Insect Molecular Biology (1996) 5(3), 153–165

The insect cytochrome oxidase I gene: evolutionary patterns and conserved primers for phylogenetic studies

D. H. Lunt, D.-X. Zhang, J. M. Szymura* and G. M. Hewitt

Population Biology Sector, School of Biological Sciences, University of East Anglia, Norwich

Abstract

Insect mitochondrial cytochrome oxidase I (COI) genes are used as a model to examine the within-gene heterogeneity of evolutionary rate and its implications for evolutionary analyses. The complete sequence (1537 bp) of the meadow grasshopper (*Chorthippus parallelus*) COI gene has been determined, and compared with eight other insect COI genes at both the DNA and amino acid sequence levels. This reveals that different regions evolve at different rates, and the patterns of sequence variability seems associated with functional constraints on the protein. The COOH-terminal was found to be significantly more variable than internal loops (I), external loops (E), transmembrane helices (M) or the NH2 terminal. The central region of COI (M5–M8) has lower levels of sequence variability, which is related to several important functional domains in this region.

Highly conserved primers which amplify regions of different variabilities have been designed to cover the entire insect COI gene. These primers have been shown to amplify COI in a wide range of species, representing all the major insect groups; some even in an arachnid. Implications of the observed evolutionary pattern for phylogenetic analysis are discussed, with particular regard to the choice of regions of suitable variability for specific phylogenetic projects.

Keywords: insect, *Chorthippus parallelus*, cytochrome oxidase I, mitochondrial DNA, conserved PCR primers, genetic marker.

*Present address: Department of Comparative Anatomy, Jagellonian University, Karasia 6, 30-060 Krakow, Poland.

Received 14 July 1995; accepted 3 November 1995. Correspondence: Professor G. M. Hewitt, Population Biology Sector, School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ.

Introduction

The study of mitochondrial DNA (mtDNA) sequences has become the method of choice in recent years for a wide range of taxonomic, population and evolutionary investigations in animals. Many aspects of the structure and evolution of mtDNA have made it a valuable evolutionary tool. These include its ease of isolation, high copy number, lack of recombination, conservation of sequence and structure across metazoa, and range of mutational rates in different regions of the molecule (reviewed by Moritz *et al.*, 1987; Harrison, 1989; Simon, 1991; Wolstenholme, 1992a). The mitochondrial gene encoding subunit I of cytochrome oxidase (COI) possesses some extra characteristics which make it particularly suitable as a molecular marker for evolutionary studies.

Firstly COI, as the terminal catalyst in the mitochondrial respiratory chain, has been relatively well studied at the biochemical level, and its size and structure appears to be conserved across all aerobic organisms investigated (Saraste, 1990). Mutational studies have been used to map the reaction centres of this subunit (Gennis, 1992) and these provide a background which enables interpretation of sequence differences in terms of gene function. Cytochrome oxidase I is involved in both electron transport and the associated translocation of protons across the membrane and it has been shown to contain a range of different types of functional domain including ligand sites, components of the proton channel, structural a-helices and interspersing hydrophilic loops (Saraste, 1990; Gennis, 1992). Amino acid residues in the reaction centres, which are highly conserved, do not dominate the entire COI molecule, allowing scope for considerable variability in some regions. Such a mix of highly conserved and variable regions so closely associated in a mitochondrial gene make the COI gene particularly useful for evolutionary studies.

Secondly, the COI gene is the largest of the three mitochondria-encoded cytochrome oxidase subunits (composed of 511 amino acids in *D. yakuba*, compared

to 228 for COII and 261 for COIII; Clary & Wolstenholme, 1985), and is one of the largest protein-coding genes in the metazoan mitochondrial genome. This enables one to amplify and sequence many more characters (nucleotides), within the same functional complex, than is possible for almost any other mitochondrial gene.

A suitable genetic marker is an essential prerequisite for success in many evolutionary studies. The crucial characteristic in the choice of such a marker is the substitutional rate of the particular region. To a large degree it is the broad spectrum of substitutional rates which accounts for the popularity of animal mtDNA as a molecular tool, since it allows resolution of both intraspecific phylogenies (e.g. Avise *et al.*, 1987) and the higher level systematics of anciently diverged taxa (e.g. Ballard *et al.*, 1992). It is well known that different genes may evolve at different rates, and the same gene may have different rates of evolution in different lineages. However, within-gene heterogeneity of evolution rate has not yet received enough attention especially in the field of lower taxonomic level phylogenetic studies. It may be misleading for many applications to consider a gene as fast or slowly evolving, because this implies a homogeneity of rate across the whole gene, which is rarely true due to the concentration of functional constraints in specific regions of the DNA sequence. Hence it is highly advantageous to have information concerning the relative substitutional rates of different gene regions, as this will allow a much more informed choice of sequence for particular phylogenetic investigations. Sequences evolving too quickly are known to lose their ability to unambiguously reveal the phylogeny of anciently diverged taxa. Similarly, choice of a sequence which is too conserved when addressing questions of intraspecific phylogeography, for example, will not provide enough informative characters to determine the requisite relationships. Thus for many studies success will depend to a large degree on the sampling of a region containing a suitable level of variability.

In this study we have used the COI gene as a model to study the within-gene heterogeneity of evolutionary rate and discuss this in terms of its implications to phylogenetic studies. The complete sequence of cytochrome oxidase I for the meadow grasshopper (*Chorthippus parallelus*) has been determined. (At the time of writing, no other Orthopteran COI sequence has previously been published, though the *Locusta migratoria* sequence, kindly provided by Paul Flook, University of Berne, Switzerland, prior to publication, is also included here; see Flook *et al.*, 1996.) Comparative analysis of this sequence with seven other published insect COI sequences has been carried out in order to identify and quantify areas of differing levels of variability. The relative rates of evolution (at the amino acid level) of these areas are considered in the context of both the structure–function model of COI and their utility in evolutionary studies. Areas of DNA sequence which are completely conserved are also shown and conserved primers designed and tested for their applicability to a wide range of insect species.

Results and Discussion

Sequence and structure of the C. parallelus *COI gene*

The COI gene of *Chorthippus parallelus* has been completely sequenced and is presented in Fig. 1 together with an alignment of eight other insect species. This sequence significantly extends our knowledge on insect COI genes, because six of the eight sequences previously available were from the same Order, Diptera. The *Chorthippus parallelus* COI gene is flanked by tRNA-Tyr at the 5′ end and tRNA-Leu at the 3′ end. Although the relocation of tRNA genes is not an uncommon event (Hauke & Gellisson, 1988; Pääbo *et al.*, 1991; Smith *et al.*, 1993), *Chorthippus* shares its positioning of these genes with many of the other insects presented here (Beard *et al.*, 1993). The complete *C. parallelus* COI nucleotide sequence was found to be 69.4% A+T, comprised of 34.1% A, 15.5% C, 15.1% G and 35.3% T. The A+T percentage of nucleotides at the third codon position is much higher (90.8% AT) than either of the other two locations (first = 57.9%; second = 60.6%). With the exception of *Apis*, these values are typical of those reported for other insects (see Table 1).

