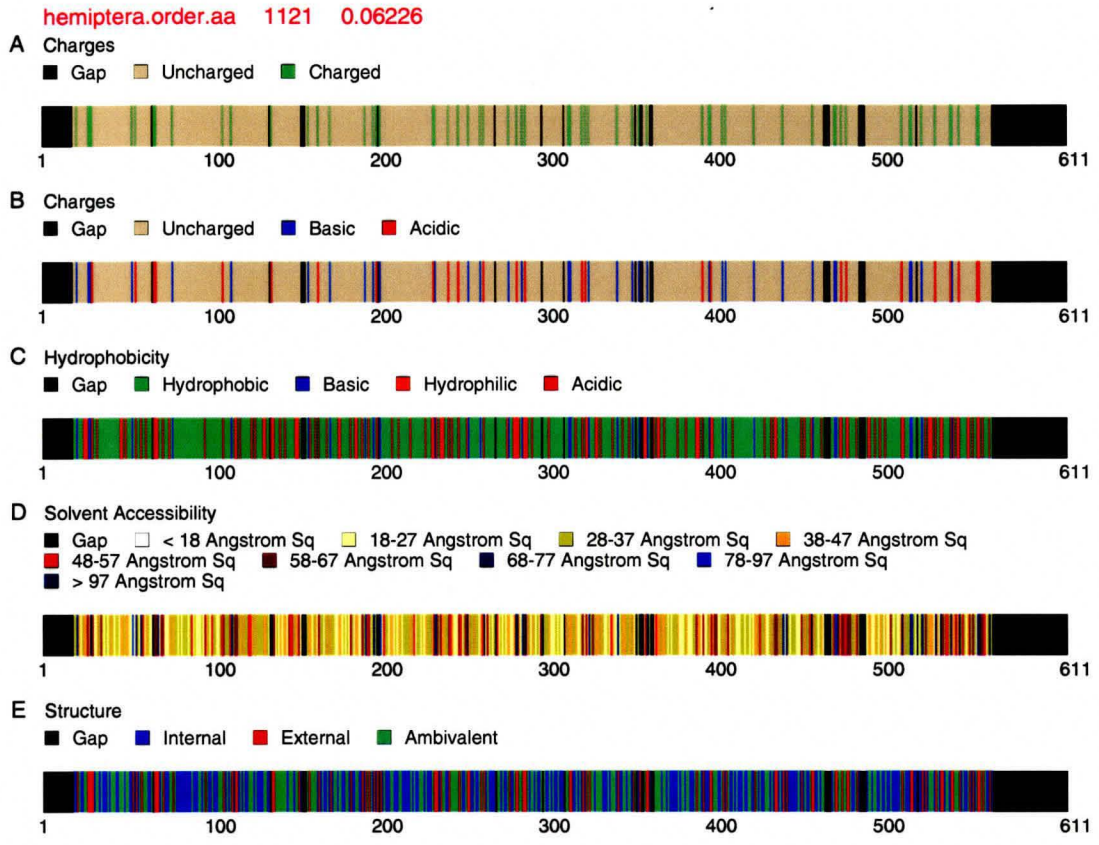

Figure 2.2: Amino acid fingerprints based on the COI gene from the order Hemiptera (true bugs). A. Charges, B. Charges (Acidic and Basic), C. Hydrophobicity, D. Solvent Accessibility and E. Structure. All fingerprints were constructed using the same input file

2.3.6 Managing Fingerprint Appearance

For publication purposes, the user has the option of manipulating several features associated with the appearance of the fingerprint. The Fingerprint assumes that the first residue in the sequence is indexed as the first position. Alternatively, the user has the option of specifying the first residue position, if it does not start at 1, and the last residue position, if not all the sites are to be represented. The Fingerprint program gives the user the option of selecting the range of sequence to be shown; the result is a "zoomed-in" view of the desired portion of the sequence. All the fingerprints in Figure 2.1 depict the nucleotide sequence in the range of 350 to 800 nucleotides. The height of the fingerprint is adjustable but must be larger than 0.1 inch. If no height is given, it is set to 1 inch by default. With the exception of the *heterogeneity* fingerprint, whose minimum height is intrinsically set to 0.5 inch, all other fingerprints shown in Figure 2.1 are shown at a height of 0.3 inch. Each label can be either hidden or displayed in the final output. While the label serves as a means of identification, labels identifying the number of sequences and the average branch length also serve as a measure of the meaningfulness of the output; these measures are located next to the input file name within the output file(s) in red (Figure 2.1). The output of the Fingerprint is, by default, written to a single PDF file. Within the output file, output from each input file is identified by the input file name. Alternatively, the user can select multiple file output in which case, output from each distinct input file is placed into its own PDF file.

