U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

PhD

Evolutionary Genomics of Genes Involved in the Environmental Adaptation of Metazoans

Daniela Martins Almeida

PhD Thesis presented to the
Faculty of Sciences of the University of Porto

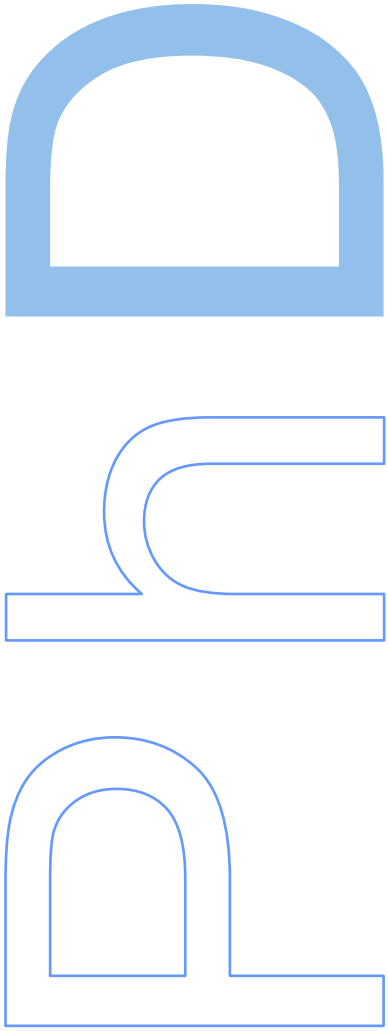
Biology
2017

Evolutionary Genomics of Genes Involved in the Environmental Adaptation of Metazoans

Daniela Martins Almeida
Biology
Faculty of Sciences
2017

Supervisor
Agostinho Antunes, Assistant Professor
Faculty of Sciences of the University of Porto

Co-supervisor
Vitor Vasconcelos, Full Professor
Faculty of Sciences of the University of Porto



Dedicated to my parents and brother.

Acknowledgements

Firstly I would like to acknowledge my supervisor, Professor Doctor Agostinho Antunes, whose insightful discussions, guidance, and encouragement proved to be fundamental contributions for the research of this thesis.

I am deeply indebted to Professor Doctor Vitor Vasconcelos for accepting to be my co-supervisor, and giving me such a warm welcome to his laboratory, always being available, and promptly providing support to my research.

My sincere gratitude to the Fundação para a Ciência e a Tecnologia (FCT) for funding the work presented in this thesis through the PhD grant SFRH/BD/79766/2011. This research was also partially supported by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT and the European Regional Development Fund in the framework of the program PT2020, by the European Structural and Investment Funds (ESIF) through the Competitiveness and Internationalization Operational Program - COMPETE 2020 and by National Funds through the FCT under the project PTDC/AAG-GLO/6887/2014 (POCI-01-0124-FEDER-016845).

I would like to thank all co-authors of each chapter for being such good examples of persistence and hard-work. It has been an honor to work with everyone at the beginning of my journey through the fantastic world of evolutionary biology.

I owe special thanks to Cameron Weadick, Romain Studer and Hidenori Nishihara researchers who were kindly available to discuss several concepts and approaches of molecular evolution, contributing to the improvement of the statistical analyses performed in this thesis and to my own progress of learning and research development.

I would like to thank Tibisay Escalona for her amazing friendship, constant mentoring and crucial scientific input, setting an extraordinary example for me on how to think about evolution. Thank you, Tiby, for always being there, specially during the writing of the thesis, and for being such a great person and inspiration.

I would like to thank my lab mates for an incredible integration and assistance during different phases of my work. In special to Emanuel Maldonado, Liliana Silva, Aldo Barreiro, João Paulo Machado, Anoop Alex, Dany Pérez, Cidália Gomes, Cristiana Moreira, Filipe Silva, Rui Borges, Siby Philip, Joana Pereira and Susana Pereira for their helpful discussions, support and motivation words that contributed a lot to reaching this stage in my scientific

formation.

To my friends that were a constant presence in all the moments, and that in one way or another walked with me, motivating and helping me to overcome this stage, that laughed and shared unforgettable moments with me. Specially, Rosalina Goulão, Ana Margarida Dantas, Joana Ribeiro, Rita Pereira, Catarina Remédios, Catarina Roque, Elis Costa and Joelma Dantas, thank you all so much. I know that all of you will be very happy to know that I have finally written my thesis.

Aos meus queridos pais e irmão, sem vocês atingir esta meta não teria sido possível. Foi uma longa caminhada que me fez crescer a vários níveis e vocês foram sempre os pilares que me guiaram, encorajaram, inspiraram, ampararam e deram força para nunca desistir de nenhum desafio, transmitindo-me todo o vosso amor e relembrando-me que todo "o trabalho tem a sua recompensa, que o conhecimento vale sempre a pena, por mais dificuldades que se deparem no nosso caminho". Muito obrigada pela vossa inesgotável paciência, por acreditarem em mim e por me ajudaram sempre a ver a luz que me mantém no caminho certo, a vossa força, coragem e determinação perante a vida.



Resumo

Na biologia evolutiva, é reconhecido que condições ambientais bióticas e abióticas distintas impõem às espécies pressões seletivas. A sobrevivência dos organismos nessas condições requer evolução adaptativa através da manutenção de mutações vantajosas por seleção natural. A identificação de regiões especificas no genoma dos organismos com assinaturas de seleção poderá contribuir para o conhecimento da história evolutiva das espécies e revelar alterações funcionais importantes em genes relevantes codificadores de proteínas. Os metazoários são um grupo ideal para se realizarem estudos de evolução molecular, uma vez que desenvolveram diversas especializações morfológicas, fisiológicas, comportamentais e ecológicas, e invadiram vários *habitats*, nomeadamente ecossistemas marinhos, de água doce e terrestre.

O objetivo principal desta tese é estudar a evolução adaptativa de genes codificadores de proteínas de dois grupos distintos de metazoários (invertebrados da Classe Cephalopoda – lulas, chocos, polvos e *Nautilus*, e vertebrados da Classe Aves – pássaros e aves) para os quais a seleção adaptativa a nível molecular é pouco conhecida, apesar da sua elevada diversidade e adaptações a ambientes extremos (temperaturas, pressões e profundidades distintas, etc.). Este foi conseguido através da utilização de diferentes modelos de evolução molecular, baseados na análise de codões por *Maximum likelihood*, que estimam a razão das taxas de substituições não-sinónimas em função das sinónimas (ω). Este parâmetro permite inferir a forma e a grandeza da pressão seletiva sobre os genes codificadores de proteínas: $0 < \omega < 1$ indica seleção negativa; $\omega = 1$ indica que as mutações são neutrais; e $\omega > 1$ é indicativo de seleção positiva ou fixação de mutações vantajosas.

A Parte I desta tese (**capítulos 2 e 3**) centra-se na análise de genomas mitocondriais completamente sequenciados da Classe Cephalopoda (metazoários invertebrados). No **capítulo 2**, foram inicialmente consideradas 17 espécies de cefalópodes, e no **capítulo 3** alargou-se a análise dos genomas mitocondriais a um total de 37 espécies. Aqui, são explorados dois grandes objetivos. Em primeiro lugar, a filogenia mitocondrial da Classe Cephalopoda (determinada com sucesso e concordante com estudos prévios) e, em segundo lugar, o teste de evidências de seleção adaptativa nos 13 genes mitocondriais que codificam proteínas (genes mitocondriais: sete desidrogenases NADH, *nd1–nd6* e *nd4l*, para o Complexo I; citocromo b, *cytb*, para o Complexo III; três citocromos c oxidases,

cox1–cox3, para o Complexo IV; e duas ATPases, *atp6* e *atp8*, para o Complexo V) para o sistema respiratório de fosforilação oxidativa (OXPHOS). Um ponto importante deste estudo foi a descoberta de seleção positiva na linhagem Octopodiformes, cefalópodes com oito braços (**capítulo 2**), ao longo das subunidades ND5 (Complexo I), CYTB (Complexo III), COX2–3 (Complexo IV) e ATP8 (Complexo V); e nas linhagens Oegopsida e Bathyteuthoidea, cefalópodes com genes mitocondriais duplicados (**capítulo 3**), ao longo das subunidades COX2 (Complexo IV) e ATP6 (Complexo V). Verificou-se que os sítios selecionados positivamente se localizam nos canais de translocação de protões ou estão a mediar interações entre as subunidades, podendo influenciar a atividade enzimática dos complexos OXPHOS e consequentemente a produção de trifosfato de adenosina. Essas mudanças podem assim representar adaptações moleculares importantes para suprir as necessidades bioenergéticos dos cefalópodes em diferentes ambientes marinhos.

A Parte II desta tese (**capítulo 4**) centra-se na análise de genomas nucleares completamente sequenciados da Classe Aves (metazoários vertebrados). Aqui, o principal objetivo é explorar evidências de seleção adaptativa numa família de genes de destoxificação (isto é, CYP2, citocromo P450 família 2), a qual é importante numa variedade de processos fisiológicos e toxicológicos em espécies de aves, sendo assim crucial para a sua sobrevivência. No total foram identificadas 12 subfamílias de CYP2 em 48 espécies, representativas dos principais clados de aves existentes, incluindo a CYP2F, CYP2G e vários genes da CYP2AF, anteriormente não relatados nesta Classe. Adicionalmente, as CYP2K e CYP2W foram encontradas em todos os grupos de aves, não sendo específicas de determinadas linhagens, como se pensava anteriormente. As principais descobertas deste estudo foram o reconhecimento de sítios selecionados positivamente nas subfamílias CYP2C, CYP2D, CYP2H, CYP2J, CYP2K e CYP2AC, e de várias restrições seletivas em algumas subfamílias de CYP2 associadas a diferentes hábitos alimentares (CYP2H e CYP2J), *habitats* (CYP2D, CYP2H, CYP2J e CYP2K) e comportamentos migratórios (CYP2D, CYP2H e CYP2J). Em geral, estes sítios encontrados na zona de ligação ao grupo prostético *heme* e nos sítios de reconhecimento de substratos da CYP2 podem ter ajudado as espécies de aves a metabolizar compostos químicos encontrados em novos ambientes, e facilitado a dispersão e sucesso evolutivo desta Classe.

O aumento de sequências genómicas disponibilizadas nas bases de dados *on-line*, constitui um desafio elevado para a tarefa árdua de analisar uma enorme quantidade de dados, um processo geralmente difícil, propenso a erros e bastante demorado, especialmente quando se analisam resultados provenientes de diferentes programas. A Parte III desta tese (**capítulo 5**) centra-se no desenvolvimento de uma ferramenta bioinformática (IMPACT_S) para ajudar a agilizar o processo de análise manual de resultados gerados por vários programas amplamente utilizados para estudar seleção (PAML, DATAMONKEY e TREESAAP). Aqui, o software é concebido para utilizar uma única interface gráfica permitindo a rápida e eficiente integração, e o mapeamento dos resíduos selecionados

positivamente na estrutura tridimensional da proteína correspondente. Isto proporciona aos utilizadores uma maneira fácil para relacionar as substituições de aminoácidos inferidas com o seu potencial efeito na estrutura e função das proteínas.

Considerados em conjunto, os resultados apresentados nesta tese fornecem evidências de seleção adaptativa em genes codificadores de proteínas das Classes Cephalopoda (mitocondriais) e Aves (nucleares). Por um lado, os resultados mostram que os desafios fisiológicos associados às adaptações ao ambiente marinho (morfológicas, metabólicas, ecológicas e genéticas) têm impulsionado a evolução do genoma mitocondrial da Classe Cephalopoda em certas linhagens, enquanto na Classe Aves, os resultados mostram que os desafios fisiológicos e toxicológicos associados às adaptações para a dispersão, dieta e o *habitat*, provavelmente têm impulsionado a evolução das enzimas nucleares CYP2, facilitando, em ambos os casos, o sucesso evolutivo destes taxa.

Finalmente, trabalhos futuros serão beneficiados por explorações experimentais do impacto de determinadas substituições amino acídicas na função e estabilidade de proteínas associadas a mecanismos metabólicos dos organismos, e compreendendo de que forma estas proteínas podem ser relevantes para a saúde humana.

Palavras-Chave: Evolução Adaptativa, Metazoários, Seleção Positiva, Genes Mitocondriais Codificadores de Proteínas, Citocromo P450 Família 2, Ferramenta de Bioinformática.

Abstract

In evolutionary biology, it is well recognized that environments with distinct biotic and abiotic conditions impose selective pressures to species. Moreover, the survival of organisms in such particular conditions requires adaptive evolution through the maintenance of advantageous mutations by natural selection. Therefore, identifying specific regions of the organisms' genome with signatures of selection can be insightful to understand the species evolutionary history and highlight important functional changes occurring in relevant protein-coding genes. Metazoans are an ideal group to perform molecular evolutionary studies, because they have numerous morphological, physiological, behavioral and ecological specializations and have invaded several habitats including marine, freshwater, and terrestrial ecosystems.

The prime aim of this thesis is to study the adaptive evolution of protein-coding genes from two contrasting metazoan groups (invertebrates of the Class Cephalopoda – squids, cuttlefishes, octopuses and *Nautilus*, and vertebrates of the Class Aves – birds) for which adaptive selection at the molecular level is poorly known, despite their high diversity and adaptations to extreme environments (distinct temperature, pressure, depth, etc.). This was accomplished by the use of different Maximum likelihood codon-based models of molecular evolution that estimate the non-synonymous to synonymous substitution rate ratio (ω). This parameter measures the form and strength of selective pressure on the protein-coding genes: $0 < \omega < 1$ indicates purifying selection; $\omega = 1$ indicates that mutations are neutral; and $\omega > 1$ is indicative of positive selection or fixation of advantageous mutations.

Part I of this thesis (**chapters 2 and 3**) focuses on the analyses of complete mitochondrial genomes of the Class Cephalopoda (metazoan invertebrates). In **chapter 2**, 17 species of cephalopods were initially considered, while in **chapter 3** the analyses of mitochondrial genomes were extended to a total of 37 species. Here, two major objectives are explored. First, the Cephalopoda mitochondrial phylogeny (successfully determined and in agreement with previous studies) and second, test for evidence of adaptive selection in the 13-mitochondrial protein-coding genes (mt genes: seven NADH dehydrogenases, *nd1–nd6* and *nd4l*, for Complex I; cytochrome b, *cytb*, for Complex III; three cytochrome c oxidases, *cox1–cox3*, for Complex IV; and two ATPases, *atp6* and *atp8*, for Complex V) for the oxidative phosphorylation (OXPHOS) respiratory system. An important finding of this study was the discovery of positive selection in the Octopodiformes lineage, cephalopods with eight arms

(**chapter 2**), along subunits ND5 (Complex I), CYTB (Complex III), COX2–3 (Complex IV) and ATP8 (Complex V); and in the Oegopsida and Bathyteuthoidea lineages, cephalopods with duplicated mt genes (**chapter 3**), along subunits COX2 (Complex IV) and ATP6 (Complex V). The positively selected sites were found to be lining the proton translocation channels or to be mediating interactions between subunits, and may possibly influence the enzymatic activity of the OXPHOS complexes and consequently the adenosine triphosphate production. Therefore, these changes may represent important molecular adaptations supplying the cephalopods' bioenergetics demand across different marine environments.

Part II of this thesis (**chapter 4**) focuses on the analyses of complete nuclear genomes of the Class Aves (metazoan vertebrates). Here, the major objective is to explore evidence of adaptive selection in a detoxification gene family (*i.e.,* CYP2, cytochrome P450 family 2), which is important in a variety of physiological and toxicological processes in the avian species, thereby crucial for their survival. A total of 12 CYP2 subfamilies were identified across 48 species, representing all major extant avian clades, including the CYP2F, CYP2G, and several CYP2AF genes, previously not reported in this Class. Additionally, CYP2K and CYP2W were detected in all the avian groups instead of being lineage-specific, as previously thought. Important findings of this study were the recognition of positively selected sites in the CYP2C, CYP2D, CYP2H, CYP2J, CYP2K, and CYP2AC subfamilies, and selective constraint variations in some CYP2 subfamilies associated with distinct feeding habits (CYP2H and CYP2J), habitats (CYP2D, CYP2H, CYP2J, and CYP2K), and migratory behaviors (CYP2D, CYP2H, and CYP2J). Overall, these sites located in the heme binding area and substrate recognition sites of CYP2, possibly assist avian species to metabolize chemical compounds found in new environments, and potentially facilitated the dispersion and evolutionary success of this Class.

With the increase of genomic sequences available in the online databases, biologists face the problem of analyzing huge amount of data, which will make the overall process hard and tedious to handle, error-prone and time-consuming, especially when using different programs. Therefore, Part III of this thesis (**chapter 5**) focuses on the development of a bioinformatics tool (IMPACT_S) to expedite the manual process of analyzing results from several programs, currently widely used to assess selection (PAML, DATAMONKEY and TREESAAP). Here, the software is designed to use a single graphical interface allowing for fast, efficient integration and mapping of the positively selected residues into the corresponding three-dimensional protein structure. This provides users with an easy way to relate among the inferred amino acid substitutions and their potential effect in the structure and function of proteins.

The results presented in this thesis provide evidence of adaptive selection in protein-coding genes of Classes Cephalopoda (mitochondrial) and Aves (nuclear). On one hand, the results show that physiological challenges associated with adaptations to the oceanic environment (morphological, metabolic, ecological and genetics) have driven the evolution of Cephalopoda mitochondrial genome in a lineage specific manner, whereas in Aves, the

results show that physiological and toxicological challenges associated with adaptations to dispersion, diet, and habitat, likely have driven the evolution of avian CYP2 nuclear enzymes, and in both cases facilitated the evolutionary success of these taxa.

Finally, future work will benefit by experimentally exploring the impact of specific amino acid substitutions in the function and stability of proteins associated with organisms' metabolic pathways, and understanding how these proteins may be relevant to human health.

Keywords: Adaptive Evolution, Metazoans, Positive Selection, Mitochondrial Protein-Coding Genes, Cytochrome P450 Family 2, Bioinformatics Tool.

Contents

Acknowledgements **v**

Resumo **ix**

Abstract **xiii**

List of Tables **xxi**

List of Figures **xxiv**

List of Abbreviations **xxvii**

1 General Introduction **1**

1.1 Background . . . 3

1.1.1 Fundamental concepts of molecular evolution . . . 4

1.1.2 Detecting molecular adaptive evolution . . . 8

1.1.2.1 Bioinformatics stages for selection analyses . . . 8

1.1.2.2 How molecular studies can assist the understanding of metazoans evolution? . . . 11

1.2 Genomes and genes included in the present study . . . 13

1.2.1 Mitochondrial DNA . . . 15

1.2.1.1 Oxidative phosphorylation pathway . . . 15

1.2.1.2 Features of mitochondrial DNA . . . 17

1.2.1.3 Evolution of mitochondrial DNA . . . 18

1.2.2 Nuclear DNA . . . 19

1.2.2.1 Features of the nuclear DNA . . . 19

1.2.2.2 Evolution of the nuclear DNA . . . 20

1.2.2.3 Cytochrome P450 family 2 . . . 21

1.3 Metazoans analyzed in the present study . . . 23

1.3.1 Class Cephalopoda . . . 23

1.3.2 Class Aves . . . 25

1.4 Thesis organization . . . 27

I ADAPTIVE EVOLUTION OF MITOCHONDRIAL GENES IN THE METAZOA: INSIGHTS FROM CLASS CEPHALOPODA (PHYLUM MOLLUSCA) 31

2 Adaptation of the Mitochondrial Genome in Cephalopods: Enhancing Proton Translocation Channels and the Subunit Interactions 33

2.1 Abstract . . . 35
2.2 Introduction . . . 35
2.3 Materials and Methods . . . 37
2.4 Results . . . 43
2.5 Discussion . . . 59

3 Phylogeny and Support for Adaptive Selection in the Driven Force of the ATP Synthesis: Evolution of Cephalopoda Mitogenomes 65

3.1 Abstract . . . 67
3.2 Introduction . . . 68
3.3 Materials and Methods . . . 70
3.4 Results . . . 76
3.5 Discussion . . . 89

II ADAPTIVE EVOLUTION OF NUCLEAR GENES IN THE METAZOA: INSIGHTS FROM CLASS AVES (PHYLUM CHORDATA) 93

4 Whole-Genome Identification, Phylogeny, and Evolution of the Cytochrome P450 Family 2 (CYP2) Subfamilies in Birds 95

4.1 Abstract . . . 97
4.2 Introduction . . . 97
4.3 Materials and Methods . . . 99
4.4 Results . . . 104
4.5 Discussion . . . 117
4.6 Conclusions . . . 121

III BIOINFORMATICS TOOL TO STREAMLINE THE ANALYSIS OF THE RESULTS FROM ESTIMATED SELECTIVE PRESSURES 123

5 IMPACT_S: Integrated Multiprogram Platform to Analyze and Combine Tests of Selection 125

5.1 Abstract . . . 127
5.2 Introduction . . . 127
5.3 Design and Implementation . . . 129

5.4 Results 134
5.5 Conclusion 140

IV GENERAL DISCUSSION AND CONCLUSION 143

6 Discussion and Concluding Remarks 145
6.1 Discussion 147
6.2 Concluding Remarks and Future Directions 151

V REFERENCES 153

References 155

VI APPENDIX 183

A Supplementary Materials 185
A.1 Appendices II – Chapter 2 185
A.2 Appendices III – Chapter 3 185
A.3 Appendices IV – Chapter 4 186
A.4 Appendices V – Chapter 5 186

List of Tables

1.1 Differences between invertebrate mitochondrial and standard genetic codes. 5
1.2 Bioinformatics analyses tools and databases used in the studies presented in chapters 2-5. 29

3.1 Statistical tests used for comparison of probabilities which assess distinct challenging phylogenetic relationships among some Decapodiformes. 81

List of Figures

1.1 Sources of genetic variation 7
1.2 Representation of the hybrid genomes from metazoans 14
1.3 Simplified view of the OXPHOS pathway 17
1.4 Scheme representing the binding of a CYP2 structure to the endoplasmic reticulum membrane 22
1.5 Exemplary images of Class Cephalopoda illustrating body shape diversity. . . 24
1.6 Exemplary images of Class Aves illustrating body shape diversity 26

2.1 Cephalopoda mitogenomic consensus phylogenetic tree 44
2.2 Negative selection reported by FEL site-by-site analyses 46
2.3 Foreground-branches tested for both branch-specific and branch-site selection models 47
2.4 Amino acid variation in functional and structural sites of Cephalopoda ND5 subunit 50
2.5 Amino acid variation in functional and structural sites of Cephalopoda ND6 subunit 53
2.6 Amino acid variation in functional and structural sites of Cephalopoda CYTB subunit 56
2.7 Amino acid variation in functional and structural sites of Cephalopoda COX2 and COX3 subunits 57

3.1 Cephalopoda mitochondrial consensus tree topology from three data sets analyzed in their original forms 78
3.2 Cephalopoda mitochondrial tree topologies from three data sets analyzed in its RY-coded forms 79
3.3 Cephalopoda mitochondrial tree topologies from three data sets analyzed in its 1^{st} and 2^{nd} forms 80
3.4 Percentage of sites under purifying selection and neutrality calculated under random-site model M1a 82
3.5 ATP6 and COX2 cladogram showing the inferred substitution history for each of the four positively selected sites identified by branch-site BEB analyses . . 84

3.6 3D structural model of Complex V showing the location of sites inferred to be evolving under positive selection according to branch-site BEB analyses . . . 87
3.7 3D structural model of Complex IV showing the location of sites inferred to be evolving under positive selection according to branch-site BEB analyses . . . 88

4.1 Identification of the CYP2 genes found in 48 avian genomes 105
4.2 Evolutionary relationships of avian CYP2 subfamilies 106
4.3 Evolutionary process of avian CYP2 subfamilies 108
4.4 SRS of CYP2 proteins 110
4.5 3D analyses of sites detected to be under positive selection in the avian CYP2C111
4.6 3D analyses of sites detected to be under positive selection in the avian CYP2D112
4.7 ω-ratio variations according to distinct avian feeding habits, habitats and migratory behaviors 115
4.8 Results of the linear discriminant function 116
4.9 Comparison of the number of CYP2 genes in birds with different migratory behaviors and feeding habits 117

5.1 PAML tab and related results 136
5.2 Datamonkey tab and related results 137
5.3 Evolutionary Pathway (Evpthwy) tab under TreeSAAP tab and related results 138
5.4 Substitutions (Substs) tab under TreeSAAP tab and related results 139
5.5 3D mapping and related functionalities under Results & 3D tab 140

List of Abbreviations

3D	Three-dimensional
AIC	Akaike Information Criterion
ATP	Adenosine triphosphate
ATP6	ATPase F0 subunit 6
ATP8	ATPase F0 subunit 8
BA	Bayesian
BEB	Bayes Empirical Bayes
bp	Base pairs
BP	Bootstrap percentages
CDS	Complete coding-sequences
COX1	Cytochrome c oxidase subunit 1
COX2	Cytochrome c oxidase subunit 2
COX3	Cytochrome c oxidase subunit 3
CRiSPs	Cysteine-rich Secretory Proteins
C-score	Confidence score
CSV	Comma Separated Values
CYP2	Cytochromes P450 family 2
CYTB	Cytochrome b
df	Degrees of freedom
dN	Non-synonymous substitution rates
DNA	Deoxyribonucleic acid
dS	Synonymous substitution rates
ESS	Effective sample size
F_o	Factor that confers sensitivity to oligomycin
F_1	Factor 1
FEL	Fixed Effects Likelihood
FUBAR	Fast Unconstrained Bayesian AppRoximation
GTR	General Time Reversible
GUI	Graphical User Interface
Iss	Observed saturation index

Iss.c	Saturation index when assuming full saturation
Iss.c Asym	Iss.c. assuming an asymmetrical topology
Iss.c Sym	Iss.c. assuming a symmetrical topology
JGI	Joint Genome Institute
LDA	Linear discriminant analysis
LHON	Leber Hereditary Optic Neuropathy
lnL	Log-likelihood
LRT	Likelihood ratio test
m	Meter
M1a	Model 1a
M2a	Model 2a
M7	Model 7
MSA	Multiple Sequence Alignment
MCMC	Markov Chain Monte Carlo
MEME	Mixed Effects Model of Evolution
ML	Maximum likelihood
mt	Mitochondrial
NCBI	National Center for Biotechnology Information
ND1	NADH dehydrogenase subunit 1
ND2	NADH dehydrogenase subunit 2
ND3	NADH dehydrogenase subunit 3
ND4	NADH dehydrogenase subunit 4
ND4L	NADH dehydrogenase subunit 4L
ND5	NADH dehydrogenase subunit 5
ND6	NADH dehydrogenase subunit 6
NNI	Nearest Neighbour Interchange
NuoA	Bacterial counterpart of the ND3
NuoH	Bacterial counterpart of the ND1
NuoJ	Bacterial counterpart of the ND6
NuoK	Bacterial counterpart of the ND4L
NuoL	Bacterial counterpart of the ND5
NuoN	Bacterial counterpart of the ND2
OXPHOS	Oxidative phosphorylation
P	Probability
PAML	Phylogenetic Analysis by Maximum Likelihood
PBR	Properties By Range
PDB	Protein Data Bank
pHi	Isoelectric point
PP	Posterior probability

PRIME	PRoperty Informed Models of Evolution
PSRF	Potential scale reduction factors
Q_i	Ubiquinone reduction site
Q_o	Quinol oxidation pocket/site
Q-site	Quinone binding site
REL	Random Effects Likelihood
ROS	Reactive oxygen species
rRNA	Ribosomal ribonucleic acid
SPR	Subtree Pruning and Regrafting
SRS	Substrate recognition sites
TENT	Total evidence nucleotide species tree
TM	Transmembrane
tRNA	Transfer ribonucleic acid
UCSC	University of California, Santa Cruz
κ	Transition/transversion ratio
$\overline{\omega}$	Average of the ω values
$\overline{\omega}_B$	Average of background ω values
$\overline{\omega}_F$	Average of foreground ω values
χ^2	Chi-square
ω	Non-synonymous to synonymous substitution rate ratio

Chapter 1

General Introduction

1.1 Background

Among the major evolutionary events in the history of life on Earth traces the “Cambrian Explosion” [1], during which multicellular organisms (*i.e.,* having cells with a nucleus, membrane, and organelles such as mitochondria), better known as metazoans, evolved from a common ancestor, and over evolutionary time gave rise to modern eukaryotic phyla (*e.g.,* Mollusca and Chordata), as reveled by the fossil evidence [1]. This rapid differentiation of body sizes and life forms (particularly bilaterians) led to the emergence of new ecological interactions and diversification. As metazoans increased in number and variety, new ecological niches were occupied within the ecosystems (*e.g.,* marine, terrestrial, freshwater), and over time affected the complexity of environments and habitats, and since then influenced the biosphere forever.

This great diversity of life across environments (*i.e.,* 36 recognized animal phyla with approximately 1,371,075 extant species [2]) known as adaptive radiation, has long attracted the attention of biologists. Indeed, from Darwin’s Theory of Evolution [3], the Modern Synthesis, to todays’ “New Biology” (*i.e.,* foundations of genomics), understanding how animals become adapted to their environments by natural selection (*i.e.,* the evolutionary mechanism responsible for driving adaptive radiation) has been the target of intensive research in the scientific community, particularly in the field of evolutionary biology. Because genes are the major contributors into the adaptability of organisms across environments it is becoming clearer that addressing natural selection and adaptive evolution requires the involvement of diverse elements of organization, from molecule to population [4].

From the molecular perspective, advances in genomics (*i.e.,* high throughput experimental tools, bioinformatics) and evolutionary genetics, have made possible the availability of full genomic sequences from diverse organisms. This provides the opportunity to examine adaptive evolution of phenotypic traits with molecular data, helping to identify changes in the sequence composition of cellular molecules (*i.e.,* DNA – deoxyribonucleic acid, RNA – ribonucleic acid, and proteins). This is possible by using the genomic sequences of species, and by comparing traits and DNA sequences from populations or species with different lifestyles and from different environments. This approach helps to identify which sequences may allow adaptation and in turn, investigate the function of a DNA sequence and its potential adaptive value for an organism [4].

Despite the breakthroughs in advanced genomic approaches there is still work to do as many organisms still remain understudied. In order to narrow this gap, the present study uses genomic data to evaluate for evidence of natural selection in the context of protein-coding genes from two metazoan groups poorly known at the molecular level: 1) Cephalopoda – species without backbones (*i.e.,* invertebrates), and 2) Aves – species with backbones (*i.e.,* vertebrates). This is accomplished by the use of current adaptive evolutionary models that aid in the interpretation of patterns of genetic variation [5–10]. Before describing such methods, first would be important to address some essential concepts of molecular evolution.

1.1.1 Fundamental concepts of molecular evolution

Molecular evolution is the area of evolutionary biology that *(i)* studies genetic variation at the nucleic acid (*e.g.,* DNA) and amino acid level (*i.e.,* rates of sequence change), *(ii)* analyzes the relative importance of adaptive and neutral changes, and *(iii)* evaluate changes in genome structure [11, 12]. The units of molecular evolution are based on genomic information, and concepts such as *i.e., gametes, genes, genome, nucleotides, codon, amino acids, genetic code, synonymous* and *non-synonymous mutations, transitions* and *transversions* need to be properly defined and are described subsequently.

Gametes and genes – The transmission of variations from parents to descendants occurs through the *gametes* (sex cells), and the *genes* are the basic units of gametic inheritance. However, over the past century, the definition of *gene* has changed dramatically, being the most accepted explanation that of “a union of genomic sequences encoding a coherent set of potentially overlapping functional products” [13].

Orthologous and paralogous genes – In comparative genomics, genes from different species are analyzed by grouping them into sets of *orthologous* (*i.e.,* genes related by speciation events) and *paralogous* (*i.e.,* genes related by intragenomic duplication events) [14].

Genome – This term refers to the complete set of genetic material that is required to build an organism, allowing for growth and development [15]. The genome can be expressed as either RNA (for viruses) or DNA (for metazoans). In the case of metazoans or eukaryotes

the basic units of DNA are the *nucleotides*, which are represented by four bases denoted by a one-letter code: the purines – adenine (A) and guanine (G); and the pyrimidines – thymine (T) and cytosine (C). Nucleotides can form triplets (*e.g.,* see Table 1.1) called *codons*. In total, there are 64 codons in the genetic code, and each codon encodes for one *amino acid* (*i.e.,* the basic building units of proteins) of 20 possible letters combinations during synthesis of proteins (*i.e.,* amino acids: A – alanine, C – cysteine, D – aspartic acid, E – glutamic acid, F – phenylalanine, G – glycine, H – histidine, I – isoleucine, K – lysine, L – leucine, M – methionine, N – asparagine, P – proline, Q – glutamine, R – arginine, S – serine, T – threonine, V – valine, W – tryptophan and Y – tyrosine) [4].

Table 1.1: Differences between invertebrate mitochondrial and standard genetic codes. Codon positions are indicated with colors, orange, red, and green for first, second, and third codon positions, respectively. Differences from standard code are highlighted in bold, followed by the corresponding amino acid in the standard code. *stop codon.

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
TTT	F	TCT	S	TAT	Y	TGT	C
TTC		TCC		TAC		TGC	
TTA	L	TCA		TAA	*	TGA	**W**, *
TTG		TCG		TAG		TGG	W
CTT		CCT	P	CAT	H	CGT	R
CTC		CCC		CAC		CGC	
CTA		CCA		CAA	Q	CGA	
CTG		CCG		CAG		CGG	
ATT	I	ACT	T	AAT	N	AGT	S
ATC		ACC		AAC		AGC	
ATA	**M**, I	ACA		AAA	K	AGA	**S**, R
ATG	M	ACG		AAG		AGG	
GTT	V	GCT	A	GAT	D	GGT	G
GTC		GCC		GAC		GGC	
GTA		GCA		GAA	E	GGA	
GTG		GCG		GAG		GGG	

Protein synthesis – Briefly, after ordered processes of transcription (from DNA to a primary RNA) and alternative splicing (removal of noncoding stretches called introns) the genetic information from DNA is converted into messenger RNA (mRNA). Then, each *codon* is recognized as a code for a specific *amino acid* through a process known as translation. Thus, the message (*i.e.,* nucleotide triplets) carried by mRNA is converted into a protein (*i.e.,* long chain of amino acids).

The *genetic code* is universal for all metazoans (standard genetic code), with exception of the mitochondrial code [4], and is often referred as degenerated since there are more than one possible codon to the same amino acid. Differences between the two genetic codes are highlighted in Table 1.1 and are used in this study (*i.e.,* invertebrate mitochondrial code for mitochondrial genetic sequences and standard code for nuclear genetic sequences).

Mutations – Are the raw material of genetic variation. They represent nucleotide changes (Figure 1.1A) in the DNA strand (*e.g.,* insertion or deletion of one or more nucleotides). When changes are from a purine by another purine or a pyrimidine by another pyrimidine, these are called *transitions*. On the other hand, if changes are from a purine by a pyrimidine or vice versa, these are called *transversions*. Mutations are also referred as *(i) synonymous* (*i.e.,* do not result in amino acid changes) or *(ii) non-synonymous* (*i.e.,* result in a distinct amino acid).

Beyond mutations, other sources causing genetic variation are illustrated in Figure 1.1. In some cases, variation is the result of *(i)* duplication events (*i.e.,* duplications of single genes, large genome sections or whole-genomes giving rise to new functions; Figure 1.1B), *(ii)* recombination (*i.e.,* new combination of parts between two different DNA strands into a single strand; Figure 1.1C), *(iii)* horizontal transfer (*i.e.,* transmission of DNA between different genomes; Figure 1.1D), and *(iv)* genetic drift (*i.e.,* stochastic fluctuation in a population due to random sampling of gametes from generation to generation) [16, 17].

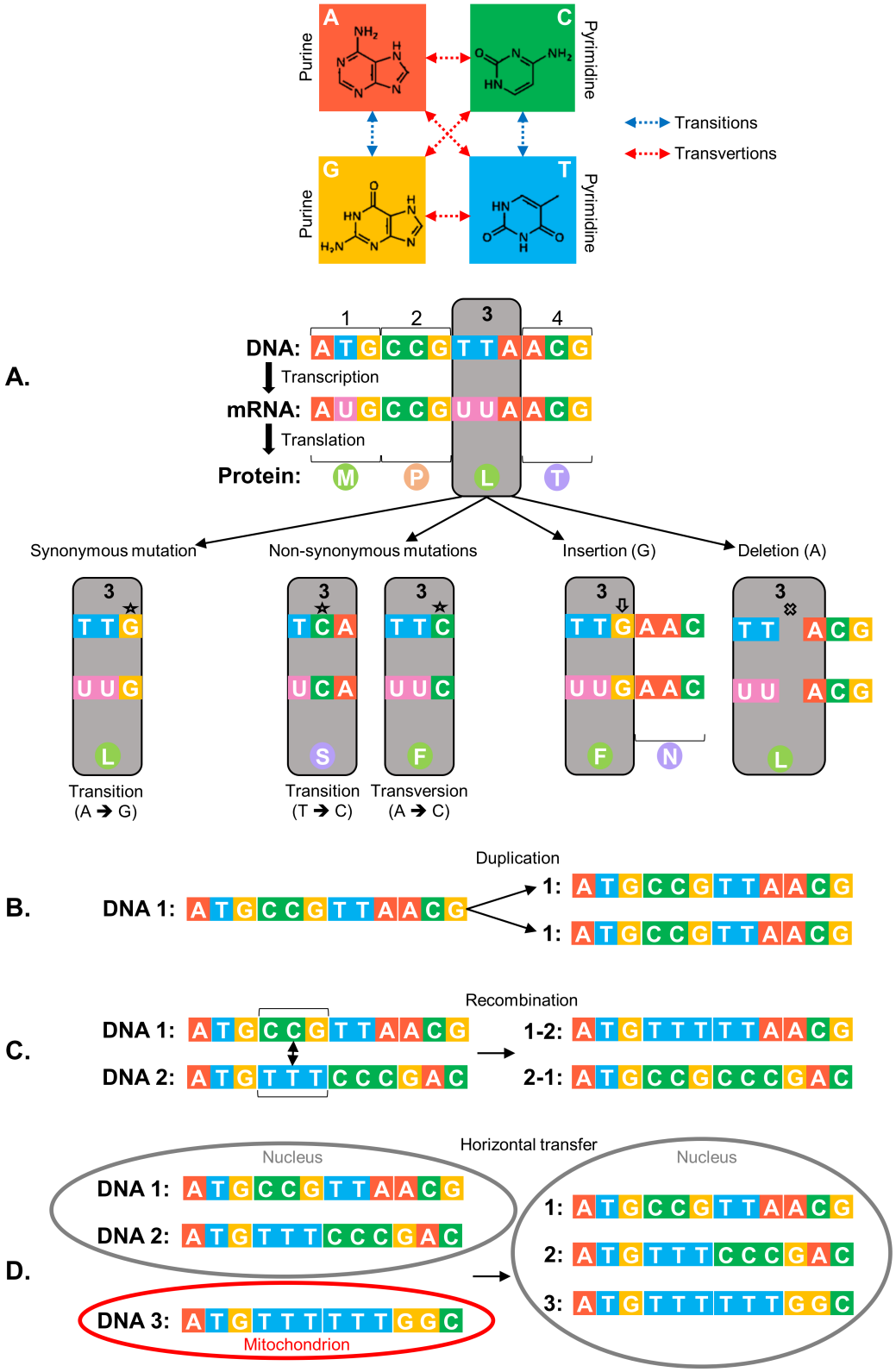


Figure 1.1: Sources of genetic variation: (A) mutations; **(B)** gene duplications; and genetic exchanges through **(C)** recombination or **(D)** horizontal gene transfer. Star point out the codon position where the nucleotide substitution took place. Arrow and "x" symbols indicate the codon position where a guanine was inserted and an adenine was deleted, respectively.

1.1.2 Detecting molecular adaptive evolution

Evolutionary biology concerns with the origin of life and the diversification and adaptation of life forms over time. Under this framework, new species (and populations) result from pre-existing ones, all sharing a common ancestor. To understand the evolution of this speciation *(i)* is important to examine patterns of genetic variation (which are the target of evolution by natural selection) and *(ii)* be able to explain the processes that may be responsible for generating such variation [17]. However, a central challenge during this examination is to discover which variation had positive fitness consequences and which are subject or not to Darwinian selection. Usually, for DNA sequences that encode proteins, changes are detected by employing statistical models of codon evolution (hereafter referred as codon models), which employ a phylogenetic framework of Maximum likelihood (ML) [5–10].