The putative initiation codon for *C. parallelus* COI is TCG and the stop codon is a single T at position 1537. Although the exact initiation and termination codons for insect COI genes are probably less clear than those for any other insect mitochondrial gene (Beard *et al.*, 1993) the codons suggested here are wholly consistent with those reported elsewhere (see Table 1). Beard *et al.* (1993) discuss the possibility that the initiation codon for *D. yakuba* is the TCG triplet immediately following the ATAA which is more usually recognized. Although TCG would fit well with many of the other insect sequences presented here, it is not supported by the other *Drosophila* sequences. Both *D. simulans* and *D. sechellia* share the ATAA motif but have undergone substitutions which alter the following TCG (Ser) triplet to CCG (Pro). The 5′ COI sequence of two other *Drosophila* species have been reported by Satta *et al.* (1987). These species commence with the tetranucleotides ATAA (*D. melanoga-*

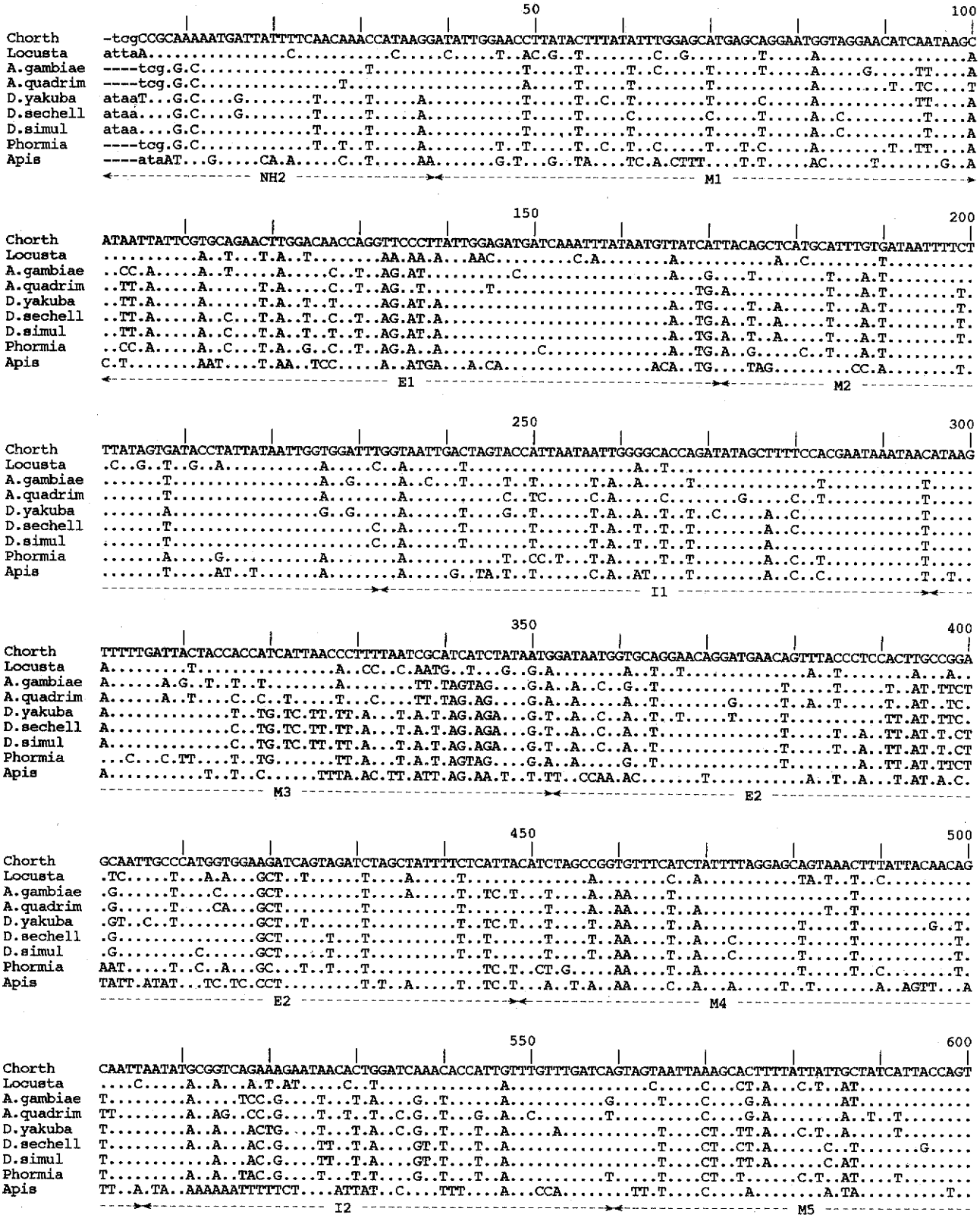
50
100
Chorth -tcgCCGCAAAAATGATTATTTTCAACAAACCATAAGGATATTGGAACCTTATACTTTATATTTGGAGCATGAGCAGGAATGGTAGGAACATCAATAAGC
Locusta attaA.....C.....C.....C.....T..AC.G..T.....C..G.....T.....A.....A
A.gambiae ----tcg.G.C.....T.....T.....T.....T..C.....T.....T.....A.....G.....TT....A
A.quadrim ----tcg.G.C.....T.....A.....T.....T.....T.....A.....T..TC....T
D.yakuba ataaT...G.C....G.....T.....T.....A.....T.....T..C..T.....T.....C.....A.....TT....A
D.sechell ataa....G.C....G.....T.....T.....A.....T.....T.....C.....C.....T.....A..C.....T....A
D.simul ataa....G.C.....T.....T.....A.....T.....T.....T.....T.....T.....A..C.....T....A
Phormia ----tcg.G.C.....T..T..T.....A.....T..T.....T..C..T..C.....T...T.C.....A.....T..TT....A
Apis ----ataAT...G.....CA.A.....C..T.....AA.....G.T...G..TA....TC.A.CTTT....T.T.....AC.....T.....G..A
NH2
M1
150
200
Chorth ATAATTATTCGTGCAGAACTTGGACAACCAGGTTCCCTTATTGGAGATGATCAAATTTATAATGTTATCATTACAGCTCATGCATTTGTGATAATTTTCT
Locusta .....A..T...T.A..T.....AA.AA.A...AAC.....C.A.....A.....A..C.....T.....
A.gambiae ..CC.A.....A..T.....A.....C..T..AG.AT.....C.....A...G....T.....T...A.T.....
A.quadrim ..TT.A.....A.....T.A.....C..T..AG..T.....T.....TG.A.....T...A.T.....T.
D.yakuba ..TT.A.....A.....T.A..T..T.....AG.AT.A.....A..TG....T..A.....T...A.T.....T.
D.sechell ..TT.A.....A..C...T.A..T..C..T..AG.AT.A.....A..TG.A..T..A.....T...A.T.....T.
D.simul ..TT.A.....A..C...T.A..T..T..T..AG.AT.A.....A..TG.A..T..A.....T...A.T.....T.
Phormia ..CC.A.....A..C...T.A..G..C..T..AG.A..A.....C.....A..TG.A..G.....C..T...A.T.....
Apis C.T.....AAT....T.AA..TCC.....A..ATGA...A.CA.....ACA..TG....TAG.....CC.A.....T.
E1
M2
250
300
Chorth TTATAGTGATACCTATTATAATTGGTGGATTTGGTAATTGACTAGTACCATTAATAATTGGGGCACCAGATATAGCTTTTCCACGAATAAATAACATAAG
Locusta .C..G..T..G..A.....A.....C..A.....T.....A..T.....
A.gambiae .....T.....A..G.....A..C...T....T..T.....T.A..A.....T.....T.....T.....
A.quadrim .....T.....A.....A.....C..TC.....C.A.....C.....G.....C..T.....T.....
D.yakuba .....A.....G..G.....A.....T....G..T.....T.A..A..T..T..C.....A..C.....T.....
D.sechell .....T.....C..A.....T.....T.....T.A..T..T..T.....A..C.....T.....
D.simul .....T.....C..A.....T.....T.....T.A..T..T..T.....A.....T.....
Phormia .....A.....G.....A.....A.....T..CC.T...T.A.....T..T.....A..C..T.....T.....
Apis .....T.....AT..T.....A.....A.....G..TA.T..T.....C.A..AT....T.....A..C..C.....T..T..
I1
350
400
Chorth TTTTTGATTACTACCACCATCATTAACCCTTTTAATCGCATCATCTATAATGGATAATGGTGCAGGAACAGGATGAACAGTTTACCCTCCACTTGCCGGA
Locusta A.....T.....A..CC..C.AATG..T...G..G.A.....A..T..T.....A..T.....A...A..
A.gambiae A.....A.G..T..T..T.....A.....TT.TAGTAG....G.A..A..C..G..T.....T.....T.....T..AT.TTCT
A.quadrim A.....A..T....C..C..T.....T...C....TT.TAG.AG....G.A..A.....A..T.....G.....T..A..T.....T..AT..TC.
D.yakuba A.....T..TG.TC.TT.TT.A...T.A.T.AG.AGA...G.T..A..C..A..T..T.....T.....T.....TT.AT.TTC.
D.sechell A.....C..TG.TC.TT.TT.A...T.A.T.AG.AGA...G.T..A..C..A..T.....T.....T..A..TT.AT.T.CT
D.simul A.....C..TG.TC.TT.TT.A...T.A.T.AG.AGA...G.T..A..C..A..T.....T.....T..A..TT.AT.T.CT
Phormia ...C...C.TT....T..TG.....TT.A...T.A.T.AGTAG....G.A..A.....G..T.....T.....T..A..TT.AT.TTCT
Apis A.....T..T..C.....TTTA.AC.TT.ATT.AG.AA.T..T.TT..CCAA.AC.....T.....A..T..A...T.AT.A.C.
M3
E2
450
500
Chorth GCAATTGCCCATGGTGGAAGATCAGTAGATCTAGCTATTTTCTCATTACATCTAGCCGGTGTTTCATCTATTTTAGGAGCAGTAAACTTTATTACAACAG
Locusta .TC.....T...A.A...GCT..T.....T....A.....T.....A.....C..A.....TA.T..T..C.....
A.gambiae .G.....T....C....GCT.....T....A.....T..TC.T...T....A..AA....T.....T.....
A.quadrim .G.....T....CA...GCT.....T.....T.....T....A..AA....T..A.....T..T.....
D.yakuba .GT..C..T.....GCT..T.....T.....T..TC.T...T....T..AA....T..A.....T.....T.....G..T.
D.sechell .G.....GCT.....T....T.....T..T.....T....T..AA....T..A...C.....T.....T.....T.
D.simul .G.....C.....GCT.....T...T.....T..T.....T....T..AA....T..A...C.....T.....T.....T.
Phormia AAT.....T..C..A...GC...T..T...T.....TC.T..CT.G.....AA....T..A.....T.....T..C.....T.
Apis TATT.ATAT...TC.TC.CCT.....T.T..A.....T..TC.T...A..T.A..AA....C..A...A.....T..T.....A..AGTT...A
E2
M4
550
600
Chorth CAATTAATATGCGGTCAGAAAGAATAACACTGGATCAAACACCATTGTTTGTTTGATCAGTAGTAATTAAAGCACTTTTATTATTGCTATCATTACCAGT
Locusta ....C.....A..A...A.T.AT.....C..T.....A.....C.....C...CT.A....C.T..AT.....
A.gambiae T.....A.....TCC.G....T...T.A....G..T.....A.....G.....T....C....G.A.....AT.....
A.quadrim TT.....A..AG..CC.G....T..T..T..C.G..T...G..A..C.....T.....C....G.A.....A..T..T.....
D.yakuba T.....A..A...ACTG....T...T.A..C.G..T...T..A.....A.....T....CT..TT.A...C.T..A.....T.....
D.sechell T.....A..A...AC.G....TT..T.A....GT.T...T..A.....T....CT..CT.A.....C..T.....G.....
D.simul T.....A...AC.G....TT..T.A....GT.T...T..A.....T....CT..TT.A.....C.AT.....
Phormia T.....A..A..TAC.G....T...T.T....G..T...T..A.....T.....T....CT..T.....C.T..AT....T.....
Apis TT..A.TA..AAAAAATTTTTCT....ATTAT..C....TTT....A...CCA.....TT.T....C....A.....A.TA.....T..
I2
M5