2.3.7 Average Branch Length

Trees are constructed using the Neighbor-joining (NJ) algorithm (Saitou and Nei, 1987) based on Kimura two-parameter (K2P) distances (Kimura, 1980). Average branch lengths are calculated as the total tree length of the NJ-tree divided by the number of branches.

2.4 Results

For illustrative purposes, Fingerprint was applied to 9195 Lepidopteran sequences that were annotated by their genus and species designations. For each sequence, the appropriate family name was determined; subsequently, the sequences were partitioned by family. Composition fingerprints revealed very similar fingerprints made distinct by subtle changes throughout the length of the sequences (e.g. Figure 2.3A). Despite the compositional similarity, the variability, heterozygosity and nucleotide diversity (e.g. Figure 2.3B) fingerprints revealed distinct patterns of variation between families.

Within each family, these three types of fingerprints were similar with respect to the

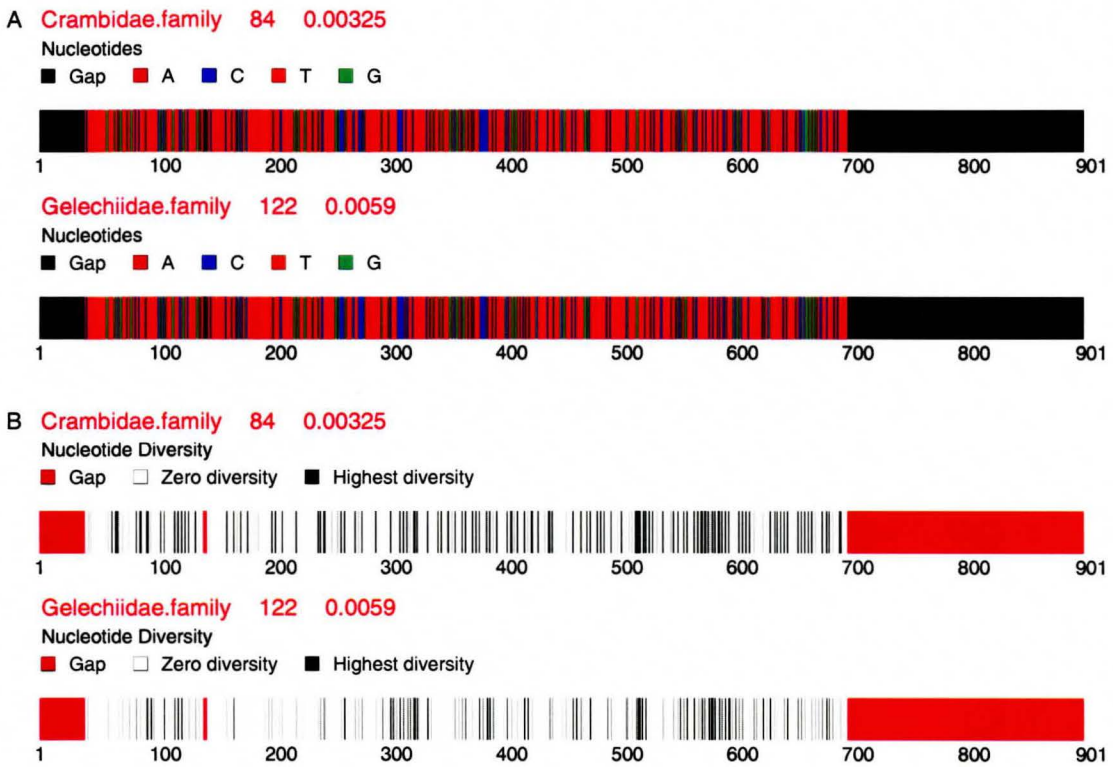

Figure 2.3: `Composition` (A) and `Nucleotide diversity` (B) fingerprints of two arbitrary Lepidopteran (butterfly) families: Crambidae and Gelechiidae

location of the sites exhibiting diversity; however, the extent of diversity differed between corresponding sites among the three representations. Generally, sites showing diversity had lower values in the heterozygosity fingerprints relative to the corresponding nucleotide diversity representations, which showed higher values for the same sites. The variability fingerprints possessed sites with values between the two extremes; this is expected since the colouring scheme is based on how variable a site is relative to sites of minimum and maximum variability.