By comparing the relative rates of non-synonymous (*dN*) to synonymous (*dS*) substitutions, codon models can identify several types of selection pressures that shape sequence evolution. This is represented by the ratio *dN/dS* (ω), where ω values greater than one ($\omega > 1$) refer to positive selection, equal to one ($\omega = 1$) means neutral evolution, and less than one ($\omega < 1$) signifies negative selection [18].

Several bioinformatics tools and softwares of codon models (*e.g.,* DATAMONKEY [19] and PAML [20]) are currently available and make possible to 1) address if there is evidence of selection on a gene; 2) identify sequence regions or individual codons under selection and its statistical significance; 3) estimate along which branch of the phylogenetic tree non-neutral selection occurred; and 4) determine if a gene evolved under similar or different selective pressures between species.

1.1.2.1 Bioinformatics stages for selection analyses

Bioinformatics tools (which combine computer science, biology, mathematics and statistics to make possible the storage, retrieval and analyses of biological data in useful time [21]) are essential pieces for the analysis of DNA or protein sequences through phylogenetic and selection analyses (*e.g.,* mutations that are involved in changing the performance of protein complexes and/or are related to certain diseases). In fact, with the advancement of high throughput sequencing techniques, the sequencing has reached a whole-genome scale, which translates into a huge storage of biological sequences made available in online

databases [22–28]. Thenceforth, bioinformatics studies have been widely performed and became an integral component of any biological research due to their ability to analyze huge amount of biological data quickly and cost-effectively, helping to make decisions before conducting an exhaustive *in vivo* experiment [29].

Several bioinformatics steps are needed to perform an evolutionary analysis of protein-coding genes (stated from 1 to 8):

1. **identification of homologous sequences and retrieval –** it is necessary to compile a robust set (data set) of homologous sequences that evolved by vertical descent from a single ancestral, *i.e.,* orthologous genes, from a group of species to be studied; if the sequences for these species are not available in databases (*e.g.,* GENBANK [27] or ENSEMBL [28]), then it is necessary to use similarity based alignment methods such as BLAST (Basic Local Alignment Search Tool [30]) to find the genes from the genomes of several species;

2. **multiple sequence alignment (MSA) and evaluation –** perform the alignment of the sequences from the previously defined data set to allow the comparison of homologous sites (*i.e.,* codons or amino acids which form columns in the alignment) between the sequences of several species; gaps in an alignment are represented as hyphens and mean deletions or insertions (in sequences that do not have a gap); to avoid breaking up the coding reading frame, the sequences must be firstly aligned as proteins and then translated back to nucleotide sequences; there are several tools to perform MSAs, such as *(i)* MUSCLE based on an iterative approach [31] and *(ii)* PROMALS3D that uses information available from sequence homologs with three-dimensional (3D) structures to construct consistency-based multiple sequence alignments [32]; then the quality of the obtained alignment must be assessed, and possibly unreliable positions filtered, with tools like GUIDANCE [33] or TCS [34], in order to minimize the false positives from the downstream analyses like inference of positive selection [35, 36];

3. **determining the nucleotide/amino acid substitution model –** the obtained alignments should be submitted to programs like JMODELTEST (in the case of codon-alignments [37]) or PROTTEST (in the case of amino acid-alignments [38]) to identify the substitution models that best explain the rate of mutations occurring in the alignment during evolution;

4. **evaluation of alignment saturation –** before moving on to phylogenetic and/or selection analyses it is crucial to evaluate if the alignment has enough signal to infer the true pattern of molecular evolution, using software like DAMBE [39, 40] that implements an index of substitution saturation (ISS) to assess whether or not sequences have experienced so many repeated substitutions that the number of substitutions between sequences has become a poor predictor of their time of divergence, that is if the alignment presents substitution saturation; thus, only sequences without or with little saturation must be used in subsequent analyses;

5. **phylogenetic tree building and evaluation –** the reconstruction of a phylogenetic tree/phylogeny, based on the substitution model previously determined in point 3), allows to form the basic hypothesis about the evolutionary relationships among a set of genes/organisms that is given by the branching pattern (topology); in this context *(i)* a root represents the common ancestral to all taxa, *(ii)* nodes represent the ancestral taxa, *(iii)* branches represent the relation between two or more taxa in terms of descent and ancestry and their lengths express the number of changes that have occurred since the split from the most recent ancestral (if a group of taxa share the same branch they have a monophyletic origin and this group is called a clade); there are several algorithms to generate phylogenies such as the character based ML [41] or Bayesian (BA) [42] approaches; the confidence in the phylogeny is given by bootstrap (BS) values in the case of ML analysis and by the posterior probabilities (PP) for the Bayesian analysis;

6. **testing for positive selection in protein-coding genes using codon models –** this step is accomplished using as input *(i)* the codon-alignment from point 2), *(ii)* a well-supported phylogenetic tree (either by BS or PP) from point 5) and *(iii)* several codon models implemented in software such as PAML [20, 43]; in a general way, a model allowing for positive selection is compared to a null model (in which positive selection is not allowed) through a statistical test (likelihood ratio test - LRT) to determine which model best fits the data; the available codon models allow evaluating selective pressure on certain sites (site-models [8, 18, 44–46]), varying along branches of the phylogenetic tree (branch-models [6, 47]), varying simultaneously among sites and branches (branch-site models [48]), or differing in pre-specified clades of the phylogeny

(clade models [49, 50]);

7. **testing for positive selection at amino acid level –** this is a complementary approach to the one previously described in the point 6) allowing to add information for the positively selected sites about their amino acid physicochemical properties; therefore it is necessary to use *(i)* software like TREESAAP [51] that integrates eight magnitude categories, ranging from 1 (most conservative) to 8 (most radical), which describe physicochemical changes owing to amino acid replacements and determines whether they deviate significantly from randomness (being under positive or negative selection) or *(ii)* the PRIME models implemented in the bioinformatics tool DATAMONKEY [19, 52] that were designed to take into account the amino acid physicochemical variation;

8. **mapping of the positively selected sites into a three-dimensional (3D) protein structural model to correlate them with the trait (phenotype) –** the use of 3D structures of proteins, as determined by X-ray crystallography, available in databases such as Protein Data Bank (PDB) or predicted structures through homology modelling using bioinformatics tools such as I-TASSER [53] and SWISS-MODEL [54–57] allow to gain insight into the function of a protein for studying and inferring the potential impact of the positively selected sites through homology comparison with better studied species; the mapping of these sites into the 3D structure can be performed and visualized using software like PYMOL [58].

Each of the above-mentioned steps is critical for understanding the evolutionary process of the data in question, and should be handled accordingly. All are based on specific mathematical models; however, it is out of scope of this study to provide a detailed explanation of them. Thus, for a complete description the reader must refer to [4, 59] and references therein.

1.1.2.2 How molecular studies can assist the understanding of metazoans evolution?

Since Darwin, metazoans have been the major focus to establish phylogenetic relationships among species and genes in order to attain a “Tree of life”. As such, several bioinformatics tools (as aforementioned) have been developed to generate and test several hypotheses about metazoan evolution [60]. In this regard, genetic information has played

a major role by allowing precision and reproducibility in studies of evolutionary biology, enabling the comparison of genes as a proxy for the traits that they encode (phenotypes), helping to understand metazoan evolution.

In the last years, the number of complete sequenced genomes (whole-genomes) from several metazoans increased in online databases (*e.g.,* UCSC Genome Browser database [61–63], JGI database [64], NCBI database [65]), due to the high-throughput technologies, such as next-generation sequencing and annotation tools [66], decreasing costs and expanded interest in "Tree of Life" projects [67]. Most of the available whole-genomes are mitochondrial [68], mainly due to their small sizes ($\sim$14 to 20 kb [69, 70]) compared to the much larger nuclear genomes (from 18.8 Mb for *Pratylenchus coffeae* to 129,907.74 Mb for *Protopterus aethiopicus* [71]). The wealth of whole-genomes annotated in databases led to several molecular evolutionary studies trying to answer how organisms survive in a changing habitat, through analyses of genes encoding crucial proteins related with energetic/thermal/metabolic response [72]. In this regard, mitochondrial protein-coding genes and genes involved in the detoxification of noxious compounds have deeply aroused the interest of biologists.

It is currently widely known that mitochondrial protein-coding genes are crucial to encode several subunits that together with nuclear encoded subunits form the machinery to produce energy (ATP - adenosine triphosphate) in eukaryotic cells, during oxidative phosphorylation (OXPHOS) [73]. Given this vital role for the survival of species, these genes and their encoded proteins have been found across different Phyla of living organisms. Despite their phylogenetic distances from each other, all of them have in common that their OXPHOS systems have similar multi-subunit enzymes (*i.e.,* complexes) allowing molecular comparative studies (*e.g.,* based on available X-ray structures), which consequently increase our knowledge about the functional impact of mutations occurring in several organisms. In this sense, some studies uncovered that mitochondrial protein-coding genes are linked to environmental adaptation of species, for example to altitude [74] and temperature [75]. However, most of the studies on mitochondrial genes showing evidences of positive selection have been restricted to vertebrates (*e.g.,* mammals [76, 77], fishes [78–80], and turtles [81]). Accordingly, there is a tremendous gap on the current knowledge about the evolutionary genomics of mitochondrial genes between vertebrates and invertebrates (particularly little is known with respect to marine invertebrates). Additionally, although it is recognized that

the interactions between the mitochondrial- and nuclear-DNA form a unit of selection and that mutations changing such interactions can have an impact in several key evolutionary processes compromising the fitness of the species, the influence of mitochondrial genomes has been deeply disregarded on the fields of ecology and evolutionary research [82].

Likewise, cytochrome P450 (CYP) genes have been found in a wide variety of metazoans [83]. These genes encode hemeproteins [84] responsible for preventing the increase of reactive oxygen species (ROS) in organisms' bodies through the metabolism of a wide variety of dangerous compounds [85] to which they are exposed in various environments, resulting in the positive selection of different molecular events. This task is efficiently fulfilled due to special characteristics of CYPs such as their high genetic diversity and catalytic versatility, allowing species to adapt and survive in distinct environments [86–89]. Despite their relevance to all metazoans, knowledge about CYP subfamilies' evolution is mostly limited to mammals [90]. Therefore, CYPs study in taxa such as birds (Class Aves), could be particularly important because these species present long life spans, have distinct feeding habits and are found in all habitats; features that together with their ability to fly (found in most birds) and related adaptations, must imply that these species have evolved a complex arsenal of CYPs to deal with oxidative damage.

Overall, an important aspect that might be influencing metazoans evolution is that, under strong environmental selective pressure, species may evolve an enhanced stress tolerance. In this regard, the mitochondrial protein-coding genes and the detoxification protein-coding genes (such as CYPs), may be crucial, since the generation of ROS and the resulting lipid peroxidation of the cell membrane lead to failure of the OXPHOS system and to the depletion of detoxification protein pools [91, 92]. For their understanding at the molecular level, it is important that atomic structures of these enzymes have been solved and have become available in online databases [26].

1.2 Genomes and genes included in the present study

The genetic information of metazoan cells is characterized by chimerism since they evolved from an endosymbiotic oxygen metabolic (aerobic) α-proteobacterium, ancestor of the mitochondria, which had been incorporated by other cell, thus obtaining protection and providing to the host the energy of the OXPHOS system that was converted by the

cells for their own use [93–96]. Therefore, the DNA of metazoans is distributed through two locations (Figure 1.2): a small part of this information is found within intracellular organelles, the mitochondria - mitochondrial genome (mtDNA), and a major part is found on the pairs of chromosomes located in the nucleus of each metazoan cell (nuclear DNA) [97].

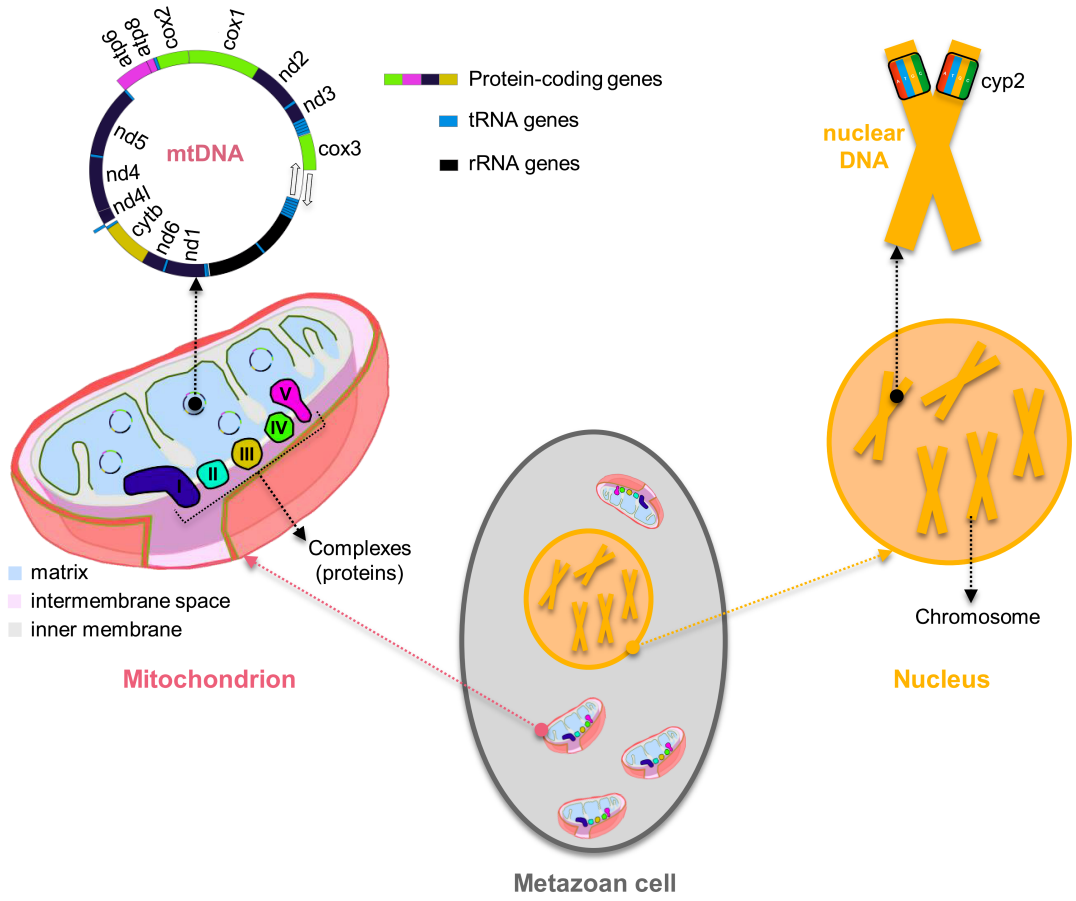


Figure 1.2: Representation of the hybrid genomes from metazoans. Inside a metazoan cell it is possible to observe the cytoplasm (grey) where a small portion of the genome (mtDNA – mitochondrial genome) remains inside mitochondria, and a nucleus (yellow) where the major part of the genome (nuclear genome) is confined. Arrows in the mtDNA molecule indicate the directions of transcription of the two strands. Protein-coding genes: *cox1-3* – cytochrome c oxidase 1-3, *cytb* – cytochrome b, *nd1-6* – NADH dehydrogenase 1-6, *nd4l* – NADH dehydrogenase 4l, *atp6* – ATPase F0 subunit 6, and *atp8* – ATPase F0 subunit 8. Protein subunits creating mitochondrial complexes (I, III, IV and V) are encoded by protein-coding genes represented with the same colour, *e.g.,* purple – Complex I: *nd1*, *nd2*, *nd3*, *nd4*, *nd4l*, *nd5*, and *nd6*. tRNA – transfer ribonucleic acid. rRNA – ribosomal ribonucleic acid. Cyp2 – cytochrome P450 family 2, as an example of a nuclear gene.

In addition to their distinct physical location inside a cell, mtDNA and nuclear DNA have several peculiar features. Thus, the organization of the following text starts from a simplistic overview of each genome, and the respective reports from literature highlighting the main research advances, as well as the existing gaps in knowledge justifying this study

on genomes and genes of relevant species previously poorly explored.

1.2.1 Mitochondrial DNA

The mtDNA is present in all Metazoa (with exception of the Loricifera [98]) and contains an almost perfectly conserved complement of "housekeeping genes", *i.e.,* genes responsible for basal cell functions, such as protein-coding genes (Figure 1.2) – cytochrome *c* oxidase 1-3 (*cox1, cox2, cox3*), cytochrome b (*cytb*), NADH dehydrogenase 1-6 (*nd1, nd2, nd3, nd4, nd5, nd6*), NADH dehydrogenase 4l (*nd4l*), ATPase F_0 subunit 6 (*atp6*) and ATPase F_0 subunit 8 (*atp8*) – for the formation of the multi-subunit enzyme complexes involved in the OXPHOS system, which is responsible for the production of up to 95% of the energy (ATP) of eukaryotic cells [99].

1.2.1.1 Oxidative phosphorylation pathway

In the OXPHOS process (Figure 1.3) two types of complexes participate: those whose subunits are encoded by both mitochondrial and nuclear genes, *i.e.,* Complexes I, III, IV and V; and one complex entirely composed of nuclear-encoded subunits, Complex II. Complexes I-IV drive the electron transport from NADH to the molecular oxygen, creating a proton (H^+) energy gradient across the inner mitochondrial membrane. Then, the Complex V uses this gradient to synthesize ATP [82]. Details about the mitochondrial-encoded subunits for each one of the OXPHOS' complexes are presented below.

Mitochondrial subunits of Complex I – The mitochondrial subunits (hereafter denoted with capital letters to distinguish them from genes) ND1, ND2, ND3, ND4, ND4L, ND5 and ND6 interact with the nuclear counterparts forming the Complex I (NADH–ubiquinone oxidoreductase). This complex is responsible for the first step in the electron transport process, performing the electron transfer between NADH and ubiquinone, coupled to the translocation of four protons across the inner mitochondrial membrane to conduce ATP production [100]. It is an L-shaped complex, the first and largest enzyme of the OXPHOS pathway, composed by a hydrophilic peripheral arm and by a hydrophobic membrane-embedded arm perpendicular to it [101]. The peripheral arm (Figure 1.3) is where the electron transfer chain events occur and projects into the mitochondrial matrix [102]. On the other hand, the membrane-embedded arm (Figure 1.3) contains seven mitochondrial

subunits [102]. The subunits ND2, ND4, and ND5 are homologous to each other and have sequence homology with Na^+/H^+ antiporters, being thus suggested to pump protons across the inner mitochondrial membrane [103]. A recent study suggested the existence of a fourth proton translocation pathway at the interface of the subunits NuoN (ND2), NuoK (ND4L) and NuoJ (ND6) in the bacterium *Escherichia coli* [101]. Regarding to the ND1 function, it remains yet unknown [101]. The Complex I provides approximately 40% of the proton flux necessary to produce ATP, being also the major source of ROS in the mitochondria [101].

Mitochondrial subunits of Complex III – The CYTB subunit integrates the third complex, cytochrome *c* reductase or cytochrome bc_1 complex (Figure 1.3) of the OXPHOS pathway. It catalyzes the reversible electron transfer from ubiquinol to cytochrome *c*, coupled to proton translocation [102].

Mitochondrial subunits of Complex IV – The subunits COX1, COX2, and COX3 integrate the Complex IV, cytochrome c oxidase (Figure 1.3), in which nuclear subunits are mainly externally located [102]. This complex catalyzes the reduction of molecular oxygen to water [104] and uses the released energy to pump protons from the mitochondrial matrix to the intermembrane space [105]. COX1 and COX2 subunits form the catalytic core of this terminal electron transfer chain complex; being directly involved in the electron transfer and proton translocation processes [76, 105]. The COX3 is not involved in proton translocation, but has an important role in the assembly and stability of a functional oxidase, and might influence proton translocation indirectly by affecting the conformation of the remaining part of the enzyme [104].

Mitochondrial subunits of Complex V – ATP6 and ATP8 subunits, integrate the Complex V (F_1F_0 complex or ATP-synthase). In this complex (Figure 1.3) takes place the final step of the respiratory chain. It uses the proton gradient generated by the Complexes I, III, and IV to produce ATP. Complex V is composed of an inner mitochondrial membrane domain (F_0) and a catalytic domain (F_1) located on the mitochondrial matrix side, linked by two stalks [106]. The ATP8 is suggested to retain a regulatory role in the assembly of the complex, being a core subunit of the F_0 domain [76]. This domain contains the proton channel through which the protons accumulated in the intermembrane space enter in the Complex V via

ATP6 subunit in direction to mitochondrial matrix. The F_1 catalytic domain contains three β subunits that are the sites of ATP synthesis.

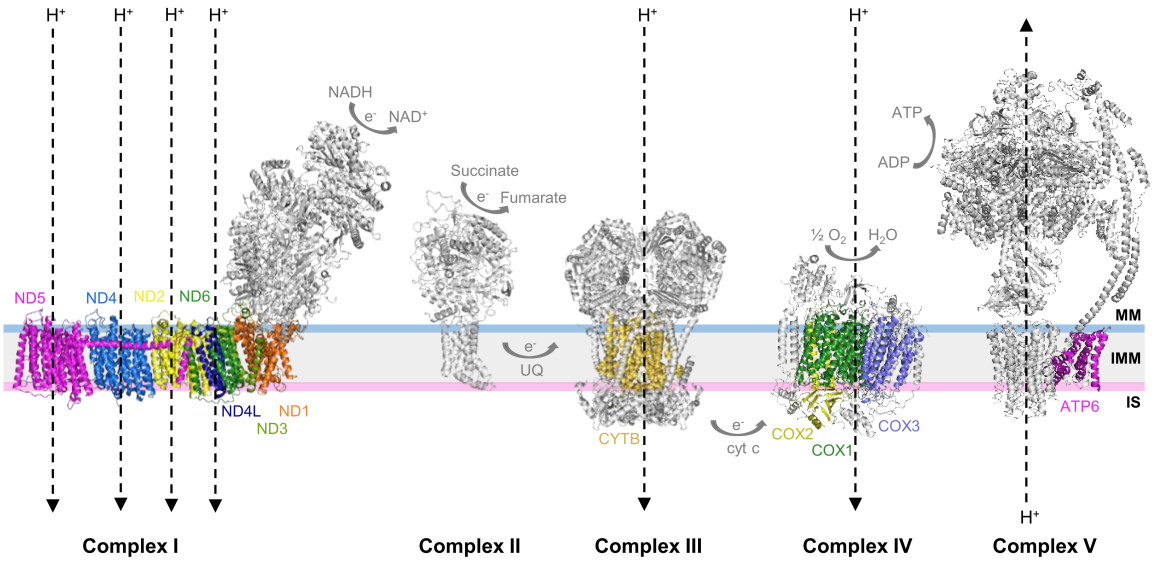


Figure 1.3: Simplified view of the OXPHOS pathway. Electron (e^-) transport between complexes starts with Complex I (NADH–ubiquinone oxidoreductase), the entry point for e^- from NADH which are used to reduce ubiquinone to ubiquinol (UQ) that is subsequently used by Complex III (cytochrome bc_1), to reduce cytochrome c (cyt c). Then, Complex IV (cytochrome *c* oxidase) uses cyt *c* to reduce molecular oxygen (O_2), the ultimate electron acceptor, to water (H_2O). e^- transfer events are used to pump protons (H^+) from mitochondrial matrix (MM) into the intermembrane space (IS) generating a proton gradient used by Complex V (F_1F_0-ATP synthase) to synthesize adenosine triphosphate (ATP). The structure of each complex is presented: Complex I from *Thermus thermophilus* (PDB: 4HEA); Complex II (succinate–quinone oxidoreductase) from *Sus scrofa* (PDB: 1ZOY); Complexes III, IV, and V from *Bos taurus* (PDBs: 1PPJ, 1V55, and 5ARA, respectively). The subunits encoded by mitochondrial protein-coding genes are highlighted in colour. The protein structure of Complex V has not been resolved in its entirety and therefore the ATP8 subunit is not shown. PDB – Protein Data Bank identifier. IMM – inner mitochondrial membrane. The figure was generated using PYMOL [58].

1.2.1.2 Features of mitochondrial DNA

Overall features of the metazoan mtDNA molecules include: 1) being smaller than nuclear genomes [70]; 2) often being compact circular molecules, although linear mitochondrial genomes have been described for some Cnidaria and Porifera species [107]; 3) having a relatively stable number of genes, typically 37 (Figure 1.2) – of which 13 protein-coding genes encode for subunits of enzymes functioning in electron transport and ATP synthesis (OXPHOS), 22 for transfer ribonucleic acid (tRNA), and 2 for ribosomal ribonucleic acid (large and small rRNA subunits) –, a number more variable in non-bilaterian species [108]; 4) presenting absence of introns and large intergenic regions (with few exceptions in Placozoa, Porifera, Cnidaria and Annelida species) [107, 109]; 5) having a large noncoding region which contains elements that control transcription of mitochondrial genes and/or replication

of these genomes (for that reason also known as control region) [69, 110]; 6) harboring well-conserved genetic functions, namely in respiration and/or OXPHOS, RNA maturation and protein import [108], 7) exhibiting a mutation rate higher than that of nuclear genomes, believed to arise due to close proximity of mtDNA to the OXPHOS system (located on the inner mitochondrial membrane), thus making mtDNA vulnerable to damage through leakage of ROS during OXPHOS process [111]; 8) having a mixture of conserved and variable sites [112]; 9) showing maternal inheritance (with exception of doubly uniparental inheritance in some species of bivalves) [113, 114]; and 10) presenting high copy number (100-1000 copies of mitochondrial genes) tightly regulated in a cell- and tissue-specific manner, depending on their energy demand [115].

1.2.1.3 Evolution of mitochondrial DNA

The evolution of mtDNA of some metazoans has been characterized by *(i)* changes in the position and strand specificity of the following genes: protein-coding genes, tRNAs, and rRNAs, as well as, by *(ii)* gene losses and/or duplications [116–118]. Duplications are often subject to pseudogenization (*i.e.,* the loss of function for a given gene that is not expressed), such as in birds (*nd6*), mollusks (*cox3*) and insects (*nd2*, *nd4*, *nd5*, *cox1*, *cox2* and *cox3*) [109].

Comparisons between mitochondrial and nuclear genomes of a given organism, from several metazoans (*e.g.,* from human and primates) revealed different evolutionary trends for the two genomes [108]. Although, several studies reported that non-synonymous mutations in mitochondrial genome accumulate at an accelerated rate relative to nuclear genomes, the mtDNA remains functional and integral to fitness [119].

The evolution of mtDNA has often been assumed as neutral with negligible selection [120]. However, more recently, it has been showed that there is significant departure from neutrality, being purifying/negative selection the predominant force in metazoan mtDNA evolution [121]. Despite being under negative selection, the ratio of non-synonymous to synonymous fixations (*dN/dS*) in mitochondrial genomes of invertebrates is higher than in vertebrates, which is consistent with two hypotheses: *(i)* the idea that more amino acid fixations have been adaptive in invertebrates, and *(ii)* a more relaxed selective constraint acting on amino acid substitutions in these species [122]. Nowadays, there are also several reports of mitochondrial protein-coding genes with evidence of diversifying/positive selection, but most

of those studies were performed in vertebrates [76, 78, 123, 124]. Some amino acid changes in the mitochondrial subunits were associated with inefficiencies in the electron transfer chain system [125], contributing to the formation of ROS. Otherwise, the variation in mitochondrial genes involved in OXPHOS pathway can directly influence metabolic performance [76], the amino acid substitutions can improve aerobic capacity and be linked with life history traits and environmental adaptations [126–128]. Moreover, functional interactions between mitochondrial- and nuclear-encoded proteins may result in co-adaptation to maintain or improve metabolic fitness of the species [73, 129].

1.2.2 Nuclear DNA

In 1974, the DNA was found to be condensed and coiled around histones (DNA-binding proteins), allowing the bulk of the genome, the nuclear DNA, to be organized within the nucleus in the form of chromosomes (Figure 1.2) – a set of linear DNA molecules [130]. The nuclear genome sizes and the haploid chromosome numbers (from $n = 3$ to $n = 39$ [131]) are variable in different metazoans, and reflect the non-uniformity of the evolutionary events that have shaped the architecture of these genomes [130]. Chromosomes are formed by chromatin in two distinct structural forms that reflect the level of activity of the cell: euchromatin and heterochromatin. The first is less condensed and therefore contains most of the active genes, whereas the second is more compacted causing inactivation of genes and containing base sequences often repeated [132]. Chromosomes have coding and non-coding regions. The genes lie in the coding regions, which contain the essential information for the synthesis of proteins [16]. The coding regions of a gene are called exons while the structural sequences are known as introns (Figure 1.2).

1.2.2.1 Features of the nuclear DNA

Overall features of the metazoan nuclear DNA include [16, 132]: 1) being linear and double-stranded; 2) having a high size ($10^8 - 10^{11}$ nucleotide pairs per haploid genome) [120]; 3) being distributed over a variable number of chromosome pairs (homologs) containing a variable number of genes; 4) possessing different associated proteins, histones that compact the DNA and others that interact in the process of gene expression, replication and DNA repair; 5) in organisms that reproduce sexually, each cell inherits a copy of DNA from each of the parents; 6) presenting homologous DNA recombination, important to the genome

stability by assuring accurate repair of double-strand breaks of DNA [133] and by allowing the correct DNA partners to recombine, promoting genetic diversity by generating new gene combinations [134]; 7) the rate of nucleotide substitution among protein-coding regions of nuclear genes is slower compared with mitochondrial protein-coding genes [135]; and 8) having a large amount of non-coding repeating DNA.

1.2.2.2 Evolution of the nuclear DNA

Comparatively to the mitochondrial genome, the nuclear genome reveals much higher plasticity [120]. In fact, the nuclear genome was subject to multiple and continuous duplication events [136, 137], creating genetic redundancy expressed at the sequence level (through duplications and retrotransposition events) [138]. These events led to variability of nuclear genome sizes, which is related to the high percentage of non-coding sequences (*e.g.,* short and long interspersed nuclear elements, long terminal repeat elements, and DNA transposons), localized in both inter and intra-genic space, and partially due to variation in coding sequence numbers (*e.g.,* duplicated genes) [139].

Metazoan nuclear genomes have several duplicated genes, called paralogous, which can originate gene families, with a variable number of members in different organisms [140, 141]. Paralogous are a source of genome functional novelties [120] and can evolve in four distinct ways: 1) neofunctionalization, where both copies may be preserved in the genome with one of them being modified, for example by positive selection or evolve rapidly free from purifying selection, with the emergence of a novel function and the other copy retaining the original function [142]; 2) subfunctionalization, where each duplicate retains a distinct aspect of the ancestral gene function, due to relaxation of purifying selection which allows accelerated evolution in both copies [143]; 3) dosage effect, when both copies are maintained for a selective advantage in synthesizing large amounts of the ancestral gene product [142]; and 4) nonfunctionalization, if one copy accumulate degenerative mutations and loses its function, being retained just as part of the sequence [144–146].

The gene families consist of groups of genes of identical or similar sequences that share important features like providing instructions for synthetize: 1) products that have a similar structure or function; 2) proteins that work together as a unit; or 3) proteins that participate in the same cellular process [139]. There are several large groups of paralogous (gene families) such as those possessing functions related to detoxification (degradation) of endogenous

or exogenous compounds, immune response, reproduction and chemosensation [14] – collectively referred to as environmental genes [147].

Some studies have reported that these families of paralogous genes often show evidence of adaptive evolution by amino acid sequence change, allowing new functions which are a result of positive selection for a specific innovation [131, 148, 149]. In addition to the observed positive selection, the expansion of functional gene families suggests that they are a standard adaptive genomic response to environmental challenges [14].

1.2.2.3 Cytochrome P450 family 2

The cytochrome P450 family 2 (CYP2) is an example of an environmental gene family crucial for the survival of several metazoans. CYP2 genes contain nine exons and produce heme-containing enzymes with around 500 amino acids (whose heme is the catalytic center of the enzyme). This gene family encodes for the phase I oxidative metabolic enzymes, one of the major components of chemical defense system of metazoans [150]. These enzymes are primarily found in the endoplasmic reticulum of liver cells (Figure 1.4), involved in the metabolism of many lipophilic endogenous (*e.g.,* toxins that are formed within cells, steroids and fatty acids) and exogenous substances (*e.g.,* xenobiotics as environmental pollutants or medications that are ingested) [151, 152].

The basis for its nomenclature is dependent on enzymes amino acid sequence homology. To belong to the same family the sequences must have more than 40% identity [83, 153, 153]. The Arabic number represents the enzyme family (*e.g.,* CYP2), divided into several subfamilies (*e.g.,* CYP2C, family 2 and subfamily C) in which, within a single subfamily, the sequence identity is greater than 55% [83, 153]. The members of this gene family present a range of substrate specificity (defined by amino acid variations in the substrate binding sites) resulting from gene duplications and subsequent divergence [154]. Some gene duplications (*e.g.,* CYP2D6 gene copies) are responsible for favorable phenotypes such as the ultra-rapid metabolizer phenotype identified in humans [155]. However, several of these detoxification-type genes also suffered frequent pseudogenization [156].

The primary structure of these enzyme sequences present the signature motifs CXG (with C residue conserved in all CYPs) located in the heme-binding domain and the EXXR (relevant for maintenance of CYP tertiary structure) in the K helix [152]. These motifs are important to assure stabilization of the core structure of CYPs. Although CYPs preserve a

basic structural fold, their substrate-binding cavities present high structural plasticity and the ability to alter shape and volume depending on the compound they accommodate. Therefore, the X-ray crystal structures of CYPs present six putative substrate recognition sites (SRS1-6 [157]), which place in position the substrate in the active-site of the enzyme (Figure 1.4). Amino acid differences between CYPs in SRSs account for differences in substrate specificity or can be related to differences in enzyme kinetics that metabolize the same substrate [152].

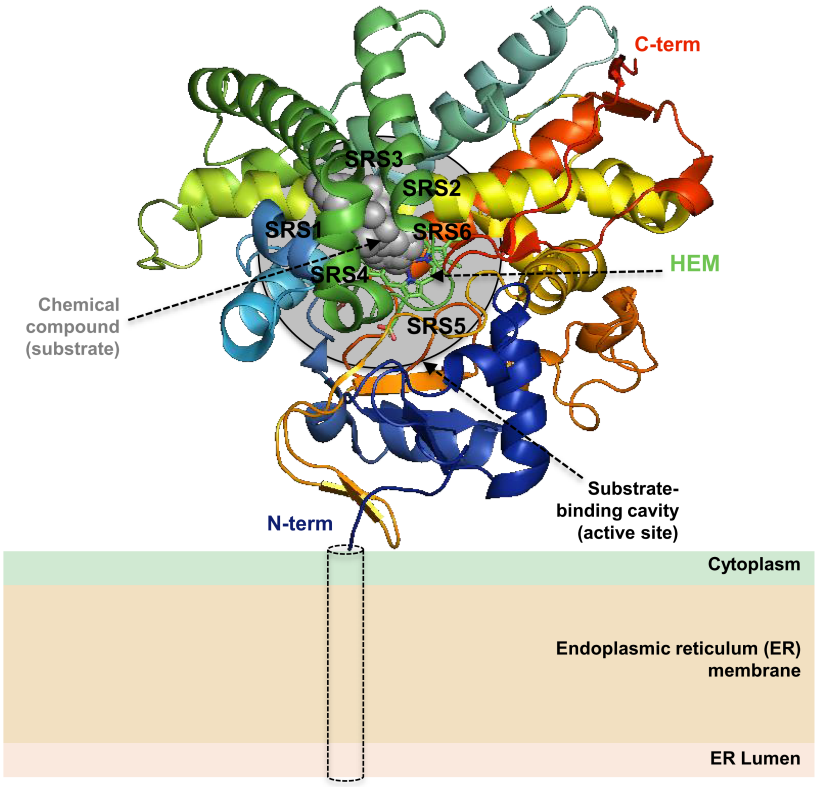


Figure 1.4: Scheme representing the binding of a CYP2 structure to the endoplasmic reticulum membrane. The X-ray crystal structure determined to this CYP2C5 (PDB: 1N6B) is represented viewed from the distal face (substrate-binding site) and does not contain the N-terminal transmembrane neither the linker region - that was here represented by the dashed cylinder to indicate the approximate location of the membrane binding interface. Secondary structural elements (α-helices, loops and β-sheets) are color-coded from N-terminal (blue) to C-terminal (red). SRS – substrate recognition sites. The figure was generated using PYMOL [58]. For a detailed identification of the α-helices (A to L) and precise location of SRSs please refer to [158].

A great expansion of the CYP2 family started around 400 million years ago and several studies suggested that they are among the most rapidly evolving genes, a characteristic needed to protect organisms from the harms when exposed to increasing toxic xenobiotic compounds present in their habitats [152]. Accordingly, coevolution of exogenous compounds with the enzymes that metabolize them indicate that CYPs are subject to positive selection

[159]. Although some studies have in fact indicated evidence of adaptive evolution in CYP2 genes, these are mostly restricted to mammals [90, 160, 161].

1.3 Metazoans analyzed in the present study

Despite the relevance of the aforementioned repertoire of genes (*i.e.,* mitochondrial protein-coding genes and CYP2) for survival, their study is restricted mostly to mammals. Aiming to extend their understanding to other metazoans, the classes Cephalopoda and Aves, were chosen to broaden our perspective of the molecular evolution of such genes in invertebrates and vertebrates, respectively. Both classes present distinctive features that will be briefly described below.

1.3.1 Class Cephalopoda

The Class Cephalopoda (Figure 1.5) comprises two subclasses of species named Coleoidea and Nautiloidea. The subclass Coleoidea comprises the following superorders: 1) Decapodiformes – that are species with ten arms, namely the squids from the major taxonomic groups Oegopsida, Bathyteuthoidea, Idiosepiidae, Myopsida, Spirulida and the bobtail squids and cuttlefishes from Sepioidea (Sepiolida and Sepiida); and 2) Octopodiformes – species with eight arms, such as octopuses and vampire squids from the Octopoda and Vampyromorpha groups, respectively. The subclass Nautiloidea (*Nautilus* and *Allonautilus* genus) contains species known as living fossils, which are the only cephalopods with an external chambered shell [162]. Species from both subclasses arose from an ancestral mollusk (from Cambrian $\sim$530 million years ago) with an external shell that was internalized as a chambered buoyancy apparatus in the presently dominant Coleoidea species. Therefore, cephalopods underwent a unique evolutionary event that was the shell internalization, being equipped with a distinctive buoyancy system for locomotion among metazoans [163].

Furthermore, cephalopods are an enigmatic group of metazoans (considered to be the smartest and most complex invertebrates), namely they evolved highly differentiated brains, advanced visual perception, camouflage strategies, diverse metabolic requirements and variations in body shapes and sizes [164]. These exclusively marine species are worldwide dispersed and possess relevant commercial importance [164, 165].