Figure 1.

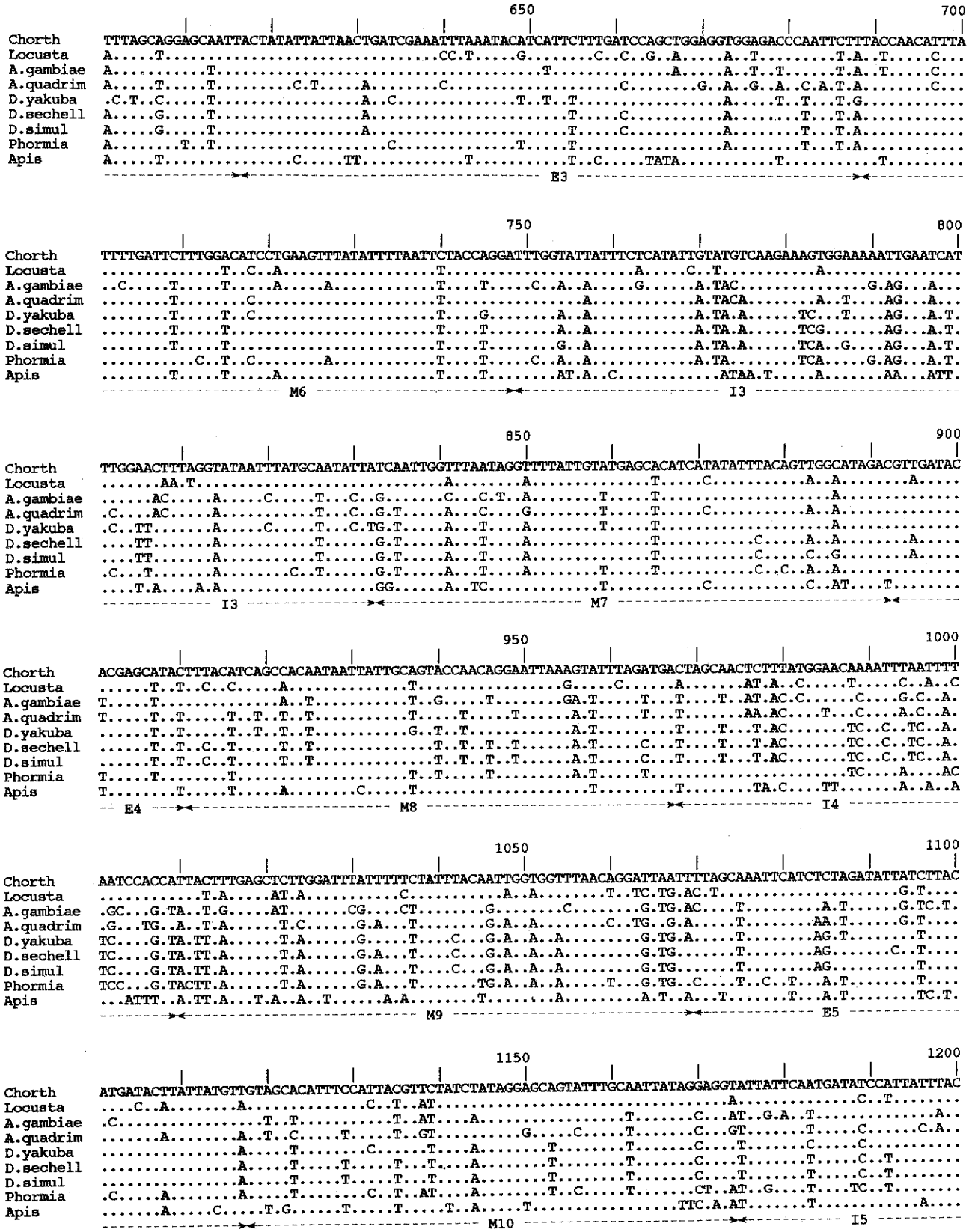

Figure 1 (*continued*)