Displaying the number of sequences and average branch length for each fingerprint proved to be worthwhile, as these values helped measure fingerprint robustness. In the lepidoptera data, fingerprints depicting little to no diversity could mean that the family of sequences are highly conserved or it could mean nothing at all, depending on the number of sequences used or their level of sequence divergence. In the cases presented, the large number of sequences would support the former interpretation. Furthermore, taking into account the average branch length helps yield further insights as to the credibility of the input data. An average branch of length of 0 or -1 would indicate that the sequences were identical copies of each other. On the other hand, if the average branch length is of reasonable value, this would suggest a family of sequences worthy of further analysis.

2.5 Discussion

Beitz's fingerprint inspiration stemmed from Fröhlich (1994). The output of Frölich's `Sequence Similarity Presenter` resembles that of our *variation* fingerprint, except that our representation provides the option of representing sites of high variability (termed by Frölich as sites that lack identity) as either white or black. The Fingerprint web server combines the standard characteristics of fingerprinting with new technological developments to produce a tool that is better equipped to accommodate the needs of the biological community. In addition to similarity, functional and variability shading, fingerprints based on composition, heterogeneity, diversity (heterozygosity and nucleotide diversity) measures and d_N/d_S ratios are now available. Fingerprint is computer- and browser-independent and easy to use. The output is compact, intuitively understandable, and is well suited for providing a quick overview of alignments consisting of one or more sequences.

The output is written in PostScript which is used to create high-quality vector-based text and graphics. Vector-based graphics do not possess unnecessary detail in visual representations of information, thus reducing file sizes, yet superior resolution is maintained because the full resolution of the display device (printer or monitor) is exploited. Since many fingerprints can be created from single and multiple file input, the output maintains a consistent appearance that is easy to reproduce.

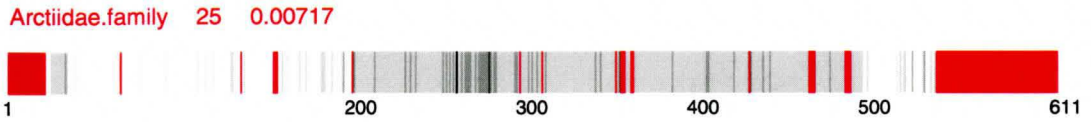

Figure 2.4: An unexpected application of Fingerprint: it is able to catch alignment errors

In addition to using the Fingerprint as a tool for identifying different types of variation, it may also be used to catch alignment errors. With reference to a fingerprint constructed using amino acid sequence data for the Lepidopteran family, Arctiidae, (Figure 2.4), it can be seen that the colour is uniform across a portion of the sequence, thus indicating the possibility of an alignment error spanning the length of that region.

However, there are some caveats to be aware of. Though, each fingerprint is accompanied by values for the number of sequences and average branch length as an indication of robustness, these values are merely two measures of fingerprint reliability. It is the responsibility of the user to follow up on the results depicted. The rate-limiting step of Fingerprint for most of the algorithms is the calculation of the average branch length. The number of pairwise distances that must be determined for the NJ-tree increase rapidly with larger datasets.

In summary, Fingerprint is effective for identifying sequence variation and for preparing high resolution, intuitive graphics for presentation.

2.6 Acknowledgments

This work was supported through funding to the Canadian Barcode of Life Network from Genome Canada through the Ontario Genomics Institute, NSERC, and other sponsors listed at www.BOLNET.ca.

Part II

CONCLUSION

It is possible that even genes with critical functions can diverge considerably across a broad range of taxa. Chapter 1 has shown that COI is subject to evolutionary change, thus proving itself suitable as the standard gene of the Barcode of Life initiative (Hebert *et al.*, 2003). We naturally expect genes that are functionally important to remain very conserved. Given the case of COI, it would be interesting to find out if there exist other genes, with indispensable roles, that are as or more evolutionary volatile as COI. More interestingly, what are the processes that would compel a vitally important gene to diverge?