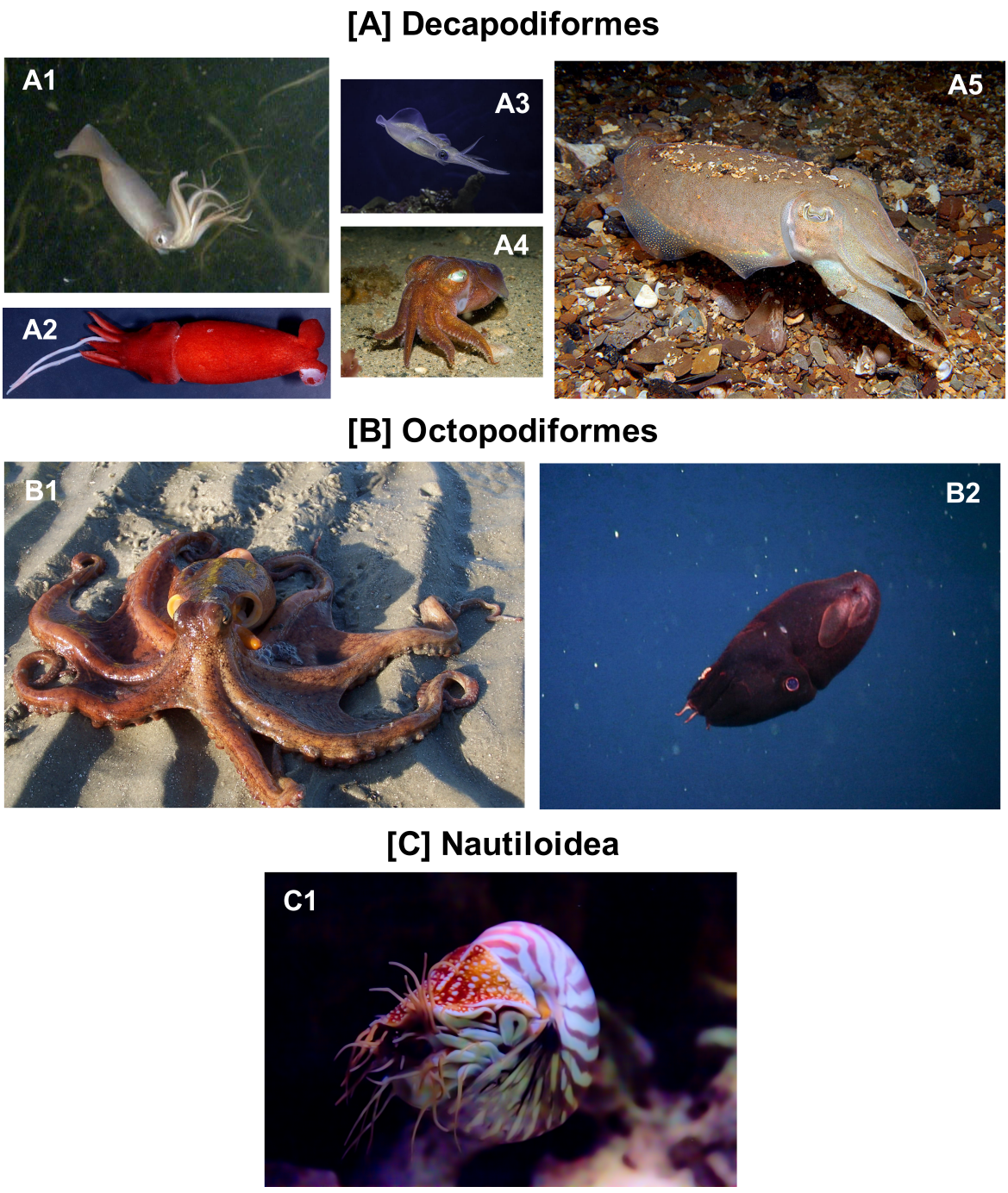


Figure 1.5: Exemplary images of Class Cephalopoda illustrating body shape diversity. Cephalopoda species representing the major taxonomic groups: Decapodiformes **(A1-4)**, Octopodiformes **(B1-2)** and Nautiloidea **(C1)**. **A1:** *Dosidicus gigas* (Oegopsida). **A2:** *Bathyteuthis abyssicola* (Bathyteuthoidea). **A3:** *Sepioteuthis lessoniana* (Myopsida). **A4:** *Semirossia sp.* (Sepiolida). **A5:** *Sepia officinalis* (Sepiida). **B1:** *Octopus vulgaris* (Octopoda). **B2:** *Vampyroteuthis infernalis* (Vampyromorpha). **C1:** *Nautilus sp* Photographs: A1 by image courtesy of NOAA/MBARI, 2006 under Public domain, source Wikimedia Commons (`https://goo.gl/vuv9kP`); A2 licensed by Darren Stevens, 2012 under a Creative Commons License – (CC BY 3.0), source Wikimedia Commons (`https://goo.gl/NSQCVV`); A3 licensed by California Academy of Sciences, 2016 under the CC BY-NC-SA 3.0, source CalPhotos (https://goo.gl/ZLbrUO); A4-5 licensed by National Museums Northern Ireland and its licensors, 2002-2015 under the CC BY-NC-SA 3.0, source Encyclopedia of Marine Life of Britain and Ireland (`https://goo.gl/qZ0Y73` and `https://goo.gl/p5OD6B`, respectively); B1 by image courtesy of Jo O'Keefe (Carolina Shores, North Carolina, USA - `http://okeefes.org`); B2 by image courtesy of NOAA/MBARI, 2006 under Public domain, source Encyclopedia of Life (`https://goo.gl/2w6T13`); C1 by image courtesy of Jon Pinder licensed under CC BY-NC-ND 2.0, source Flickr (`https://goo.gl/KgTTk2`).

They thrive at all depths of oceans and across the widest ranges of environmental conditions (*i.e.,* variations in temperature, salinity, oxygen, pressure) [166]. An extraordinary example of extreme environmental conditions exists in the deep-sea (depths greater than 200m [167]), such as the oxygen minimum zones characterized by very low oxygen availability [168]. Although these environments are inhospitable to most marine species, cephalopods are among the metazoans that can tolerate these extreme environmental conditions [168, 169].

In addition to their morphological novelties, cephalopods also possess remarkable characteristics at the molecular level that clearly distinguish them from other members of the Animal kingdom, including 1) the presence of duplicated mitochondrial genes that are identical and/or nearly identical in nucleotide sequence (*cox1-cox3*, *atp6* and *atp8* from squids of the major taxonomic groups Oegopsida and Bathyteuthoidea [170]), a feature that contrasts with the occurrence of pseudogenization documented in other metazoans [109], and 2) their mtDNA gene arrangement, with gene order varying extensively even across closely related lineages [116, 171], contrasting with the highly conserved gene content and gene order in vertebrates [171].

All these remarkable biological features have left footprints on their genomes, making them an ideal group to answer molecular evolutionary questions, specifically associated with bioenergetics, namely how accumulation of mutations in the mtDNA can be influenced by life histories and cellular energy requirements and thereby how positively it affects fitness.

1.3.2 Class Aves

The species of the vertebrate Class Aves arose arround 150 million years ago, being documented in the fossil record as birdlike dinosaurs' forms [172, 173]. Currently, the $\sim$ 10,500 contemporary bird species make this Class the most species-rich of all tetrapod vertebrates [174]. Avian species are distributed in three major evolutionary groups (Figure 1.6): Neoaves (contains most of extant birds, such as flamingos, owls, passerines, hummingbirds and penguins), Galloanserae (waterfowl – such as ducks, geese and swans, that spend most of their non-flying time on water and landfowl – heavy-bodied ground-feeding birds which includes chickens and turkeys) and Palaeognathae (that is species-poor, represented by some tinamous with poor ability to fly and flightless ratites) [174]. Given the incredible number of species makes this Class unique among metazoans since the evolutionary history of birds is an outcome of an astonishing rapid radiation and worldwide dispersion [175].

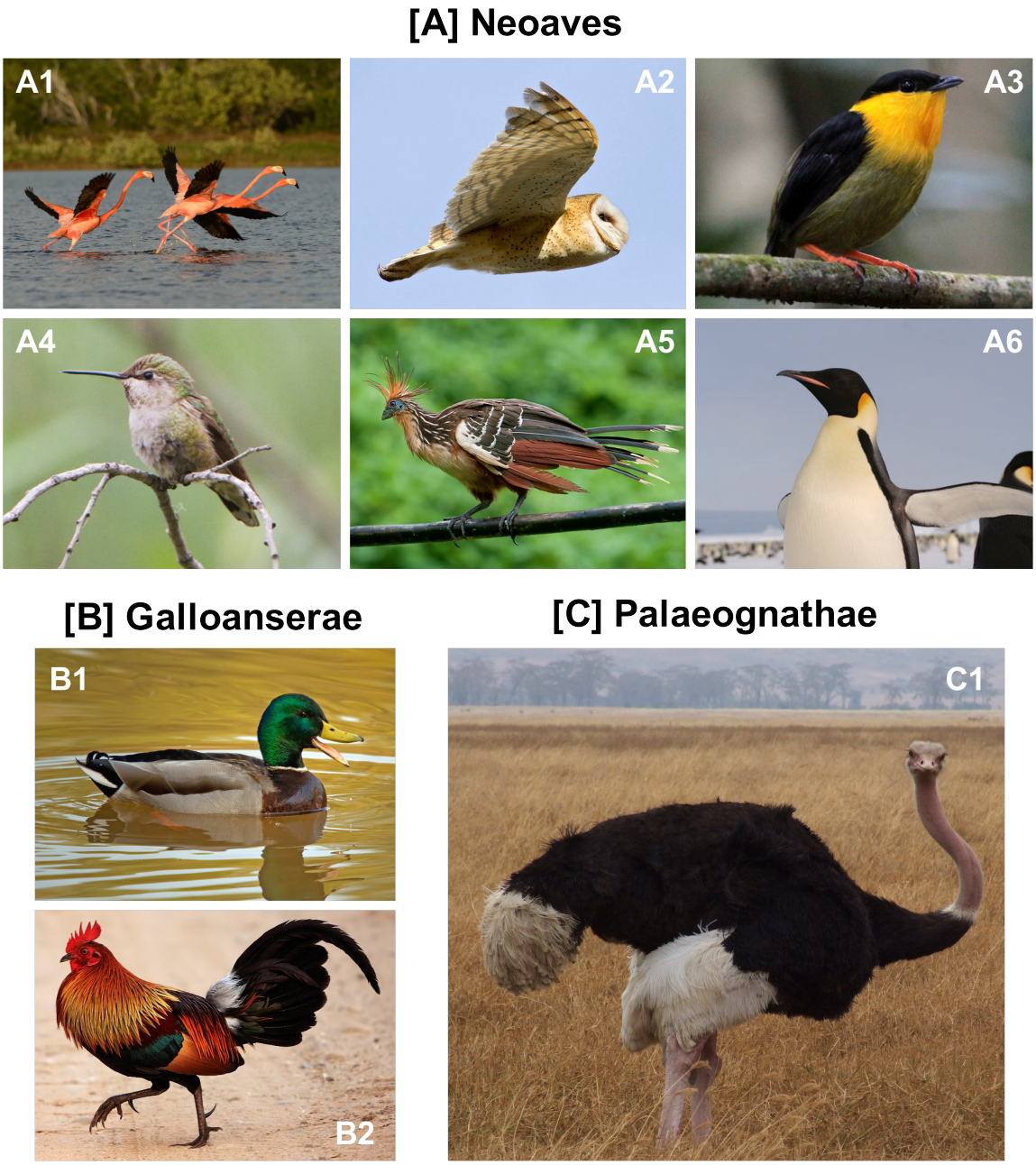


Figure 1.6: Exemplary images of Class Aves illustrating body shape diversity. Avian species representing the major taxonomic groups: Neoaves (**A1-6**), Galloanserae (**B1-2**) and Palaeognathae (**C1**). **A1:** *Phoenicopterus ruber*. **A2:** *Tyto alba*. **A3:** *Manacus vitellinus*. **A4:** *Calypte anna*. **A5:** *Opisthocomus hoazin*. **A6:** *Aptenodytes forsteri*. **B1:** *Anas platyrhynchos*. **B2:** *Gallus gallus*. **C1:** *Struthio camelus*. Photographs: A1 licensed by Amado Demesa, 2015 under a Creative Commons License – (CC BY-SA 2.0), source Flickr (`https://goo.gl/tmVXji`); A2 licensed by Bill Bouton, 2011 under the CC BY-SA 2.0, source Flickr (`https://goo.gl/cH70ZZ`); A3 licensed by Carmelo López Abad, 2015 under the CC BY-NC 4.0, source iNaturalist.org (`https://goo.gl/4WiATN`); A4 licensed by BJ Stacey, 2013 under the CC BY-NC 4.0, source iNaturalist.org (`https://goo.gl/OMyIeW`); A5 licensed by Cláudio Dias Timm, 2012 under the CC BY-NC-SA 2.0, source Flickr (`https://goo.gl/brrW3L`); A6 licensed by Dbush, 2005 under the CC BY-SA 3.0, source Wikimedia Commons (`https://goo.gl/eRFtjr`); B1 licensed by Richard Schuerger, 2010 under the CC BY-NC-SA 2.0, source Flickr (`https://goo.gl/qcl066`); B2 licensed by Subramanya C K, 2012 under CC BY-SA 3.0, source Wikimedia Commons (`https://goo.gl/0Z1U9C`); C1 licensed by Nicor, 2009 under CC BY-SA 3.0, source Wikimedia Commons (`https://goo.gl/P63YX8`). Photographs from all panels, except A2, were cropped from originals.

Additionally, avian species have traits that are either unique to them or are shared with only a limited number of other lineages of metazoans: 1) they evolved feathers, which allows most birds to fly, and whose subsequent morphological diversification is associated with different avian lifestyles [176]; 2) they have bills and endetulism (that is the phenotype of being toothless) [174]; 3) their bones are lightweight [176]; 4) they have evolved an amazing diversity of dietary specializations (namely diet-related enzymes [174]); 5) they have the most advanced vertebrate visual system [174]; 6) they have vocal communication and learning [176]; 7) they present migratory behaviors, travelling long or short distances [176]; 8) they show remarkable slow aging rates and long life spans for their sizes [177]; 9) despite of their incredible morphological diversity they have the smallest genomes among amniotes [15], as they contain lower levels of repeat elements compared to other tetrapod vertebrates [174], 10) they present a stable karyotype ($\sim$30 pairs of microchromosomes), a feature not found in any other vertebrate group, which is remarkable in the light of their rapid radiation [174]; and finally 11) they exhibit female heterogamety (ZW [175]).

All these special features of avian species have attracted considerable interest, among the scientific community, making them relevant models to answer evolutionary and ecological questions, and by investigating links between their genomic variation and phenotypic diversity. In particular, the identification of sites within genes under positive selection underlying lineage-specific traits (*e.g.,* different habitats, food habits, migratory behaviors) can increase our knowledge about metazoans' adaptations and their environments.

1.4 Thesis organization

Evolution by natural selection is the central theme of this thesis. Therefore, the major goal is to test for evidence of adaptive evolution, by evaluating whether changes (patterns of genetic variation) in genes represent positive Darwinian selection. This was accomplished via comparative evolutionary analyses using genes from both mitochondrial and nuclear genomes. In the case of the mitochondrial protein-coding sequences the data set was retrieved from different sources: MITOZOA [68, 178] and GENBANK [27, 179], while the nuclear avian genomes were retrieved from Ensembl [28] and the Avian Phylogenomics Project in GigaScience database [180].

This thesis consists of a general introduction (**chapter 1**), four journal papers (**chapters**

2-5), and a general discussion and concluding remarks (**chapter 6**). The databases and software tools for specific bioinformatics analyses used in this thesis are listed in Table 1.2. Detailed descriptions about methods are provided in the **chapters 2-5**.

This thesis is divided into three major parts:

- **Part I (chapters 2-3)** relates to adaptive evolution of mitochondrial protein-coding genes from the Class Cephalopoda (Phylum Mollusca), previously poorly studied, aiming to clarify some elusive taxa relationships and obtain insights about which selective forces are shaping their evolution. This is accomplished by investigating the codons encoding amino acids involved in the OXPHOS process and its possible association to species fitness;

- **Part II (chapter 4)** relates to adaptive evolution of nuclear genes from the Class Aves (Phylum Chordata), as part of the Avian Genome Consortium [174], aiming to characterize cytochrome P450 family 2 (CYP2) sub-families' genes and investigate the selection forces shaping these genes, in order to link between its genomic variations and different bird's life history traits (namely, migration and feeding habits);

- **Part III (chapter 5)** proposes a bioinformatics tool to streamline the analysis of the outputs from softwares widely used to investigate evidences of selection, by integrating, filtering and combining rapidly and efficiently the results, and then allowing the mapping of the positively selected sites into the 3D protein structure, using a single user-friendly interface.

Table 1.2: **Bioinformatics analyses tools and databases used in the studies presented in chapters 2-5.**

Bioinformatics Analyses	Tools/packages/databases	Chapters
Sequence collection	MITOZOA, GENBANK, Ensembl, GigaScience	2, 3, 4
Gene search	tBLASTn, BLASTx	4
Multiple sequence alignment	MUSCLE, GENEIOUS, PROMALS3D	2, 3, 4
Alignment concatenation	GENEIOUS	2
Alignment quality	GUIDANCE, TCS	3
Nucleotide substitution model	JMODELTEST	2, 3, 4
Nucleotide saturation	DAMBE	2, 3, 4
Phylogenetic analyses	PHYML, MRBAYES	2, 3, 4, 5
Phylogenetic chains/parameters convergence	TRACER, R	2, 3
Statistical evaluation of phylogenetic relationships	CONSEL	3
Ultrametric phylogeny	R8S	3, 4
Phylogeny partitioning supported by evolutionary model	Phytools and geiger R packages	3, 4
Phylogenetic visualization and analysis	ARCHAEOPTERYX	5
Selection analyses	PAML, DATAMONKEY, TREESAAP	2, 3, 4, 5
Protein structural analyses	PDB, I-TASSER, TMHMM, SWISS-MODEL, PYMOL, JMOL, GNUPLOT	2, 3, 4, 5
Recombination or gene conversion analyses	RDP4	4
Ancestral reconstruction analyses	COUNT, MEGA	3, 4
Statistical analyses on trait associations	MASS R package	4

The chapters included in each one of the aforementioned parts (**chapter 2-5**) represent the research that I carried out during my PhD program where I studied the *"Evolutionary genomics of genes involved in the environmental adaptation of metazoans"*. Each chapter has its own abstract, introduction, materials and methods, results, discussion and/or conclusion sections.

Part I

ADAPTIVE EVOLUTION OF MITOCHONDRIAL GENES IN THE METAZOA: INSIGHTS FROM CLASS CEPHALOPODA (PHYLUM MOLLUSCA)

Chapter 2

Adaptation of the Mitochondrial Genome in Cephalopods: Enhancing Proton Translocation Channels and the Subunit Interactions

Published in

Almeida D, Maldonado E, Vasconcelos V, Antunes A

PLoS ONE, 2015, 10(8): e0135405.

DOI: 10.1371/journal.pone.0135405

2.1 Abstract

Mitochondrial protein-coding genes (mt genes) encode subunits forming complexes of crucial cellular pathways, including those involved in the vital process of oxidative phosphorylation (OXPHOS). Despite the vital role of the mitochondrial genome (mt genome) in the survival of organisms, little is known with respect to its adaptive implications within marine invertebrates. The molluscan Class Cephalopoda is represented by a marine group of species known to occupy contrasting environments ranging from the intertidal to the deep sea, having distinct metabolic requirements, varied body shapes and highly advanced visual and nervous systems that make them highly competitive and successful worldwide predators. Thus, cephalopods are valuable models for testing natural selection acting on their mitochondrial subunits (mt subunits). Here, we used concatenated mt genes from 17 fully sequenced mt genomes of diverse cephalopod species to generate a robust mitochondrial phylogeny for the Class Cephalopoda. We followed an integrative approach considering several branches of interest – covering cephalopods with distinct morphologies, metabolic rates and habitats – to identify sites under positive selection and localize them in the respective protein alignment and/or three-dimensional (3D) structure of the mt subunits. Our results revealed significant adaptive variation in several mt subunits involved in the energy production pathway of cephalopods: ND5 and ND6 from Complex I, CYTB from Complex III, COX2 and COX3 from Complex IV, and in ATP8 from Complex V. Furthermore, we identified relevant sites involved in protein-interactions, lining proton translocation channels, as well as disease/deficiencies related sites in the aforementioned complexes. A particular case, revealed by this study, is the involvement of some positively selected sites, found in Octopoda lineage in lining proton translocation channels (site 74 from ND5) and in interactions between subunits (site 507 from ND5) of Complex I.

2.2 Introduction

The mitochondrial subunits of the respiratory chain complexes (mt subunits) – cytochrome c oxidase subunit 1-3 (COX1, COX2, COX3), cytochrome b (CYTB), NADH dehydrogenase subunit 1-6 (ND1, ND2, ND3, ND4, ND5, ND6), NADH dehydrogenase subunit 4L (ND4L), ATPase F0 subunit 6 (ATP6) and ATPase F0 subunit 8 (ATP8) – encoded by 13 mitochondrial

genes (mt genes) in most metazoans, are involved in several key evolutionary processes of eukaryotes having a major role in the production of energy [73]. The mt subunits interact with nuclear-encoded proteins establishing four of the five multi-subunit enzyme complexes – Complex I (ND1, ND2, ND3, ND4, ND4L, ND5 and ND6), III (CYTB), IV (COX1, COX2 and COX3) and V (ATP6 and ATP8) – involved in the respiratory chain of aerobic cells. Complex II consists entirely of nuclear-encoded proteins. Globally, Complexes I, II, III and IV are responsible for the mitochondrial OXPHOS pathway and the Complex V uses the generated energy gradient to synthesize adenosine triphosphate (ATP) [82].

The effect of any amino acid substitution can vary greatly depending on its location in the protein structure. Thus, site-specific amino acid substitutions that do not confer structural changes in the protein can affect protein function, such as the changes in binding sites. Some amino acid changes in mt subunits cause inefficiencies in the electron transfer chain system, contributing to the increase of reactive oxygen species (ROS) that can lead to the disruption of OXPHOS [76, 91, 92]. On the other hand, the amino acid substitutions can also improve or decrease aerobic capacity and be linked with life history traits and environmental adaptation [126–128].

Under strong selective pressure, organisms may evolve adaptations to increase their survival rate under certain ecological conditions. This can be achieved by changes in protein structure through amino acid substitutions, and/or by altering the amount of protein expression through modifications in the transcriptional regulation [181]. Dowling and colleagues highlighted the key importance of the mitochondrial-nuclear interaction as a unit of selection and the consequences of mitochondrial encoded fitness effects on several key evolutionary processes [82]. Furthermore, mt genes have been proposed to be connected with species adaptation to altitude [74] and temperature [75]. However, most of the studies on mt gene mutations that affect phenotypic variation, with positive selection shaping their evolution, have been conducted on vertebrates (mammals [76, 77] and fish [78]). Thus, there is little information with respect to the adaptation of the mt genomes of invertebrates, particularly marine invertebrates.

Cephalopoda (octopus, squid, cuttlefish, *Nautilus*) is the Class of the Mollusca phylum with the third highest number of species ($\sim$ 700 extant species) presenting remarkable morphological and physiological innovations [164]. Overtime, these marine invertebrate species have evolved several biological features such as a highly advanced visual and

nervous system, camouflage strategies, diversified body shapes, sizes and metabolic rates, making them highly competitive and efficient predators, which contribute to their successful worldwide dispersion [164]. Thus, cephalopods cover a wide range of marine depths, some species are found from the sea surface to 100m in depth (*Sepioteuthis lessoniana*) [182] and others between 600 to 1200m in depth (*Vampyroteuthis infernalis*) [183].

All these features make cephalopods an attractive group of species to study signatures of selection of the mitochondrial-encoded proteins. Therefore, we consider 17 mt genomes from cephalopod species – representing eight major taxonomic groups from the Class Cephalopoda – to evaluate the role of selection in the mt genes that may affect the adaptive features of cephalopods. We perform detailed molecular evolution analyses – in particular, branch [6, 47] and branch-site [48] specific likelihood analyses – to test several branches of interest related with species habitats (distinct depths), metabolic rates and physical traits (number of arms), where positive selection may be significantly influencing the evolution of cephalopod mt genes. We identify the location of positively selected sites in the Cephalopoda protein sequence alignments. Then, we analyze the amino acid physicochemical properties of these sites and discuss them in the context of well-studied respiratory chain complexes (through structure-based homology comparisons), in order to assess whether these properties might have changed in functionally important regions.

2.3 Materials and Methods

Sequence collection

The complete coding-sequences (CDS) of the 13-mt genes from 17 fully sequenced cephalopod mt genomes, covering the major taxonomic groups Oegopsida, Myopsida, Bathyteuthoidea, Sepiolidae, Sepiidae, Vampyromorpha, Octopoda and Nautiloidea (Appendices II: Table S1), have been retrieved from GENBANK [27] and MITOZOA [68, 178] databases.

Alignment and phylogenetic analyses

Since some Cephalopoda mt genomes have duplicated protein-coding genes (Appendices II: Table S2) – a feature only found in Oegopsida and Bathyteuthoidea species (Appendices II: Table S2) [184, 185]: with *cox1*, *cox2*, *atp6* and *atp8* genes duplicated

in *Dosidicus gigas* (NC_009734), *Sthenoteuthis oualaniensis* (EU660577), *Todarodes pacificus* (NC_006354), *Watasenia scintillans* (NC_007893), *Architeuthis dux* (FJ429092) and *Bathyteuthis abyssicola* (AP012225) and *cox3* that is also duplicated in all the above species except for *B. abyssicola* (Appendices II: Table S2) – the codon based CDS alignment of the aforementioned genes allowed the verification that the duplicated genes are highly conserved; some are exact copies and others present slight nucleotide substitutions (Appendices II: Table S2).

Thus, in order to find the best way to deal with duplicated genes to assess the phylogenetic relationships among the species of the Class Cephalopoda, we considered two data sets: one of them with 13 protein-coding genes where a sequence of a single copy of duplicated genes was included (the left side duplicated genes; see Appendices II: Figure S1) and other containing 13 protein-coding genes with the other copy of duplicated genes (the right side duplicated genes; see Appendices II: Figure S1). For each one of the mentioned data sets, individual codon based CDS alignments of the mt genes were performed using MUSCLE [31] integrated in the SEAVIEW software version 4.4.0 [186]. All the alignments were manually inspected and concatenated in GENEIOUS software version 5.6.7 [187].

The substitution nucleotide model that best fit each one of the Cephalopoda data sets was calculated using the Akaike Information Criterion (AIC), implemented in JMODELTEST software version 2.1.1 [37], starting with 11 substitution schemes and using the fixed BIONJ-JC option base tree for likelihood calculations. For both data sets, the selected model was the General Time Reversible (GTR) model, with a proportion of invariable sites (I) and heterogeneity of substitution rates among sites modeled following a gamma distribution (G). Additionally, we performed the Xia *et al.* statistic test [188], implemented in the DAMBE software version 5.3.31 [40], on each codon position of our data sets, in order to check for evidences of saturation bias that could derail phylogenetic/adaptive analyses. This test compares the saturation index expected when assuming full saturation (Iss.c, critic value) with the observed saturation index (Iss), assuming symmetrical (Iss.c Sym) and asymmetrical (Iss.c Asym) topologies, which did not showed significant evidence of saturation ($p < 0.05$) at the first and second codon positions. For the third codon positions, this test suggested that they are only marginally useful for phylogenetic reconstruction when the true tree topology is symmetrical and useless if the true tree topology is asymmetrical [188].

For each data set, phylogenies were estimated using the Maximum likelihood (ML) and

Bayesian (BA) inference methods. ML phylogenetic trees were built in PHYML software version 3.0 [41], with 1000 bootstrap replicates and the parameter 'best of Nearest Neighbour Interchange (NNI) and Subtree Pruning and Regrafting (SPR)' branch search algorithms. BA trees were inferred in MRBAYES software version 3.2 [42]. Two runs with two chains of the Markov Chain Monte Carlo (MCMC) method were performed for 10,000,000 generations, with sampling trees every 100 generations. The following criteria were used to test convergence of the chains: the standard deviation of split frequencies was less than 0.01 and the potential scale reduction factors (PSRF) were close to 1.0 for all parameters. Additionally, we used TRACER software version 1.6 [189] to assess if the chains have converged to its stationary distribution for all the parameters, and to better determine the appropriate burn-in (2,500,000 of the sampled trees were discarded). The effective sample size (ESS) for all the parameters was greater than 800 (ESS for the log-likelihood = 4232).

To assess the effect of possible saturation biases at the third codon position nucleotides on tree topology, we created modified data sets (RY coded alignments) in which all third codon positions were coded as purines (adenine and guanine: R) or pyrimidines (cytosine and thymine: Y) [190–193]. The RY coding of the alignments were performed according to Meiklejohn et al. (2014) [192]. These RY coded alignments were analyzed under the same parameters previously described for ML phylogenetic analyses.

Selection analyses

The effect of natural selection on the evolution of the 13-mt genes was assessed by comparing the number of non-synonymous substitutions per non-synonymous sites (*dN*) with that of synonymous substitutions per synonymous sites (*dS*), where negative/purifying selection is characterized by a *dN/dS* (ω) ratio < 1 and positive/diversifying selection by a $\omega > 1$, using: site, branch and branch-site specific likelihood analyses. For all the selection analyses the *atp6, atp8, cytb, cox1, cox2, cox3, nd1, nd2, nd3, nd4, nd4l, nd5* and *nd6* individual alignments and the cephalopod tree topology (unrooted), previously obtained in this study, were submitted to the CODEML program from the PAML 4.7 package [20, 43].

Site models use a statistical distribution to account for variation of the ω ratio among sites (codons) [18, 194]. In these analyses likelihood ratio tests (LRTs) compare the Maximum likelihoods of a null model that does not allow for any codons with $\omega > 1$ (M0, M1a, M7, M8a) against an alternative model that allows sites with $\omega > 1$ (M3, M2a, M8) [43], to find which

model fits the data significantly better. LRTs were conducted by comparing the following nested models: M0 (one ratio: assumes a constant ω ratio for all sites) *vs.* M3 (discrete: assumes 3 site classes for ω ratio estimated from the data) [195–198], M1a (nearly neutral: one ω class between 0 and 1, and one class of $\omega = 1$) *vs.* M2a (positive selection: same as M1a model plus an extra class of $\omega > 1$) [8, 18, 45], M7 (beta: assumes a beta distribution for $0 \leq \omega \geq 1$) *vs.* M8 (beta&ω: same as the M7 model plus an extra class of $\omega > 1$) [194] and M8a (beta&$\omega = 1$: same as M7 plus an extra class of $\omega = 1$) *vs.* M8 (beta&ω) [46]. Positive selection was indicated when a freely estimated ω parameter was greater than 1 and the LRT was significant.

Branch models test for different evolutionary rates in specific branches of the phylogeny [6, 47], comparing the average of foreground ω values (branches of interest, $\overline{\omega}_F$) with the average of background ω values (other branches, $\overline{\omega}_B$), using as null hypothesis $\overline{\omega}_F = \overline{\omega}_B$ and alternative hypothesis $\overline{\omega}_F \neq \overline{\omega}_B$. Following this test, if the alternative (unconstrained) model was found as a better fit to the data, then it needs to be tested against another null (constrained) model where $\overline{\omega}_F = 1$, in order to check for the prevalence of the previous selection result.

The branch-site specific likelihood analyses (modified branch-site model A) test for episodic positive selection on particular residues along the foreground-lineages [48], allowing ω to vary, both, among sites in the protein and across branches on the tree. For this test, the alternative model assumes that only the foreground-lineages (branches of interest) may have experienced positive selection, admitting three ω_F values ($0 < \omega_F < 1$, $\omega_F = 1$ and $\omega_F > 1$), and the null model constraints $\omega_F = 1$.

In branch and branch-site specific likelihood analyses the branches of the phylogenetic tree are divided *a priori* into foreground- (branch of interest) and background-branches (the remaining branches). In this study we tested if there are positive selection in any of 8 foreground-branches (**named A to H**). These analyses consider only one foreground branch at a time, thus we performed 8 analyses employing branch and branch-site approaches. More specifically, we tested for positive selection in: (**A**) genes of species living in deep waters; (**B**) squids that present high metabolic rates; (**C**) the Loliginidae family, having the highest metabolic rates relatively to all other lineages of the tree; (**D**) the Sepiidae group, with high metabolic rate and species living in shallow waters; (**E**) Octopoda, which has intermediate metabolic rates and species living in shallow waters; (**F**) Decapodiformes,

cephalopods with 10 arms and high metabolic rates; (**G**) Octopodiformes, cephalopods with 8 arms and low or intermediate metabolic rates; (**H**) *Nautilus macromphalus*, with low metabolic rate and dwelling in deep. Appendices II: Table S1 contains more details about the species features mentioned here.

For all these analyses we applied the F3x4 codon model [9] allowing for ML estimation of κ (transition/transversion ratio) and ω. All the models were run several times, adjusting the initial κ and ω values in order to avoid local optima. For all the model comparisons, the hypothesis decision was calculated as the double of the difference between the alternative and null model log-likelihood (2ΔlnL) and assuming that the null distribution of these results can be approximated by a chi-square (χ^2) distribution (*p*-value < 0.05). The number of degrees of freedom (*df*) was calculated as the difference of the number of estimated parameters between the nested models [8, 194]. Finally, when it was revealed that the alternative model fit the data better, the codon sites under positive selection were identified using the Bayes Empirical Bayes (BEB) calculation, which analyzes the posterior probabilities (PP) for these sites [45].

We also used the DATAMONKEY web server [19, 52] to assess codons under purifying selection [43, 45], implementing the fixed effects likelihood model (FEL, site-by-site analysis) [10] and to obtain further evidences of episodic diversifying selection employing the mixed effects model of evolution (MEME; branch-site analysis) [199], both inferred at *p*-value < 0.05 significance level. FEL model compares the instantaneous *dS* and *dN*, based on a codon-substitution model, without assuming a prior *dN/dS* [10]. This model was chosen since it is a rigorous method and tends to be less conservative and to have less false positives/negatives than other methods in data sets of intermediate size [19]. MEME model is capable of identifying instances of both, episodic and pervasive positive selection allowing the distribution of ω to vary across lineages at individual sites [199].

TREESAAP software [51] – designed to classify the impact of amino acid replacements on local physicochemical properties in eight magnitude categories, from the most conservative (1) to the most radical (8) [200, 201] – was employed to infer about the strong positive selection for a given physicochemical amino acid property. Thus, only changes of great magnitude (7 and 8 categories) at the *p*-value < 0.001 (Z-score $> |3.09|$) significance level were considered.

Protein structure analyses

In order to understand how the sites under positive selection, influence the structure and/or function of the mt subunits, we located them in the respective Cephalopoda protein sequence alignment. To perform this, we used the Alignment Filter tool of the IMPACT_S software [202]. These sites were identified (*p*-value < 0.05) by the BEB (CODEML branch-site model test) and MEME (DATAMONKEY episodic diversifying selection) analyses. We also mapped and analyzed the results displayed by the TREESAAP analyses (*p*-value < 0.001), for the same mt subunits, in order to get additional insight about the amino acid properties of common sites, identified by the previous methodological approaches.

Furthermore, to complement our analyses with the spatial position of these sites in a 3D plane, and since the X-ray crystal structures of the mitochondrial encoded proteins from cephalopods are not available on the Protein Data Bank (PDB) [26], we modeled its 3D structures for the *Octopus vulgaris* species, from ND5 (YP_112440.1), ND6 (YP_112444.1), CYTB (YP_112443.1), COX2 (YP_112437.1), COX3 (YP_112433.1) and ATP8 (YP_112438.1), using the I-TASSER server with default parameters [53]. This species was chosen due to its worldwide dispersion and because it is one of the cephalopod species widely studied and known by most of the world population. We have only modeled the 3D structure of proteins for which evidence of positive selection was detected. I-TASSER is a platform that allows the automatic generation of high-quality predictions of 3D structure and biological function of proteins from their amino acid sequences, using PDB templates or by *ab initio* modeling approaches, where the assembly follows several Monte Carlo simulations [53]. The quality of the generated model is evaluated considering a confidence score (C-score) value, which is calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations [53]. Also a score scale is attributed to the structural similarity between two structures (T-score) in order to guarantee that the resulting topology is not random. Following the author instructions to evaluate the reported scores [53], the most likely structure for each of the six mt subunits modeled was retrieved from I-TASSER. Then, we performed the superimposition of available X-ray crystal structures – ND5 (*Escherichia coli* PDB: 3RKO), CYTB (*Bos taurus* PDB: 1PPJ and *Saccharomyces cerevisiae* PDB: 1P84), COX2 (*B. taurus* PDB: 1V54), COX3 (*Rhodobacter sphaeroides* PDB: 1M56) – with their corresponding 3D structures of *O. vulgaris*, in which the positively

selected sites were mapped. The superimposition, visualization and manipulation of the 3D structures were performed with PYMOL software version 1.5.0.4 [58]. When, the obtainment of a reliable 3D model was not satisfactory, which were the cases of the ND6 and ATP8, insight into the evolution of the mt subunits in cephalopods was obtained by inspection of the amino acid substitutions in their protein sequence alignments. Additional information about the topology and orientation of the transmembrane (TM) domains of these subunits was obtained using the TMHMM server version 2.0, which is a membrane protein topology prediction method based on a hidden Markov model [203].

Since the mt subunits of the respiratory chain complexes are highly conserved among species (sharing the same cofactors and performing the same functions) [78], we searched if the positively selected sites, present in the sequence alignments of cephalopods, are related with important functional and/or mutational sites reported in the literature (*e.g., Homo sapiens*, *E. coli*, *B. taurus*, *S. cerevisiae* and *R. sphaeroides*) [101, 204–223]. Therefore, through profile alignments (using GENEIOUS software version 5.6.7 [187]), we performed homology analyses of the mt subunits from our Cephalopoda data sets with their well-studied counterparts in other species. The details of these analyses for each one of the mt subunits, with positively selected sites, are described in the following tables: ND5 (Appendices II: Table S3), ND6 (Appendices II: Table S4), CYTB (Appendices II: Table S5), COX2 (Appendices II: Table S6) and COX3 (Appendices II: Table S7).

2.4 Results

Phylogenetic analyses

Regardless of the data set used (with either copy of duplicated genes and their RY coded versions), the Cephalopoda tree topology obtained was the same (Figure 2.1). The ML and BA analyses for the 17 cephalopods consistently supported the Decapodiformes clade, with bootstrap percentages (BP) of 86% (from the original alignments), 99% from the RY coded alignments (99%RY) and Bayesian PP of 1 (Figure 2.1). With respect to the intra-relationships of Decapodiformes, these analyses strongly supported a close relationship between the Oegopsida species *D. gigas*, *S. oualaniensis* and *T. pacificus* (BP of 99%, BP of 99%RY, and PP of 1), contrasting with the relationships of the species *W. scintillans* and *A. dux* that are not strongly supported (Figure 2.1). Moreover, these analyses also

revealed a high support to the species relationships within the Myopsida clade (BP of 99%, BP of 100%RY, and PP of 1), as well as, to the sister relationship of Sepiidae to all other Decapodiformes (BP of 86%, BP of 99%RY, and PP of 1) (Figure 2.1). The topology obtained also suggested that *Semirossia patagonica* is sister taxon to the clade containing Myopsida, Bathyteuthoidea and Oegopsida (BP of 82%, BP of 91%RY, and PP of 1) (Figure 2.1). Furthermore, the Octopoda clade was highly supported (BP of 100%, BP of 100%RY, and PP of 1) and the monophyly of Octopoda and Vampyromorpha was suggested (Figure 2.1).

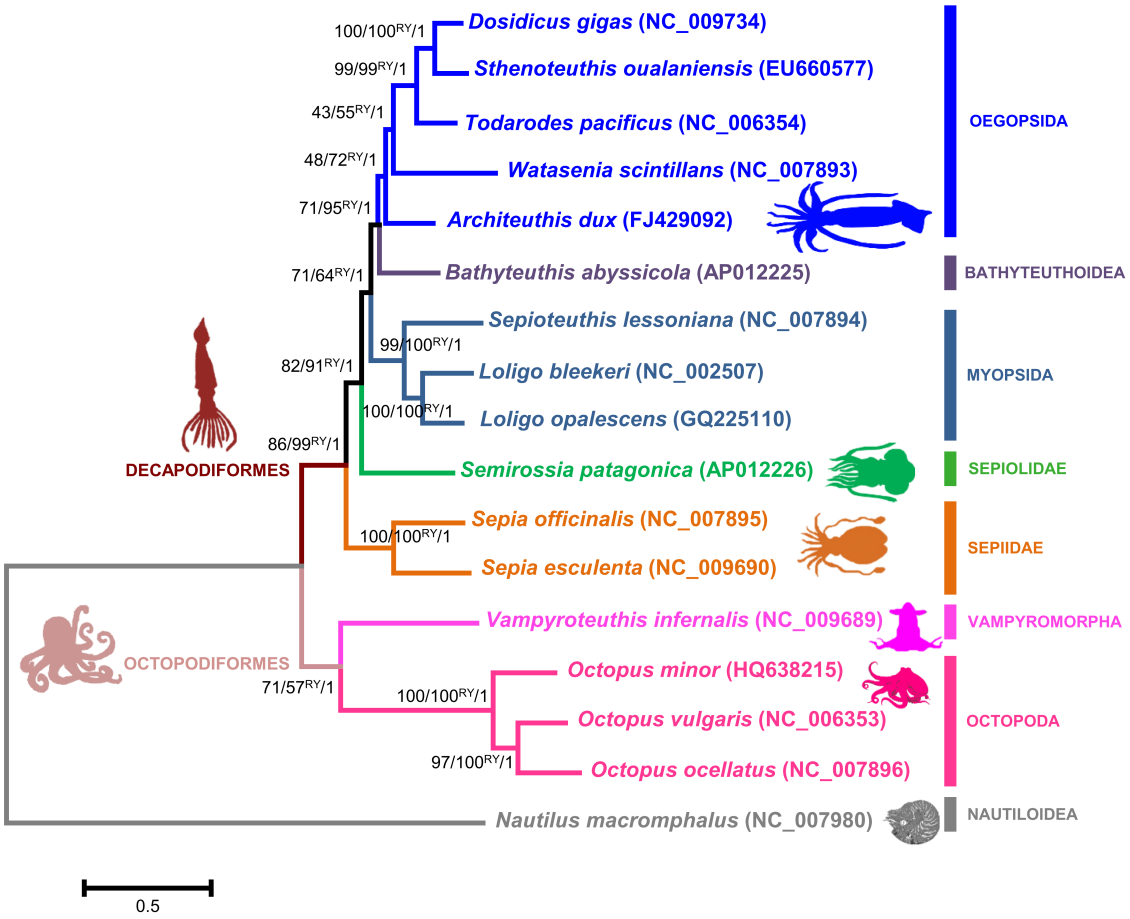


Figure 2.1: Cephalopoda mitogenomic consensus phylogenetic tree. The topology shown corresponds to the Maximum likelihood (ML) tree. The bootstrap probabilities (%) supporting each node were estimated with ML analyses using PHYML software version 3.0 (1000 bootstrap replicates, GTR+G+I model and best of NNI&SPR branch search algorithm) and are shown by the first and second values next to the branches. The RY symbol indicates that these values resulted from the RY coded alignments. The posterior probability supporting each node was estimated from Bayesian analyses using MRBAYES (GTR+G+I model; 10,000,000 generations; a sample frequency of 100 and a burn-in corresponding to 25% of the sampled trees) and corresponds to the third value next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Superorders – Decapodiformes (10 arms) and Octopodiformes (8 arms). The species are designated by their scientific names followed by the respective accession numbers. On the right, are indicated the names of the major taxonomic groups of cephalopods.