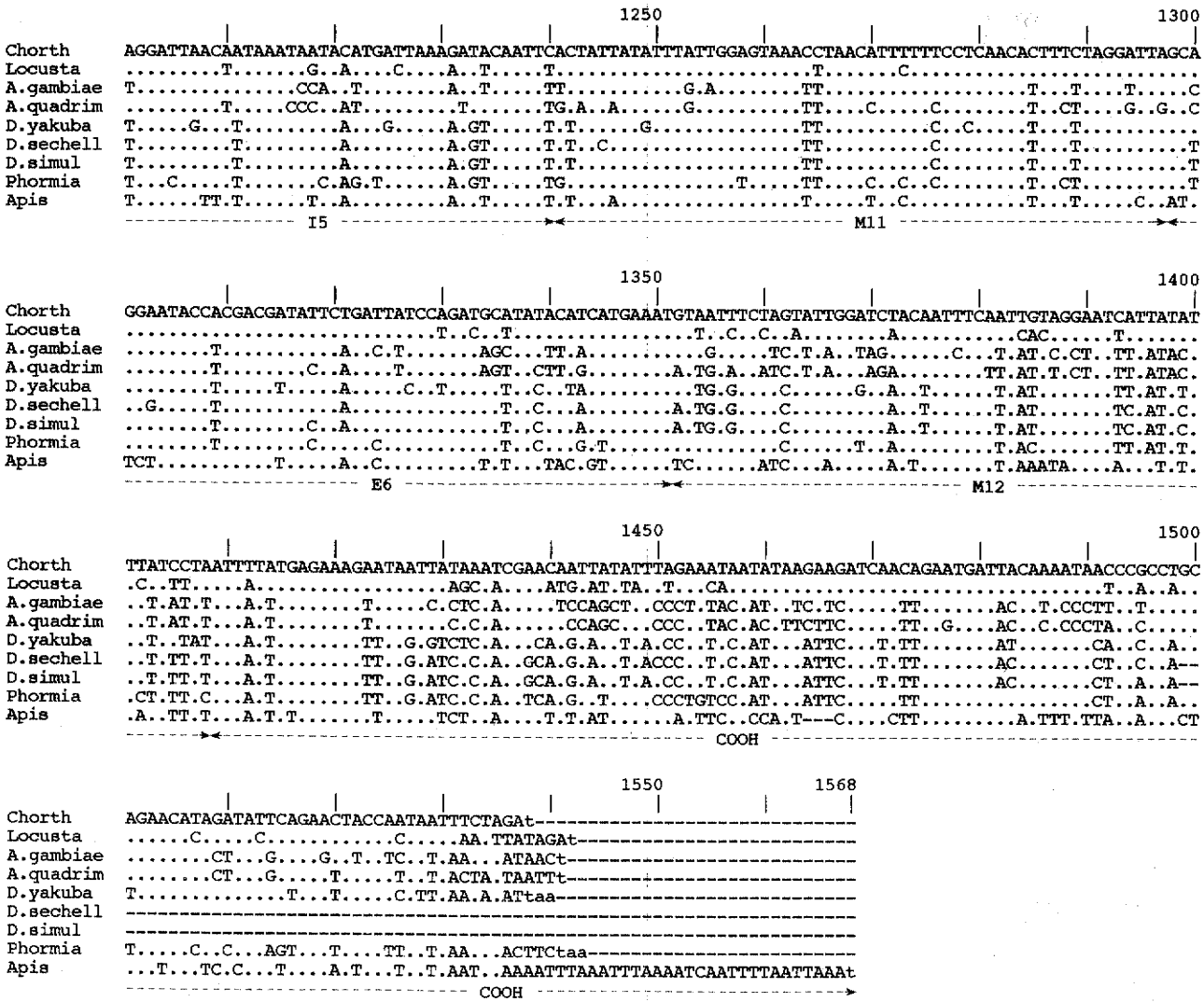

Figure 1. DNA alignment of the COI gene for nine species of insect. Identity to *C. parallelus* is indicated by a period, a deletion by a dash. (Putative) termination and initiation codons are displayed in lower case. The twenty-five structural regions are indicated (see text and Fig. 3 for details).

ster) and GTAA (*D. mauritiana*), indicating that differences in initiation codon can be present even between sibling species. The termination codons used for insect COI genes are shown in Table 1. Many organisms terminate with a single T, or TA, immediately adjacent to the tRNA gene, and it is known that complete (TAA) termination codons can be produced by post-transcriptional polyadenylation (Ojala *et al.*, 1981). *Drosophila* and *Phormia*, however, show the complete TAA termination codon common in many other mitochondrial genes.

Translation of this sequence with the invertebrate mitochondrial genetic code (Clary & Wolstenholme, 1985) gives a protein sequence with a mixture of residues conserved across all the studied species and residue positions of differing levels of variability [see below and Fig. 2. Note that the first amino acid residues for all COI proteins discussed here are defined as methionines regardless of the initiation codons, as suggested by Wolstenholme (1992b)]. No insertion/deletion events are apparent between *C. parallelus* and eight of the nine other insects. *Apis*, however, shows a deletion of 3 bp at position 1464 (Fig. 1). The position indicated here differs by three codons from that given by Crozier & Crozier (1993) but seems to give a better overall alignment. This deletion falls within the COOH-terminal region and may not be constrained with respect to either size, structure or amino acid function to the same extent as would a deletion of non-terminal residues. These observations agree with the expectation of conservation of size and structure in functionally constrained systems. Figure 3

Table 1. Summary of data concerning the COI gene for nine species of insect. The sequences for *D. simulans* and *D. sechellia* terminate prematurely.

Ref	Organism	Order	Init codon	Term codon	Length (bp)	% AT	Accession number
1	*C. parallelus*	Orthoptera	TCG	T	1537	69.4	
2	*Locusta migratoria*	Orthoptera	ATTA	T	1542	69.1	X80245
3	*Anopheles gambiae*	Diptera	TCG	T	1537	68.6	L20934
4	*A.quadrimaculatis*	Diptera	TCG	T	1537	68.1	L04272
5	*Drosophila yakuba*	Diptera	ATAA	TAA	1540	69.9	X03240
6	*D. sechellia*	Diptera	ATAA	-	(1498)	70.2	M57908
7	*D. simulans*	Diptera	ATAA	-	(1498)	70.7	M57911
8	*Phormia regina*	Diptera	TCG	TAA	1539	68.3	L14946
9	*Apis mellifera*	Hymenoptera	ATA	T	1561	75.9	L06178

References: (1) This paper; (2) Flook *et al.*, 1996; (3) Beard *et al.*, 1993; (4) Mitchell *et al.*, 1993; (5) Clary & Wolstenholme, 1985; (6 and 7) Satta & Takahata, 1990; (8) Sperling *et al.*, EMBL database access L14946 (unpublished); (9) Crozier & Crozier, 1993.

shows a two-dimensional structural model of the COI protein, with functionally essential (boxed) and variable (filled circles) residues in insects highlighted, respectively. It is clear that the distribution of variable residues is not random along the molecule (see below for further discussion).

Mode and tempo of evolution of the insect COI gene

The COI amino acid sequence was divided into twenty-five regions comprising five structural classes [twelve transmembrane helices (M1–M12), six external loops (E1–E6), five internal loops (I1-I5), carboxy (COOH) and amino (NH2) terminals] as shown in Figs 2 and 3. In order to test the null hypothesis that there is no difference between the average amino acid variability, per site, between the five structural classes, a Kruskal-Wallis analysis was employed which led us to reject this hypothesis with an associated probability level of <0.0001. When the average variability per residue site was calculated for the five structural classes the COOH terminal was found to be significantly more variable than any other region ($\leqslant 5\%$ significance level). The observed mean levels of variability were not significantly different between the amino terminal, internal loops, external loops or transmembrane regions (Table 2). This difference reflects the highly variable nature of COOH terminal amino acid sequences and agrees with Liu & Bechenbach (1992) who report that for an alignment of the insect COII gene the COOH terminal also appears to be the most variable region.

Pooling the data into classes in this way however will lose much information if there are large differences within these classes. Figures 4 and 5 show the mean variability per region (individual loops, transmembrane stretches or terminal regions) and there can be seen to be large differences in the mean variability of different regions of the same structural class.

Transmembrane helices M2, M6, M7 and M10 provide the metal ligands to interact with the two haem groups and copper atom which are essential for the activity of COI (Gennis, 1992). These regions can be seen to account for four of the seven highly conserved transmembrane helices. The fifth of these conserved transmembrane helices is M8, which is suggested to be involved with the cytochrome oxidase proton-conduction channel. This region contains three polar residues (Thr-352, Thr-359 and Lys-362) which are completely conserved among all organisms so far studied, and which are thought to be essential for this translocation activity (Gennis, 1992). Transmembrane helices M5 and M11 are also very conserved.