We've discussed in length what processes may be responsible for generating sequence diversity in COI. And it is evident from our discussion that change may not be the result of one factor but a composite of many. Whether it is in the context of the combination of evolutionary processes contributing to genetic variation, or coadaptation of structural contacts, or adaptation to hosts, or to the local environment, it is often the sum of the components that make up the system and the interactions between them that should be under scrutiny rather than the individual parts. This is best expressed by John Donne in 1962 (with a modern twist): No man (or woman) is an island.

From an ecological perspective, this is especially true with regard to the diversity of roles of insects in society. Insects are of economic and ecological importance. They provide us with genetic (Grimaldi and Engel, 2005), medical (maggots), and economically (honey and silk) important resources. They are predominantly responsible for the stability of the ecosystem given their roles in recycling of organic matter, pollination, and insect population control. Thus, maintaining a wide diversity of insects is crucial.

As shown in chapter 1, COI is certainly capable of specimen identification in insects given its level of genetic heterogeneity. In chapter 2, the same result was graphically verified in Lepidopteran families using `Fingerprint`. Besides verifying genetic diversity and identifying putative molecular markers, the tool may be a cost-friendly alternative to DNA sequencing when designing primers; a researcher incurs no cost to screen highly variant candidates. Furthermore, such a tool would be useful in diagnostic and forensic research, which depends on unique sequences.

Overall, both COI and the `Fingerprint` are viable tools for quantifying global biodiversity. Coming back to the theme of interconnection, from mito-nuclear coadaptations to complex interactions within the ecosystem, conservation of global biodiversity is imperative as we are affected by devastations to the Earth's biological diversity.

Part III

REFERENCES

Bibliography

Alvarez, N., M. Hossaert-McKey, J.-Y. Rasplus, D. McKey, L. Mercier, L. Soldati, A. Aebi, T. Shani, and B. Benrey (2005). Sibling species of bean bruchids: a morphological and phylogenetic study of Acanthoscelides obtectus Say and Acanthoscelides obvelatus Bridwell. *Journal of Zoological Systematics & Evolutionary Research 43(1)*, 29–37.

Anisimova, M., J. P. Bielawski, and Z. Yang (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol. 18*, 1585–1592.

Anisimova, M., J. P. Bielawski, and Z. Yang (2002). Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol Biol Evol. 19*, 950–958.

Ball, S. L. and P. D. Hebert (2005). Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *Journal of North American Benthological Society 24(3)*, 508–524.

Barton, G. J. (1993). ALSCRIPT: a tool to format multiple sequence alignments. *Protein Engineering Design Protein & Selection 6*, 37–40.

Bazin, E., S. Glemin, and N. Galtier (2006). Population size does not influence mitochondrial genetic diversity in animals. *Science 312*, 570–572.

Beitz, E. (2000). TEXshade: shading and labeling of multiple sequence alignments using LATEX2 epsilon. *Bioinformatics 16*, 135–139.

Beldade, P. and P. M. Brakefield (2002). The genetics and evo-devo of butterfly wing patterns. *Nat Rev Genet. 3*, 442–452.

Bernays, E. A., E. A. Jarzzembowski, and S. B. Malcolm (1991). Evolution of Insect Morphology in Relation to Plants [and Discussion]. *Philosophical Transactions: Biological Sciences 333(1267)*, 257–264.

Bland, J. M. and D. G. Altman (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal 310*, 1073.

Bustamante, C. D., J. P. Townsend, and D. L. Hartl (2000). Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Molecular Biology and Evolution 17*, 301–308.

Capaldi, R. A. . (1990). Structure and function of cytochrome c oxidase. *Annual Review of Biochemistry 59*, 569–596.

Castro, L. R., A. D. Austin, and M. Dowton (2002). Contrasting rates of mitochondrial molecular evolution in parasitic Diptera and Hymenoptera. *Mol Biol Evol. 19*, 1100–1113.

Caterino, M. S., S. Cho, and F. A. Sperling (2000). The current state of insect molecular systematics: a thriving Tower of Babel. *Annu Rev Entomol. 45*, 1–54.

Chapco, W. (2002). A molecular phylogenetic analysis of the grasshopper genus melanoplus stal (orthoptera: Arcididae) - an update. *Journal of Orthoptera Research 11(1)*, 1–9.