Purifying selection in Cephalopoda mitochondrial genes

Employing site-specific analyses – site models (CODEML) and FEL analyses (DATAMONKEY web server) – we were able to infer the global molecular evolutionary rates acting in each one of the 13-mt genes of our Cephalopoda data set. In the site model analyses from CODEML, for all the mt genes, the null model was rejected only in the M0 *vs.* M3 test (*p*-value < 0.05) (Appendices II: Table S8). This means that for all the analyzed mt genes, the $\overline{\omega}$ ratio differs among sites. These results also indicated that the 13-mt genes of cephalopods are globally evolving under negative constraints, with a few percentage evolving under neutrality. However, since these models are not able to report which codons are under negative selection, we also performed FEL site-by-site analyses (*p*-value < 0.05). According with these analyses, the three mt subunits (COX1, COX2 and COX3) belonging to Complex IV presented the highest percentage of codons under negative selection – with COX1 subunit standing out with 89% of codons under negative selection (Figure 2.2A) – as well as, the greatest efficient purifying selection (smaller $\overline{\omega}$ value) (Figure 2.2B). CYTB from Complex III also revealed a high value of purifying selection contrasting with the subunits from Complexes I and V, which showed more relaxed selection (Figure 2.2B). In particular, the subunits having the highest percentage of codons under negative selection present the following order: COX1 $>$ COX3 $>$ COX2 $>$ ATP6 $>$ ND1 $>$ CYTB $>$ ND3 $>$ ND5 $>$ ND4 $>$ ND6 $>$ ND2 $>$ ND4L $>$ ATP8 (Figure 2.2A). Additionally, it was possible to verify that the subunits under stronger purifying selection (smaller $\overline{\omega}$) display the following order: COX1 $>$ COX2 $>$ CYTB $>$ COX3 $>$ ND1 $>$ ATP6 $>$ ND3 $>$ ND5 $>$ ND4L $>$ ND2 $>$ ND4 $>$ ND6 $>$ ATP8 (Figure 2.2B). Both analyses confirmed, similarly to vertebrate species, the great influence of purifying selection in the 13-mt subunits of our Cephalopoda data set (Figure 2.2 and Appendices II: Table S8).

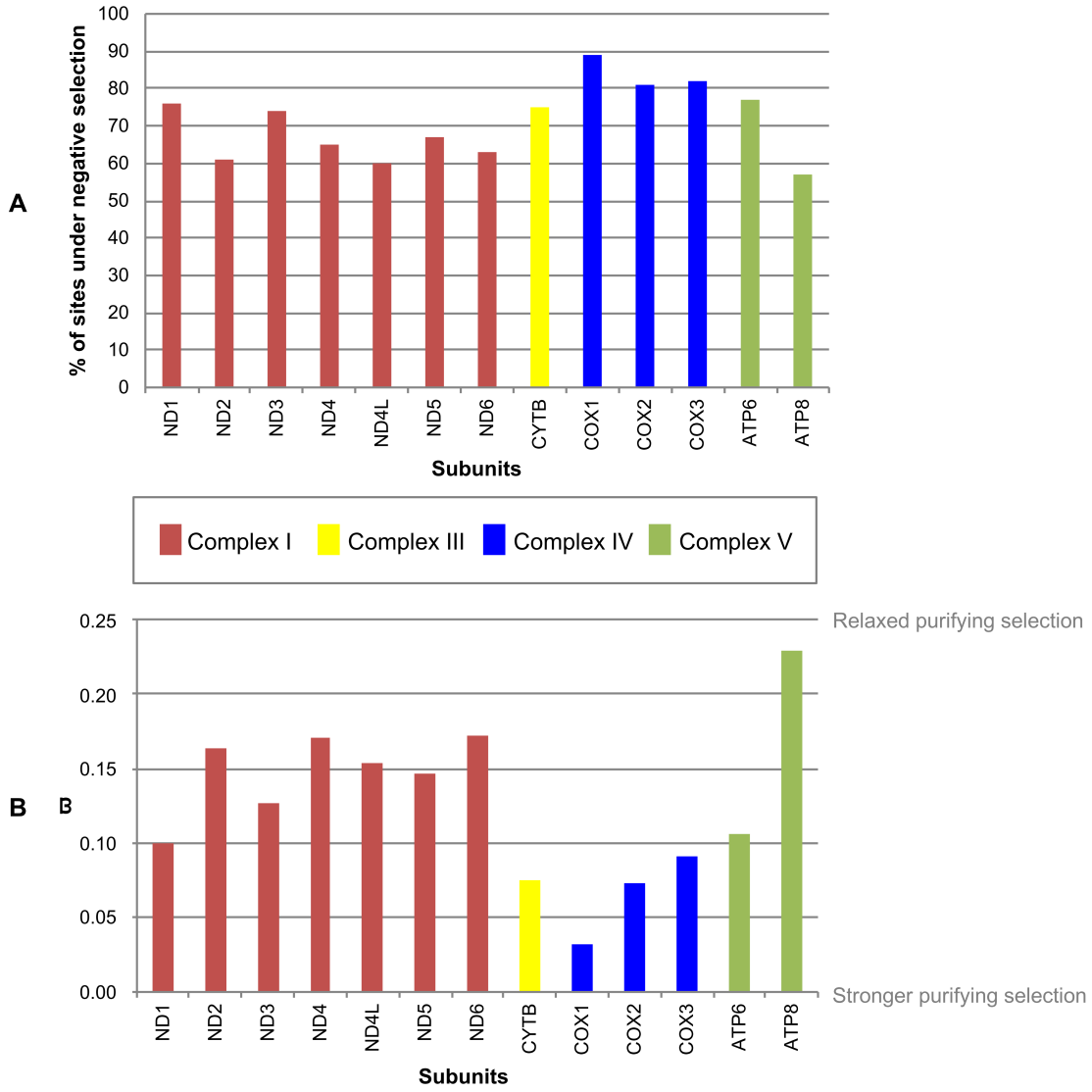


Figure 2.2: Negative selection reported by FEL site-by-site analyses. (**A**) Percentage of codons under negative selection in cephalopod mitochondrial subunits. (**B**) Cephalopoda $\overline{\omega}$ estimates for each one of the mitochondrial subunits.

Evidence of natural selection in Cephalopoda mitochondrial genes

From the employed branch (CODEML) and branch-site specific (CODEML from PAML and MEME from DATAMONKEY web-server) analyses, considering several branches of interest (foreground-branches), only the branch-site specific analyses (*p*-value < 0.05) uncovered events of positive selection along the ML mitochondrial phylogeny of cephalopods (Figure 2.3).

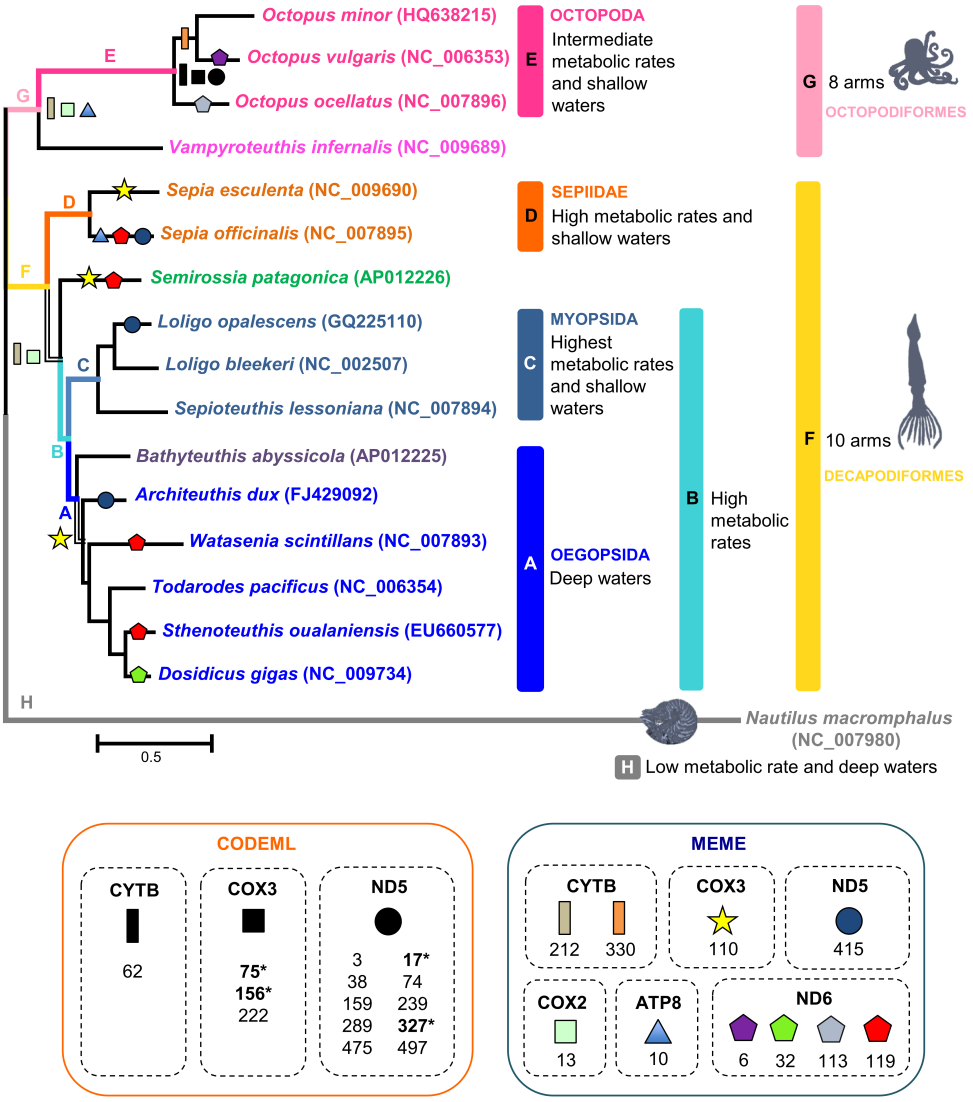


Figure 2.3: Foreground-branches tested for both branch-specific and branch-site selection models. Identification of the Maximum likelihood tree branches to test the adaptive evolution for each one of the 13-mitochondrial genes in 17 representative species from the molluscan Class Cephalopoda. The letters indicate the branches of interest (foreground-branches: named **A to H**). We performed 8 tests, where only one of the branches pointed by the letters was considered at a time; all other branches are corresponding to background-lineages for the analyzed gene. The sites represented correspond to positively selected sites by the employed approaches (CODEML and MEME). The numbers in bold, with an asterisk mark, represent sites that obtained posterior probabilities $\geq 99\%$ (p-value < 0.05) and the other numbers, posterior probabilities $\geq 95\%$ (p-value < 0.05), respectively.

Branch analyses

For each one of the foreground-branches tested (**A to H**) (Figure 2.3), branch model analyses from CODEML (p-value < 0.05) revealed that globally (Appendices II: Table S9): (**A**) the mt genes of species living in deep waters, compared with species inhabiting different

depths, are evolving under negative selection, with the exception of *cox2* and *cox3* which appeared to be evolving under relaxed selective constraints; (**B**) the mt genes of squid species (with high metabolic rates) when compared with the remaining groups (distinct metabolic rates and depths) are also evolving under negative selection, with the exception of the *cox2*, *cox3*, *cytb* and *nd1*, evolving under relaxed selective constraints; (**C**) the Loliginidae family, with the highest metabolic rate, presents the *cox1*, *cox2*, *cox3* and *nd5* genes evolving under relaxed selective constraints and the remaining nine genes, evolving under negative selection, at the same rate, in all the lineages of cephalopods; (**D**) in Sepiidae group (high metabolic rate and species living in shallow waters) all the mt genes are evolving under negative selection, at the same rate of the other cephalopods, with the exception of the *cox1*, *cox2*, *cox3*, *cytb* and *nd5* genes that are evolving under relaxed selective constraints; (**E**) in Octopoda species (intermediate metabolic rates and shallow waters) all the genes are under negative selection (similar rates to the other species), except the *cox1* that is evolving neutrally; (**F**) in Decapodiformes all the genes are evolving under negative selection (the same rate of Octopodiformes), except for *cox1* and *cox3* evolving neutrally; (**G**) in Octopodiformes all the genes are evolving under negative selection, except the *cox1* that is evolving neutrally; (**H**) in the *N. macromphalus* species, all the mt genes are evolving under negative selection, at the same rate of all the other cephalopods (branches), with the exception of the genes *cox1* and *cytb* which have evolutionary rates ($\overline{\omega}cox1 = 0.036$ and $\overline{\omega}cytb = 0.002$) distinct from the other cephalopods.

Branch-site analyses

From the branch-site model tests of CODEML (Appendices II: Table S10), the BEB analyses (PP $\geq 95\%$) were able to identify a few codons under positive selection (*p*-value < 0.05) for the genes *nd5, cytb* and *cox3* in the ancestral lineage of the Octopoda group (foreground-branch E) (Figure 2.3 and Appendices II: Table S10). Additionally, MEME analyses (PP $\geq 95\%$) also reported sites with evidence of episodic diversifying selection (*p*-value < 0.05) in the following genes: *nd5, nd6, cytb, cox2, cox3* and *atp8* (Figure 2.3). It is worth highlighting that MEME analyses revealed the presence of codons under positive selection in *cytb, cox2* and *atp8* genes in the ancestral lineage of Octopodiformes (Figure 2.3 - foreground-branch G). Assuming as reference the protein sequences of the *O. vulgaris*, the positively selected sites were the following (Figure 2.3): from CODEML *(i)* ND5 subunit -

3, 17, 38, 74, 159, 239, 289, 327, 475, and 497, CYTB subunit - 62, and COX3 subunit - 75, 156, and 222; and from MEME *(ii)* ND5 subunit - 415, ND6 subunit - 6, 32, 113, and 119, CYTB subunit - 212 and 330, COX2 subunit - 13, COX3 subunit - 110, and ATP8 subunit - 10 (Figure 2.3).

Mapping of the positively selected and other functional relevant sites

The implemented methodological approaches revealed sites under positive selection in only six of the 13 protein-coding genes considered in this study: *nd5, nd6, cytb, cox2, cox3* and *atp8* (*p*-value < 0.05).

ND5: The sites identified as positively selected by branch-site analyses (*p*-value < 0.05) were mapped in the Cephalopoda ND5 protein sequence alignment (Cephalopoda ND5 data set) (Figure 2.4A) and into the predicted *O. vulgaris* 3D protein-structure (Figures 2.4B and 2.4C), allowing to localize the sites 3, 17, 74, 159, and 497 (CODEML) in TM helices, the site 38 in a β sheet (CODEML), and the sites 239, 289, 327, 475 (CODEML) and 415 (MEME) in loop regions (Figure 2.4). The profile alignment of the Cephalopoda ND5 data set with the structure-based alignment from a recent study (using 30 representatives from all kingdoms of life: vertebrates, plants, algae, bacteria, yeasts, fungi and cnidarians) [101] (Appendices II: Table S3), indicated that in cephalopods this protein has 15 TM helices. This was also illustrated by the superimposition of the predicted *O. vulgaris* 3D protein-structure with the X-ray 3D structure of the NuoL (a counterpart of the ND5), from the bacterium *E. coli* (PDB: 3RKO), which revealed that the TM missing in cephalopods corresponds to the TM1 of the bacterium (Figure 2.4). Among the mentioned sites, only the 74 and the 497 presented a substantial level of conservation (Figure 2.4A) in the Cephalopoda ND5 data set. Noteworthy that the Octopoda species have a positively selected cysteine (C) amino acid in the position 74 of TM2, instead of a methionine (M) amino acid (M74C), which is present in the remaining cephalopods, with exception of the *N. macromphalus* (Figure 2.4A). Thus, Octopoda species have a polar uncharged amino acid, soluble in water, instead of a non-polar hydrophobic one, present in the other cephalopods. Our homology comparisons (Figure 2.4A) suggested that this site – homologous of the site H100 in *E. coli* [101] (Appendices II: Table S3) – is part of the proton translocation channel, across the inner mitochondrial membrane to enable ATP production, being one of the three amino acids that form its main link to the mitochondrial matrix in cephalopods (Figure 2.4B).

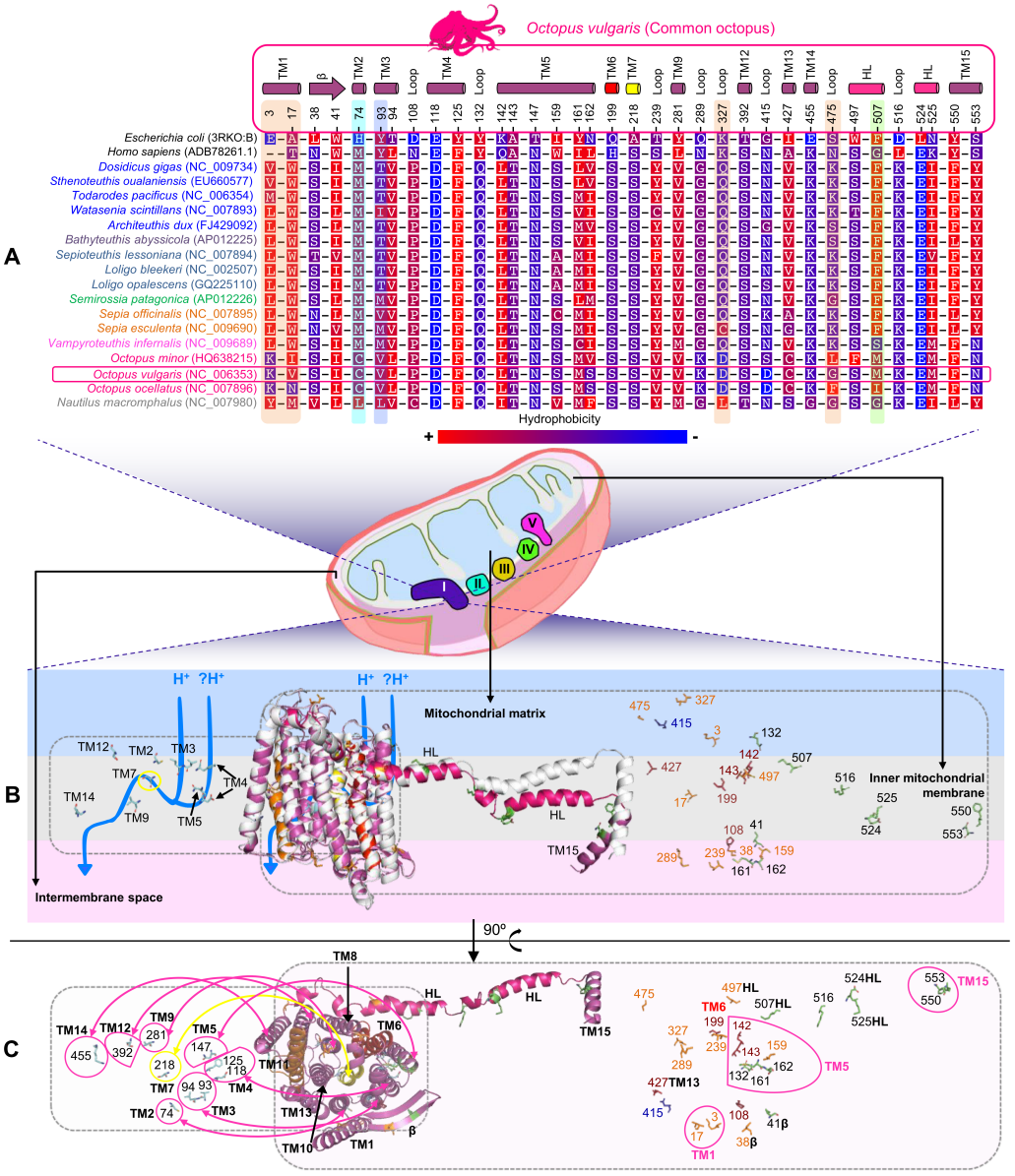


Figure 2.4: Amino acid variation in functional and structural sites of Cephalopoda ND5 subunit. (**A**) Amino acid alignment for sites *(i)* under positive selection – 3, 17, 38, 74, 159, 239, 289, 327, 475, 497 (CODEML branch-site model test) and 415 (MEME), *(ii)* sites involved in interactions between subunits – 41, 132, 161, 162, 507, 516, 524, 525, 550, 553 *(iii)* sites lining proton translocation channels – 74, 93, 94, 118, 125, 147, 218, 281, 392, 455 and *(i)* described mutations – 108, 142, 143, 199 and 427. All the sites have as reference number, the site position from the *Octopus vulgaris* ND5 protein-sequence, highlighted with a magenta rectangle shape. Sites highlighted by: *(i)* an orange shape – identified by CODEML branch-site model test and employing TREESAAP analysis; *(ii)* cyan shape – site lining proton translocation channel and reported by CODEML branch-site model test; *(iii)* blue shape – site lining proton translocation channel and positively selected by TREESAAP; *(iv)* green shape – site involved in subunit interactions (ND5-ND4) and also positively selected by TREESAAP. Background amino acid colors represent a hydrophobicity scheme from the most hydrophobic (red) to the most hydrophilic (blue). (**B**) Superimposition of 3D structures of ND5 from *Octopus vulgaris* (magenta) and NuoL from *Escherichia coli* (white), viewed from inner mitochondrial membrane – map of all the previously mentioned sites. At the left are projected the sites (cyan sticks) lining the proton translocation channels (blue arrow): H^+ – putative channel and $?H^+$ – suggested channel at the interfaces of ND5 and ND4 subunits. At the right, the projected sites are involved in interactions between subunits (green sticks), described mutations (red sticks) and the sites under positive selection (CODEML – orange sticks; MEME – blue sticks). (**C**) View from the intermembrane space. TM – transmembrane helices: helices that are probably involved in conformational changes are shown in red (TM6 – discontinuous helix), orange (TM11 – discontinuous helix) and yellow (TM7 – discontinuous helix). Connecting elements: β–β-hairpins and HL–HL helix (hot pink).

The following sites – 3 (TM1), 159 (TM5), and 239 (loop) – presented average amino acid conservation (Figure 2.4A). Additionally, TREESAAP analysis (*p*-value < 0.001) also reported that the site 3 presents the amino acid property polarity under positive selection. The sites 17 (TM1) and 475 (loop) showed more variable amino acid changes (Figure 2.4A). Accordingly, TREESAAP analysis (*p*-value < 0.001) revealed the following amino acid properties under positive selection: polarity at the codon 17 and chromatographic index, hydropathy and surrounding hydrophobicity at the codon 475. The codon 17 is variable (Figure 2.4A and Appendices II: Table S3) among the 30 representative species of all kingdoms of life [101], with contrasting hydrophobicity and polarity. In our Cephalopoda ND5 data set the hydrophobicity of the amino acids in the referred position is well conserved (exception *Octopus ocellatus*) (Figure 2.4A). Furthermore, the tryptophan (W17) amino acid is present in all Decapodiformes, which could be suggestive of an important adaptive role in the referred group (Figure 2.4A). The remaining positively selected sites, 38 (β sheet), 289 (loop), and 327 (loop), present average conservation (Figure 2.4A) and the TREESAAP analysis (*p*-value < 0.001) was able to identify the polarity property as positively selected at the codon 327. The amino acid present in the site 415 (loop) is highly variable, being mostly hydrophilic (Figure 2.4).

Additionally, through homology comparisons based on the sites described by Efremov & Sazanov (2011) (for *E. coli* and/or *H. sapiens*) [101], which are involved in interactions between subunits, mutations and residues lining proton translocation channels, we were able to determine the sites that present different amino acids in homolog positions of cephalopods (Figure 2.4 and Appendices II: Table S3). It is worth highlighting two of these sites, 93 and 507, to which the TREESAAP (*p*-value < 0.001) revealed amino acid properties under positive selection. The codon 93 (homologous to Y119 of *E. coli*) is referred as a possible link of the putative proton translocation channel at subunits interface [101] (Figure 2.4A and Appendices II: Table S3). This suggested function could be maintained in Oegopsida (with exception of *W. scintillans* species), Bathyteuthoidea and Myopsida, since at this position these species have a polar threonine (T93) amino acid that may participate in hydrogen bonds (Figure 2.4A). However, in the case of the remaining cephalopods that have non-polar amino acids (M, V-valine and L-leucine) (Figure 2.4A), incapable of forming hydrogen bonds, this interface link could be compromised. The TREESAAP analysis (*p*-value < 0.001) revealed, for this site, the following properties under positive selection: solvent accessible

reduction ratio and surrounding hydrophobicity; being in accordance with the previously described amino acid substitutions. The codon 507 is located in a structural element (Figure 2.4 - HL helix) – which extends along nearly the entire length of the hydrophobic membrane-embedded domain of Complex I, composed by the seven NADH dehydrogenase mt subunits – that may act as a connecting element (Appendices II: Figure S2), coordinating conformational changes. For this site, the TREESAAP analysis (*p*-value < 0.001) indicated the amino acid properties bulkiness, solvent accessible reduction ratio and surrounding hydrophobicity, as positively selected.

ND6: For the ND6 subunit was not possible to acquire a satisfactory 3D model. Therefore, insight into the evolution of this mt subunit in cephalopods was obtained by inspection of the amino acid substitutions, positively selected by MEME (*p*-value < 0.05) and TREESAAP (*p*-value < 0.001), observed in the sequence alignment of the cephalopods (Figure 2.5). Furthermore, sites relevant for the proton translocation and stability of the complex [101] were also considered (Figure 2.5 and Appendices II: Table S4). Employing the MEME analysis (*p*-value < 0.05) was possible to detect the following sites – 6 (TM1), 32 (TM2), 113 (loop), and 119 (loop) – evolving under positive selection in Cephalopoda species. The site 6 presents amino acids varying between V, M, L and tyrosine (Y) (Figure 2.5). All of these amino acids have hydrophobic side chains, however only the Y amino acid can form hydrogen bonds. The profile alignment of the Cephalopoda ND6 data set with the structure-based alignment containing the NuoJ (*E. coli*: counterpart of Cephalopoda ND6) [101] (Appendices II: Table S4) suggests that the site 32 is adjacent to amino acids close to a fourth proton translocation channel (Figure 2.5 - cyan shape) at the interface of subunits ND2, ND4L, ND6 and ND3 [101]. At this site, the most drastic amino acid change occurs in the *A. dux* and *S. patagonica*, both with a serine (S) amino acid. This is a small hydrophilic amino acid capable of forming hydrogen bonds, contrasting with the V, M, L and isoleucine (I) of the remaining cephalopods (Figure 2.5). Observing the Figure 2.5 it is possible to see that the site 113 (loop), is mainly dominated by hydrophilic amino acids with the exception of four species of Decapodiformes, that have a Y, phenylalanine (F) or M in this position. Considering the predicted topology (Appendices II: Figure S3), this site is facing the mitochondrial matrix. The site 119 seems to be a hyper variable site (Figure 2.5 - green shape), comprising amino acids with distinct properties and sizes (cephalopods: S, F, asparagine-N, Y, C, or glycine-G). Accordingly, TREESAAP analysis (*p*-value < 0.001) also revealed the following amino acid

properties as being under strong positive selection: hydropathy, solvent accessible reduction ratio, polarity and chromatographic index.

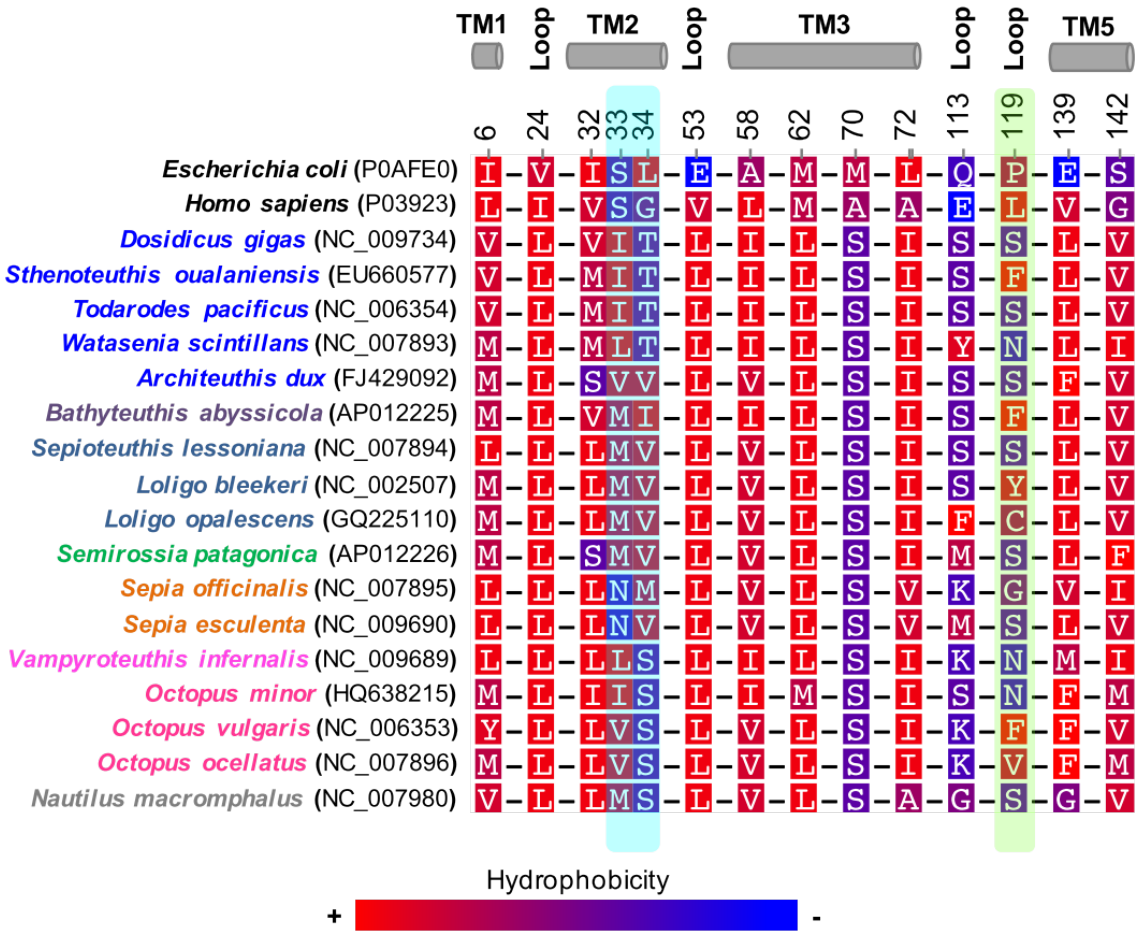


Figure 2.5: Amino acid variation in functional and structural sites of Cephalopoda ND6 subunit. Amino acid alignment for sites: *(i)* under positive selection – 6, 32, 113 and 119 (MEME); *(ii)* at the surface – 24 and 72; *(iii)* lining proton translocation channels – 33, 34, 53, 58, 62, 70, 139 and 142. All the sites have as reference number the site position from the *Octopus vulgaris* ND6 protein-sequence. The sites lining proton translocation channel also identified by TREESAAP are highlighted with a cyan shape and with a green shape the sites positively selected by MEME and also by TREESAAP. Background amino acid colors represent a hydrophobicity scheme from red – the most hydrophobic to blue – the most hydrophilic. TM – transmembrane helices. TM and loops – according to the topology TMHMM prediction.

Furthermore, our analysis suggested that the site L24, conserved in all the cephalopods (Figure 2.5), is a counterpart of the site 26 of *H. sapiens* (Appendices II: Table S4), for which a described mutation I26M was associated with [208]: loss of vision, caused by the Leber Hereditary Optic Neuropathy (LHON) disease and severe Complex I deficiency. Its homolog site in *E. coli* (V26) (Figure 2.5 and Appendices II: Table S4) is located at the surface and is suggested to be near the probable quinone binding site (Q-site), formed at the interface of

subunits NuoH (ND1), NuoJ (ND6) and NuoA (ND3) with the hydrophilic peripheral domain of Complex I [224]. We also found that the sites 33 and 34 (Figure 2.5 - cyan shape), homologs of S35 and L36 of *E. coli* (Appendices II: Table S4), can be lining a proton translocation channel [101]. These sites were also reported by TREESAAP, which indicated the following amino acid properties, as being under strong positive selection (*p*-value < 0.001): *(i)* site 33 - polarity and *(ii)* site 34 - surrounding hydrophobicity and solvent accessible reduction ratio. Additionally, it was described the mutation G36S in *H. sapiens*, which occurs in a site homolog of the codon 34 in cephalopods (Appendices II: Table S4). This mutation was associated with the LHON disease [222] and consists in the replacement of a non-polar residue (G) by an amino acid with a polar uncharged side chain (S).

CYTB: The positively selected sites found by CODEML branch-site model test (62) and MEME analyses (212 and 330) (*p*-value < 0.05), were mapped in the corresponding alignment (Figure 2.6A) and into the predicted 3D protein-structure (Figure 2.6B). The mapping of these sites allowed to verify that the site 62 is well conserved in cephalopods, except for the *Sepia esculenta*, Octopoda species and the *N. macromphalus* (Figure 2.6A - orange shape). Overall, despite the amino acid replacements, the hydrophobicity of this site is maintained. The exception is the Nautiloidea which presents in this site an amino acid, glutamic acid (E), highly hydrophilic. Observing the CYTB 3D protein-structure (Figure 2.6B - orange shape) it was also possible infer the location of this site in a helix named *ab*, outside the membrane and parallel to the membrane plane. This is an amphipathic helix and thus the amino acid replacements verified do not cause detrimental structural changes, being possibly advantageous for these species. Both sites, 212 and 330, have amino acid substitutions conserved in subgroups of cephalopods, with distinct metabolic rates, all keeping the hydrophobicity of the sites (Figure 2.6A - blue shapes). The 3D protein-structure revealed that the site 212 is located in a loop region facing the mitochondrial matrix and the site 330 is located in the TM helix G (Figure 2.6B - blue shapes). Furthermore, we mapped in the corresponding alignment (Figure 2.6A) relevant functional binding sites, obtained through homology comparisons with well-studied CYTB X-ray crystal structures from *B. taurus* [210, 223] and *S. cerevisiae* [221], as well as, mutations related with *H. sapiens* exercise intolerance [215]. These analyses demonstrated the high degree of conservation among CYTB sequences in cephalopods (Figure 2.6A), even at sites for which have been reported exercise intolerance associated mutations in humans [215] (Figure 2.6A - mapped

in maroon). Examples of extreme conservation are the following relevant sites, inferred in cephalopods: F129, Y132, E272, and Y279 (Figure 2.6A and Appendices II: Table S5). The site 129 presumably stabilizes the hydrophobic tail of the ubiquinol, proposed to center quinol oxidation (Q_o) pocket [221]. The hydroxyl group of the tyrosine at the site 132, located between E272 and the prosthetic group low-potential *b* heme (b_L), is thought to stabilize the proton transfer pathway [221]. The site 272, located in the transition between the TM F1 and a loop region, is proposed to be a ligand for ubiquinol and to accept protons released during ubiquinol oxidation [221]. The residue 279 is proposed to pre-orient the substrate [221].

Accordingly, TREESAAP did not showed properties under positive selection (*p*-value < 0.001) for any of these sites. However, by inspection of other relevant sites in the sequence alignment, which are forming the Q_o site, ubiquinone reduction (Q_i) site and the b_L binding sites [210, 223], we found amino acid substitutions in some of the mentioned sites (Figure 2.6A - marked with arrows). The *V. infernalis* has two atypical amino acid replacements at positions 28 and 36 (Figure 2.6A - yellow arrows) that are conserved in all the other members of Class Cephalopoda. Both residues are located at the Q_i site (Figure 2.6 and Appendices II: Table S5), suggesting that these amino acid replacements will affect the binding of ligands at this site. Furthermore, the sites 18 and 195 (both from Q_i site) evidenced amino acid changes conserved in subgroups of cephalopods, where the hydrophobicity is changed (Figure 2.6A - yellow arrows). The site 161 belonging to the Q_o site (Appendices II: Table S5), has a L conserved in all the Octopodiformes, alternating between an I/V in the other cephalopods, or M in *N. macromphalus* (Figure 2.6A - orange arrow). The *Octopus minor* species presents an atypical amino acid replacement, in cephalopods, in the sites 107 and 207 (Figure 2.6A - black arrows), both belonging to the prosthetic group high-potential *b* heme (b_H) binding site (Appendices II: Table S5). Similarly, the *A. dux* is the only species having a G instead of S at the site 206 (b_H binding site) (Figure 2.6A - black arrow). The sites 99 and 198 (Figure 2.6A - black arrows), both b_H binding sites, present amino acid substitutions keeping the hydrophobicity of the sites.

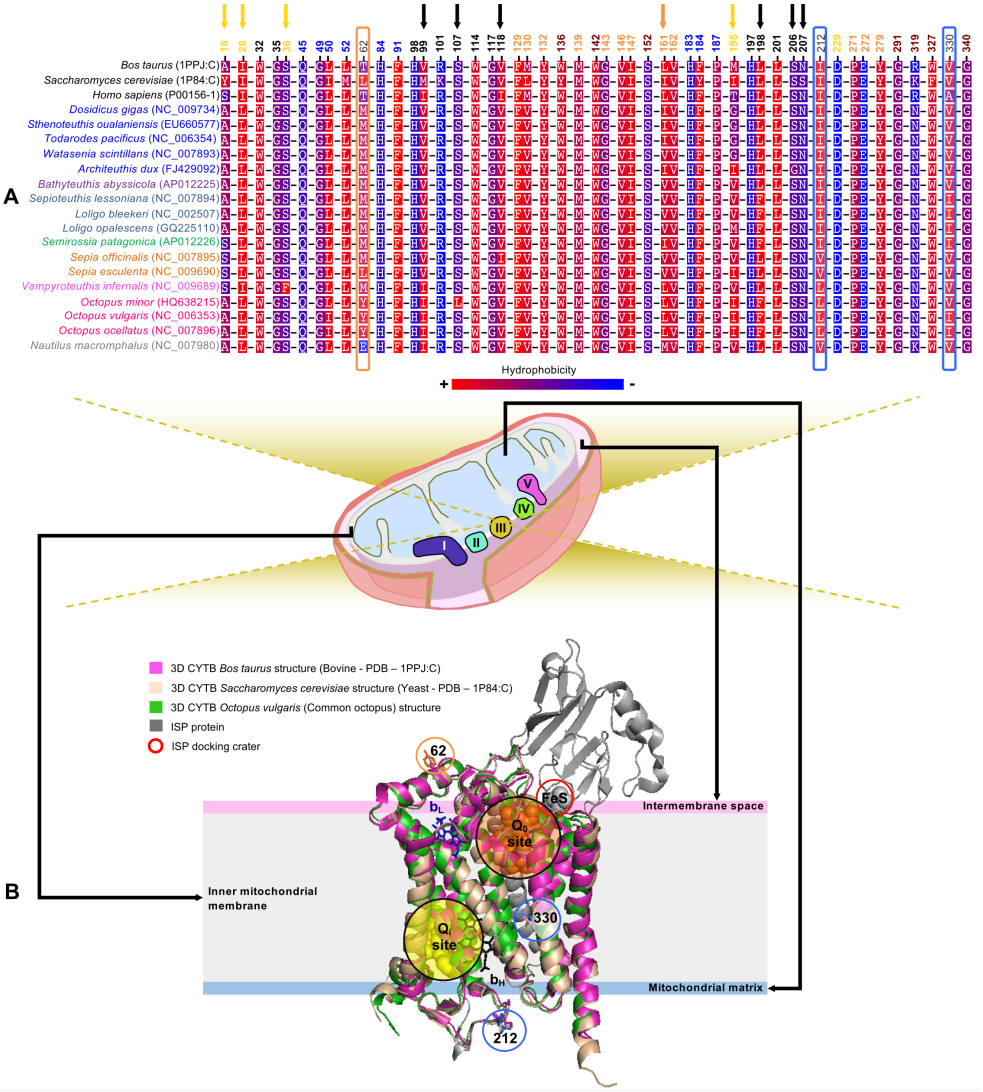


Figure 2.6: Amino acid variation in functional and structural sites of Cephalopoda CYTB subunit. (**A**) Amino acid alignment for sites *(i)* under positive selection – orange shape (CODEML branch-site model test) and blue shape (MEME), *(ii)* mutations related with human exercise intolerance – maroon sites, *(iii)* Q_o site binding sites – orange sites, *(iv)* Q_i site binding sites – yellow sites, *(v)* b_L binding sites – blue sites and *(vi)* b_H binding sites – black sites. The amino acid substitutions in some of the mentioned sites are marked with arrows. (**B**) Superimposition of 3D structures with the positively selected sites mapped as sticks (orange for CODEML and blue for MEME). All the sites have as reference number, the site position from the *Octopus vulgaris*.