External loops E3, E5, and especially E4, seem to be very conserved (Fig. 4). Although the functional role played by these interhelical loops is unclear, E5 is thought to lie very close to heme-A in the association to which Tyr-414 has been suggested to play an important role (Holm *et al.*, 1987; Gennis, 1992).

Table 2. Mean variability, per amino acid position, for the different structural classes of COI.

Region	Size (amino acids)	Standard deviation	SE mean	Mean variability
NH_2 terminal	13	0.947	0.263	1.692
Internal loops	103	0.908	0.089	1.689
External loops	120	0.788	0.072	1.467
Transmembrane	232	0.755	0.050	1.461
COOH terminal	30	1.306	0.239	2.500

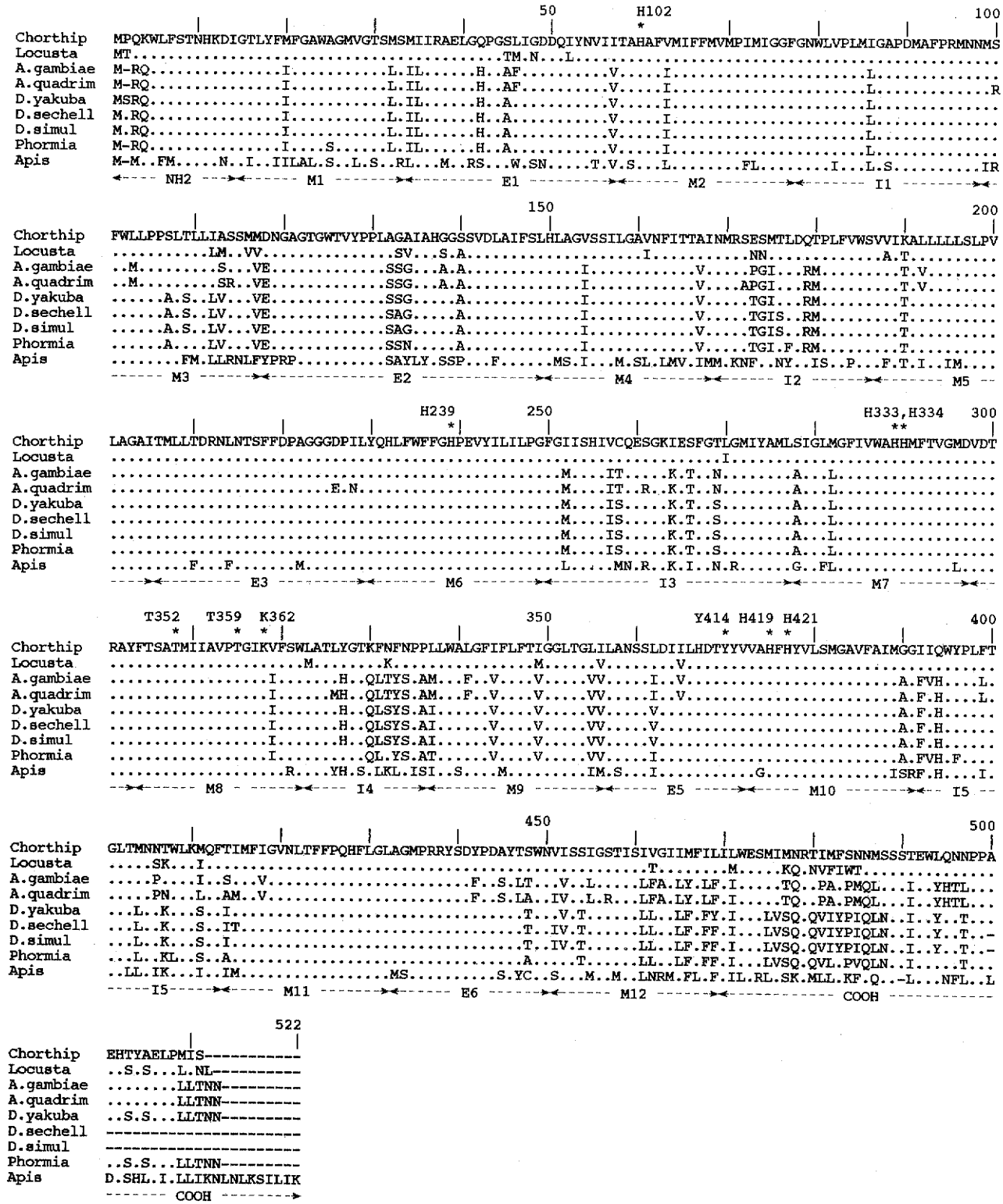

Figure 2. Alignment of COI amino acid sequences for nine species of insect. Identity to *C. parallelus* sequence is denoted by period, a deletion by a dash. Asterisks denote universally conserved residues, or those with functional significance discussed in the text. The twenty-five structural regions are indicated (see text and Fig. 3 for details). Note that the first amino acid residues for all COI proteins discussed here are defined as methionines regardless of the initiation codons, as suggested by Wolstenholme (1992).

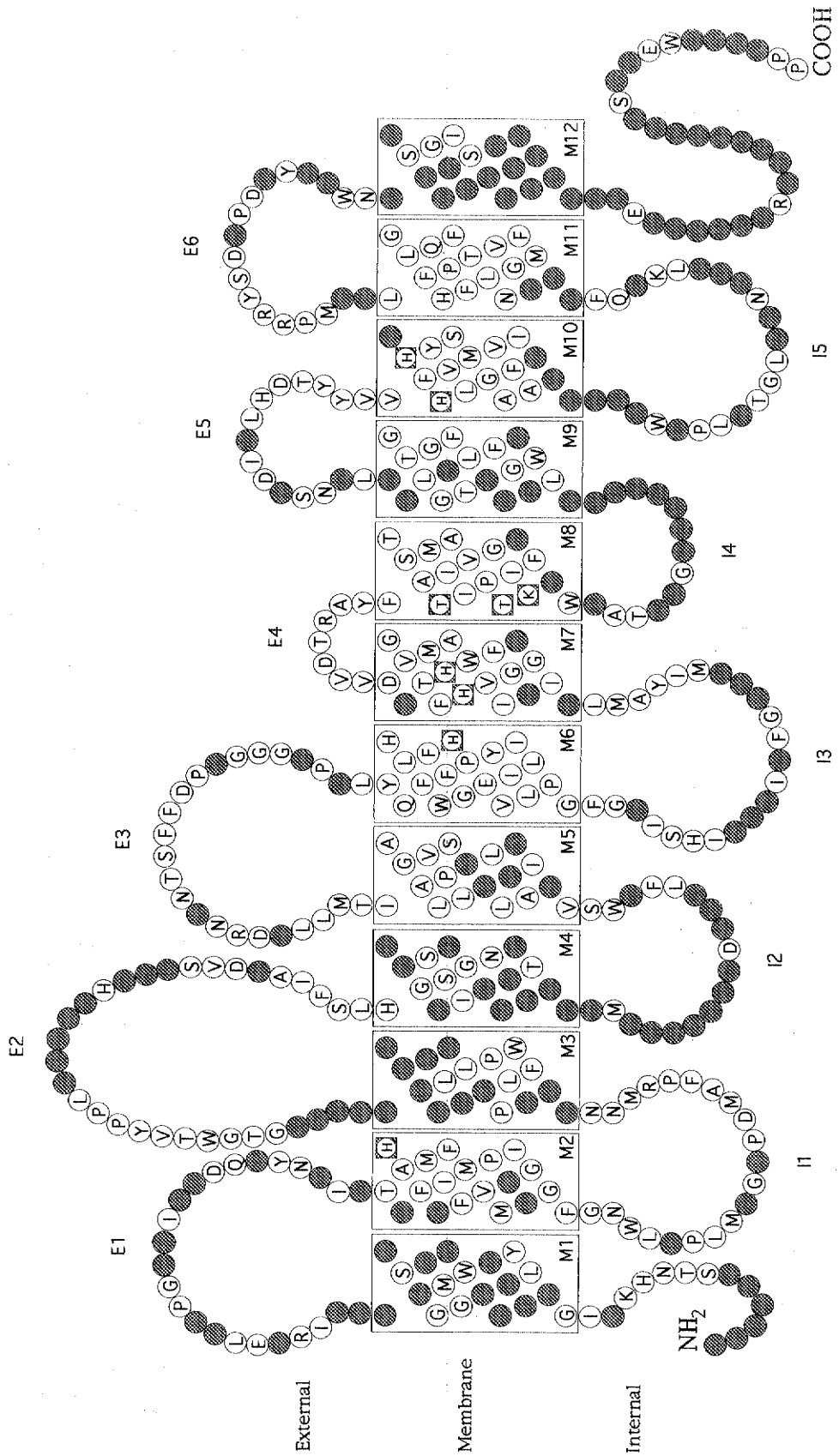

Figure 3. A two-dimensional model of the insect COI gene. The twenty-five structural regions, i.e. twelve membrane-spanning helices (M1–M12), six external loops (E1–E6), five internal loops (I1–I5), carboxy (COOH) and amino (NH_2) termini are shown. Boxed residues have special functional significance, and filled circles represent residues observed to be variable between insect species.