Chapman, A. (2005). *Numbers of Living Species in Australia and the World.* Australian Biological Resources Study.

Clamp, M., J. Cuff, S. M. Searle, and G. J. Barton (2004). The Jalview Java alignment editor. *Bioinformatics 20*, 426–427.

Clare, E. L., B. K. Lim, M. D. Engstrom, J. L. Eger, and P. D. Hebert (2007). DNA barcoding of Neotropical bats: species identification and discovery within Guyana. *Molecular Ecology Notes 7(2)*, 184–190.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner (2004). WebLogo: a sequence logo generator. *Genome Research 14*, 1188–1190.

Cywinska, A., F. F. Hunter, and P. D. Hebert (2006). Identifying Canadian mosquito species through DNA barcodes. *Med Vet Entomol. 20*, 413–424.

Dowling, D. K., K. C. Abiega, and G. Arnqvist (2007). Temperature-specific outcomes of cytoplasmic-nuclear interactions on egg-to-adult development time in seed beetles. *Evolution Int J Org Evolution. 61*, 194–201.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32*, 1792–1797.

Edmands, S. and R. S. Burton (1999). Cytochrome c oxidase activity in interpopulation hybrids of a marine copepod: a test for nuclear-nuclear or nuclear-cytoplasmic coadaptation. *Evolution 53(6)*, 1972–1978.

Felsenstein, J. (1989). PHYLIP - Phylogeny Infernce Package (Version 3.2). *Cladistics 5*, 164–166.

Fricke, C. and G. Arnqvist (2007). Rapid adaptation to a novel host in a seed beetle (Callosobruchus maculatus): the role of sexual selection. *Evolution Int J Org Evolution. 61*, 440–454.

Fröhlich, K. U. (1994). Sequence Similarity Presenter: a tool for the graphic display of similarities of long sequences for use in presentations. *Comput Appl Biosci 10*, 179–183.

Galtier, N., M. Gouy, and C. Gautier (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci 12*, 543–548.

Goldberg, A., D. E. Wildman, T. R. Schmidt, M. Huttemann, M. Goodman, M. L. Weiss, and L. I. Grossman (2003). Adaptive evolution of cytochrome c oxidase subunit VIII in anthropoid primates. *Proc Natl Acad Sci U S A. 100*, 5873–5878.

Goldman, N., J. L. Thorne, and D. T. Jones (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics 149*, 445–458.

Goldman, N. and Z. Yang (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol. 11*, 725–736.

Gouet, P., E. Courcelle, D. I. Stuart, and F. Metoz (1999). ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics 15*, 305–308.

Grimaldi, D. and M. Engel (2005). *Evolution of the Insects*. Cambridge University Press.

Hajibabaei, M., D. H. Janzen, J. M. Burns, W. Hallwachs, and P. D. Hebert (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proc Natl Acad Sci U S A. 103*, 968–971.

Hebert, P. D., A. Cywinska, S. L. Ball, and J. R. deWaard (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences 270*, 313–321.

Hebert, P. D., E. H. Penton, J. M. Burns, D. H. Janzen, and W. Hallwachs (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences 101*, 14812–14817.

Hebert, P. D., S. Ratnasingham, and J. R. deWaard (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences 270*, S96–S99.

Hebert, P. D., M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis (2004). Identification of Birds through DNA Barcodes. *PLoS Biol. 2(10)*, 1657–1663.

Henikoff, S. and J. G. Henikoff (1994). Protein family classification based on searching a database of blocks. *Genomics. 19*, 97–107.

Hogg, I. D. and P. Hebert (2004). Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. *Canadian Journal of Zoology 82(5)*, 749–754.

Hwang, J. S., H. J. Go, T. W. Goo, E. Y. Yun, K. H. Choi, S. I. Seong, S. M. Lee, B. H. Lee, I. Kim, T. Chun, and S. W. Kang (2005). The analysis of differentially expressed novel transcripts in diapausing and diapause-activated eggs of Bombyx mori. *Arch Insect Biochem Physiol. 59*, 197–201.

Jamnongluk, W., V. Baimai, and P. Kittayapong (2003). Molecular evolution of tephritid fruit flies in the genus Bactrocera based on the cytochrome oxidase I gene. *Genetica. 119*, 19–25.