COX2: From the analyses performed only the site 13 presented evidences to be evolving under positive selection, according to MEME (p-value < 0.05) and to TREESAAP (p-value < 0.001) results. The inspection of the Cephalopoda COX2 protein sequence alignment revealed the presence of a conserved amino acid (C) in Oegopsida, Bathyteuthoidea, Myopsida and Sepiolidae groups at this site (Figure 2.7A). At the same position the other cephalopods have an amino acid, with polar uncharged side chain, S or N (Figure 2.7A).

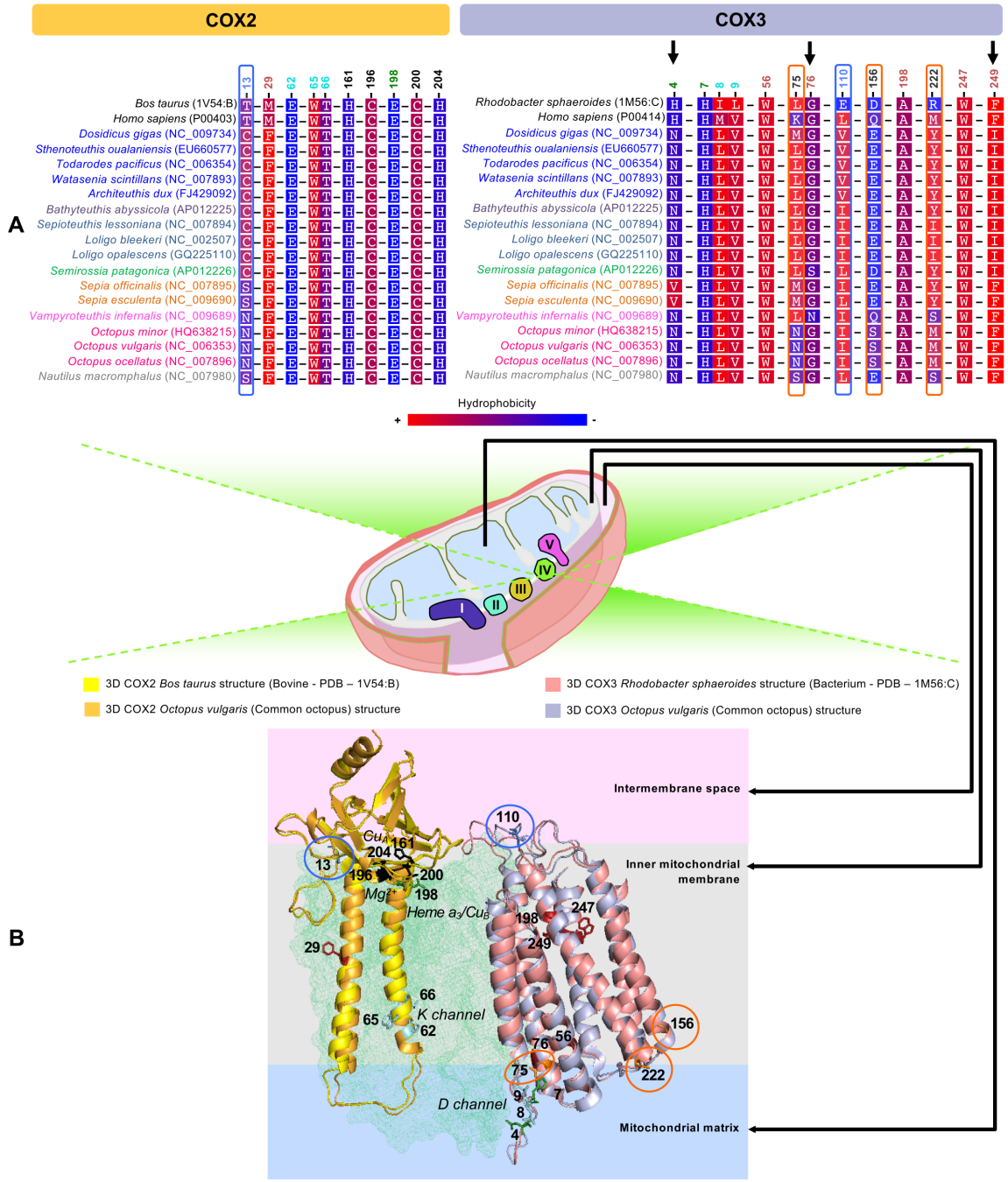


Figure 2.7: Amino acid variation in functional and structural sites of Cephalopoda COX2 and COX3 subunits. (**A**) COX2: Amino acid alignment for *(i)* sites under positive selection – blue shape (MEME), *(ii)* mutations related with human exercise intolerance – red sites, *(iii)* Cu_A metal binding sites – black sites, *(iv)* sites involved in proton coupling mechanism – cyan sites and *(v)* metal binding site for Mg^{2+} – green site. COX3: Amino acid alignment for *(i)* sites under positive selection – orange shape (CODEML branch-site model test) and blue shape (MEME), *(ii)* sites in the proximity of the entrance of D-channel – green sites, *(iii)* residues that protect the entrance of the D-channel from direct solvent exposure – cyan sites and *(iv)* sites related with diseases in humans – red sites. The amino acid substitutions in some of the mentioned sites are marked with arrows. (**B**) Superimposition of 3D structures with the sites mapped as sticks (colors according to Figure 2.7A). All the sites have as reference number, the site position from the *Octopus vulgaris*.

Accordingly, for this site, TREESAAP showed the amino acid property polarity under positive selection (*p*-value < 0.001). The mapping of this site into the predicted 3D protein-structure revealed that it is located in a loop region facing the intermembrane space (Figure 2.7B).

The homology comparisons performed (Appendices II: Table S6) showed that the well conserved residues E62, W65, and T66 of cephalopods (Figure 2.7 - cyan sites) correspond to residues of particular interest in relation to proton coupling mechanism, since they form hydrophilic channels that provide the necessary routes for intra protein proton movements [213]. These analyses also pointed that the Cu_A metal binding sites H161, C196, C200, and H204 (Appendices II: Table S6) are conserved in all the cephalopods (Figure 2.7 - black sites), highlighting its functional importance. Furthermore, we mapped the site 29, to which was reported a mutation (M29K) that causes exercise intolerance in humans [219] (Appendices II: Table S6), showing a conserved amino acid (F) in all the cephalopods (Figure 2.7A). We also mapped the E198, metal binding site for Mg^{2+} [213] (Appendices II: Table S6), which is equally conserved in all the cephalopods.

COX3: The mapping of the positively selected sites (*p*-value < 0.05) into the predicted 3D protein-structure showed that the sites 75 and 156 are located in TM helices, and the sites 110 (intermembrane space) and 222 (mitochondrial matrix) in loop regions, respectively (Figure 2.7B). TREESAAP indicated as positively selected (*p*-value < 0.001) the following amino acid properties: polarity for the site 75, and solvent accessible reduction ratio for the site 222. Through homology comparisons (Appendices II: Table S7), we observed that the site 75 is adjacent to a site (76) to which a disease related mutation (G78S - LHON) was described in humans (Figure 2.7) [213]. The considerable polarity change, observed in the site 75 of the species *O. minor* (L75N), *O. vulgaris* (L75N), *O. ocellatus* (L75N) and *N. macromphalus* (L75S) (Figure 2.7A), was also reported by TREESAAP (*p*-value < 0.001). These sites could be important for the stabilization of the Complex IV thereby enhancing ATP production proper to each species with different habitats and metabolic requirements. We also mapped other relevant sites, such as L8 and V9 (Appendices II: Table S7), conserved in all the cephalopods. These hydrophobic residues protect the entrance of the D-channel in COX1 from direct solvent exposure [204]. The homology analyses (Appendices II: Table S7) revealed that the sites 4 and 7 in cephalopods are counterparts of the sites H7 and H10 of the *R. sphaeroides*, respectively. These sites are also in the proximity of the entrance of D-channel. A recent study indicated that mutations in these sites (H7Q and H10Q) indirectly

lower the activity of the complex, by slowing the uptake of protons through the D-channel, and that their histidine residues stabilize the interactions between COX1 and COX3 [204]. The inspection of the Cephalopoda COX3 protein sequence alignment showed that the histidine of the site 7 is conserved in all the cephalopods (Figure 2.7A). However, the site 4 presents an asparagine (N4) conserved in all the cephalopods with exception of the Sepiidae species that instead possess a valine (V4) (Figure 2.7A). Furthermore, TREESAAP reported the amino acid properties hydropathy and solvent accessible reduction ratio for the site 4. Therefore, these substitutions may have impact at the stabilization of the complex. From the sites corresponding to sites related with diseases in humans (Appendices II: Table S7) we observed that they are mostly conserved in cephalopod species (Figure 2.7A - red sites). The exceptions occur at sites 76 and 249, both located in TM helices (Figure 2.7B). At the site 76 it is possible to observe a mutation G76S in *S. patagonica* and G76N in the *V. infernalis* (Figure 2.7A). At the site 249 a mutation F249I occurs in the groups Oegopsida, Bathyteuthoidea, Myopsida and Sepiolidae (Figure 2.7A).

ATP8: For the ATP8 subunit it was not possible to obtain a satisfactory 3D model. However, a topological assignment (Appendices II: Figure S4) showed that the positively selected site 10 (MEME method, *p*-value < 0.05) is located in the transition of a loop, located in the intermembrane space, to a TM helix. This site presents the hydrophobic residues L10 and I10; the former in the Oegopsida, Bathyteuthoidea, Myopsida, *S. patagonica*, *S. esculenta* and *N. macromphalus*, and the latter in the remaining cephalopods, respectively.

2.5 Discussion

Overall, all the employed approaches for phylogenetic analyses – considering the concatenation of 13-mt genes of 17 representative species from the Cephalopoda major taxonomic groups – generated a robust mitochondrial phylogeny for the Class Cephalopoda. Additional support can be inferred for the few cases where only a moderate support was obtained for the species relationship (*e.g., W. scintillans* and *A. dux* intra-relationships within the Oegopsida clade), through the observation that all the achieved tree topologies where equal. Furthermore, the mitochondrial topology (Figure 2.1) globally agrees with a recent phylogeny of 188 cephalopod species (which used mitochondrial and nuclear genes) [190]. The clades defined in the mitochondrial phylogeny are also corroborated when

some biological and ecological characteristics are considered (number of arms, metabolic rates and depths) (Figure 2.3). In fact, we can observe that the Octopodiformes (8 arms) and Decapodiformes (10 arms) form the two major clades. Within these major clades the species grouped into several sub clades, which have cephalopods with identical metabolic rates and inhabiting similar ocean depths (Figure 2.3). Consequently, we expected that similar selective pressures could be shaping the evolution of mitochondrial genes of species within the same clade.

Our site model results (*p*-value < 0.05, Appendices II: Table S8) indicated that the 13-mt genes of cephalopods are globally evolving under negative constraints (purifying selection), with a small percentage of codons evolving under neutrality, reinforcing their crucial and conserved role for the body energy production in cephalopods. This is also in agreement with the general trend of the mt genome evolution in vertebrates [225].

Given the high percentage of sites in the mt genes evolving under negative selection, evidences of pervasive positive selection (affecting few sites in particular linages) may be masked by the majority sign of negative selection in the branch-specific analyses. This is because these types of analyses are more conservative than branch-site model tests, thus justifying their distinct results.

From the several foreground-branches (**A to H**) – considering groups of cephalopods with different metabolic rates, depths, and number of arms – branch-site model tests revealed sites under positive selection in particular lineages (Figure 2.3). It is worth highlighting the presence of a few codons under positive selection (*p*-value < 0.05) in the genes *nd5*, *cytb*, and *cox3* (CODEML) in the ancestral lineage of the Octopoda group (**foreground-branch E**) (Figure 2.3 and Appendices II: Table S10), and also in *cytb*, *cox2*, and *atp8* genes (MEME) in the ancestral lineage of Octopodiformes (**foreground-branch G**) (Figure 2.3).

We have placed special attention to the codons with the hydrophobicity property under positive selection, since OXPHOS complexes are intricately linked to the mitochondrial inner membrane [226]. Thus, we would expect this property to be under particularly high purifying selection. Indeed, previous evidence covering four phylogenetic distant animal clades – mammals, birds, insects and nematodes – revealed that the negative ω value per clade was on average 3.9 times lower in mitochondrial OXPHOS genes (mt OXPHOS) than in nuclear OXPHOS genes, even restricting the mitochondrial-nuclear comparison to the same OXPHOS pathway [226]. This study refers with statistical support, that genes with high

levels of hydrophobicity and expression, are those under the strongest selective constraint [226]. Moreover, as in the case of mt OXPHOS subunits, other studies revealed that highly expressed proteins have particularly detrimental effects when they are hydrophobic at the same time [227, 228]. Hence, this property can be important in protein-protein (mitochondrial-mitochondrial and mitochondrial-nuclear) interactions, reinforcing the role of these subunits in the OXPHOS pathway [226]. Additionally, a change in this functional property can compromise the folding of a subunit, or even interfere in the stability of the entire complex. If this situation occurs, the non-polar residues could invade other subunits and membrane areas causing dysfunctional protein-protein interactions and the disruption of membranes [226]. The inverse case is also possible, the occurrence of some variations that could support a more efficient process, conferring adaptive advantage.

Our results suggest that positive selection may have had particular effect on the subunits ND5 and ND6 of cephalopod species. This can be striking in the ND5 subunit, which is a highly hydrophobic protein, conserved from prokaryotes to eukaryotes, and critical for transducing conformational energy to proton-pumping elements in the distal module of the membrane arm of the Complex I [101]. According to a recent study, the lack of the TM1 helix in ND5 only occurs in arthropods and nematodes [101]. Our study also provides evidence of the lack of the TM1 helix in cephalopods. Furthermore, we also revealed for the first time that the positively selected site 74 is also part of the proton translocation channel in cephalopods. This relevant site was positively selected in the Octopoda lineage (**foreground-branch E**, *p*-value < 0.05), suggesting an adaptive advantage to these species with intermediate metabolic rates and inhabiting shallow waters. Regarding the ND6 subunit in cephalopods, the amino acid at the site 24 of this subunit – counterpart of the site 26 of *H. sapiens*, associated with loss of vision [208] – is highly conserved. Therefore, we suggest that the existence of a L instead of a I/V (reported in *H. sapiens* and *E. coli*, respectively) should have no negative impact in the vision of Cephalopoda species, because hydrophobicity is maintained at the mutated site in cephalopods (Figure 2.5). However, further studies are required to understand if the activity of the complex is changed, since this site is near the Q-site of Complex I. Furthermore, we observed the presence of distinct amino acid substitutions occurring in the sites 33 and 34 – suggested to be lining the proton translocation channel [101] (Appendices II: Table S4) – that although changing the hydrophobicity of these sites, are conserved in subgroups of cephalopods. Through homology studies (Appendices

II: Table S4), we verified that a mutation (G36S) in the human counterpart of the cephalopod site 34 was associated with the LHON disease [222]. Similarly, in cephalopods we observed (Figure 2.5) that all the Octopodiformes and the Nautiloidea have a S in their counterpart site (34S). Furthermore, in the same position, most of the cephalopods living in deep waters and without cornea covering the eyes – *D. gigas*, *S. oualaniensis*, *T. pacificus* and *W. scintillans* – have also an amino acid with a polar uncharged side chain (34T). On the other hand, species with cornea – Myopsida, Sepiolidae and Sepiidae (Figure 2.5) – have an amino acid with a hydrophobic non-polar side chain (V or M). Thus, the presence of a polar uncharged amino acid (34T or 34S) may have some impact in the vision of the referred species. Consequently, this can have a critical effect in Octopodiformes species since they have cornea and live in shallow waters, where predators abound.

Overall, our results showed that most of the sites related with mutations described in humans, such as exercise intolerance, have conserved amino acids in all the cephalopods (Figures 2.6A and 2.7A). Similarly, the homology comparisons showed that residues of particular interest, as those involved in proton coupling mechanisms (Figure 2.7A), are mostly well conserved in all the cephalopods, highlighting its vital functional importance.

Since the beginning of this work, more mt genomes from cephalopods (NC_017746, NC_017749, NC_020348, NC_021146, NC_022466, NC_022467, NC_022468, NC_022693, NC_022959 and NC_023257) became available, however these are from the same major taxonomic groups represented in our data set. Thus, although these new mt genomes were not included in our analyses, the concatenation of mt genes from fully sequenced mt genomes of 17 species still representative of the Class Cephalopoda, covering the eight taxonomic groups Oegopsida, Bathyteuthoidea, Myopsida, Sepiolidae, Sepiidae, Vampyromorpha, Octopoda and Nautiloidea, and provided a coherent Cephalopoda mitochondrial phylogenetic tree.

Furthermore, we were able to assess the adaptive evolution of cephalopod mt genomes, through the identification of negative and positive selection. We considered several foreground-branches, in order to better comprehend the adaptive evolution of cephalopod lineages, well adapted to contrasting environments, namely diverse depths, and consequently temperatures, pressures, light, food, etc. In particular, the mapping of the identified positively selected sites in the Cephalopoda ND5 data set, along with homology analyses, allowed to conclude that some of these sites are: *(i)* lining proton translocation channels such as site 74TM2

(ND5 - CODEML), and *(ii)* involved in interactions between subunits, namely site 507HL (ND5 - TREESAAP). Additional relevant sites were identified through inspection of multiple sequence alignments of subunits from cephalopods with their counterparts (bacteria, yeast, bovine and/or human), and/or through its superimposition. Our study provides valuable insight into the adaptive evolution of the mt genome in cephalopods, and its implications for their molecular mechanism of OXPHOS.

Chapter 3

Phylogeny and Support for Adaptive Selection in the Driven Force of the ATP Synthesis: Evolution of Cephalopoda Mitogenomes

Submitted

3.1 Abstract

The genomic changes of the mitochondrial genome of Cephalopoda provides an opportunity to explore their phylogenetic history and to unravel events of adaptive mitochondrial DNA (mtDNA) evolution within Oegopsida and Bathyteuthoidea (OB), an enigmatic Decapodiformes clade, whose ancestral is postulated to be the first, among squids, to transit from bottom feeders or demersal to open ocean feeder or pelagic lifestyle. OB mtDNA genomes also accommodates functional duplicate copies of *atp6, atp8*, and *cox1-3* genes, a feature absent in other cephalopods. Here, we analyzed the protein-coding genes from the mtDNA genomes of 37 Cephalopoda species. Firstly, we used the best orthologous gene copy to examine phylogenetic reconstruction, and compared three data sets (L, R, OG), each one classified by three forms ("original", RY-coded, 1^{st} and 2^{nd} codon positions, respectively). Then, the form and strength of selection, operating on protein-coding genes, were explored through a variety of codon substitution models, namely, random-site, branch-site and clade models. Overall, the OG "original" form better fit the topology of cephalopods major clades, and was in agreement with previous phylogenetic studies. Using the latter tree, the branch-site test along the ancestral branch of OB, revealed patterns for positive selection in two genes involved in the oxidative phosphorylation pathway (OXPHOS) – *atp6* (sites 46 and 194) and *cox2* (sites 129 and 132); curiously belonging to the set of genes with duplicate copies. Interestingly, homologues of the identified sites are located in relevant regions of the protein structure, which were predicted to influence the enzymatic activity of the OXPHOS complexes IV and V, thereby regulating the ATP synthesis. This suggests that metabolic and mitochondrial functional changes have likely been enhanced by natural selection to insure OB bioenergetics efficiency across the pelagic environment.

3.2 Introduction

The Earth biomes sustain an incredible diversity of life with a tremendous array of body plans, providing the foundation to explore how organisms and their genes adapted and evolved to cope phenotypically with the heterogeneity of environments. Therefore, identifying the molecular basis of these traits will assist in future understanding of the mechanism underlying metazoan evolutionary biodiversity [164, 229]. Towards this end, mollusks of the Class Cephalopoda (*i.e.,* squids, cuttlefishes, octopods, vampire squids, and *Nautilus*) possess exceptional adaptive features that are unique among invertebrates. With 800 known extant species, this ancient and diverse group evolved 530 million years ago [163] and split into two major groups: 1) Nautiloidea (*Nautilus* and *Allonautilus*) or armored shell species, and the most diversified 2) Coleoidea or soft-bodied species, with or without internal shell, with ten-armed (*i.e.,* Decapodiformes: Spirulida, Sepioidea, Myopsida, Oegopsida, Idiosepiidae, Bathyteuthoidea) and eight-armed (*i.e.,* Octopodiformes: Octopoda, Vampyromorpha) taxa, (http://tolweb.org/Cephalopoda, last accessed January 7, 2016) [162]. Furthermore, cephalopods are found living from polar to tropical seas [184, 230–232], and were successful to colonize from neritic (*e.g.,* Myopsida and *Sepia*) to pelagic ocean waters (*e.g.,* Oegopsida, Bathyteuthoidea, and Octopoda) [184, 230–232]. These dramatic ecological and geographical variations, with contrasting habitats of markedly unique properties (*i.e.,* temperatures, light intensity, oxygen, salinity concentrations, and pressure) likely exerted strong selective pressures on morphology, physiology and behavior that contributed to the current phenotypic innovations [163, 164]. Despite their interesting attributes, the genesis and evolution of these novel traits are still poorly understood. In Cephalopoda, the existing genomic research has primarily focused on mtDNA phylogenetic reconstruction, although this work still remains controversial and requires further analysis [170, 184, 190, 193, 233–235]. Few studies have identified mtDNA protein-coding genes (hereafter referred to as 'mtDNA genes') with duplicates [170, 184, 190, 236] and one previous study on mtDNA genes demonstrated positive selection in energy metabolism associated with morphology and habitat [236]. Recently, more efforts have been directed towards increasing the number of mtDNA coding gene sequences, which has been updated from 24 to a total of 37 species (http://www.ncbi.nlm.nih.gov/, last accessed January 8, 2016; [27]). This new data set provides an opportunity to test *(i)* whether the inclusion of new lineages

can help to clarify the weak support among some of the lineages of the current mitochondrial phylogeny, and *(ii)* to continue exploring patterns of mtDNA adaptive selection, considering an earlier work by Almeida *et al.* (2015) [236] and past studies from other marine organisms showing successful adaptive selection, such as in several fishes [78–80], sea turtles [81], and cetaceans [76, 77].

Given that mtDNA evolution may be an important driver in the diversification of life in Cephalopoda and in the oceanic environments in general, in this study we investigated various important aspects of the cephalopods mitochondrial genome evolution across 37 species. Firstly, we assessed whether the new data set improved phylogenetic reconstruction and relationships among taxa, particularly in Oegopsida, as well as in the Octopoda and Vampyromorpha, three lineages that to date have been poorly resolved [170, 184, 193, 233–236]. This was performed by phylogenetic estimation methods and a novel approach that considered the inclusion of the best orthologous mtDNA gene copy with regard to those species without duplicated genes. Secondly, we examined patterns of positive selection on the ancestral branch leading to Oegopsida and Bathyteuthoidea, which was selected *a priori* because in Decapodiformes this branch represents the most ancient transition from demersal (*i.e.,* bottom feeders) to the pelagic (*i.e.,* open sea feeders) lifestyle [190, 237] and the acquisition of duplicated mtDNA genes (*i.e., atp6* and *atp8*, *cox1-3*), non-existent in all other cephalopods [170, 184, 236, 238]. Duplicated genes have been proposed important in the evolution of the cephalopod body plan and are commonly considered a major source of new gene functions, as revealed in other organisms (*e.g.,* mollusk-bivalves, cnidarians, insects, amphibians and reptiles [109]). Therefore, the pattern of mtDNA gene rearrangements found in Oegopsida and Bathyteuthoidea may be informative and an important component of the mitogenome molecular evolution and squid adaptation in the pelagic zone. Here, we hypothesize that during this transition, Oegopsida and Bathyteuthoidea may have required an increase in energetic demands, thus selection on mtDNA genes was necessary for adaptation to the pelagic environment in order to meet the metabolic requirements. Our analyses showed evidence of signatures of positive Darwinian selection among Oegopsida and Bathyteuthoidea mtDNA genes, including gene duplicates, affirming the adaptive evolution of the Cephalopoda mitochondrial genomes.

3.3 Materials and Methods

Mitochondrial protein-coding gene sequences

The current study includes the complete nucleotide coding-sequences (CDS) of 13 mtDNA genes (*atp6, atp8, cox1, cox2, cox3, cytb, nd1, nd2, nd3, nd4, nd4l, nd5* and *nd6*) of 20 fully sequenced mitochondrial genomes of cephalopods in addition to the 17 cephalopod species used in a previous work [236] for a total of 37 cephalopod species obtained from GENBANK (http://www.ncbi.nlm.nih.gov/genbank/, last accessed January 8, 2016) [27]. Mitochondrial DNA gene arrangement is shown in Appendices III: Figure S1. Details about taxonomy (*i.e.,* species scientific name, common name, Cephalopoda major taxa and family), accession numbers, as well as, cephalopods average ocean depths are provided in Appendices III: Table S1.

Mitochondrial protein-coding gene data sets

For mitochondrial phylogenetic tree estimation, three distinct data sets were considered. The first two were prepared according to Almeida, *et al.* (2015) [236]: 1) "L-data set" contains species without duplicated mtDNA genes (herein referred as single genes) and species with the left-side (L) duplicated genes (Appendices III: Table S2) and 2) "R-data set" contains species with single genes and species with the right-side (R) duplicated genes (Appendices III: Table S3). The third, "OG-data set", holds 13 orthologous groups (OG) of mtDNA genes (Appendices III: Table S4) assigned by a reliable algorithm, implemented in OMA stand-alone software version 0.99t, using default parameters [239, 240]. The L- and R-data sets were analyzed, given that previous published cephalopod phylogenies based on mtDNA genes, used empirical criteria to select each of two duplicated genes. The latter was justified on the basis of the copies having high rate of conserved nucleotides or absence of nucleotide substitutions (Appendices III: Table S5) [170, 184, 190, 236]. The OG-data set was examined in order to understand whether using an algorithm criterion to help select the best orthologous gene groups could lead to an improvement of the Cephalopoda mitochondrial phylogeny. A similar approach was implemented in a recent work on centipedes [241].

L- and R- data sets individual alignments

The CDS were codon-based aligned using MUSCLE algorithm through the GUIDANCE web server facility (http://guidance.tau.ac.il/ver2/, last accessed January 15, 2016) [33]. All the individual-gene alignments were highly robust and reliable according to the high alignment scores (range from 0.98 to 1.00) revealed by GUIDANCE (whose maximum score is 1.00); therefore, no positions or sequences were masked or removed.

OG-data set individual alignments

For each gene, the sequences were first translated into proteins in SEAVIEW software version 4.4.0 [186] and then aligned with PROMALS3D web server utility using default parameters (http://prodata.swmed.edu/promals3d/promals3d.php, last accessed January 18, 2016; [32]). All the individual-gene alignments were evaluated applying the Transitive Consistency Score (TCS) web server (http://tcoffee.crg.cat/tcs, last accessed January 18, 2016) [34], which revealed high alignment scores (range from 961 to 999; the maximum alignment TCS is 1000); therefore, filtering was not necessary. Then, each of the protein alignments were back-translated into codon-alignments through the RevTrans web server (http://www.cbs.dtu.dk/services/RevTrans/, last accessed January 19, 2016) [242].

Concatenated data sets

GENEIOUS software version 5.6.7 [187] was used to combine all 13 individual-genes, previously aligned as described above, resulting in a total of three concatenated data sets: “Conc_L” (11,682 nucleotides), “Conc_R” (11,712 nucleotides) and “Conc_OG” (11,712 nucleotides).

Phylogenetic analyses of concatenated data sets

The three codon positions (1^{st}, 2^{nd} and 3^{rd}) of each of the three concatenated data sets (Conc_L, Conc_R and Conc_OG) were evaluated for saturation bias, using the Xia *et al.* statistic test [188] as implemented in DAMBE software version 6.1.7 [40]. All the above-mentioned revealed evidence of some saturation in its 3^{rd} codon position (details in Appendices III: Table S6). Since this could compromise phylogenetic topology and

downstream analyses, we proceeded to verify the potential effect of the 3^{rd} codon in the following manner: phylogenetic reconstruction 1) keeping the three codon positions – "original form" alignment; 2) "RY-coded form" (where 3^{rd} codon positions were coded as R for all purines and Y for all pyrimidines), and 3) removal of the 3^{rd} codon positions – "1^{st} and 2^{nd} form" alignment. For all cases, the best nucleotide substitution model was determined to be the General Time Reversible (GTR), with a proportion of invariable sites (I) and heterogeneity of substitution rates among sites modelled following a gamma distribution (G) using the Akaike Information Criterion (AIC) in JMODELTEST software version 2.1.7 [37] with default parameters. Mitochondrial phylogenies were estimated using the GTR+G+I model through Maximum likelihood (ML) and Bayesian (BA) algorithms.

ML phylogenetic trees were built in PHYML software version 3.0 [41], with 1000 bootstrap replicates and the parameter 'best of Nearest Neighbor Interchange (NNI) and Subtree Pruning and Regrafting (SPR)' branch search algorithms. Since the bootstrap replicates required to achieve a certain accuracy level is highly data set dependent [243], we plotted the parameters estimated by each bootstrap to ensure that convergence was attained (Appendices III: Figures S2-S10 for more details).

BA trees were inferred in MRBAYES software version 3.2.6 [42]. Two runs with four chains of the Markov Chain Monte Carlo (MCMC) method were performed for 5,000,000 generations, with sampling trees every 100 generations. The following criteria were used to test convergence of the chains: the standard deviation of split frequencies was less than 0.01 and the potential scale reduction factors (PSRF) were close to 1.0 for all parameters. Additionally, the convergence of the chains to its stationary distribution for all the parameters was evaluated using TRACER software version 1.6 (http://tree.bio.ed.ac.uk/software/tracer2013, last accessed March 23, 2016) [189], which also helped to determine the appropriate burn-in (1,250,000 of the sampled trees were discarded). The effective sample size (ESS) for all the parameters was greater than 1608 (ESS for the log-likelihood: Conc_L-data set "original form" = 8791, "RY-coded form" = 10123, and "1^{st} and 2^{nd} form" = 12640; Conc_R-data set "original form" = 8974, "RY-coded" = 10497, and "1^{st} and 2^{nd} form" = 12013; Conc_OG-data set "original form" = 9363, "RY-coded form" = 11225, and "1^{st} and 2^{nd} form" = 11030, respectively).

To statistically evaluate the phylogenetic relationships among possible cephalopod clades, which were suggested by the topologies (obtained as described above), the following

tests were performed: AU – approximately unbiased [244], NP – non-scaled bootstrap percentage [245], and KH – Kishino-Hasegawa [246], using CONSEL software version 0.20 [245] with default parameters, and considering a significance level of α = 0.05.

Molecular evolutionary analyses of the OG-data sets

Darwinian selection analyses

All the analyses were performed using the ML phylogeny obtained from the Conc_OG data set in its "original form" (see Results - Congruent mitochondrial phylogeny).

The form and strength of selection, operating on individual-genes and concatenated-gene alignments, were explored through a variety of codon substitution models (codon models), namely, random-site, branch-site and clade models; which estimate the non-synonymous (*dN*) to synonymous (*dS*) substitution rate ratios (ω = *dN/dS*: where ω greater than, equal to, and less than one is indicative of positive selection, neutrality and negative selection, respectively) using the CODEML program from the PAML software version 4.7 package [43].

Random-site models allow the ω ratio to vary among sites in the alignments in different ways: alternative models M2a, and M8 allow sites to have ω greater than one, meaning adaptive evolution [8]; while null models such as M1a, M7, and M8a do not allow sites with ω greater than one. In these analyses, likelihood ratio tests (LRTs) were conducted by comparing the alternative models with the null models: M2a *vs.* M1a [8, 18, 45], M8 *vs.* M7 [44], and M8 *vs.* M8a [46]; to infer which models best fit the data.

Both, branch-site [48] and clade models [49, 50] allow for ω variation among sites in the alignments and between branches of interest (foreground-lineages) and remaining branches (background-lineages), or between clades, based on *a priori* partitioning of the phylogeny [7] which takes into account a hypothesis of where adaptive evolution may have occurred [50].

Briefly, the alternative model A (Alt) of the branch-site model, comprises four site classes (SC0, SC1, SC2a, and SC2b), which allow the following ω variations among the pre-specified branches: SC0 and SC1 correspond to codons that experience selection consistently across all branches of the phylogeny, evolving under either purifying selection (SC0: $0 < \omega 0 < 1$) or neutrality (SC1: $\omega 1 = 1$); while SC2a corresponds to codons that experience purifying selection or neutrality on the background ($0 < \omega 2a = \omega 0 < 1$) and neutrality or positive selection on the foreground ($\omega 2a = \omega 2b \geq 1$); and lastly SC2b admits neutrality ($\omega 2b =$

$\omega 1 = 1$) on the background and neutrality or positive selection ($\omega 2b = \omega 2a \geq 1$) on the foreground [48]. Its corresponding null model (Null) is obtained by constraining the $\omega 2$ to be equal one. In this analysis, the LRT was conducted by comparing the alternative model with the null model: Alt *vs.* Null, thus testing for the presence of positive selection.

The alternative hypothesis of the Clade model C (CmC) comprises three site classes (SC0, SC1, and SC2): SC0 and SC1 correspond to codons which evolve consistently across the entire phylogeny, either through purifying selection ($0 < \omega 0 < 1$) or neutrality ($\omega 1 = 1$); while the SC2 accommodates codons that experience divergent selection in the two pre-specified partitions (*i.e.,* $\omega 2 > 0$ for the background-clade, and $\omega 3 > 0$ for the foreground-clade). Its corresponding null model, M2a_rel, is obtained by constraining the $\omega 2$ to be greater than zero for all branches of the phylogeny [247]. In this analysis, the LRT was conducted by comparing the alternative model with the null model: CmC *vs.* M2a_rel.

For all codon models described above, the alignments gaps and ambiguities were first removed (cleandata = 1) to avoid false positives [248, 249]. The M0 model, which admits uniform selective pressures among sites and branches [5], was run several times using the unrooted ML topology, through the adjusting of the initial ω and κ (transition-to-transversion substitution rate ratio) values in order to avoid possible local-likelihood peaks [250]. This way, were also obtained the starting branch-lengths necessary for run all codon models analyses. To ensure the reliability of the parameters estimates (ω, κ, etc.) the M0 model, was inspected for the branch-specific estimates of the *dS* in order to investigate if the synonymous sites are saturated [251]. The estimated values ranged from 0.01 to 3.25, with the majority of sites and median below one, suggesting that the OG-data sets (individual and Conc_OG alignments) were not strongly affected by synonymous saturation (Yang Z., personal communication). All the codon models analyses were run multiple times from different starting values of ω and κ (depending on the particular model) to help ensure that local optima were avoided [247]. For all models comparisons, via LRTs, the hypothesis decision threshold was calculated by doubling the difference between the alternative and null model log-likelihood (2ΔlnL) and assuming that the null distribution of these results could be approximated by a chi-square (χ^2) distribution (*p*-value < 0.05). The number of degrees of freedom (*df*) was calculated as the difference in the number of estimated parameters between the models [8, 44].

Whenever the LRT was significant (*p*-value < 0.05) under the alternative models, the codon sites under positive selection or evolving divergently were identified using the Bayes

Empirical Bayes (BEB) calculation, which analyze the posterior probabilities (PP) for these sites [45]. In the current study, only the sites with PP ≥ 0.70 were considered.

Tree partitioning supported by evolutionary model

In order to define the partition scheme along the phylogeny where adaptive evolution may have occurred, we first performed an ultrametric tree using the R8S 1.8 software according to the author's instructions [252]; then we used an stochastic character mapping of each trait (presence-foreground or absence-background of duplicated genes) across the ultrametric phylogeny following the method of Bollback [253], as implemented in phytools [254] and geiger [255] R packages, using R 3.2.2 software [256] (Appendices III: Figure S11 A); and finally we obtained the partitioning of the unrooted phylogeny according to these mapping results (Appendices III: Figures S11 B and C). This allowed us to pre-specify the ancestral-branch leading to the "Oegopsida and Bathyteuthoidea" clades (*i.e.,* foreground), both for the branch-site (Appendices III: Figure S11 B) and clade model (Appendices III: Figure S11 C) analyses.

Physicochemical amino acid properties

Sites found by BEB from branch-site and/or clade models were assessed for site-specific physicochemical amino acid properties that are being preserved (conserved properties) or modified (changing properties) through the evolutionary process, using the PRoperty Informed Models of Evolution (PRIME) method (Pond, unpublished work), implemented in the DATAMONKEY webserver (http://www.datamonkey.org/, last accessed June 8, 2016) [19,52]. We considered the two predefined sets of five physicochemical amino acid properties, *i.e.*: *(i)* CC–chemical composition of the side chain, P–residue polarity, V–volume of the residue side chain, pHi–isoelectric point of the side chain and H–hydropathy [257], and *(ii)* P–polarity index, SS–secondary structure factor, V–volume, μ–refractivity and pHi–isoelectric point [258]. Since the estimates of amino acid exchangeabilities implemented by PRIME are based on multiple tests performed on the same residue site, the number of false positives reported at a site is prevented through the Bonferroni correction (http://hyphy.org/w/index.php/PRIME). Therefore, the sites were selected based on a corrected *p*-value of less than 0.05.

Three-dimensional structural analyses

The mitochondrial subunits of the respiratory chain complexes are highly conserved from mammals to bacteria [78, 259, 260]. Therefore, the potential impact of Cephalopoda positively selected sites (*p*-value < 0.05, found only for *atp6* and *cox2*), can be inferred through homology comparison with better studied species. This was accomplished by aligning our protein OG-data sets with its homologous proteins for which there are a three-dimensional (3D) structure [261] and/or well-studied functional sites [76, 262–265], and/or related human diseases at MITOMAP database (http://www.MITOMAP.org, last accessed June 16, 2016) [125]. For the referred comparison, we added to the ATP6 and COX2 alignments the following homologous sequences: ATP6 – *Homo sapiens* (P00846) and *Bos taurus* (P00847), and COX2 – *H. sapiens* (P00403) and *B. taurus* (P68530). Then the Protein Data Bank (PDB) files from bovine, "5ARA" (ATP6) and "1OCC" (COX2), were used to assess the spatial locations of the homologous identified sites relative to known functional domains of Complex subunits. For all proteins, the 3D structures were visualized and annotated using PYMOL software version 1.5.0.4 [58].