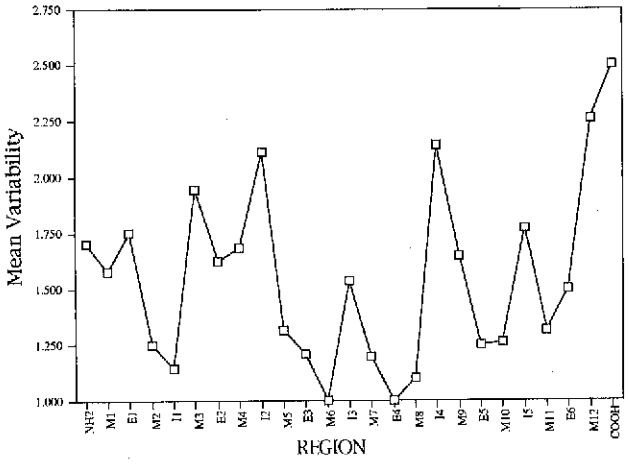

Figure 4. A graph showing the mean amino acid variability for the twenty-five structural regions of the insect COI gene. The variability is expressed as the average number of amino acids per site observed in a given region.

Internal loop 1 is highly conserved, in contrast to the other four internal loops which are relatively variable (Fig. 4). This pattern of conservation is also apparent in a similar alignment of twenty-two animal COI genes (unpublished data). A functional role for internal loop 1 is as yet unclear from current structural models of COI, though this data indicates that its function may be relatively important.

The foregoing discussion of protein variability begs the question: does amino acid variability reflect DNA variability? In this situation it probably does. The evidence for this assertion is twofold. Firstly, a certain correspondence between amino acid substitutions and nucleotide substitutions is expected despite of the degeneracy of the genetic code, since it is DNA sequence which determines amino acid sequence. (However, the form of this relationship does not follow a simple linear function, as the degenerate nature of genetic codes makes it possible for DNA variation to either match or outweigh protein variation, and it also falls under the influence of the phylogenetic distances of organisms compared.) Secondly, the relationship between patterns of amino acid variability and nucleotide variability can clearly be seen by comparing Figs 1 and 2, with the COOH terminal coding region being the most variable part, followed by regions coding for E1, M3, E2, I2, I4, M9 and M12 structural regions. Finally, there is empirical evidence for the level of protein variability being a good predictor of DNA variability. The distribution of polymorphic sites in the COI gene between *C. parallelus* individuals in a population survey (Lunt, 1994) has been observed to match very closely to the variability predictions made by this study. Furthermore, Howland & Hewitt (1995) have sequenced 400 bp of the COI gene in Coleopteran species with a variety of divergence levels. Their results show adjacent regions of DNA sequence conservation and variability, which correspond to the loop and transmembrane structures in exactly the same pattern as reported here.

These studies show that the expectations of shared patterns of DNA and protein variability are well founded, and that the regions identified in this study will almost certainly express the expected level of variability in most organisms. Given this pattern of conservation of sequence and structure, one is able to choose a level of variability to suit one's particular phylogenetic investigation.

Conserved primers and their potential for use in evolutionary studies

Primers which are conserved across many insect groups would be extremely useful in many types of evolutionary studies. The variability observed between insect COI sequences has been quantified by position and conserved primers designed to exploit

Table 3. Details of the ten conserved primers designed to cover the whole insect COI gene. Positions given are those in Fig. 1. Primers UEA1 and UEA10 are found in the tRNA-Tyr and tRNA-Leu genes respectively. Their positions relative to the COI gene (in the *D. yakuba* sequence; Clary & Wolstenholme, 1985) are 56 bp (Tyr) and +8 bp (Leu).

Primer	Strand	Position (3' base)	Length (bp)	Sequence
UEA1	Sense	tRNATyr	26	GAATAATTCCCATAAATAGATTTACA
UEA2	Antisense	375	26	TCAAGATAAAGGAGGATAAACAGTTC
UEA3	Sense	294	24	TATAGCATTCCCACGAATAAATAA
UEA4	Antisense	618	24	AATTTCGGTCAGTTAATAATATAG
UEA5	Sense	621	24	AGTTTTAGCAGGAGCAATTACTAT
UEA6	Antisense	966	29	TTAATWCCWGTWGGNACNGCAATRATTAT
UEA7	Sense	900	24	TACAGTTGGAATAGACGTTGATAC
UEA8	Antisense	1266	24	AAAAATGTTGAGGGAAAAATGTTA
UEA9	Sense	1284	26	GTAAACCTAACATTTTTTCCTCAACA
UEA10	Antisense	tRNALeu	25	TCCAATGCACTAATCTGCCATATTA

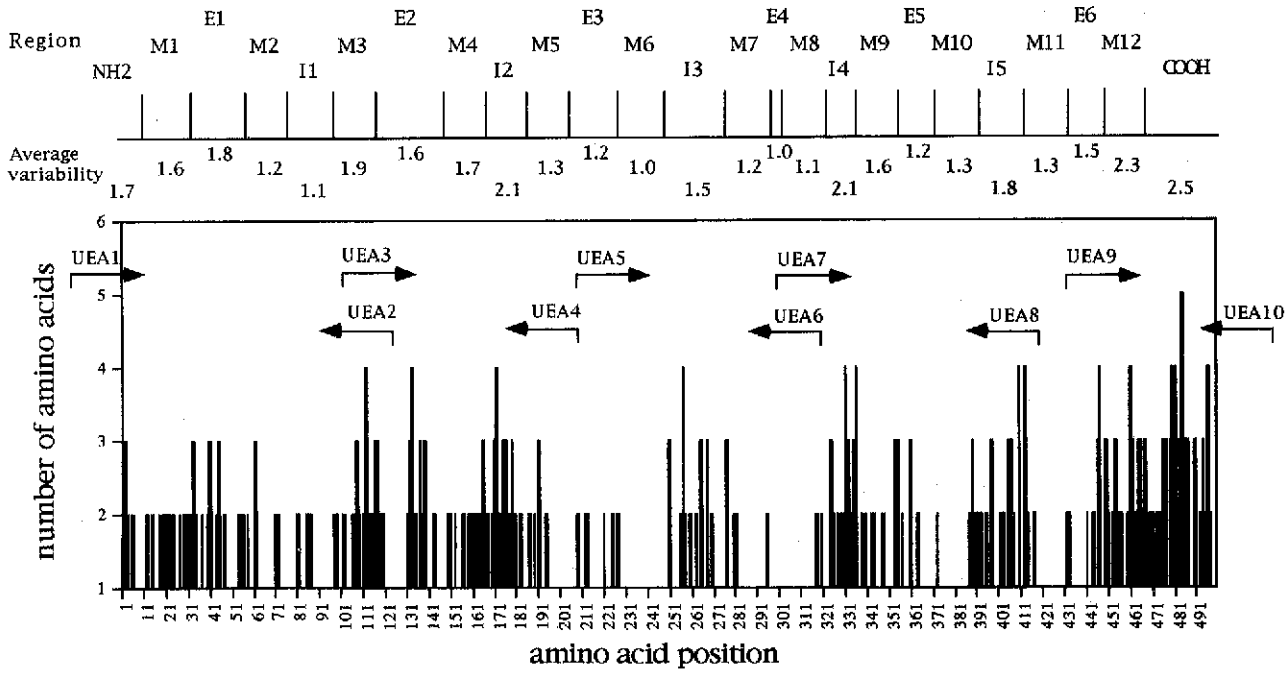

Figure 5. An illustration of the distribution of amino acid variability (number of different amino acids per residue site) along the insect COI gene. The twenty-five structural regions, and their mean levels of variability, are shown above the graph. The position of the ten conserved primers are given by arrows, primers UEA1 and UEA10 are in the tRNA-Tyr and tRNA-Leu genes respectively.