Katewa, S. D. and J. W. Ballard (2007). Sympatric Drosophila simulans flies with distinct mtDNA show difference in mitochondrial respiration and electron transport. *Insect Biochem Mol Biol. 37*, 213–222.

Keller, G. P., D. M. Windsor, J. M. Saucedo, and J. H. Werren (2004). Reproductive effects and geographical distributions of two Wolbachia strains infecting the Neotropical beetle, Chelymorpha alternans (Chrysomelidae, Cassidinae). *Molecular Ecology 13*, 2405–2420.

Kerr, K., M. Y. Stoeckle, C. J. Dove, L. A. Weigt, C. M. Francis, and P. D. Hebert (2007). Comprehensive DNA barcode coverage of North American birds. *Molecular Ecology Notes 7(4)*, 535–543.

Kimura, M. . (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution 16*, 111–120.

Knowles, L. L. (2000). Tests of pleistocene speciation in montane grasshoppers (genus Melanoplus) from the sky islands of western North America. *Evolution Int J Org Evolution. 54*, 1337–1348.

Kourti, A. (2006). Mitochondrial DNA restriction map and cytochrome c oxidase subunits I and II sequence divergence of corn stalk borer Sesamia nonagrioides (Lepidoptera: Noctuidae). *Biochem Genet. 44*, 321–332.

Kress, W. J., K. J. Wurdack, E. A. Zimmer, L. A. Weigt, and D. H. Janzen (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences 102*, 8369–8374.

Laffin, R., D. Langor, and F. Sperling (2004). Population structure and gene flow in the white pine weevil, Pissodes strobi (Coleoptera: Curculionidae). *Annals of the Entomological Society of America 97(5)*, 949–956.

Langor, D. W. and F. A. Sperling (1997). Mitochondrial DNA sequence divergence in weevils of the Pissodes strobi species complex (Coleoptera:Curculionidae). *Insect Mol Biol. 6*, 255–265.

Li, W. and D. Graur (1991). *Fundamentals of Molecular Evolution.* Sinauer Associates.

Ludwig, B. ., E. . Bender, S. . Arnold, M. . Huttemann, I. . Lee, and B. . Kadenbach (2001). Cytochrome C oxidase and the regulation of oxidative phosphorylation. *Chembiochem 2*, 392–403.

Machado, C. A., E. Jousselin, F. Kjellberg, S. G. Compton, and E. A. Herre (2001). Phylogenetic relationships, historical biogeography and character evolution of fig-pollinating wasps. *Proc Biol Sci. 268*, 685–694.

Martinez-Navarro, E. M., J. Galian, and J. Serrano (2005). Phylogeny and molecular evolution of the tribe Harpalini (Coleoptera, Carabidae) inferred from mitochondrial cytochrome-oxidase I. *Mol Phylogenet Evol. 35*, 127–146.

Mayden, R. L. (1997). *A hierachy of species concepts: The denouement in the saga of the species problem.*, pp. 381–424. London: Chapman and Hall.

Nei, M. and T. Gojobori (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol. 3*, 418–426.

Perrot-Minnot, M. J., L. R. Guo, and J. H. Werren (1996). Single and double infections with Wolbachia in the parasitic wasp Nasonia vitripennis: effects on compatibility. *Genetics. 143*, 961–972.

Pfenninger, M., C. Nowak, C. Kley, D. Steinke, and B. Streit (2007). Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic Chironomus (Diptera) species. *Mol Ecol. 16*, 1957–1968.

Prestwich, K. N. (1994). The energetics of acoustic signaling in anurans and insects. *American Zoologist 34(6)*, 625–643.

Rand, D. M., R. A. Haney, and A. J. Fry (2004). Cytonuclear coevolution: the genomics of cooperation. *Trends Ecol Evol. 19*, 645–653.

Reinhold, K. (1999). Energetically costly behaviour and the evolution of resting metabolic rate in insects. *Functional Ecology 13*, 217–224.

Remigio, E. A. and P. D. Hebert (2003). Testing the utility of partial COI sequences for phylogenetic estimates of gastropod relationships. *Molecular Phylogenetics and Evolution 29*, 641–647.

Rubinoff, D. (2006). Utility of mitochondrial DNA barcodes in species conservation. *Conserv Biol. 20*, 1026–1033.