3.4 Results

Congruent mitochondrial phylogeny among Cephalopoda major taxonomic groups

Overall, the recovered mitochondrial tree topologies (Figures 3.1–3.3 – shown as cladograms) were identical to each other, with high support bootstrap percentages (BP) and posterior probabilities (PP), regardless of the data sets employed (*i.e.,* L, R and OG), its forms (*i.e.,* "original" – Figure 3.1, "RY-coded" – Figure 3.2, "1^{st} and 2^{nd}" – Figure 3.3), and the phylogenetic algorithm methods (*i.e.,* ML and BA) used. All the aforementioned approaches corroborated two major clades (Figures 3.1–3.3): *(i)* Decapodiformes – recovered five major taxonomic groups (Oegopsida, Bathyteuthoidea, Myopsida, Sepiolida and Sepiida); and *(ii)* Octopodiformes – recovered two major taxonomic groups (Vampyromorpha and Octopoda). In more detail, data sets L, R and OG recovered in its: *(i)* "original" form, exactly the same topology using both algorithms (Figure 3.1, ML = BA); *(ii)* "RY-coded" form, the

general topology of Figure 3.1, but with some minor intra-clade species rearrangements as highlighted in Figure 3.2; *(iii)* "1^{st} and 2^{nd}" form, the overall topology of Figure 3.1, but with some minor intra-clade species rearrangements (Figure 3.3, ML = BA). A closer look to Figure 3.1 reveals that the two major clades (Decapodiformes and Octopodiformes) are consistently supported by ML and BA, and both generated high values of BP and PP, respectively. For the ML analysis, Decapodiformes score BP values of 94% (for either L or OG-data sets) and 96% (R-data set) (Figure 3.1). In Octopodiformes, BP attained values of 73% (L-data set), 79% (OG-data set) and 74% (R-data set) (Figure 3.1). Similarly, all BA were well-resolved, and both major clades reached PP values of one (Figure 3.1). When comparing Figures 3.1 and 3.2, the ML analysis displayed minor changes (red branches) for the positions of *Uroteuthis edulis* and *U. chinensis*, as well as, between *Sepia esculenta*, *S. lycidas*, *S. aculeata* and *S. pharaonis*, whereas BA differ for the position of *Sthenoteuthis oualaniensis* and *Dosidicus gigas*. In contrast, a separate comparison between Figures 3.1 and 3.3, obtained different rearrangements (red branches) for the positions of *Watasenia scintillans* and *Architeuthis dux*, as well as, between *S. esculenta*, *S. lycidas*, *S. aculeata* and *S. pharaonis*. All aforementioned rearrangements were within Decapodiformes, while Octopodiformes remain in identical placement (Figure 3.1).

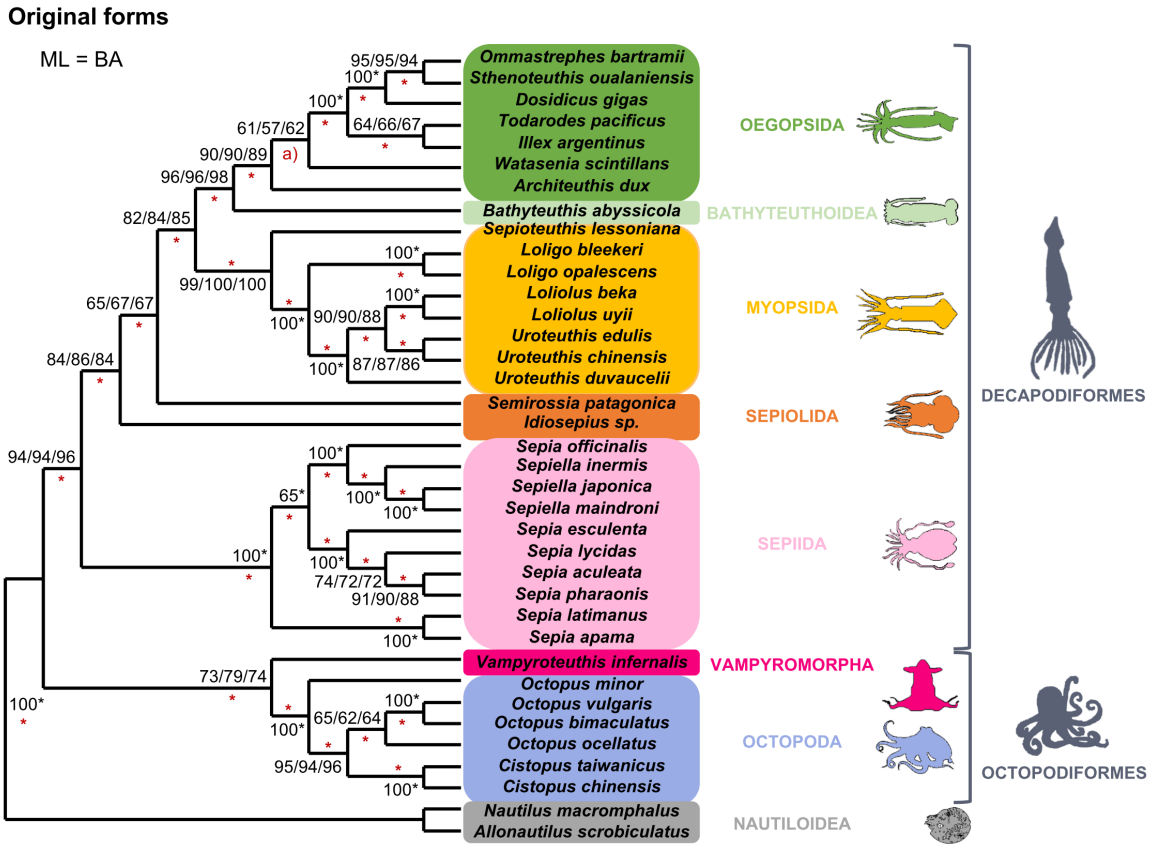


Figure 3.1: Cephalopoda mitochondrial consensus tree topology from three data sets analyzed in their original forms. Cladogram representing the same tree topology obtained by both Maximum likelihood and Bayesian (ML=BA) analyses, using each one of the three data sets, *i.e.,* L/OG/R, in its original forms. The values supporting each node are shown next to the branches as follows: ML bootstrap percentages – one value, *e.g.,* 100*, meaning that the three data sets revealed the same support value; three values, *e.g.,* 95/95/94, corresponding to L/OG/R data sets, respectively; BA posterior probabilities – lower case a) that corresponds to 0.91/0.93/0.92 for data sets in the same order as above-mentioned, and * asterisk marks in red indicating that all the BA posterior probabilities were equal to 1.00. Species are designated by their scientific names with background colors highlighting the major taxonomic groups of cephalopods. Orders – Decapodiformes (ten-arms) and Octopodiformes (eight-arms).

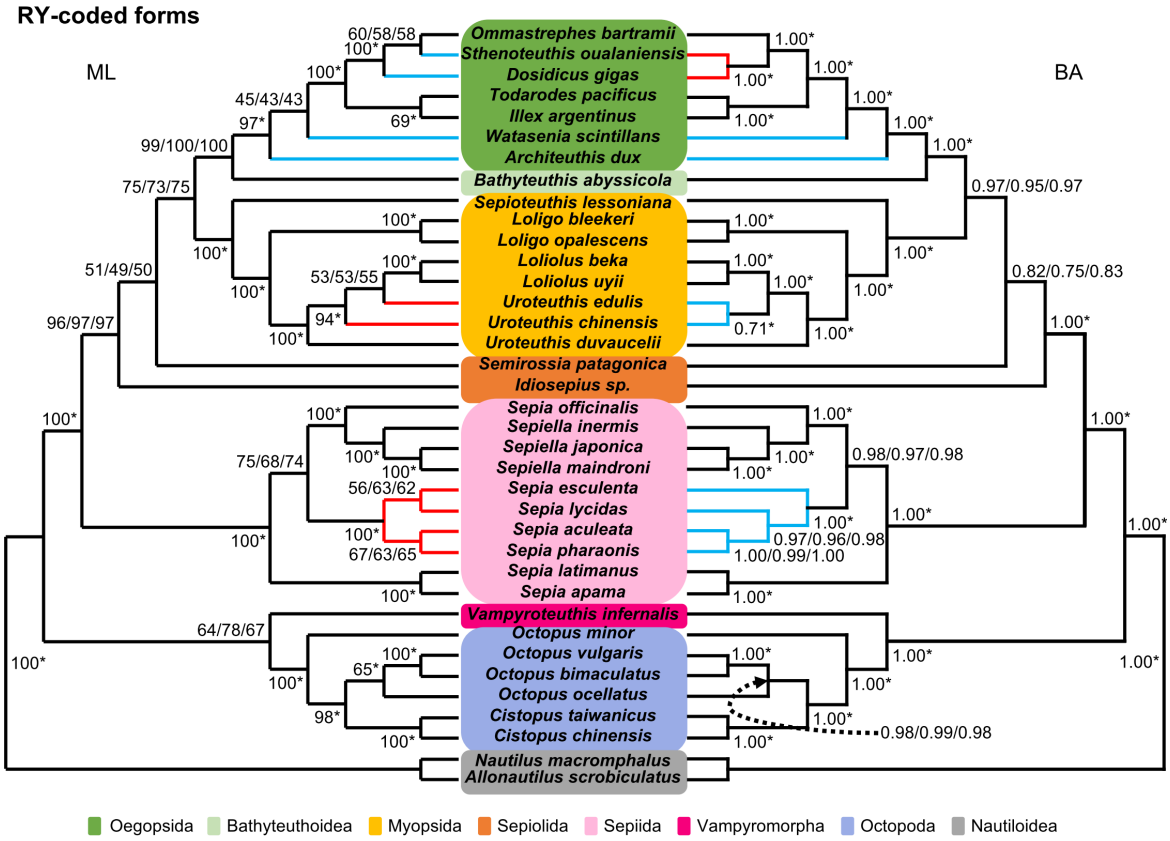


Figure 3.2: Cephalopoda mitochondrial tree topologies from three data sets analyzed in its RY-coded forms. Cladograms showing the tree topologies obtained by both Maximum likelihood (ML – at the left) and Bayesian (BA – at the right) analyses, using each one of the three data sets, *i.e.,* L/OG/R, in its RY-coded forms. The values supporting each node are shown next to the branches as follows: ML bootstrap percentages – one value, *e.g.,* 100*, meaning that the three data sets revealed the same support value, or three values, *e.g.,* 60/58/58, corresponding to L/OG/R data sets, respectively; BA posterior probabilities – one value, *e.g.,* 1.00*, corresponding to the same support for all data sets; or three values 0.97/0.95/0.97 for data sets in the same order as above-mentioned. Species are designated by their scientific names with background colors highlighting the major taxonomic groups of cephalopods. Colorful branches: blue – indicating similar species positions and red – indicating distinct rearrangements, when comparing both topologies here displayed with Figure 3.1.

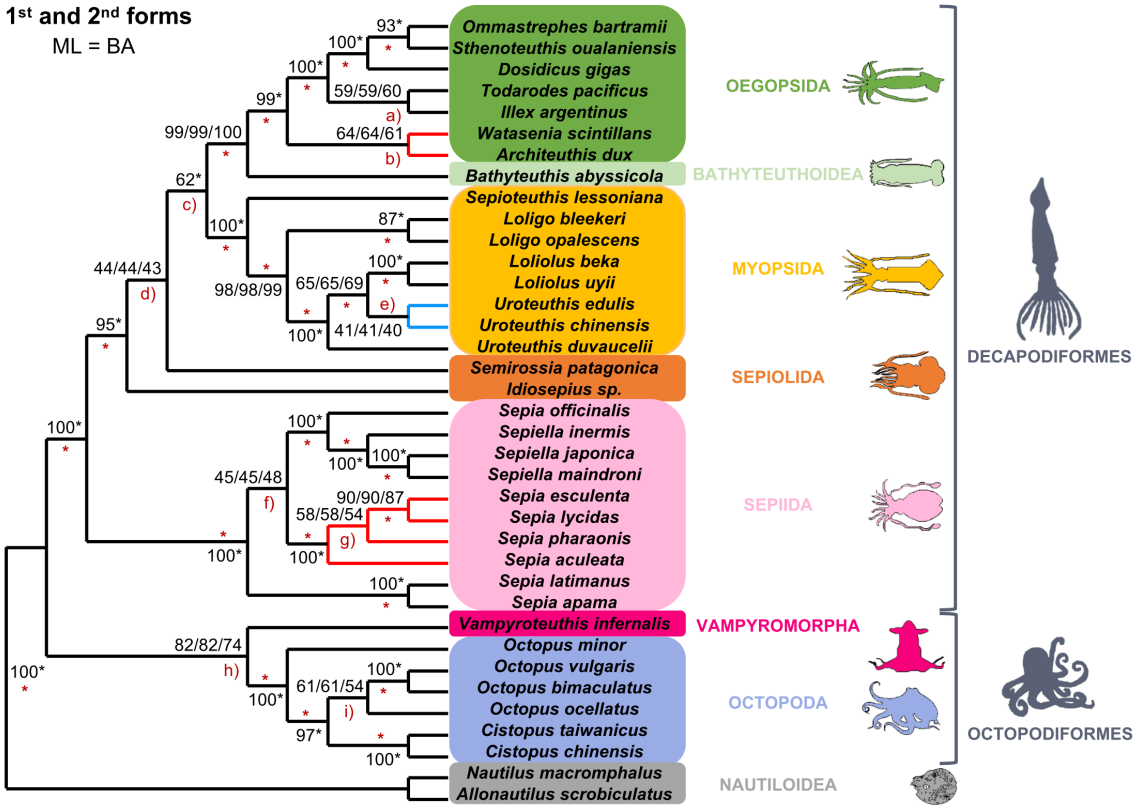


Figure 3.3: Cephalopoda mitochondrial tree topologies from three data sets analyzed in its 1^{st} and 2^{nd} forms. Cladogram representing the same tree topology obtained by both Maximum likelihood and Bayesian (ML=BA) analyses, using each one of the three data sets, *i.e.,* L/OG/R, in its 1^{st} and 2^{nd} forms. ML bootstrap percentages – one value, *e.g.,* 93*, meaning that the three data sets revealed the same support value, or three values, *e.g.,* 59/59/60, corresponding to L/OG/R data sets, respectively; BA posterior probabilities – lower cases that correspond to a) 1.00/0.98/1.00, b) 0.99/1.00/0.99, c) 0.83/0.86/0.83, d) 0.50/0.76/0.51, e) 0.89/0.93/0.89, f) 0.81/0.80/0.81, g) 0.89/0.93/0.89, h) 0.99/0.97/0.99 and i) 0.95/0.90/0.96, for data sets in the same order as above-mentioned; and * asterisk marks in red indicating that all the BA posterior probabilities were equal to 1.00. Species are designated by their scientific names with background colors highlighting the major taxonomic groups of cephalopods. Colorful branches: blue – indicating similar species positions and red – indicating distinct rearrangements, when comparing the topology here displayed with Figure 3.1.

For the above-suggested Decapodiformes position rearrangements, we also evaluated which one suggested the most probable phylogenetic topology or hypothesis. This was performed through the statistical tests AU, NP and KH (see Materials and Methods) as shown in Table 3.1.

Table 3.1: Statistical tests used for comparison of probabilities which assess distinct challenging phylogenetic relationships among some Decapodiformes. The statistical (*P*-probability) values that best support the cephalopods relationships are shown in bold.

Phylogenetic hypotheses		AU	NP	KH
Figure 3.1	*Sthenoteuthis oualaniensis* is close to *Ommastrephes bartramii*	**0.937**	**0.985**	**0.973**
Figure 3.2 (BA)	*Sthenoteuthis oualaniensis* is close to *Dosidicus gigas*	0.05	0.001	0.024
Figures 3.1 and 3.2	*Watasenia scintillans* is close to the (*Ommastrephes bartramii* + *Sthenoteuthis oualaniensis* + *Dosidicus gigas* + *Todarodes pacificus* + *Illex argentinus*) clade	**0.714**	**0.704**	**0.703**
Figure 3.3	*Watasenia scintillans* is close to *Architeuthis dux*	0.394	0.254	0.297
Figures 3.1, **3.2** (BA), and **3.3**	*Uroteuthis edulis* is close to *Uroteuthis chinensis*	**0.929**	**0.943**	**0.938**
Figure 3.2 (ML)	*Uroteuthis edulis* is close to the (*Loliolus beka* + *Loliolus uyii*) clade	0.122	0.03	0.062
Figures 3.1 and 3.2 (BA)	*Sepia lycidas* is close to the (*Sepia aculeata* + *Sepia pharaonis*) clade	**0.733**	**0.743**	**0.772**
Figures 3.1 (ML) and **3.3**	*Sepia lycidas* is close to *Sepia esculenta*	0.294	0.243	0.228
Figure 3.3	*Sepia pharaonis* is close to the (*Sepia esculenta* + *Sepia lycidas*) clade	0.214	0.089	0.21
Figure 3.2 (ML)	*Sepia pharaonis* is close to *Sepia aculeata*	**0.817**	**0.902**	**0.79**

These tests identified strong probability (*P*) values for the following taxa relationships: *(i) S. oualaniensis* as a sister-taxon to *Ommastrephes bartramii*, supported by *P* values greater than 0.93; *(ii) W. scintillans* as a sister-taxon to the clade composed by *O. bartramii*, *S. oualaniensis*, *D. gigas*, *Todarodes pacificus*, and *Illex argentinus*, supported by *P* values greater than 0.70; *(iii) U. edulis* as a sister-taxon to *U. chinensis*, supported by *P* values greater than 0.92; and finally *(iv) S. lycidas* as a sister-taxon to the clade composed by *S. aculeate* and *S. pharaonis*, supported by *P* values greater than 0.73. Taken together, the tree topology that overall best described cephalopod major clades relationships was that obtained from data sets in their original forms (Figure 3.1). Thus, the mitochondrial tree from the "original" OG data set was chosen and used for the mtDNA adaptive selection analyses that follow next in this study.

Random-site models reinforced that purifying selection is the predominant force shaping the evolution of mitochondrial protein-coding genes in cephalopods

The LRT random-site model analyses (p-value < 0.05) tested whether long-term positive selection occurred within a codon site in multiple cephalopods. However, results revealed no significant evidence of adaptive evolution in any of the mitochondrial subunits encoded by the 13 mitochondrial genes among the 37 Cephalopoda species (Appendices III: Table S7). When M2a-M1a models were compared for each of the 13 individual-gene alignments and for the concatenated-gene data set (Appendices III: Table S7 – Conc_OG), the M1a model could not be rejected (Appendices III: Table S7 – p-values between 0.995 and 1.000), supporting negative or purifying selection as the predominant force shaping the evolution of these genes, and with a few codons evolving under neutrality (Figure 3.4).

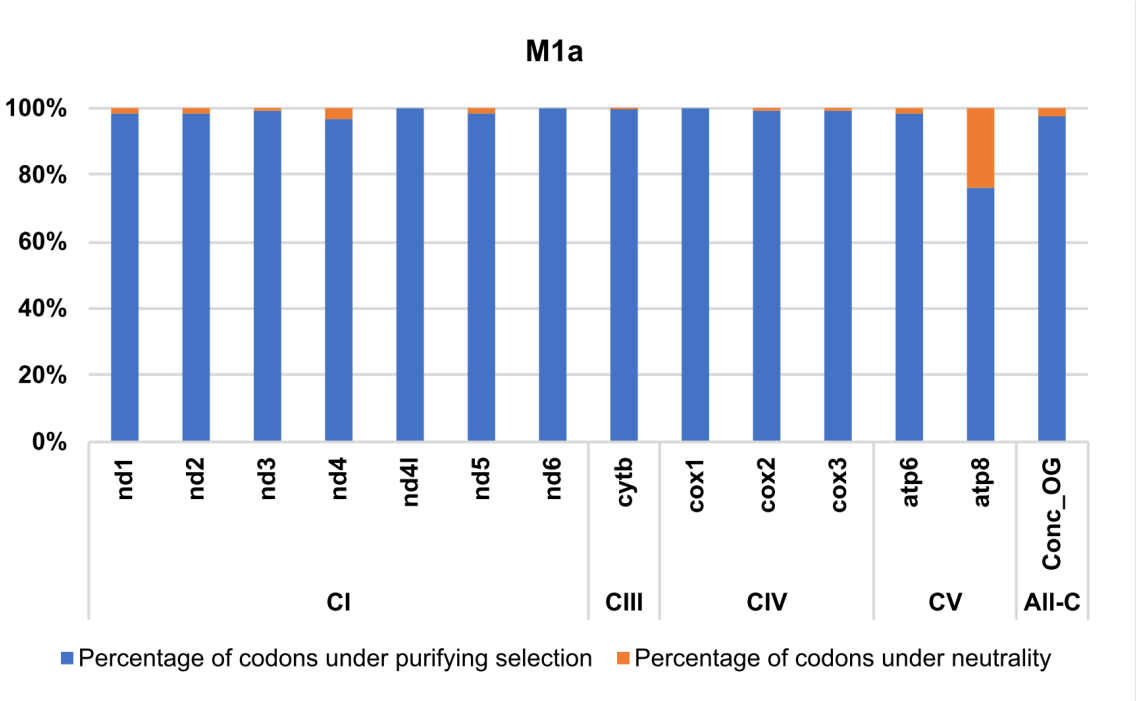


Figure 3.4: **Percentage of sites under purifying selection and neutrality calculated under random-site model M1a.** Mitochondrial protein-coding genes: *nd1-6, cytb, cox1-3, atp6, atp8*, and Conc_OG (concatenated OG-data set). Complexes are abbreviated as: CI - Complex I, CIII - Complex III, CIV - Complex IV, CV - Complex V, and All-C - Complexes I, III, IV, and V.

Similar trend was observed when the M7 and M8 models were compared, except for *nd4* and concatenated alignments, and both identified a signal of positively selected sites with significant *p*-values of less than 0.05. However, a confirmatory test model of positive selection (M8 *vs.* M8a) clarified that this result is really under relaxed purifying selection (Appendices III: Table S7). Thus we conclude that no significant finding of positive selection was provided by any random-site model.

Branch-site models revealed positively selected sites on the ancestral-lineage leading to Oegopsida and Bathyteuthoidea clades

Despite the strong background of purifying selection detected in these genes, the branch-site model was able to reveal episodes of positive selection acting only at certain sites on the ancestral-branch (foreground-lineage) that leads to the Oegopsida and Bathyteuthoidea clades (Appendices III: Table S8). This test recovered significant evidence (Appendices III: Table S8 – *p*-values < 0.05) of positively selected sites within two genes, that are, *atp6* (sites 46 and 194) and *cox2* (sites 129 and 132). The numbering of these sites refer to the genome of *Sepiella maindroni* (NC_028731). The analysis with the concatenated-alignment also revealed to be significant, but BEB was not sensitive enough to identify sites under selective constraint (Appendices III: Table S8 – Conc_OG).

Ancestral ATP6 and COX2 amino acid sequences were inferred at each node in order to reconstruct the evolutionary history of these sites (using MEGA software version 5.2.2 [266]). Through this reconstruction along the phylogeny we notice various substitutions in the ancestral branch of Oegopsida-Bathyteuthoidea and in some of the terminal branches of these clades (Figure 3.5). For instance, one amino acid substitution was observed in the ancestral branch in ATP6 site 46, which introduced a larger residue replacement (S46T). Further, we see another substitution along the branch terminating in the species *S. oualaniensis* (T46N); however, this amino acid change was conservative. For site 194 of the same protein, we also observed one replacement in the ancestral branch, which introduced a polar uncharged amino acid substitution (L194T), while three substitutions were observed in branches terminating in species *O. bartramii* (T194A – small size-hydrophobic non-polar), *S. oualaniensis*, *D. gigas*, *T. pacificus*, *I. argentinus* (T194M – medium size-hydrophobic non-polar), and *A. dux* (T194S – small size-weak acidic-polar uncharged), respectively. For

COX2 site 129, we observed one substitution in the ancestral branch, which introduced a negatively charged amino acid (N129E), and one substitution in species *A. dux* (E129D), but this last replacement was conservative. Finally, for site 132 of COX2, one substitution was introduced in the ancestral branch (D132N), changing the amino acid property to polar uncharged and further maintained in the terminal branches.

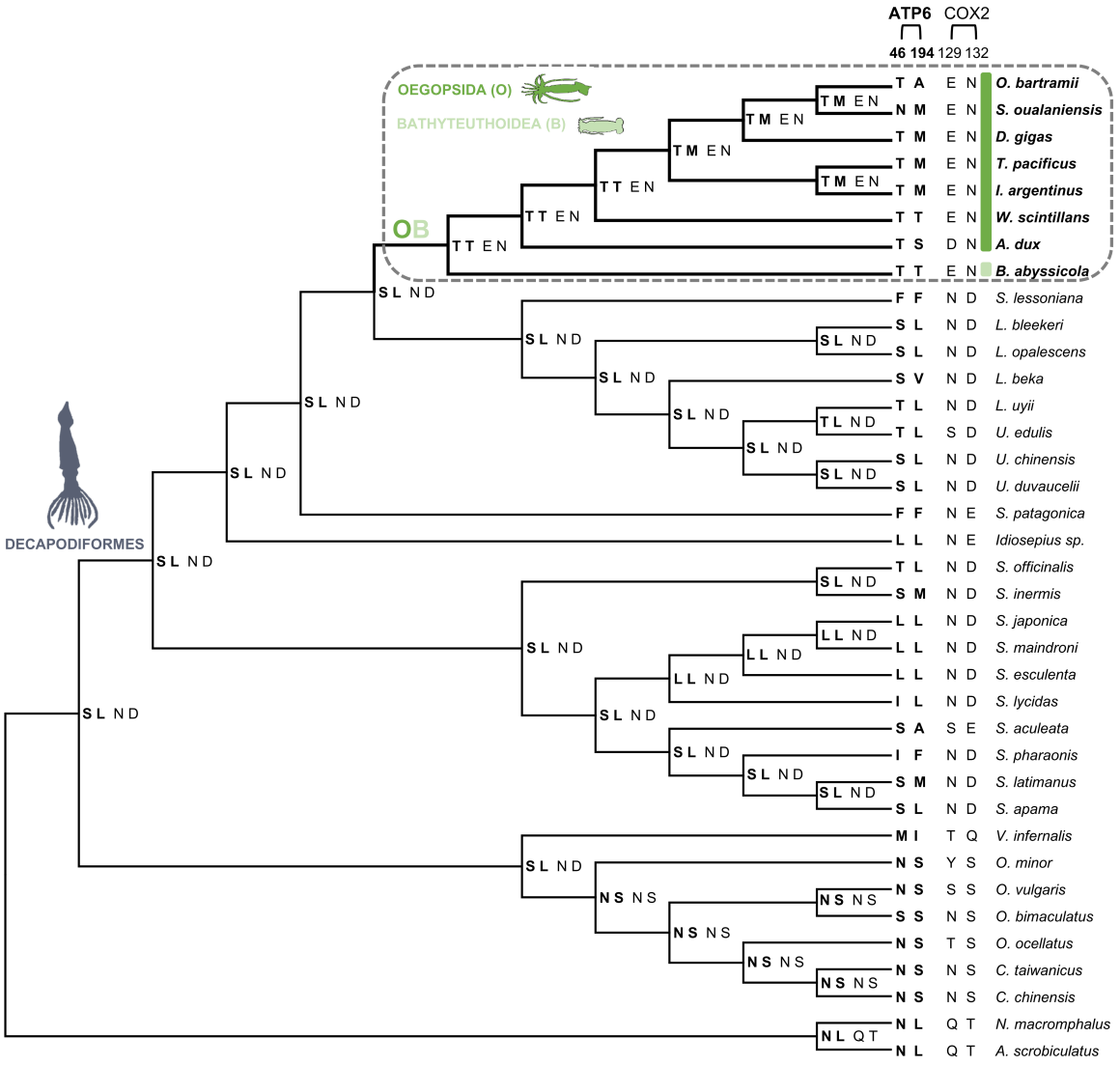


Figure 3.5: ATP6 and COX2 cladogram showing the inferred substitution history for each of the four positively selected sites identified by branch-site BEB analyses. Amino acids (AA) are indicated using one-letter code, next to each node and tips (species), in the following order: first two AA in bold corresponding to ATP6 sites (46 and 194, respectively) and the latter two AA corresponding to COX2 sites (129 and 132, respectively). Numbering follow cephalopod *Sepiella maindroni* (corresponding to bovine: ATP6 – T44 and L190, and COX2 – K129 and E132, respectively. Colorful vertical bars indicate the two major taxonomic groups of cephalopods Oegopsida and Bathyteuthoidea (OB), respectively. OB – foreground branch tested (Decapodiformes ancestral transition from demersal to pelagic environment, and unique species with duplicated genes).

Clade model C analyses revealed divergently evolving sites on the Oegopsida and Bathyteuthoidea clades

The alternative CmC model was found to be a better fit than the M2a_rel on the entire Oegopsida and Bathyteuthoidea clades (*p*-value < 0.05), for both individual-genes (*atp6, cytb, nd1, nd4, nd4l, nd5*, and *nd6*), and for the concatenated alignment (Conc_OG) (Appendices III: Table S9). Accordingly, BEB identified several sites that were divergently evolving, except for *nd6*, to which no site was reported (Appendices III: Table S9). For all aforementioned alignments, the ω ratio for the divergent selection SC2 was less than one, and values were extremely low, ranging from 0.001 to 0.073 (Appendices III: Table S9). Further, in all cases the ω ratio remained almost the same ($b\omega 2 \approx f\omega 3$), *i.e.,* neither increase nor decrease, from the background ($b\omega 2$) to the foreground ($f\omega 3$) (Appendices III: Table S9). Therefore, despite the significant improvement over the M2a_rel, these results most likely reflect strong purifying selection or simply that the CmC model has difficulties to detect any positive selective constraint within this data set.

Physicochemical amino acid properties of sites positively and divergently evolving

Some of the BEB-identified sites (PP ≥ 0.70) were common between branch-site and CmC models, as well as, between branch-site or CmC with PRIME analyses: *(i)* two sites (46 and 194; Appendices III: Tables S8 and S9) in *atp6* overlapped between branch-site and CmC, and are reported by PRIME analyses as having conserved and no changes in physicochemical properties (Appendices III: Table S10); *(ii)* one site (129) from *cox2*-branch-site model, was highlighted in PRIME as conserved for pHi but with physicochemical changing properties that resulted in a shift in volume; all other remaining sites *(iii)* with changing (*nd1* – 305, and *nd4* – 170, 186, 294) and/or conserved (μ, pHi, V, SS, P and CC – Appendices III: Table S10) physicochemical properties were only detected in CmC. Overall, PRIME suggested that adaptive evolution may have affected five sites (see sites with changing properties – Appendices III: Table S10).

ATP6 and COX2 three-dimensional mapping of positively selected sites

Observing the 3D structure of bovine Complex V-ATP synthase (Figure 3.6) – includes a

mitochondrial matrix F1 catalytic-region and a transmembrane (TM) F0 region (accessible from intermembrane space) – we were able to locate the BEB-identified sites in ATP6 (*i.e.,* a-subunit in the F0 region) as follows: *(i)* site T44 (site 46 of cephalopods) is located within the external α-helix 2, which is found in the matrix side of F0 region and links TM1 or α-helix 1 with the membrane-inserted (MI) helical-hairpin that is composed by α-helices 3 and 4; *(ii)* site L190 (site 194 of cephalopods), is located in the N-terminus of α-helix 6 and facing the intermembrane space.

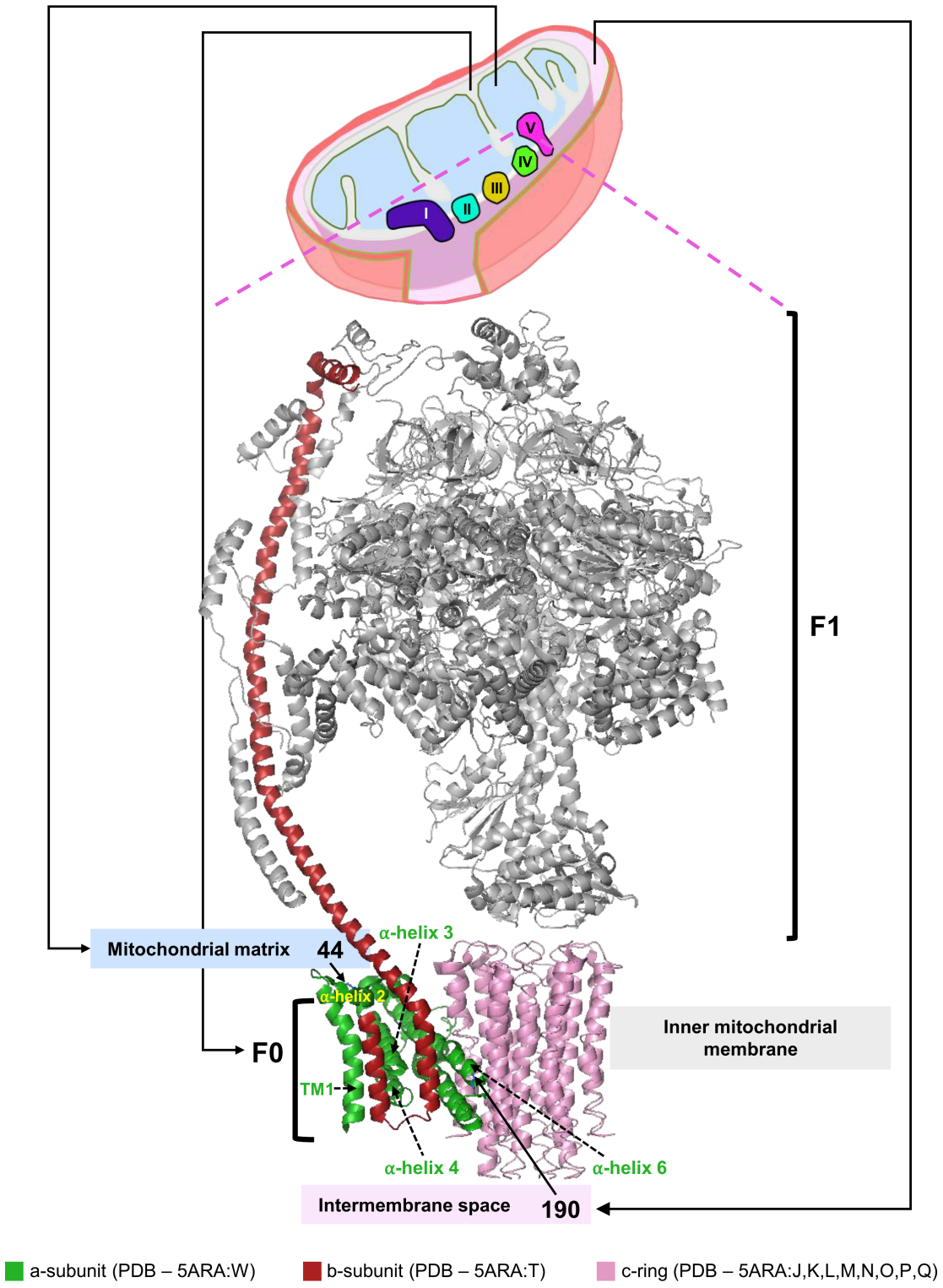


Figure 3.6: 3D structural model of Complex V showing the location of sites inferred to be evolving under positive selection according to branch-site BEB analyses. Only sites with PP ≥ 0.70 are shown. The protein structure and numbering follow bovine ATP6. TM1 – transmembrane 1 or α-helix 1. F1 – factor 1. F0 – factor that confers sensitivity to oligomycin (F0 forms a proton turbine). C-ring – rotate against the a-subunit (ATP6) forming a proton turbine.

Observing the 3D structure of bovine Complex IV-*cytochrome c oxidase* – includes three core subunits COX1, COX2 and COX3 – we found that the sites K129 and E132 (same numbering in cephalopods) were located within a loop in the intermembrane space (Figure 3.7).

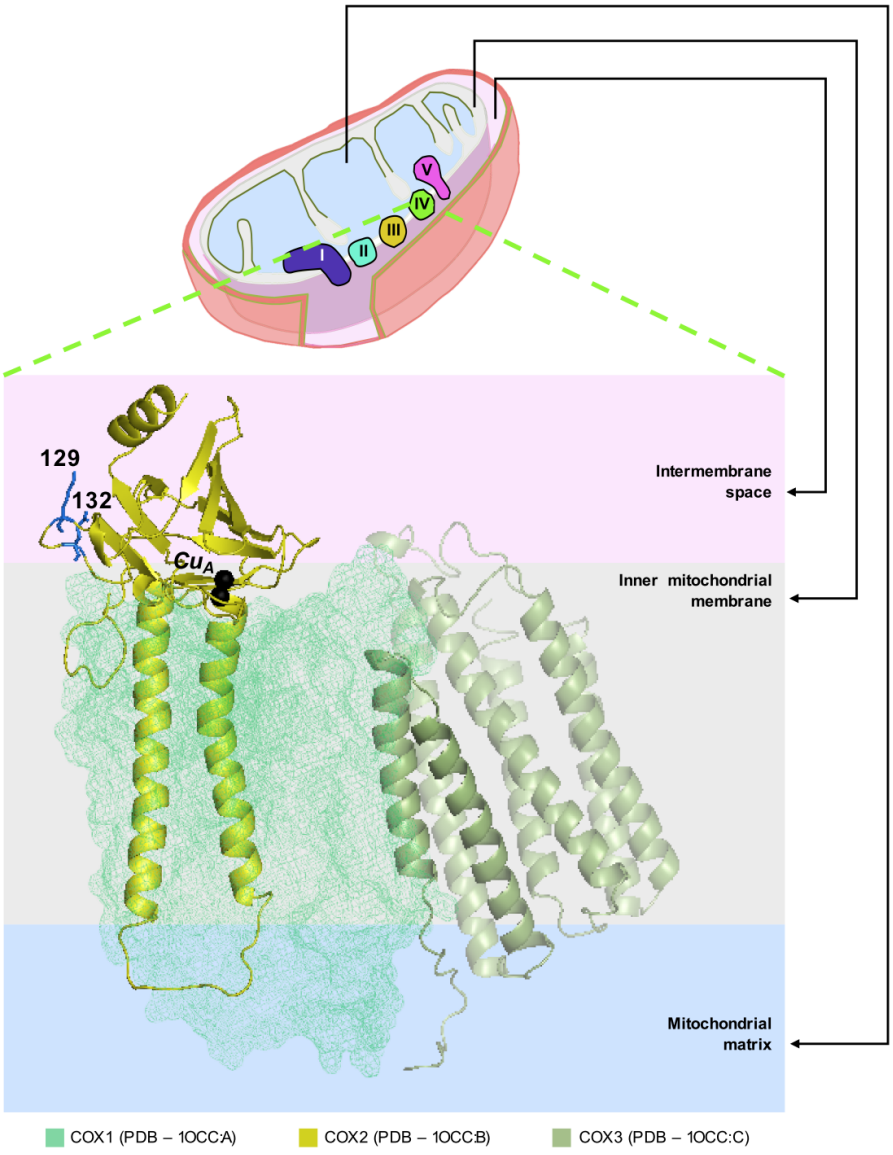


Figure 3.7: 3D structural model of Complex IV showing the location of sites inferred to be evolving under positive selection according to branch-site BEB analyses. Only sites with PP ≥ 0.70 are shown. The protein structure and numbering follow bovine COX2. Cu_A – binuclear copper center (black). Adapted from [236].

3.5 Discussion

The current study – consisting of concatenated 13-complete mitochondrial genes from 37 species of major taxonomic Cephalopoda groups – provides new insights of an enigmatic group of pelagic squids (Oegopsida and Bathyteuthoidea, hereafter named OB), whose ancestor experienced a transition from demersal to the pelagic environment. This important adaptation conferred an extraordinary ability for exercising distinct bioenergetics mechanisms that aid them to tolerate distinct environmental conditions (*e.g.,* variation in oxygen concentrations) [190, 267]. Moreover, they hold mtDNA duplicated genes (*atp6, atp8, cox1, cox2, cox3*), a feature not found in other Cephalopoda lineages [185, 238, 268]. The alignment of both copies (from the same cephalopod species) corroborate previous studies [185,238,268], still indicating that the sequences of duplicates are near-identical (Appendices III: Table S5); a result in agreement with the idea that both copies are functional [238]. Furthermore, the maintenance of these copies possibly meets high expression demands of OB species [269].

We used several methods (ML, BA and codon substitution models) to explore *(i)* phylogenetic reconstruction of Cephalopoda (considering three distinct data sets Conc_L, Conc_R and Conc_OG, each one in three distinct forms – "original", RY-coded, and 1^{st} and 2^{nd} codon positions), while taking into account the best orthologous (OG) mtDNA gene copy, and *(ii)* adaptive evolution of the OB.