areas of differing variability. From the multiple alignment of the DNA sequences of nine insect COI genes presented in Fig. 1, it can be seen that several areas are very generally conserved and the same, or slightly modified, primers may be applicable to many organisms. Primers were thus designed to cover the whole of the insect COI gene and to be positioned to aid the sequencing of regions of different variability. The primers are described in Table 3 and their positions shown in Fig. 5. Primer UEA10, located in the tRNA-Leu gene, although identified independently in this study, turns out to be the general insect primer 'PAT' designed in the lab of R. Harrison (Cornell University, pers. comm.).

To test the universality of the conserved primers identified in this study, PCR amplifications have been carried out for nine insect taxa. These taxa cover the main divisions of the class Insecta, from wingless insects of the order Thysanura (silverfish, *Lepisma saccharina*, and firebrat, *Thermobia domestica*), to winged insects of the orders Odonata (damselfly, *Calopteryx splendens*), Orthoptera (desert locust, *Schistocerca gregaria* and meadow grasshopper *Chorthippus parallelus*), Hemiptera (pea aphid, *Acyrthosiphon pisum*), Coleoptera (beetle, *Carabus vidaceous*), Diptera (fruit fly, *Drosophila melanogaster*) and Hymenoptera (bumble bee, *Bombus lapidarius*). Figure 6 shows the amplification product obtained using the primer pair UEA3–UEA8. These primers cover a large part (1018 bp) of the COI gene from the I1 region to M11 region (see Figs 3 and 5). A single main band of the expected size was produced in all insect taxa described above with the exception of the pea aphid, *A. pisum* (no product under the assay conditions, although other primers amplify the COI gene in this aphid). Direct terminal sequence analysis of the amplified PCR fragments confirms that they are the COI gene and that all sequences are taxa-specific, thus excluding the possibility of cross-contamination (terminal sequences of the PCR fragments for these species are available from the authors). A preliminary phylogenetic analysis of these sequences further confirmed the authenticity of the amplified bands. PCR amplifications with other primer combinations indicate that most of the primers described here work well in many of these different taxa. Degenerate forms of some of these primers have also been designed based on the sequence alignment in Fig. 1. We are currently investigating further the universality of all these primers and results of this analysis will be reported later. Initial results indicate that the primers identified in this study are quite broadly conserved, with primers UEA7, UEA9 and UEA10 working well even between the superclasses Insecta and Arach-

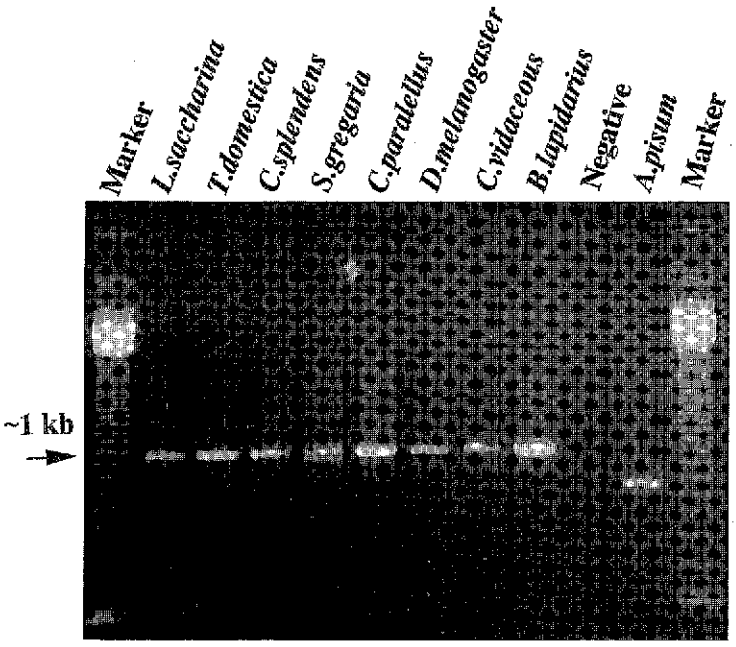

Figure 6. PCR products amplified from nine different arthropod taxa using primer pairs UEA3–UEA8 or UEA7–UEA10. A specific band of ~1 kb was amplified from all taxa except the pea aphid, *A. pisum*, using the primer pair UEA3–UEA8. The ~0.7 kb PCR product for *A. pisum* showed in the photograph was obtained using primer pair UEA7–UEA10. The marker is a 123 bp size ladder (GIBCO BRL).

nida, which are thought to have diverged during the Devonian period at least 400 million years ago (Pearse *et al.*, 1987).

As a general guide to the applicability of these primers, regions amplified by primer combinations UEA5–UEA6, UEA5–UEA8 or UEA7–UEA8 could be suitable for higher-level evolutionary studies (e.g. at genus or family level); regions amplified with primer pairs UEA3–UEA4 or UEA7–UEA10 should be more variable and thus suitable for lower-level analyses (such as study of intraspecific variation, and the phylogenetics of closely related species). The use of the last two primer pairs (UEA3–UEA4 and UEA7–UEA10) in population analysis of beetles, grasshoppers and other insects in our laboratory suggests that these regions are probably variable enough for revealing intraspecific polymorphisms (unpublished data, G.M.H. *et al.*).

In this report we show that a detailed examination of the evolutionary patterns of a DNA region could provide valuable guidelines for its effective use as a molecular marker in phylogenetic studies. While this paper was in review, Simon *et al.* (1994) published a most useful compendium of conserved primers covering the whole insect mitochondrial genome. Increasing practices employing such primers will certainly reveal more information on the evolutionary patterns of other mitochondrial genes, which will in turn help us to assess their usefulness for addressing phylogenetic questions of different taxonomic levels.

Experimental procedures

Insect COI gene sequences

The sequence of the *C. parallelus* COI gene was obtained as part of the complete sequence of a 6.4 kb mtDNA *Hin*dIII fragment as described in Zhang *et al.* (1995). Both strands of the COI gene have been sequenced. Other insect DNA sequences were taken from the GenBank and EMBL databases, with the exception of the *Locusta migratoria* sequence which was kindly supplied by P. Flook prior to publication. The accession numbers of these sequences are L20934 (*Anopheles gamblae*), L04272 (*Anopheles quadrimaculatis*), L14946 (*Phormia regina*), L06178 (*Apis mellifera*), X03240 (*Drosophila yakuba*), M57908 (*Drosophila sechellia*), M57911 (*Drosophila simulans*) and X80245 (*Locusta migratoria*) (see Table 1 for individual references). All sequences were aligned by the Clustal V method (Higgins & Sharp, 1989) and translated with the Invertebrate mitochondrial genetic code using programs implanted in the LASERGENE computer software package (DNASTAR). The aligned DNA sequences were then examined manually using amino acid sequences and codon positions as references, producing the alignment shown in Fig. 1 which seems to be quite robust.

Analysis of variability and statistical tests

The aligned amino acid sequence was divided into twenty-five regions comprising five structural classes (twelve transmembrane helices, six external loops, five internal loops, carboxy and amino terminals; Fig. 3). The points of transition between these regions were taken from Gennis (1992). The number of different amino acids observed at each position of the protein alignment was recorded and the variability level expressed as the average number of amino acids per site observed in a

given region. A spreadsheet written (by J. B. Lunt and D. H. Lunt) to score the variability across such alignments is available from the authors. This analysis was limited to the 498 homologous positions between the ends of the shortest sequence. Insertion and deletion events were scored equally to the possession of a novel amino acid.