Sackton, T. B., R. A. Haney, and D. M. Rand (2003). Cytonuclear coadaptation in Drosophila: disruption of cytochrome c oxidase activity in backcross genotypes. *Evolution Int J Org Evolution. 57*, 2315–2325.

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

Saunders, G. W. (2005). Applying DNA barcoding to red macroalgae: a preliminary appraisal holds promise for future applications. *Philosophical Transactions of the Royal Society of London, Series B Biological Sciences. 360*, 1879–1888.

Seifert, K. A., R. A. Samson, J. R. Dewaard, J. Houbraken, C. A. Levesque, J. M. Moncalvo, G. Louis-Seize, and P. D. Hebert (2007). Prospects for fungus identification using CO1 DNA barcodes, with Penicillium as a test case. *Proc Natl Acad Sci U S A. 104*, 3901–3906.

Shoemaker, D. D., K. G. Ross, L. Keller, E. L. Vargo, and J. H. Werren (2000). Wolbachia infections in native and introduced populations of fire ants (Solenopsis spp.). *Insect Mol Biol. 9*, 661–673.

Smith, M. A., B. L. Fisher, and P. D. Hebert (2005). DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society of London, Series B Biological Sciences. 360*, 1825–1834.

Smith, M. A., D. M. Wood, D. H. Janzen, W. Hallwachs, and P. D. Hebert (2007). DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae) are not all generalists. *Proc Natl Acad Sci U S A. 104*, 4967–4972.

Smith, M. A., N. E. Woodley, D. H. Janzen, W. Hallwachs, and P. D. Hebert (2006). DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). *Proc Natl Acad Sci U S A. 103*, 3657–3662.

Stevens, L. and M. J. Wade (1990). Cytoplasmically inherited reproductive incompatibility in Tribolium flour beetles: the rate of spread and effect on population size. *Genetics. 124*, 367–372.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research 25*, 4876–4882.

Torda, A. E., J. B. Procter, and T. Huber (2004). Wurst: a protein threading server with a structural scoring function, sequence profiles and optimized substitution matrices. *Nucleic Acids Research 32*, W532–W535.

Tsukihara, T., H. Aoyama, E. Yamashita, T. Tomizaki, H. Yamaguchi, K. Shinzawa-Itoh, R. Nakashima, R. Yaono, and S. Yoshikawa (1995). Structures of metal sites of oxidized bovine heart cytochrome c oxidase at 2.8 A. *Science 269*, 1069–1074.

Tsukihara, T., H. Aoyama, E. Yamashita, T. Tomizaki, H. Yamaguchi, K. Shinzawa-Itoh, R. Nakashima, R. Yaono, and S. Yoshikawa (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. *Science 272*, 1136–1144.

Ward, R. D., T. S. Zemlak, B. H. Innes, P. R. Last, and P. D. Hebert (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society of London, Series B Biological Sciences. 360*, 1847–1857.

Webster, A. J., R. J. Payne, and M. Pagel (2003). Molecular phylogenies link rates of evolution and speciation. *Science. 301(5632)*, 478.

Weiblen, G. D. (2001). Phylogenetic relationships of fig wasps pollinating functionally dioecious Ficus based on mitochondrial DNA sequences and morphology. *Syst Biol. 50*, 243–267.

Witt, J. D., D. L. Threloff, and P. D. Hebert (2006). DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation. *Mol Ecol. 15*, 3073–3082.

Yan, G., D. D. Chadee, and D. W. Severson (1998). Evidence for genetic hitchhiking effect associated with insecticide resistance in Aedes aegypti. *Genetics. 148*, 793–800.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci 13*, 555–556.

Yang, Z. (2001). *Handbook of statistical genetics*, pp. 327–350. Wiley.

Yang, Z. (2002). Inference of selection from multiple species alignments. *Curr Opin Genet Dev. 12*, 688–694.

Yang, Z. and J. P. Bielawski (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol Evol. 15*, 496–503.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics 155*, 431–449.

Yang, Z. and W. J. Swanson (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol. 19*, 49–57.

Zhang, D. X. and H. G. M. (1996). Nuclear integrations: challenges for mitochondrial dna markers. *Trends in Ecology and Evolution 11(6)*, 247–251.