Overall, phylogenetic reconstructions resulted in a well resolved and largely congruent tree topology (Figures 3.1–3.3). Small disagreement between ML and BA analyses occurred solely on the location of the genera *Uroteuthis*, *Sepia*, species *S. oualaniensis* and *D. gigas* when all third codon positions of data sets were coded as R (purines: adenine and guanine) and Y (pyrimidines: thymine and cytosine), *i.e.,* RY-coded (Figure 3.2). All the data sets in their original (Figure 3.1) and 1^{st} and 2^{nd} (Figure 3.3) forms generated congruent topologies between ML and BA approaches, whereby differences in position of *W. scintillans*, *A. dux* and genus *Sepia* were possibly caused by loss of codon information. We also notice that *(i)* the topology produced using each one of the three data sets (Conc_L, Conc_OG or Conc_R) were always the same and *(ii)* the support values conferred to each node were also high regardless of the strategy used (ML, BA and data set forms); this suggests that for the cephalopods in study, the inclusion of the best duplicated homologous sequences (orthologues) is not a major issue in phylogenetic analyses using concatenated mitochondrial protein-coding genes.

Furthermore, based on the above-mentioned and in the statistical tests applied (AU, NP and KH), the tree topology that largely best described cephalopod major clades relationships was that obtained from data sets in their original forms (Figure 3.1). Our topology is broadly consistent with other published studies of cephalopods (*e.g.,* [190]), showing major taxonomic groups as monophyletic: Oegopsida, Myopsida and, Sepiida. Regarding highly debated cephalopod positions, our study recovered: *(i) Idiosepius sp.* as a sister taxon to Oegopsida, similarly to a previous study performed by [270]; *(ii)* support to the monophyly of Vampyromorpha and Octopoda (forming Octopodiformes clade), a result consistent with strong morphological and embryological evidence revealed by [271]; *(iii)* a sister-taxon relationship between OB in accordance with previous molecular [190, 193, 272, 273] and morphological [274] studies. Thus, we concluded that the recommended topology to use for further analyses is the one provided by the complete-codon data set information.

Our molecular evolutionary analysis which was investigated through the use of codon substitution models (random-sites, branch-site and clade model) and physicochemical properties confirm that purifying selection is the dominant force shaping the evolution of Cephalopoda mtDNA, which is expected given its crucial and conserved role for body energy production in cephalopods [236] and metazoans in general [261]. However, evidence of positive selection was also revealed along the ancestral branch leading to OB, as well as, along the Oegopsida phylogeny. This was identified primarily by the branch-site model, while no indication of selective constraints was suggested by the random-site model. We also found that although the clade model was able to suggest divergent selective pressures among the partitions, and that few of these sites were shared with the branch-site model, the results were insufficient to support overall differences between the groups. Nonetheless, CmC results were only validated when overlapped with those of branch-site test. Within the branch-site analyses we observed signatures of strong positive selection in two genes, *atp6* and *cox2*, which encode protein subunits belonging to Complex V and IV, respectively. Interestingly, these genes are two of the set of genes with duplicated copies found in OB. The positively selected sites from *atp6* were also supported by CmC. The BEB identified sites along *atp6* and *cox2* genes resulted in amino acid substitutions that generally were non-conservative. In this sense, the ancestral reconstruction analyses (Figure 3.5) along the ancestral branch of OB revealed at ATP6 one site (S46T) that experience size-changing replacements, while the second site (L194T) underwent polarity-changing substitutions. For

COX2, two sites (N129E and D132N) indicated charges-changing substitutions (Figure 3.5). Additionally, some derived amino acids along the Oegopsida phylogeny were conservative, except for site 194-ATP6 which underwent additional stringent substitutions (*i.e.,* polarity, size or a combination of both). This was observed along the ancestral of Oegopsida (T194M), and in two of the terminal branches (M194S – *A. dux*, M194A – *O. bartramii*; Figure 3.5). We notice that site 194 resides in a structural region of the F0 a-subunit that is postulated to play a critical functional role in proton translocation and rotor performance (Figure 3.6). Moreover, the last site is located within the flexible and highly tilted TM α-helix 6, which is also in contact to the rotatory c-ring helices through the a-subunit interface, while in close proximity to the tilted TM α-helix 5 where the proton moving couple mechanism is found [262, 265]. The TM α-helix 6 is also important in humans, since a mutation in position L217R or P is associated with Leigh syndrome, a severe disease that decreases the rate of ATP production [275]. Based on this scenario, we infer the possibility of cooperative conformational changes between TMs α-helix 5 and 6 given that mutations on site 194 seem to represent important replacements for efficient ATP6 protein function in OB, and consequently for conserving the overall functionality of Complex V. Similar positions (ATP6-43 and 195), under selective constraint, have been indicated for freshwater eels, and are considered key sites linked to eel migrations from freshwater or coastal regions to the ocean [276] and in mammals sites ATP6-43 and 52 have been reported as well under selection [76]. We also notice that the two BEB sites (129 and 132) of COX2 occupy a functional section of Complex IV (Figure 3.7) that is highly conserved, namely the negative region that surrounds the area that is directly involved in the electron transfer from cytochrome c to the binuclear copper (Cu_A) center [277]. Mutations in this region are known to cause exercise intolerance in humans (*e.g.,* L135P). This implies that the amino acids substitution from our BEB selected sites seem not to dramatically alter the structure of the region, such that, they are compatible with function, as it has been experimentally supported by some amino acid replacements along the conserved regions of COX2 [263]. Although it is difficult to speculate with confidence on the potential effects of these specific substitutions found, both in ATP6 and COX2, we predict a possible effect on protein, likely in relation to metabolic efficiency. We infer that these mutations reshaped Complex IV and V of OB clades, thereby harboring these mutant proteins, modified the response of OB metabolic and energetic recovery rates. Clearly, OB will have experienced significant changes in their metabolic needs as they occupied different

niches along the pelagic oceanic environment and likely impose diversifying selection on *cox2* and *atp6* sequences.

Overall, our study yields new insights into the adaptation of Oegopsida-Bathyteuthoidea clade, a special group of fast-moving pelagic cephalopod-squids. Since these species migrate vast distances through the open ocean (*e.g.,* as part of its life cycle) and through the waters column (diel migration – to find food and/or avoid predators) we suggest that these squids take advantage of having *(i)* mtDNA duplicated copies and *(ii)* enhanced or compensatory amino acid substitutions, in order to sustain bioenergetics efficiency across ocean oxygen fluctuating conditions [278], to perform sudden locomotive swimming-jet propelled and to attain fast growth – all of which are associated with extremely high metabolic rates and energy requirements, compared to all other marine invertebrates [267]. Certainly, more Oegopsida sampling and further mtDNA sequencing are needed to better understand the many interesting features of this group, such as *(i)* the functional evolutionary history of cephalopods mitochondrial duplicated genes and *(ii)* the functional impact of the reported positively-selected sites in the OXPHOS pathway. In fact, although, at the present it is clear that mtDNA genome duplications do occur in other metazoan (*e.g.,* some birds – degenerated copy of the duplicated *nd6*, and *cytb* reduced in size), the complete-duplications refer only to noncoding control regions [269]. Thus, further studies are necessary to understand the evolutionary mechanism for the origin and maintenance of several duplicated mtDNA protein-coding genes in Oegopsida-Bathyteuthoidea clade, as well as, the potential costs and benefits of having a second copy of these genes. Additionally, we strongly recommend experimental evidence of the functional implications of some of the sites substitutions found in this study, given the positive correlation between Complex V flexibility and the rapid rate of enzymatic activity [279]. Specifically, we refer to position ATP6-variation L194T; this will help clarify the potential role of selection acting on the rotatory c-ring helices (*i.e.,* proton translocation across half channels near the interface between a-subunit and c-ring) and how this may relate to other adapted aquatic organisms.

Part II

ADAPTIVE EVOLUTION OF NUCLEAR GENES IN THE METAZOA: INSIGHTS FROM CLASS AVES (PHYLUM CHORDATA)

Chapter 4

Whole-Genome Identification, Phylogeny, and Evolution of the Cytochrome P450 Family 2 (CYP2) Subfamilies in Birds

Published in

Almeida D*, Maldonado E*, Khan I*, Silva L, Gilbert MTP, Zhang G, Jarvis ED, O'Brien SJ, Johnson WE, Antunes A

Genome Biology and Evolution, 2016, 8(4):1115-1131.

DOI: 10.1093/gbe/evw041

* - These authors contributed equally

4.1 Abstract

The cytochrome P450 (CYP) superfamily defends organisms from endogenous and noxious environmental compounds, and thus is crucial for survival. However, beyond mammals the molecular evolution of CYP2 subfamilies is poorly understood. Here, we characterized the CYP2 family across 48 avian whole-genomes representing all major extant bird clades. Overall, 12 CYP2 subfamilies were identified, including the first description of the CYP2F, CYP2G, and several CYP2AF genes in avian genomes. Some of the CYP2 genes previously described as being lineage-specific, such as CYP2K and CYP2W, are ubiquitous to all avian groups. Furthermore, we identified a large number of CYP2J copies, which have been associated previously with water reabsorption. We detected positive selection in the avian CYP2C, CYP2D, CYP2H, CYP2J, CYP2K, and CYP2AC subfamilies. Moreover, we identified new substrate recognition sites (SRS0, SRS2_SRS3, and SRS3.1) and heme binding areas that influence CYP2 structure and function of functional importance as under significant positive selection. Some of the positively selected sites in avian CYP2D are located within the same SRS1 region that was previously linked with the metabolism of plant toxins. Additionally, we find that selective constraint variations in some avian CYP2 subfamilies are consistently associated with different feeding habits (CYP2H and CYP2J), habitats (CYP2D, CYP2H, CYP2J, and CYP2K), and migratory behaviors (CYP2D, CYP2H, and CYP2J). Overall, our findings indicate that there has been active enzyme site selection on CYP2 subfamilies and differential selection associated with different life history traits among birds.

4.2 Introduction

Cytochrome P450 (CYP) genes encode hemeproteins [84] in a wide variety of organisms [83]. CYPs confer protection against reactive oxygen species that form in organisms after exposure to toxins and other environmental contaminants, including the drugs and carcinogenic compounds present in food. They are mainly expressed in the liver endoplasmic reticulum, but are also highly expressed in the small intestine and olfactory mucosa, suggesting that they have tissue-specific roles [280]. CYP enzymes are involved in phase I of detoxification [281] and are typically membrane-bound [282]. Usually they act as terminal

oxidases in multicomponent electron-transfer chains called P450-containing monooxygenase systems [283] and contribute to the inactivation and excretion of several endogenous and exogenous noxious metabolites via urine or bile [281]. The large variety and number of xenobiotics constantly encountered by species offer numerous challenges. To recognize and efficiently metabolize the array of common and novel substrates [281], CYPs have evolved multiple gene families consisting of several members with a diverse range of substrate specificities and regulation pathways [86]. These genes are suggested to be among the fastest-evolving gene systems [281] and have been linked with migratory behaviors, adaptations to novel habitats [86–89], more-efficient water retention [284], and food selection [285]. Their role in drug interactions and processing are also of interest to the pharmaceutical industry [286]. CYPs are vital in mechanisms of resistance to natural and synthetic compounds that potentially interfere with normal growth, development, and reproduction through their role in the processing of endogenous substrates [287]. CYPs unique features, including high genetic diversity, broad substrate specificity, and catalytic versatility, enable them to deal with a wide variety of substrates [85], fostering adaptation to and survival in new environments (*e.g.,* migratory species, invasive species and resistance to pest controls drugs).

Since CYPs are a gene superfamily, their nomenclature system is based on a hierarchical clustering of genes into families and subfamilies [288]. CYP families are named by number (*e.g.,* CYP2), the subfamilies by capital letters (*e.g.,* CYP2C), and the specific genes by a second number (*e.g.,* CYP2C8). By convention, members of new CYP families must share greater than 40% amino acid identity, while members of subfamilies must share greater than 55% amino acid identity [83, 153, 288].

In vertebrates, the CYP2 family (29 subfamilies) is one of the largest and most diverse and has the least-conserved nucleotide sequences [288, 289]. Typically, these genes have nine exons and are approximately 1500 base pairs (bp) long. The CYP2U and CYP2R subfamilies are considered to be the most basal [288,290]. In spite of their diverse and critical roles, understanding of the relationships among CYP2 subfamilies beyond the mammals is limited [90]. Prior to our analyses, eight CYP2 subfamilies had been characterized (CYP2AB, CYP2AC, CYP2C, CYP2D, CYP2J, CYP2R, CYP2U, and CYP2W) in a limited number of bird species (chicken–*Gallus gallus*, turkey–*Meleagris gallopavo*, and zebra finch–*Taeniopygia guttata*) [291]. Because avian species have diverse feeding habits,

specific adaptations, and a worldwide distribution, they are exposed to a wide variety of compounds (environmental chemicals) and likely have developed an array of novel xenobiotic-metabolizing mechanisms [291]. To test this idea, here we assessed the evolutionary history of the avian CYP2 family by conducting detailed analyses of gene content, adaptive evolution, and phylogenetic patterns across the whole-genomes of 48 bird species from 36 orders from the recently conducted avian Phylogenomics Project [174, 292], including species from the three major avian evolutionary groups: Palaeognathae, Galloanserae, and Neoaves.

4.3 Materials and Methods

CYP2 gene sequences

To characterize avian CYP2 genomics evolutionary diversity, we employed tBLASTn searches [30] on 48 sequenced avian genomes from the Avian Phylogenomics Project in the GigaScience Database [180] and other sources [293, 294] using as query sequences individual CYP2 subfamily protein sequences annotated in Ensembl [28] (release 75) for chicken (*G. gallus*), turkey (*M. gallopavo*), anole lizard (*Anolis carolinenesis*), frog (*Xenopus tropicalis*), zebrafish (*Danio rerio*), and human (*Homo sapiens*). From the CYP2 sequences retrieved, only nucleotide sequences with more than 1125 bp and high-identity (*e*-value $< 1e^{-5}$) were considered for further analyses. We then submitted these sequences from all 48 avian species to a BLASTx search (NCBI), which searches the protein database of all vertebrate species in NCBI using a translated nucleotide query optimized to find highly similar sequences, to accurately characterize the avian CYP2 subfamilies.

CYP2 phylogenetic analysis and filtering of gene subfamily data sets

We performed a codon-based alignment of all identified avian nucleotide CYP2 sequences (genes and pseudogenes) along with some reference sequences of CYP2 subfamilies found in public databases (Ensembl release 75: http://www.ensembl.org and NCBI: http://www.ncbi.nlm.nih.gov, last accessed March 2014) for anole lizard, Chinese alligator (*Alligator sinensis*), green turtle (*Chelonia mydas*), frog, zebrafish, human, chicken, medium ground-finch (*Geospiza fortis*), common ostrich (*Struthio camelus australis*), and cormorant (*Phalacrocorax carbo*). Sequences were aligned in a “global data set” using MUSCLE [31]

as integrated in the SEAVIEW 4.4.0 software package [186]. This alignment was tested for saturation bias using the Xia *et al.* statistic test [188] implemented in DAMBE 5.3.31 [40].

To assess the adequacy of current consensus avian CYP2 subfamily nomenclature, a Maximum likelihood (ML) phylogeny was estimated using our "global data set", which showed no significant evidence of saturation (*p*-value < 0.05). This ML phylogeny assumed a General Time Reversible (GTR) evolutionary model, with a proportion of invariable sites (I) and heterogeneity of substitution rates among sites modeled following a gamma distribution (G), as determined by JMODELTEST 2.1.1 [37]. The ML phylogeny was estimated using PHYML 3.0 [41] with 100 bootstrap replicates and the Nearest Neighbor Interchange (NNI) branch search algorithm.

Based on this ML phylogeny, we retrieved the CYP2 nucleotide sequences from each well-defined clade in order to guarantee high identity among sequences that were compiled to create the "CYP2 subfamily data sets". Each subfamily data set was inspected closely to ensure there was only one CYP2 sequence per avian species and to remove pseudogenes and databases reference sequences. Several MUSCLE [31] alignments were constructed to corroborate previous avian CYP2 gene classifications, a process which led to the identification of smaller (incomplete) sequences, which were removed from their respective data sets. Avian CYP2 sequences with evidence of recombination or gene conversion events (Bonferroni corrected *p*-values < 0.05) were also removed from the data sets. Recombination was assessed with the RDP4 software package using default settings and seven algorithms (RDP, GENECONV, Chimaera, MaxChi, SiScan, BootScan, and 3Seq) [295]. Following this approach we obtained 17 "final avian CYP2 subfamily data sets": "CYP2AC", "CYP2AF", "CYP2C", "CYP2D", "CYP2H", "CYP2J", "CYP2J_1", "CYP2J_2", "CYP2J_3", "CYP2J_4", "CYP2J_5", "CYP2K_1", "CYP2K_2", "CYP2R", "CYP2U", "CYP2W_1", and "CYP2W_2", which were used for selection analyses. Appendices IV: Table S1 details which species are represented in each data set.

Ancestral reconstruction analysis of avian CYP2 subfamilies

To elucidate the evolutionary process of avian CYP2 subfamilies, we performed ancestral reconstructions using the COUNT software and employing default parameters [296]. We used only the CYP2 subfamilies as identified above (threshold > 1125 bp), and the total evidence nucleotide species tree (TENT) of the Avian Phylogenomics Project [292], which

was converted into an ultrametric format with the R8S 1.8 software following the author's instructions [252]. In this approach, the numerical gene profiles (number of genes present in each avian species per subfamily) were first converted into binary format (1–present or 0–absent) and the data were posteriorly analyzed using the Dollo parsimony model [297].

Selection analyses and reassessment of the CYP2 substrate recognition sites

We estimated the nature and strength of the evolutionary selection pressures at the molecular level by assessing ratios of nonsynonymous (*dN*) to synonymous (*dS*) substitution rates, or omega (ω = *dN/dS*), where ω greater than, equal to and less than 1 is indicative of positive, neutral and negative selection, respectively. Strong negative selection ($\omega < 1$) pressures generally prevent the accumulation of amino acid changes in the regions of proteins that are essential for its structure and/or function [161]. In contrast, novel functionalities are often driven through positive selection ($\omega > 1$) favoring amino acid replacements in protein-coding genes [298].

We started by employing site-models [18, 194] and branch-specific [6, 47] likelihood analyses. For both analyses, we submitted the codon-based alignments of the "final avian CYP2 subfamily data sets" (Appendices IV: Table S1) with the respective unrooted avian TENT [292] to the CODEML program from the PAML 4.7 package [20, 43].

We considered different codon substitution models (site-models) which allow the ω ratio to vary along sequences in different ways: *(i)* null models – M0 model that admits uniform selective pressure among sites and M1a, M7, and M8a models that do not allow sites with $\omega > 1$ and *(ii)* alternative models – M3 model which assumes variable selective pressures among sites and M2a and M8 models which allow sites with $\omega > 1$ [8]. In these analyses, likelihood ratio tests (LRTs) were conducted by comparing the null models with the alternative models: M0 *vs.* M3 [195–198], M1a *vs.* M2a [8, 18, 45], M7 *vs.* M8 [194], and M8a *vs.* M8 [46] to infer which models best fit the data. Whenever the LRT was significant (*p*-value < 0.05) under the models M2a and/or M8, the codon sites under positive selection were identified using the Bayes Empirical Bayes (BEB) calculation, which analyzes the posterior probabilities (PP) for these sites [45]. We only considered positively selected sites with PP $> 95\%$.

CYP2 subfamily data sets with evidence of positively selected sites (from site-models) were also submitted to branch-specific likelihood analyses [6,47] to assess if its ω-ratio varied

significantly among distinct avian groups (branches of interest–foreground lineages) of the phylogeny. In these analyses, alternative branch models (with multiple ω-ratios for foreground lineages) were tested against simpler null models (which assume that all branches in the phylogeny are evolving at the same rate). The foreground lineages for alternative models were specified *a priori* based on the following categories: feeding habits (carnivorous, herbivorous, and/or omnivorous birds), habitat (dry, moist, and/or semi-moist) and migration (migratory and non-migratory). Appendices IV: Table S2 contains the correspondence between each avian species and the above-mentioned traits. In order to guarantee a robust grouping of branches into several partitions, where the strength of selection may be different (alternative models), we: *(i)* generated stochastic character maps for each trait across the previously obtained ultrametric TENT [292], following the method of Bollback [253], as implemented in phytools [254] and geiger [255] R packages, using R 3.2.2 software [256]; *(ii)* labeled the unrooted TENT according to these mapping results; and *(iii)* trimmed the resulting labeled TENT to retain only the avian species represented in each one of the corresponding "final avian CYP2 subfamily data set" (Appendices IV: Figures S1-S11).

For both site-models and branch-specific selection analyses we applied the F3x4 codon model [9] allowing for ML estimation of κ (transition/transversion ratio) and ω. All the models were run several times, adjusting the initial κ and ω values in order to avoid possible local-likelihood peaks. For all model comparisons, the hypothesis decision threshold was calculated by doubling the difference between the alternative and null model log-likelihood (2ΔlnL) and assuming that the null distribution of these results could be approximated by a chi-square (χ^2) distribution (*p*-value < 0.05). The number of degrees of freedom (*df*) was calculated as the difference in the number of estimated parameters between the models [8, 44]. We used the IMPACT_S software to automate these calculations [202].

To search for site-specific amino acid properties that are being preserved (conserved properties) or modified (changing properties) through the evolutionary process, we used the PRoperty Informed Models of Evolution (PRIME) method (Pond, unpublished work) as implemented in the DATAMONKEY webserver [19, 52]. The PRIME method considers two predefined sets of five physicochemical amino acid properties. These include five empirically measured amino acid properties proposed by Conant *et al.* (2007): *(i)* chemical composition [201] of the side chain [CC], *(ii)* residue polarity [P], *(iii)* volume [V] of the residue side chain, *(iv)* isoelectric point [pHi] of the side chain, and *(v)* hydropathy [H] [257]

and five composite properties proposed by Atchley *et al.* (2005): *(i)* polarity index [P], *(ii)* secondary structure factor [SS] [201], *(iii)* volume [V], *(iv)* refractivity [μ], and *(v)* isoelectric point [pHi] [258]. The estimates of amino acid exchangeabilities implemented by this method are based on multiple tests performed on the same residue site. Therefore this method includes the Bonferroni correction to control the number of false positives reported at a site. From the sites reported by this approach, we only considered and analyzed those that were coincident with previously identified sites with significant evidence of positive selection by BEB.

CYP2 enzymes have substrate recognition sites (SRS), where the amino acids are close to the ligands and thus influence substrate recognition and/or binding [157] and induce chemical and structural variations that are reflected on the size, shape, and chemical features of substrates and products [161]. To map the positively selected sites onto the three-dimensional (3D) structure of the CYP2 subfamilies and to facilitate visualization of sites hypothesized to be under important SRS, we first considered the available information about six CYP2 SRS [157]. These six SRS were described by Gotoh in 1992 (SRS*) from the alignment of mammalian CYP (1, 2 and 3) sequences with the CYP101A1 sequence from the bacterium *Pseudomonas putida*, whose substrate-binding sites were identified by X-ray crystallography of a substrate-bound form [299]. To update these six regions, we first searched the Protein Data Bank (PDB) database (release April 8, 2014) to obtain CYP2 X-ray crystal structures with the appropriate ligand annotations, to further perform their amino acid alignment (MUSCLE) [31] with a consensus sequence showing the six Gotoh's SRS. Through consensus alignments–using the GENEIOUS 5.6.7 consensus align option [187] – of the sequence containing the annotations of updated SRS with each one of the avian CYP2 subfamily alignments, we verified if their positively selected sites were within important SRS. We then performed homology modeling of the 3D structure of the avian CYP2, which showed evidence of positive selection, using the SWISS-MODEL webserver [55, 57]. If the avian predicted models were not reliable, we only mapped the positively selected sites onto the 3D structure when such models (from the same subfamily) were available in the PDB database for other vertebrates (*i.e.,* human CYP2C and CYP2D PDB codes: 2VN0 and 3TDA, respectively). The superimposition, visualization and manipulation of the 3D structures were performed with PYMOL 1.5.0.4 software [58].

Statistical analyses on trait associations

In order to understand if the distribution of the number of CYP2 genes could be used to differentiate bird species according to their migratory/non-migratory behavior, we performed a linear discriminant analysis (LDA). The classification variable was the migratory behavior of the species, with two classes (migratory/non-migratory) and the independent variables were the ten CYP2 subfamilies (CYP2F and CYP2G were excluded as they are outliers). The percent of correct predictions in each class was evaluated by cross-validation. These analyses were performed with the lda function from MASS R package, using R 3.2.2 software [256].

4.4 Results

CYP2 genes have diverse paralogs depending on subfamily

The BLAST analyses performed in the 48 avian genomes identified 642 CYP2 (including genes and pseudogenes) with sequence identity with other gene members available varying between 61% and 100%. Following the current nomenclature system [83, 153, 288], these BLAST results identified 12 CYP2 subfamilies in birds: CYP2AC, CYP2AF, CYP2C, CYP2D, CYP2F, CYP2G, CYP2H, CYP2J, CYP2K, CYP2R, CYP2U, and CYP2W (Figure 4.1).

Subfamilies CYP2C (38 genes and 4 pseudogenes across 40 species), CYP2D (36 genes and 1 pseudogene across 37 species), CYP2H (40 genes and 1 pseudogene across 39 species), CYP2J (205 genes and 13 pseudogenes across 48 species), CYP2K (84 genes and 3 pseudogenes across 47 species), CYP2R (43 genes across 43 species), CYP2U (43 genes across 43 species), and CYP2W (77 genes and 2 pseudogenes across 46 species) were widely present in birds (Figure 4.1). We also found members of the CYP2AC subfamily in birds (22 genes and 2 pseudogenes across 24 species). Our analyses also revealed several previously undescribed subfamilies in birds: CYP2F in the grey-crowned crane genome, CYP2G in the chimney swift, and 26 CYP2AF genes in 26 avian species (Figure 4.1). Overall, most subfamilies only had one paralog across species (CYP2C, 2D, 2H, 2R, 2U, 2AC, and 2AF), but the CYP2K and CYP2W had two paralogs each, whereas the CYP2J had between 1 and 7 paralogs per species (Figure 4.1).

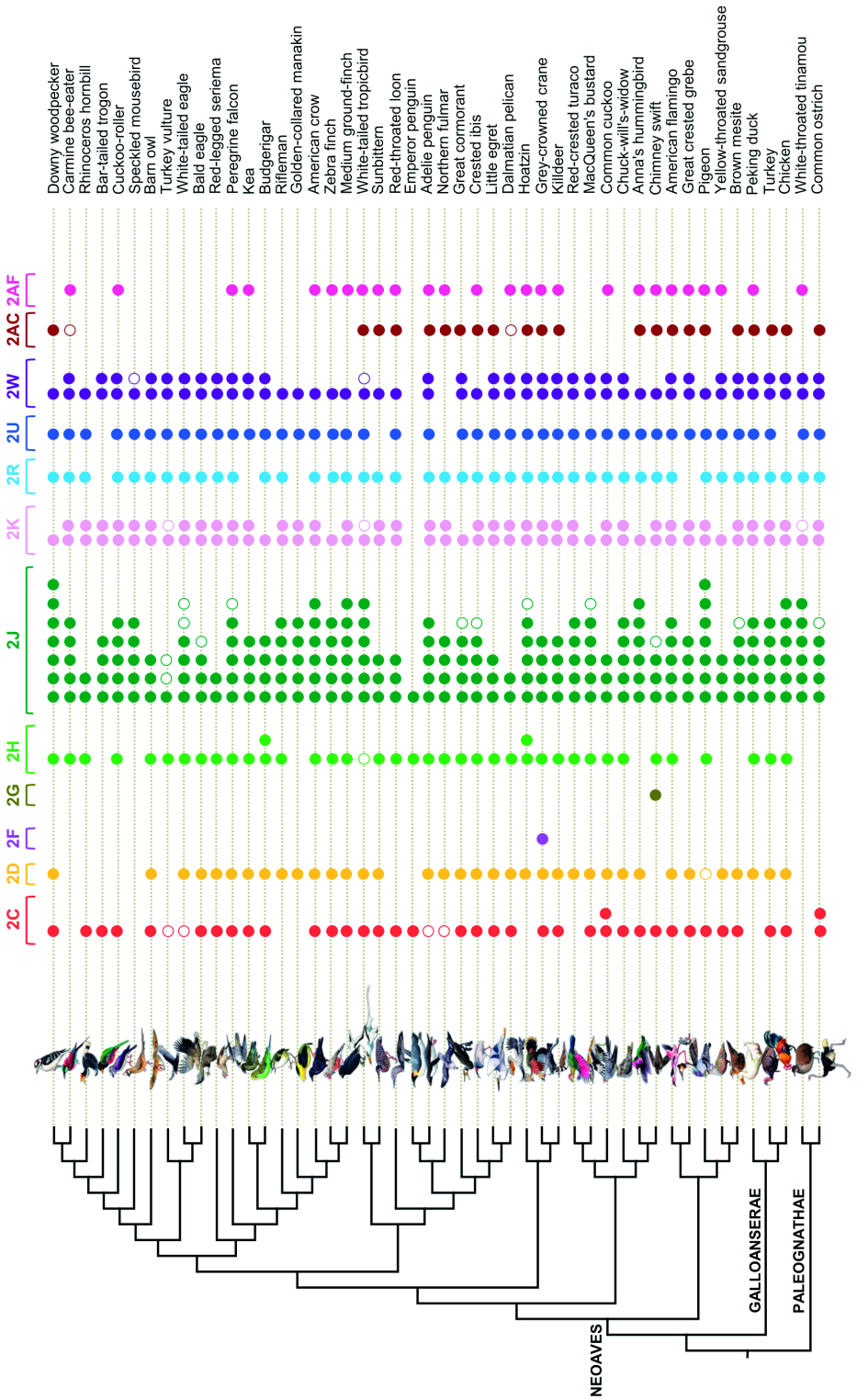


Figure 4.1: Identification of the CYP2 genes found in 48 avian genomes. At the bottom is shown a representation of the avian TENT tree [292]. Middle is full and open circles symbolizing the genes and pseudogenes, respectively. Left row list the ends of the CYP2 gene subfamily names. Top row list the avian species common names.

The ML phylogeny from the “global alignment” replicated the currently accepted basal nomenclature (based on BLAST analyses) with high node bootstrap support (73–100%) for each CYP2 subfamily clade (Figure 4.2).

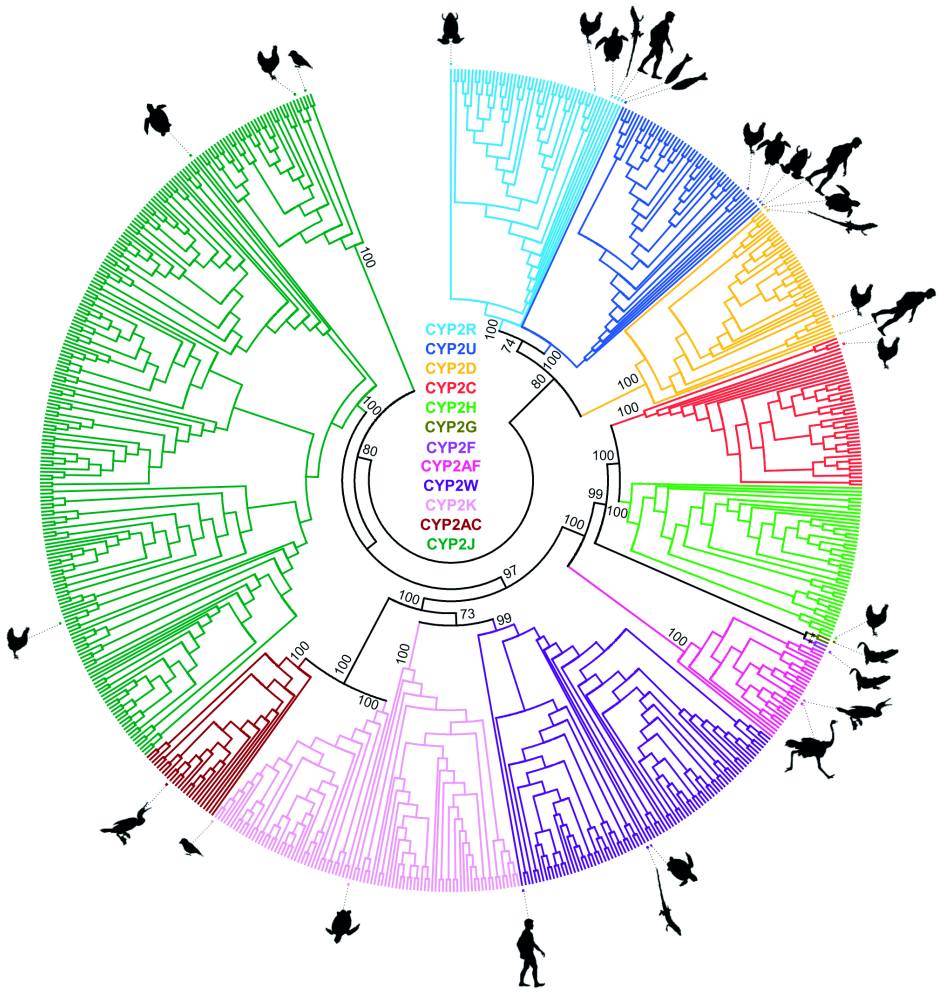


Figure 4.2: Evolutionary relationships of avian CYP2 subfamilies. The phylogenetic tree was built in PHYML 3.0 software using the ML method, with 100 bootstrap replicates and the NNI branch search algorithm. The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the 642 avian CYP2 nucleotide sequences from 48 avian genomes and 31 CYP2 nucleotide sequences from reference species available in the public databases. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test is shown next to the branches or represented by an asterisk mark (* indicating 100% support) for each one of the CYP2 subfamily clades. Values less than 50% support are not shown.

However, the support within each subfamily across species were not as clearly resolved ($< 50\%$ node bootstrap support). This is consistent with the findings of [292], where most individual gene trees of birds do not have enough phylogenetic resolution or have a large amount of incomplete lineage sorting such that no gene tree completely matches the

genome-scale species tree. Therefore, our analyses of the avian CYP2 subfamily evolution and selection analyses were done using the more-robust genome-scale TENT tree as reference [292].

Evolutionary process of CYP2 subfamilies among avian lineages

The ancestral reconstruction (Figure 4.3), performed by COUNT software [296], suggested that: *(i)* the most recent common ancestor of modern birds must have had elements of CYP2C, 2J, 2K, 2R, 2U, 2W, 2AC, and 2AF and then, over time, several genes might have been lost, mainly during the evolution of Neoaves; *(ii)* CYP2D and CYP2H subfamilies are likely to have been lost in the Paleognathae lineage. Moreover, we detected a very large number of CYP2AC and CYP2AF subfamily genes that were lost, in sharp contrast with CYP2J, which is the most conserved subfamily across birds. Interestingly, the emperor penguin lost seven subfamilies during its evolution, in contrast with its close relative, the Adelie penguin, which lost only one (Figure 4.3).

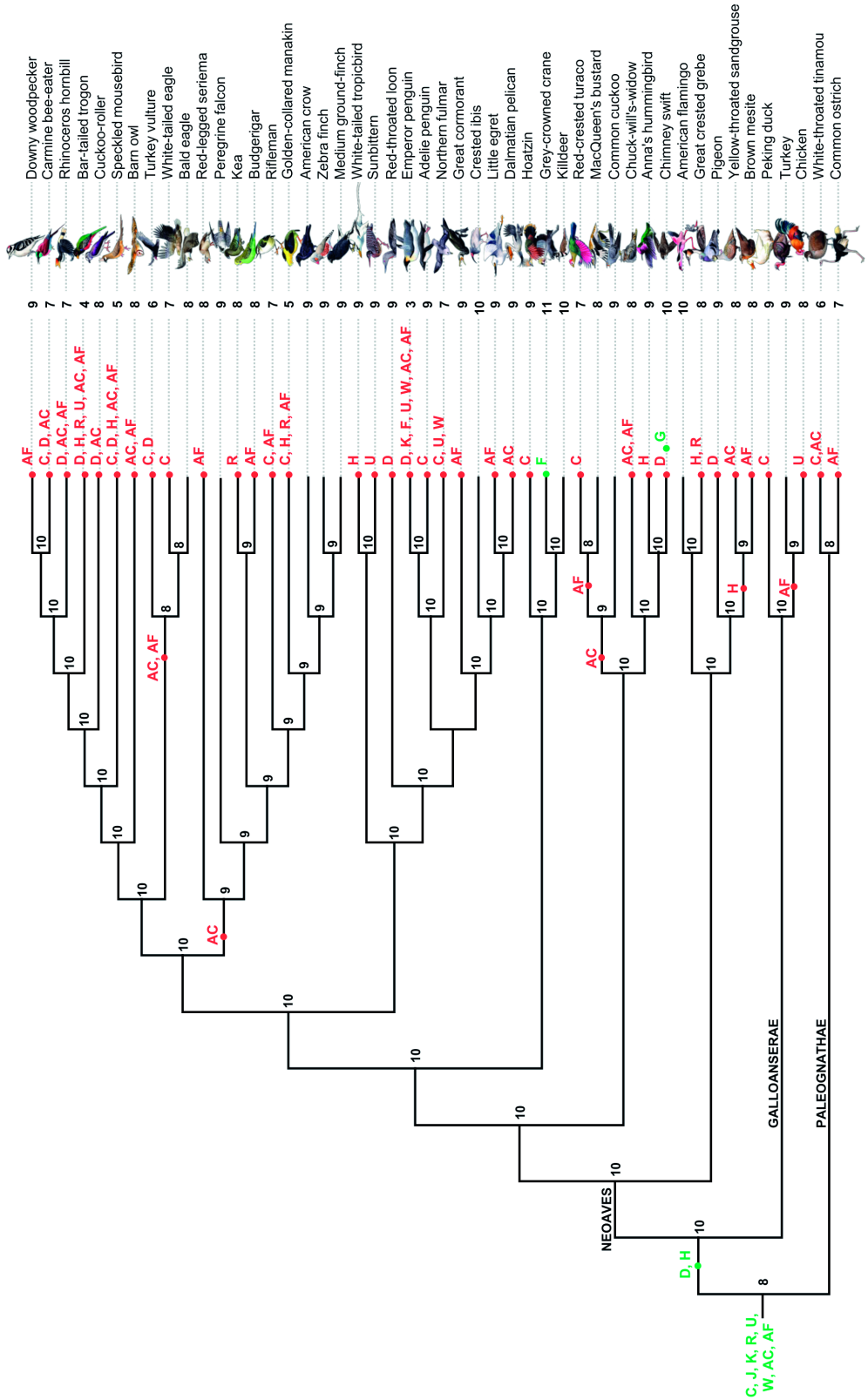
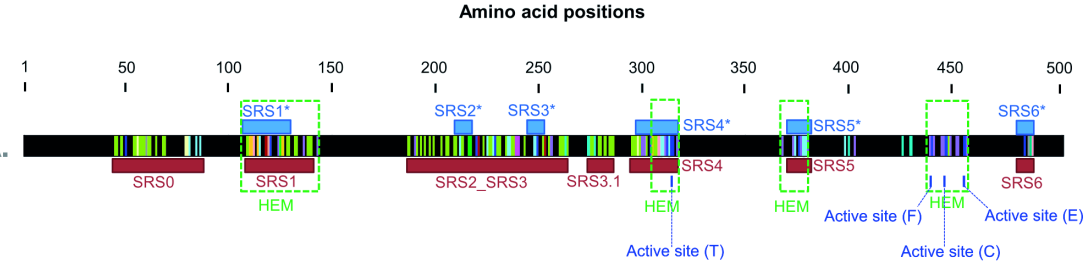


Figure 4.3: **Evolutionary process of avian CYP2 subfamilies.** Green circles indicate the gain of the corresponding CYP2 subfamilies while red circles indicate their loss. Numbers along the phylogenetic tree denote the number of CYP2 subfamilies present in each moment of the avian evolution.