Statistical tests were carried out using the StatView SE v1.03 (Abacus Concepts Inc.) software package. A Kruskal-Wallis test (analysis of variance by ranks) was performed on the data sets to test the null hypothesis (H_o): there is no difference between the average amino acid variability, per site, between the five structural classes. In the event of this *Ho* being rejected, an analysis would be performed to determine between which of the samples significant differences occur. This analysis followed the method described by Zar (1984).

Conserved primers, PCR amplification and direct sequencing

Primers for amplifying and sequencing the whole of the COI gene were designed using the Oligo 4.0 (National Biosciences Inc.) software package following the guidelines given by Rychlik (1992). Nine insect taxa, which cover the main divisions of the superclass Insecta, were used to test the primers identified in this study, viz: *Lepisma saccharina* (silverfish, Apterygota, order Thysanura), *Thermobia domestica* (firebrat, Apterygota, order Thysanura), *Calopteryx splendens* (damselfly, Pterygota, order Odonata), *Schistocerca gregaria* (locust, Pterygota, order Orthoptera), *Chorthippus parallelus* (grasshopper, Pterygota, order Orthoptera), *Acyrthosiphon pisum* (aphid, Pterygota, order Homoptera), *Drosophila melanogaster* (fruitfly, Pterygota, order Diptera), *Carabus vidaceous* (beetle, Pterygota, order Coleoptera) and *Bombus lapidarius* (bee, Pterygota, order Hymenoptera). An arachnid (spider, *Tegenaria domesticus*) was also included to gauge the broader generality of these primers.

DNA was purified from individual insects using a phenol/chloroform based extraction as described by Zhang *et al.* (1995). PCR was carried out in a 50 μl reaction containing 1.5–2.0 mM $MgCl_2$, 200 μM dNTP, 0.15 μM of each primer, and 2 units of *Taq* polymerase (Promega) in 1 x reaction buffer (50 mM KCl, 10 mM Tris-HCl, 0.1% Triton X-100, pH 9.0 at 25°C, Promega). Following an initial denaturation at 94°C for 5 min, thirty to forty cycles were performed in a DNA Thermal Cycler 480 (Perkin Elmer Cetus), each consisting of melting at 95°C for 40 s, annealing at 48–55°C for 1 min, and extension at 72°C for 40 s to 1 min 40 s. 5 μl of PCR product were used for direct DNA sequencing using the Sequenase PCR Products Sequencing Kit (USB-Amersham) following the manufacturer's protocol.

Acknowledgements

We acknowledge the help of Pamy Nöldner, Marcus Rowcliffe, Nick Watmough, John Noble-Nesbit, Graham Hopkins and John Lunt. We also thank three anonymous reviewers for their valuable comments. We are grateful to Paul Flook for providing us with the *Locusta* COI sequence prior to publication. This work was supported by grants from the S.E.R.C. and E.U.

References

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel J.E., Reeb, C.A. and Saunders, N.C. (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annu Rev Ecol Syst* **18:** 489–522.

Ballard, J.W., Olsen, G.J., Faith, D.P., Odgers, W.A., Rowell, D.M. and Atkinson, P.W. (1992) Evidence from 12s ribosomal RNA sequences that Onychophorans are modified arthropods. *Science* **258:** 1345–1348.

Beard, C.B., Hamm, D.M. and Collins, F.H. (1993) The mitochondrial genome of the mosquito *Anopheles gambiae*: DNA sequence, genome organization, and comparisons with mitochondrial sequences of other insects. *Insect Mol Biol* **2:** 103–124.

Clary, D.O. and Wolstenholme, D.R. (1985) The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *J Mol Evol* **22:** 252–271.

Crozier, R.H. and Crozier, Y.C. (1993) The mitochondrial genome of the Honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics* **133:** 97–117.

Flook, P.K., Rowell, C.H.F. and Gellissen G. (1996) The sequence, organisation and evolution of the *Locusta migratoria* mitochondrial genome. *J Mol Evol*, in press.

Gennis, R.B. (1992) Site-directed mutagenesis studies on subunit I of the aa3-type cytochrome c oxidase of *Rhodobacter sphaeroides*: a brief review of progress to date. *Biochim Biophys Acta* **1101:** 184–187.

Harrison, R.G. (1989) Animal mtDNA as a genetic marker in population and evolutionary biology. *Trends Ecol Evol* **4:** 6–11.

Hauke, H.-R. and Gellisson, G. (1988) Different mitochondrial gene orders among insects: exchanged tRNA gene positions in the COII/COIII region between an orthopteran and dipteran species. *Curr Genetics* **14:** 471–476.

Higgins, D.G. and Sharp, P.M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *Computer Applic Biosci* **5:** 151–153.

Holm, L., Saraste, M. and Wilkstrom, M. (1987) Structural models of the redox centers in cytochrome-oxidase. *EMBO J* **6:** 2819–2823.

Howland, D.E. and Hewitt, G.M. (1995) Phylogeny of the Coleoptera based on mitochondrial cytochrome oxidase I sequence data. *Insect Mol Biol* **4:** 203–215.

Liu, H. and Beckenbach, A.T. (1992) Evolution of the mitochondrial cytochrome oxidase II gene among ten orders of insects. *Mol Phylogen Evol* **1:** 41–52.

Lunt, D.H. (1994) MtDNA differentiation across Europe in the meadow grasshopper *Chorthippus parallelus* (Orthoptera: Acrididae). Ph.D. thesis, University of East Anglia, Norwich.

Moritz, C., Dowling, T.E. and Brown, W.M. (1987) Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annu Rev Ecol Syst* **18:** 269–292.

Ojala, D., Montoya, J. and Attardi, G. (1981) tRNA punctuation model of RNA processing in human mitochondria. *Nature* **290:** 470–474.

Pääbo, S., Thomas, W.K., Whitfield, K.M., Kumazawa, Y. and Wilson, A.C. (1991) Rearrangements of mitochondrial transfer-RNA genes in marsupials. *J Mol Evol* **33:** 426–430.

Pearse, V., Pearse, J., Buchsbaum, M. and Buchsbaum, R. (1987) *Living Invertebrates*. The Boxwood Press, Pacific Grove, California.

Rychlik, W. (1992) *Oligo Version 4.0: Reference Manual*. National Biosciences, Inc., Plymouth, Minnesota.

Saraste, M. (1990) Structural features of cytochrome oxidase. *Q Rev Biophys* **23:** 331–366.

Satta, Y., Ishiwa, H. and Chigusa, S.I. (1987) Analysis of nucleotide substitutions of mitochondrial DNAs in *Drosophila melanogaster* and its sibling species. *J Mol Biol* **4:** 638–650.

Simon, C. (1991) Molecular systematics at the species boundary: exploiting conserved and variable regions of the mitochondrial genome of animals via direct sequencing from amplified DNA. *Molecular Techniques in Taxonomy* (Hewitt, G.M., Johnston, A.W.B., Young, J.P.W., eds), pp. 33–71. Springer, Berlin.

Simon, C., Frati, F., Beckenbach, A., Crespi, B., Liu, H. and Flook, P. (1994) Evolution, weighting, and phylogenetic utility of mitochondrial gene-sequences and a compilation of conserved polymerase chain-reaction primers. *Ann Ent Soc Amer* **87:** 651–701.

Smith, M.J., Arndt, A., Gorski, S. and Fajber, E. (1993) The phylogeny of echinoderm classes based on mitochondrial gene arrangements. *J Mol Evol* **36:** 545–554.

Wolstenholme, D.R. (1992a) Animal mitochondrial DNA: structure and evolution. *Int Rev Cytol* **141:** 173–216.

Wolstenholme, D.R. (1992b) Genetic novelties in mitochondrial genomes of multicellular animals. *Curr Op Genet Devel* **2:** 918–925.

Zar, J.H. (1984) *Biostatistical Analysis*. Prentice-Hall, London.

Zhang, D.-X., Szymura, J.M. and Hewitt, G.M. (1995) Evolution and structural conservation of the control region of insect mitochondrial DNA. *J Mol Evol* **40:** 382–391.