Active sites of CYP2 enzymes have been differentially selected

The PAML LRT site-model analyses of the filtered 17 final CYP2 data sets' without pseudogenes, revealed significant evidence of positive selection in six of the 12 avian CYP2 subfamilies: *(i)* CYP2C, *(ii)* CYP2D, *(iii)* CYP2H, *(iv)* "CYP2J" ("CYP2J_1", "CYP2J_2", "CYP2J_3", "CYP2J_4", and "CYP2J_5" data sets), *(v)* CYP2K ("CYP2K_1" and "CYP2K_2" data sets), and *(vi)* CYP2AC. To determine the impact of positive selection at these sites, we performed a detailed molecular analysis based on the 3D structure of the CYP2 proteins, including SRS* sites identified by Gotoh (1992). Search of the PDB database (release April 8, 2014) identified 61 available 3D CYP2 (A, B, C, D, E, and R subfamilies) structures and their sequences. Our comparative analyses (Appendices IV: Figure S12) of these PDB sequences and their respective ligand annotations with a consensus sequence with the six SRS* revealed several SRS areas (updated SRS): SRS0 (new), SRS1, SRS2_SRS3 (new, resulting from the fusion of SRS2* and SRS3*), SRS3.1 (new, between the SRS3* and SRS4*), SRS4, SRS5, and SRS6 (Figure 4.4). Comparative analyses of updated SRS with each one of the avian CYP2 subfamily alignments allowed the identification of several positively selected sites within important SRS. Below we highlight the SRS sites for each of six subfamilies with evidence of positive selection (the numbering of the sites is based on their corresponding amino acid sequences shown in Appendices IV: Table S5).

A.



B.

Binding regions	Start position	End position	Length
SRS0	45	86	42
SRS1	109	141	33
SRS1*	108	129	22
SRS2*	210	217	8
SRS2_SRS3	187	264	78
SRS3*	245	252	8
SRS3.1	274	286	13
SRS4	295	317	23
SRS4*	298	316	19
SRS5	371	381	11
SRS5*	371	381	11
SRS6	482	489	8
SRS6*	482	489	8
HEM	439	457	19
HEM	109	143	35
HEM	306	317	12
HEM	368	380	13
Active site (C)	447	447	1
Active site (E)	456	456	1
Active site (F)	440	440	1
Active site (T)	313	313	1

Figure 4.4: SRS of CYP2 proteins. The SRS are inferred from comparative analyses of amino acid sequences from available 3D CYP2 structures (Appendices IV: Figure S12) with the six SRS previously identified by Gotoh (1992). (**A**) Schematic representation of the consensus sequence resulting from the alignment of 61 CYP2 amino acid sequences (Appendices IV: Figure S12), containing ligand annotations, with the annotations of the six SRS (in blue: SRS1*, SRS2*, SRS3*, SRS4*, SRS5*, and SRS6*) inferred by Gotoh (1992). The sites where these substrates bind are represented by several colors, according with the number of different ligands. Red boxes represent new defined SRS, based on the available information about CYP2 ligands interacting with the 3D structure of CYP2. The high rate of binding sites, distributed by well-defined regions, allowed us to define seven distinct regions that were named SRS0, SRS1, SRS2_SRS3 (resulting from the fusion of the Gotoh's SRS2 and SRS3), SRS3.1, SRS4, SRS5, and SRS6 (all in red), in order to keep the nomenclature previously used by the referred author. The heme binding regions are denoted by HEM. (**B**) Identification of the amino acid boundaries and length of the referred SRS.

CYP2C: Model M2a indicated that approximately 2% of the sites were under positive selection ($\omega 2 = 4.159$) whereas model M8 showed that approximately 3% were under positive selection ($\omega = 3.240$) (Appendices IV: Table S3). These included the following seven sites: 239, 254, 281, 333, 369, 379, and 453 (Appendices IV: Table S4). Three of these are located within SRS: sites 239 and 254 are located within SRS2_SRS3 (Figure 4.5A) and site 369 is within SRS5 (Figure 4.5B), a recognized heme binding area (HEM). The PRIME analyses suggested that several of the positively selected sites detected by PAML [43] would have amino acid changing properties that could affect the chemical composition, secondary structure, isoelectric point and refractivity of the CYP2C sequences (Appendices IV: Table S4). This is the case of sites 239 and 281 (Figure 4.5A), 369 (Figure 4.5B), and 453 (Figure 4.5C) (Appendices IV: Table S4).

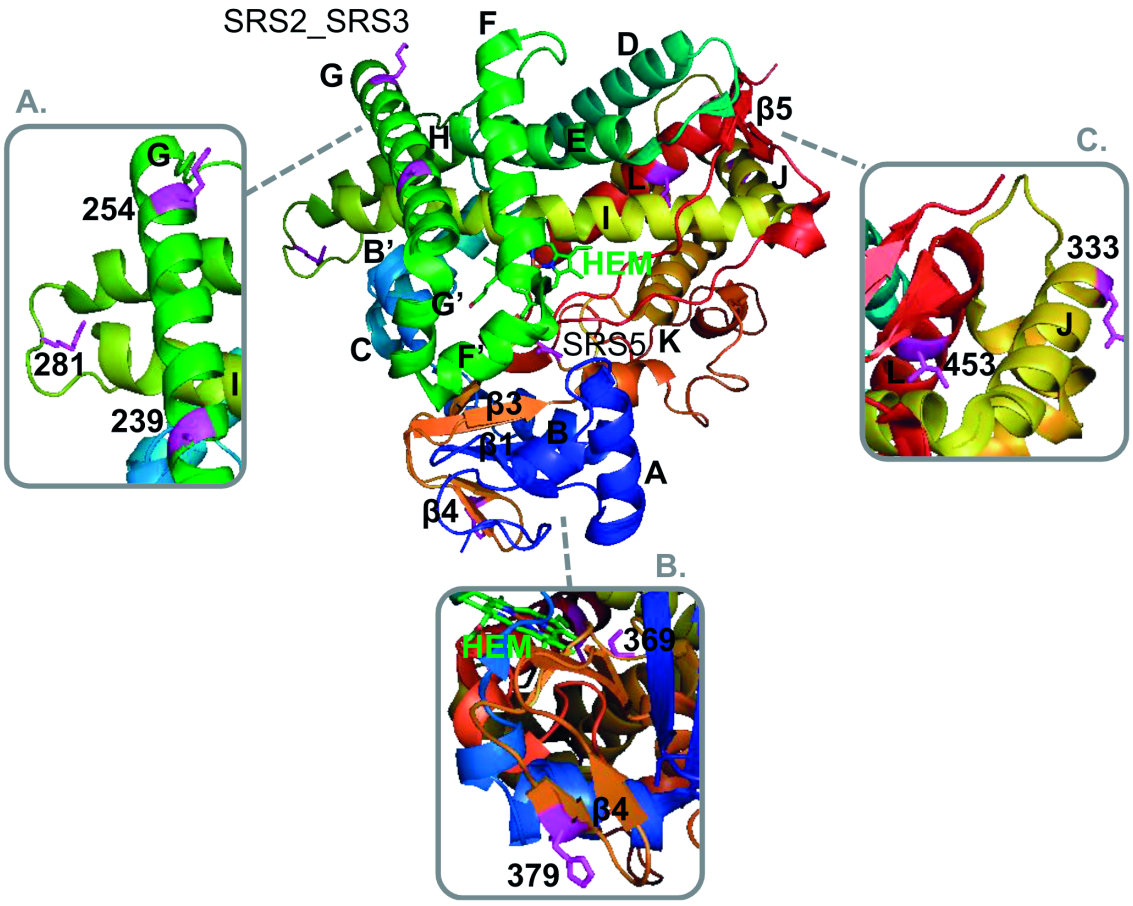


Figure 4.5: 3D analyses of sites detected to be under positive selection in the avian CYP2C. The CYP2C 3D structure predicted in this study is superimposed to its CYP2C8 template (2VN0 human). The positively selected sites are shown as pink sticks, indicating the corresponding amino acid. The regions corresponding to the helices are named with the corresponding capital letter and the β-structures are named with a β followed by a number. HEM represents the heme group. Two SRS regions are represented: SRS2_SRS3 (top) and SRS5 (bottom).

CYP2D: Model M2a indicated that approximately 5% of the sites were under positive selection ($\omega 2 = 2.983$) whereas model M8 indicated 8% ($\omega = 2.504$) (Appendices IV: Table S6). Of the seven sites that were identified to be under positive selection by both methods (54, 74, 123, 236, 240, 359, and 437– Appendices IV: Table S5, line 2), two (54 and 74) were located within the SRS1 and HEM regions (Figure 4.6), which have been linked with CYP2D catalytic activity. The PRIME results showed that some of the positively selected sites located outside the active site areas (that were also detected by PAML analyses) also would likely change the properties of the selected amino acid and thus affect the hydropathy (123), polarity (236), and volume (359) of this enzyme (Appendices IV: Table S7).

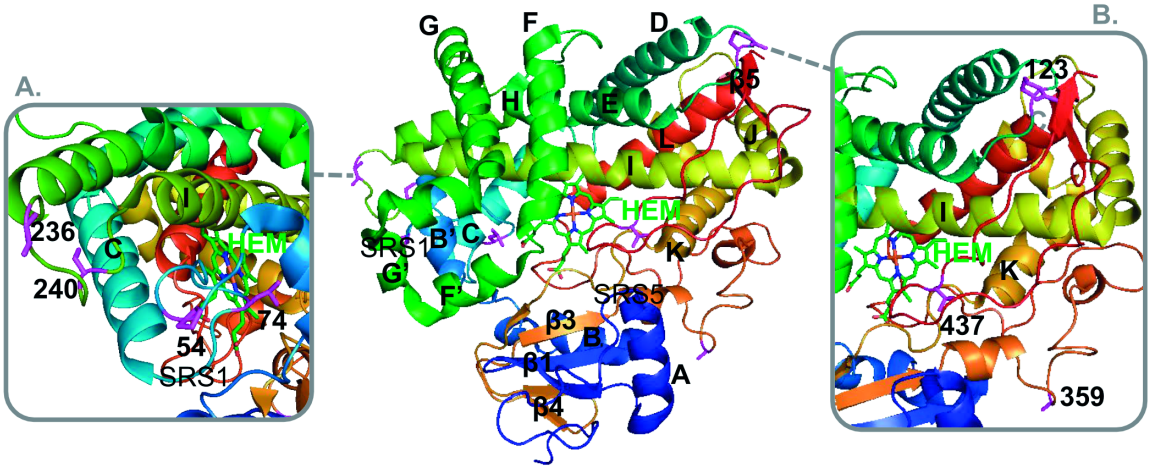


Figure 4.6: **3D analyses of sites detected to be under positive selection in the avian CYP2D.** The CYP2D 3D structure from PDB (3TDA human) was used to map the positively selected sites found. The sites are shown as pink sticks. The regions corresponding to the helices are named with the corresponding capital letter and the β-structures are named with a β followed by a number. HEM represents the heme group. One SRS region is represented, the SRS1 (on the left).

CYP2H: Model M2a indicated that approximately 12% of the sites were under positive selection ($\omega 2 = 2.503$) whereas model M8 indicated 16% ($\omega = 2.191$) (Appendices IV: Table S8). The BEB analysis from both models identified 21 positively selected sites (Appendices IV: Table S9). Sites 38, 45, and 71 are located within the newly defined SRS0 (Appendices IV: Table S9) and site 102 is within the SRS1 and HEM regions (Appendices IV: Table S9). Sites 212, 227, 228, and 248 were located within the newly determined SRS2_SRS3 region (Appendices IV: Table S9) and sites 236 and 240 were in SRS2_SRS3 which matches SRS3* (Appendices IV: Table S9 and Figure 4.4). Site 305 corresponds with the SRS4 and HEM

binding site. Site 365 is located in the HEM binding region and also in SRS5 (Appendices IV: Table S9). Finally, site 370 is within SRS5. Some of these sites also had amino acid changing properties that could affect the polarity, hydropathy, isoelectric point, chemical composition, volume, secondary structure, and refractivity of the CYP2H sequences (Appendices IV: Table S9).

CYP2J: For "CYP2J_1", model M2a indicated approximately 7% of sites under positive selection ($\omega 2 = 3.591$) whereas model M8 indicated 9% ($\omega = 3.162$) (Appendices IV: Table S10). Both models identified positively selected sites located within SRS0 (22, 43, and 46), SRS1 and HEM (73, 75 and 76) and SRS2_SRS3 (182, 185, 186, 199, 204, 207, and 208) (Appendices IV: Table S11). These sites accounted for approximately 76% of the positively selected sites that were found. Changes at some of these sites likely changed properties such as chemical composition (43), hydropathy (182), and volume (185) (Appendices IV: Table S11). For "CYP2J_2", model M2a indicated approximately 2% of sites under positive selection ($\omega 2 = 3.952$) whereas model M8 indicated 5% ($\omega = 2.346$) (Appendices IV: Table S12). Both models identified only two sites (58 and 59) under positive selection located within the SRS0. For site 59, the PRIME analysis suggests that the mutation would cause a shift in polarity. For "CYP2J_3", only model M8 indicated approximately 10% of sites under positive selection ($\omega = 1.360$ – Appendices IV: Table S13). Only one site (21) was positively selected and no amino acid property was selected. For "CYP2J_4" model M2a indicated approximately 2% of sites under positive selection ($\omega 2 = 3.033$) whereas model M8 indicated 5% ($\omega = 2.038$) (Appendices IV: Table S14). Two (201 and 250) of the five positively selected sites, identified by both models, were located within SRS2_SRS3 (Appendices IV: Table S15). For this last site, refractivity was the amino acid changing property (Appendices IV: Table S15). Finally, for "CYP2J_5" only model M8 indicated approximately 1% of sites under positive selection ($\omega = 2.124$ – Appendices IV: Table S16). Two sites (65 and 196) were selected. However, only site 196 was located in the SRS2_SRS3, and selective changes would have affected its isoelectric point.

CYP2K: For "CYP2K_1", both models indicated approximately 3% of sites under positive selection (M2a: $\omega 2 = 3.679$ and M8: $\omega = 3.088$ – Appendices IV: Table S17) and nine sites were identified (Appendices IV: Table S18). From these, four sites were within important regions of the enzyme: 104 (SRS1 and HEM) and 217, 230, and 240 (SRS2_SRS3). For

these sites, the following changing properties were identified: refractivity (104), volume (217), and isoelectric point (230 and 240) (Appendices IV: Table S18). For "CYP2K_2" model M2a indicated approximately 1% of the sites under positive selection ($\omega 2 = 3.341$) whereas model M8 indicated 3% ($\omega = 1.515$) (Appendices IV: Table S19), especially sites 277 and 321.

CYP2AC: Model M2a indicated approximately 12% of sites under positive selection ($\omega 2 = 1.628$) whereas model M8 indicated 17% ($\omega = 1.517$) (Appendices IV: Table S20). Only one site (92) was identified with PP $> 95\%$ (model M8). This site was located in the SRS1 and HEM regions and the property volume was pointed as acting in this site.

The remaining subfamilies CYP2R (Appendices IV: Table S21), CYP2U (Appendices IV: Table S22), CYP2W (data sets "CYP2W_1" – Appendices IV: Table S23 and "CYP2W_2" – Appendices IV: Table S24), CYP2AF (Appendices IV: Table S25), and CYP2J (only data set named as "CYP2J" – Appendices IV: Table S26) showed no significant evidence of positive selection.

Selection of CYP2 has been branch-specific in species with shared traits

The branch-specific analyses, applied to CYP2 subfamilies with evidence of positive selection (from site-models), revealed that "CYP2H" (ωcarnivorous $= 0.570$, ωomnivorous $= 0.490$, and ωherbivorous $= 0.555$) and "CYP2J_2" (ωcarnivorous $= 0.265$, ωomnivorous $= 0.356$, and ωherbivorous $= 0.393$) are evolving differently in birds with distinct carnivorous, omnivorous, and herbivorous feeding habits (Figure 4.7). The strength of selection is also variable among birds occupying distinct habitats (moist, semi-moist and/or dry) in six of the CYP2 subfamily data sets ("CYP2D", "CYP2H", "CYP2J_2", "CYP2J_3", "CYP2J_5", and "CYP2K_1") (Figure 4.7). "CYP2D", "CYP2H", and "CYP2J_3" are evolving differently in migratory and non-migratory birds (Appendices IV: Figures S1-S11). Only the "CYP2AC"($\omega = 0.353$), "CYP2C"($\omega = 0.362$), "CYP2J_1"($\omega = 0.450$), "CYP2J_4"($\omega = 0.252$), and "CYP2K_2"($\omega = 0.178$) subfamily data sets are evolving at the same rate in all bird groups regardless of their distinct feeding habits, habitats and migratory behaviors (Figure 4.7).

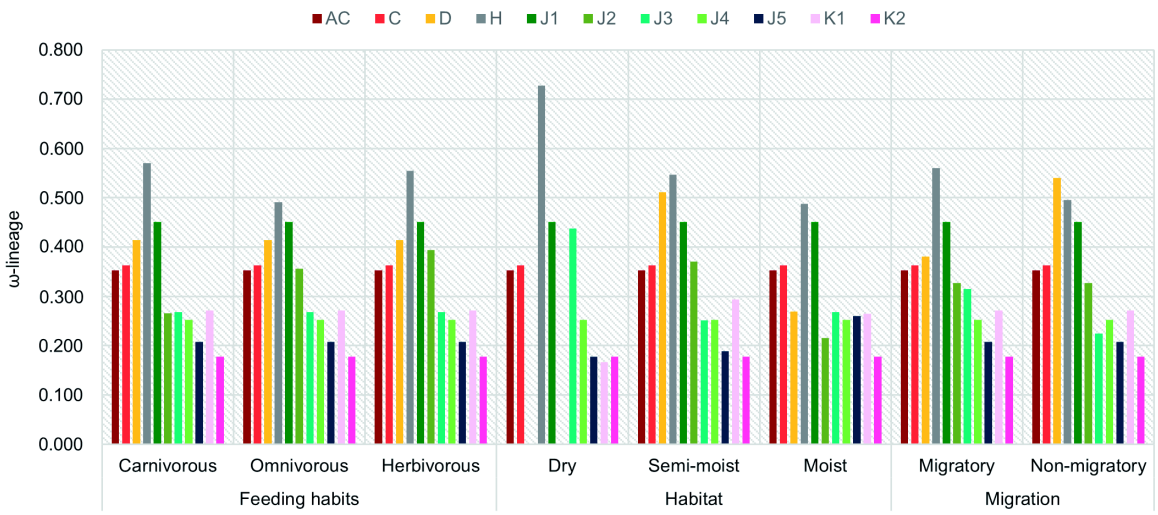


Figure 4.7: ω-ratio variations according to distinct avian feeding habits, habitats and migratory behaviors. ω-ratio variations according to distinct avian feeding habits, habitats and migratory behaviors. The ω values represented arise from the hypothesis that best fits each CYP2 subfamily data set according to the branch-specific LRT analyses (p-value < 0.05) (see Appendices IV: Figures S1-S11 for more methodological details). CYP2 subfamily code abbreviations indicate each one of the data sets used: AC–"CYP2AC", C–"CYP2C", D–"CYP2D", H–"CYP2H", J1–"CYP2J_1", J2–"CYP2J_2", J3–"CYP2J_3", J4–"CYP2J_4", J5–"CYP2J_5", K1–"CYP2K_1", and K2–"CYP2K_2".

CYP2 gene subfamily numbers vary according to migration and feeding habits

The LDA constructed a single discriminant function (Figure 4.8). The success of classification, estimated by cross-validation, was low (58% global, 67% migratory, and 48% non-migratory). This was due to the large degree of similarity among the number of CYP2 genes present in the two classes (migratory and non-migratory), as is apparent from the average variable scores per group (Figure 4.8A). The minor differences found between the two classes were due to CYP2 genes scoring in the extremes of the discriminant function, with migratory bird scores tending slightly towards negative values whereas non-migratory are closer to the positive end of the function (Figure 4.8B). However, a Kolmogorov–Smirnoff test applied to the frequencies of scores from the linear discriminant function (Figure 4.8B) showed a significant difference between the distribution of each class (p-value < 0.01). Appendices IV: Table S27 shows that among all these genes, CYP2D, CYP2U, CYP2H, and CYP2AF are the main ones responsible for negative scores (migratory) whereas CYP2K is responsible for positive scores (non-migratory).

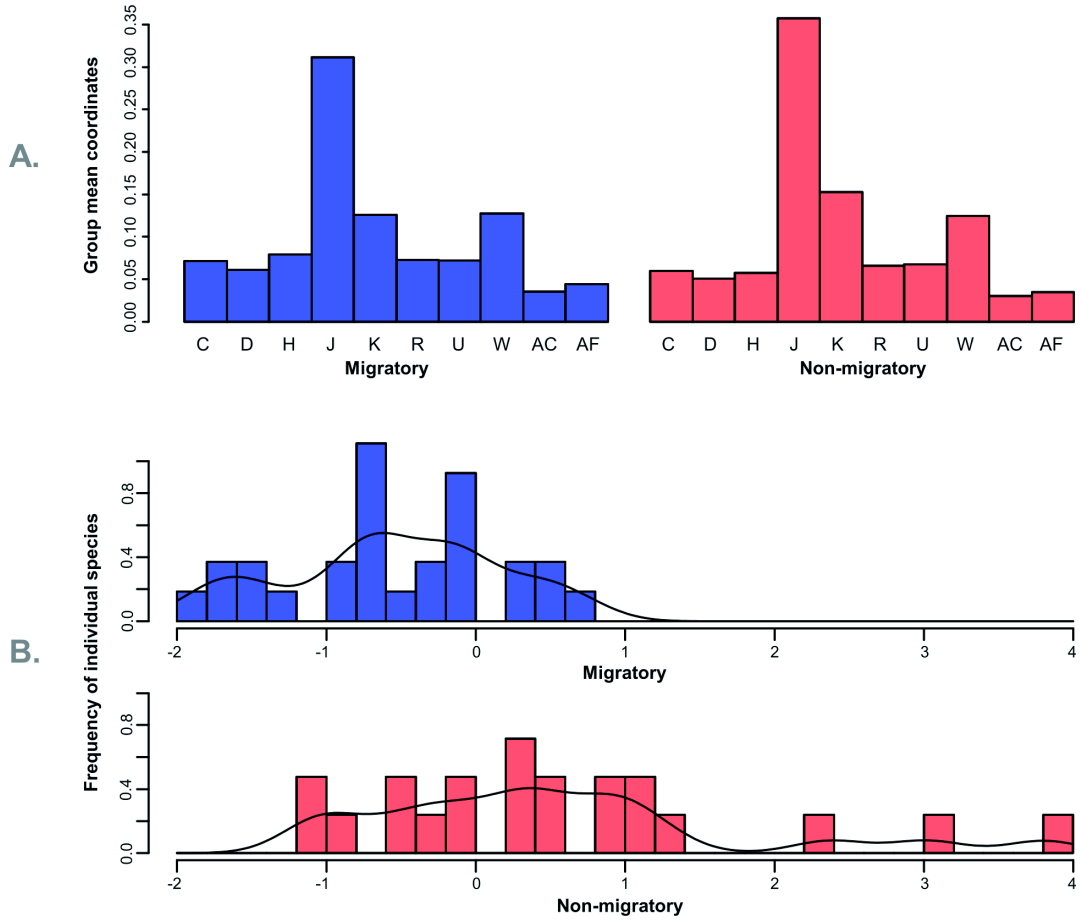


Figure 4.8: Results of the linear discriminant function. (**A**) Average scores per class. (**B**) Individual species scores per class. The black line represents a nonparametric fit of frequency distributions.

Furthermore, differences in the number of CYP2 genes were also observed when analyzing feeding habits in association with the migratory behavior of birds (Figure 4.9). Migratory carnivores and herbivores have less CYP2 genes than migratory omnivores, but these differences were only significant between migratory carnivores and migratory omnivores (Mann–Whitney test, *p*-value < 0.05). Non-migratory carnivores have less CYP2 genes than non-migratory omnivores and herbivores, but none of these differences resulted to be significant. Globally, omnivores have a higher number of CYP2 genes than specialist birds (carnivorous and herbivorous birds), except for non-migratory herbivores, which have a number of CYP2 genes similar to non-migratory omnivores (Figure 4.9). The only significant differences were detected between migratory omnivores and both types of carnivores (Mann–Whitney test, *p*-value < 0.05) (Figure 4.9).

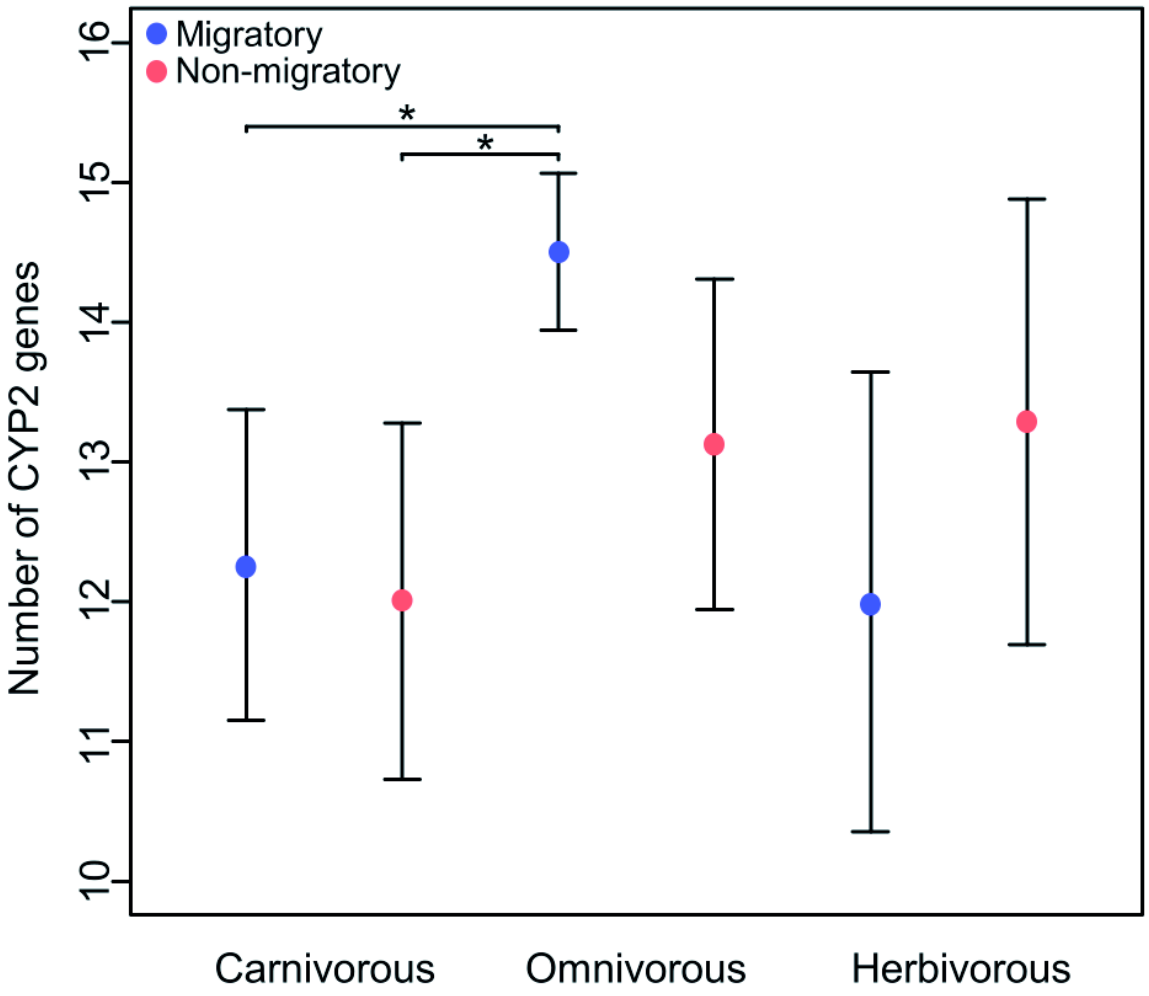


Figure 4.9: Comparison of the number of CYP2 genes in birds with different migratory behaviors and feeding habits. Plotted are the mean $\pm 95\%$ confidence interval. Asterisks indicate significant differences in pairwise comparisons performed by Mann–Whitney tests. Species trait classifications are in Appendices IV: Table S2.

4.5 Discussion

The avian CYP2 subfamilies corresponded well with previous classifications [291], including the finding that all avian species have only a single gene member in the CYP2D, CYP2R, and CYP2U subfamilies (Figure 4.1). CYP2J, CYP2W, and CYP2K were the largest subfamilies in the 48 avian genomes analyzed, with two to seven duplicated genes (Figure 4.1). Although it had previously been suggested that the CYP2W1 and CYP2W2 genes

had duplicated only in the Galloanserae lineage [291], our results covering all the three main avian groups (Palaeognathae, Galloanserae, and Neoaves) clarified that there have been several CYP2W duplication events (Figure 4.1). Our identification of avian CYP2F, CYP2G, and CYP2AF subfamilies indicates that the CYP2F and CYP2G subfamilies are not mammalian-specific as was previously suggested [90] and that the CYP2AF has not been lost in the avian lineage, contrarily to what has been previously hypothesized [291].

Some of our most striking results are the presence or absence of one or many CYP families in closely related species. The Adelie penguin and the closely related emperor penguin have similar habitats (both live exclusively in the Antarctic region), feeding habits (carnivores), and lifestyles (colonial, social, specialized for swimming) [300] but the number of CYP2 subfamilies between them was striking (nine and only three, respectively – Figure 4.3). This is an intriguing scenario because it was expected that these species facing similar selective pressures would present a similar number of CYP2 subfamilies. However, according to our approach the emperor penguin appears to have lost many subfamilies that the Adelie penguin still has and shares with other more distantly related bird species. In order to exclude possible false-negatives due to the strict criteria of our identification approach, we performed exploratory tBLASTn searches against the emperor penguin genome with less stringent criteria and using the Adelie penguin CYP2 sequences as query. We retrieved some partial sequences (ranging between 177 and 849 bp) possibly representing the CYP2C, CYP2D, CYP2H, CYP2J, CYP2K, CYP2W, and CYP2U subfamilies. However, due to their short lengths (as opposed to complete CYP2 genes that encompass approximately 1500 bp), the accurate assignment to a CYP subfamily becomes compromised. Thus, such short sequences were not considered for further detailed evolutionary analyses. Conversely, CYP2F is only present in grey-crowned crane and CYP2G in chimney swift. This would be consistent with other studies. For example, a study of 200 humans found that a functional CYP2G allele was also uncommon in humans (detected in only 11.6% of the individuals) [301]. We cannot exclude the possibility that these genes are present in the gaps of avian genome assemblies, or even that evidence of their presence in other birds can be missed, as demonstrated above, by our threshold of requiring at least 1125 bp when searching for avian CYP2 nucleotide sequences.

Our selection analyses results are consistent with some findings in avian and non-avian species. The residue at position 369 of the avian “CYP2C” data set was within

SRS5 and corresponds to site 364 in *Oryctolagus cuniculus* (rabbit) CYP2C3v, where the mutation T364S has been linked with changes in progesterone region selectivity [302]. The chicken CYP2C gene is activated by the Chicken Xenobiotic Receptor, supporting its role in xenobiotic metabolism [303]. Thus, it is plausible that the positively selected sites (239, 281, 369, and 453) leading to changes in amino acid properties might have provided an important adaptation by facilitating the efficient inactivation and removal of several xenobiotic compounds in birds.

CYP2D is present in several mammalian species (*e.g.,* rodents, primates, rabbit, and horse), and has been linked with feeding habits and with metabolizing plant toxins such as alkaloids [304]. Therefore, the positively selected sites 54 and 74 found in the avian CYP2D subfamily, located within the SRS1 and HEM functional regions, could be particularly advantageous for an efficient dietary detoxification [305]. Our lineage-specific analyses of CYP2D suggest similar impacts among birds with distinct feeding habits. The effect of changing amino acid properties for some of the positively selected sites (123, 236, 359) located outside of the active site areas of the avian CYP2D subfamily could possibly be related with global protein folding or substrate recognition [161].

CYP2H enzymes are involved in reactions of epoxygenation [306] and they are the major phenobarbital-inducible enzymes in the chicken liver [307]. Agonist for the CYP2H are the drugs dexamethasone and metyrapone and also the compounds okadaic acid, pregnenolone16 alpha-carbonitrile, and squalestatin1 [308]. The substrate of these enzymes is arachidonic acid [306], which implies that positively selected sites located within the SRS1 and HEM regions of the avian CYP2H subfamily (site 102), within the SRS4 and HEM binding site (305), and in the HEM binding region and in the SRS5 (site 365) might have an impact on the activity of this enzyme in birds. Since CYP2H has evolved differently among different avian groups, it is most likely that the changes have distinct adaptive relevance possibly involving feeding habits, habitats, and migratory behaviors.

The greater number of gene duplications and large variability in the amino acid patterns among the different copies of CYP2J genes among birds could also be related to different habitats and feeding adaptations. Similar duplications events have been detected in bactrian camels, which are hypothesized to be linked with the importance of CYP2J in the conversion of arachidonic acid into 19(S)-HETE – a potential vasodilator of renal preglomerular vessels – that stimulates water reabsorption [284]. CYP2J also has epoxygenase activity and

can convert arachidonic acid into epoxyeicosatrienoic acids that have antihypertensive vasodilatory properties [309]. Thus, it has been hypothesized that an increased number of CYP2J copies would increase water absorption and could influence survival in dry conditions [284]. The large number of avian CYP2J copies and the extraordinary high degree of positive selection detected in the SRS suggest that this subfamily might be important to adaptation to distinct habitats by using water more efficiently. The genes from the "CYP2J_2", "CYP2J_3", and "CYP2J_5" subfamily data sets could have a particular role in this process as we found: *(i)* distinct ω for "CYP2J_2" among carnivorous, omnivorous, herbivorous, land (semi-moist), and water (moist) birds; *(ii)* distinct ω for "CYP2J_3" among land (dry and semi-moist), water (moist), migratory, and non-migratory birds, and *(iii)* distinct ω for "CYP2J_5" among land (dry and semi-moist) and water (moist) birds.

Although we have identified sites under positive selection in the avian CYP2K and CYP2AC subfamilies, further studies are necessary to infer the impact of these substitutions, since their function remains little known.

The absence of positive selection in CYP2R and CYP2U subfamilies could be explained by their essential role in the metabolism of vitamin D and arachidonic acid. These genes have conserved synteny between birds and humans [291]. Similarly, the remaining subfamilies without positive selection (CYP2W and CYP2AF) also can have essential functions in the metabolism of endogenous compounds, being strongly adapted to their substrates.

For the LDA analyses of CYP2 genes, despite the low success of classification of genes, this analysis provides an opportunity for the future, when a greater number of avian genomes will become available, because it suggests that some CYP2 subfamilies, including CYP2D, CYP2U, CYP2H, and CYP2AF, are more related to migratory species. Our analysis also revealed that the variation in the number of CYP2 genes is related to different feeding habits and migratory behaviors of birds. This suggests that the higher number of CYP2 genes found in avian migratory omnivores might be related with their need to adapt to a wider variety of environments and food resources, and thus the higher exposure to several toxins, would require a more efficient detoxification capacity [310].

While our study provides significant advances to our understanding of avian CYP2 evolution, the currently available genome scaffold lengths limit the number of CYP2 sequences that can be confidently classified. Ongoing efforts to increase the scaffold lengths of bird genomes will enhance our understanding of avian CYP2 identification and adaptation.

4.6 Conclusions

To our knowledge, this is the first study of avian CYP2 subfamilies that includes representatives of the three avian evolutionary groups: Palaeognathae, Galloanserae, and Neoaves [180]. We identified 12 CYP2 subfamilies in 48 avian genomes and showed that some of the CYP2 genes that were previously described as being lineage-specific, such as CYP2K and CYP2W, are present in representatives of all the avian groups. Additionally, we demonstrated the presence of the CYP2F and CYP2G subfamilies in some avian genomes. From our comparative analyses we updated our knowledge of the SRS of CYPs, and identified several new regions (SRS0, SRS2_SRS3, and SRS3.1). We identified several significant signatures of positive selection in the six avian CYP2 subfamilies (CYP2C, CYP2D, CYP2H, CYP2J, CYP2K, and CYP2AC), some of which are located in relevant SRS- and heme-binding areas (HEM) that influence CYP2 structure and function. The six CYP2 subfamilies that showed positive selection had sites under positive selection in HEM and in one or both SRS1 and SRS3. Of the six, only the CYP2C and CYP2H subfamilies had sites under positive selection in SRS5, suggesting that these two subfamilies may be under similar evolutionary pressures in this enzyme region, that allow them to phenotypically adapt and acquire similar substrate affinities. The positive selected sites in these avian CYP2 subfamilies likely have helped them adapt to distinct chemical compounds in new habitats with distinct food resources, and facilitate the dispersion and evolutionary success of birds.

Part III

BIOINFORMATICS TOOL TO STREAMLINE THE ANALYSIS OF THE RESULTS FROM ESTIMATED SELECTIVE PRESSURES

Chapter 5

IMPACT_S: Integrated Multiprogram Platform to Analyze and Combine Tests of Selection

Published in

Maldonado E*, Sunagar K*, Almeida D*, Vasconcelos V, Antunes A

PLoS ONE, 2014, 9(10): e96243.

DOI: 10.1371/journal.pone.0096243

* - These authors contributed equally

5.1 Abstract

Among the major goals of research in evolutionary biology are the identification of genes targeted by natural selection and understanding how various regimes of evolution affect the fitness of an organism. In particular, adaptive evolution enables organisms to adapt to changing ecological factors such as diet, temperature, habitat, predatory pressures and prey abundance. An integrative approach is crucial for the identification of non-synonymous mutations that introduce radical changes in protein biochemistry and thus in turn influence the structure and function of proteins. Performing such analyses manually is often a time-consuming process, due to the large number of statistical files generated from multiple approaches, especially when assessing numerous taxa and/or large datasets. We present IMPACT_S, an easy-to-use Graphical User Interface (GUI) software, which rapidly and effectively integrates, filters and combines results from three widely used programs for assessing the influence of selection: CODEML (PAML package), DATAMONKEY and TREESAAP. It enables the identification and tabulation of sites detected by these programs as evolving under the influence of positive, neutral and/or negative selection in protein-coding genes. IMPACT_S further facilitates the automatic mapping of these sites onto the three-dimensional structures of proteins. Other useful tools incorporated in IMPACT_S include JMOL, ARCHAEOPTERYX, GNUPLOT, PHYML, a built-in SWISS-MODEL interface and a PDB downloader. The relevance and functionality of IMPACT_S is shown through a case study on the toxicoferan-reptilian Cysteine-rich Secretory Proteins (CRiSPs). IMPACT_S is a platform-independent software released under GPLv3 license, freely available online from http://impact-s.sourceforge.net.

5.2 Introduction

The nature and strength of evolutionary selection pressures can be estimated at the molecular level, as a non-synonymous to synonymous substitution rate ratio omega (ω = *dN/dS*), where ω greater than, equal to and less than 1 is indicative of positive, neutral and negative selection, respectively [44]. This approach often fails to detect subtle adaptations that only affect certain regions of the protein and/or take place over a very short period of evolutionary time [199]. Moreover, the evaluation of selective pressures solely at the

nucleotide level and the assumption that all mutations affect the fitness of the organism equally could be misleading. Although non-synonymous substitutions introduce variations in coding regions, a novel amino acid could have identical or similar biochemical and/or structural properties to that of the ancestral residue. Such substitutions are unlikely to influence the structure or the function of the protein and hence are least likely to affect the fitness of an organism. Thus, it is important to discern the nature of mutations to precisely understand the evolution of a protein. By employing mapping strategies of mutational sites onto the three-dimensional (3D) structure of the protein, it is possible to gain further insights into how its structure and/or function are affected by the changes in certain residues or regions [76]. For example, using a similar integrative approach, we and others have demonstrated that most predatory venom-components in a diversity of animal lineages adopt Rapid Accumulation of Variations in Exposed Residues of Toxins [311] and accumulate mutations on the molecular surface under the influence of positive Darwinian selection, while preserving the key functional and structural residues that stabilize the overall structure of the protein [311–320]. The mutation of the molecular surface may increase the toxin's ability to target novel molecular receptors and aid in evading immune response upon injection into prey animals [315]. Phenomena like this can be easily detected by mapping sites under positive and negative selection on 3D structures of proteins. Mapping of sites onto multiple sequence alignments of protein-coding genes is also beneficial, as it enables the identification of differential evolution of domains by revealing mutational hotspots [312, 317, 318]. Hence, to efficiently assess the influence of natural selection on protein-coding genes and to accurately identify regions under various regimes of natural selection, it is essential to adopt a complementary approach [298]. Although several software applications have been proposed to independently evaluate selection pressures at the codon-level [10, 43, 51, 52], an integrative approach, which additionally evaluates the strength and radicalness of non-synonymous substitutions at the amino acid-level, is still missing. In addition, the manual integration of results from various approaches is time-consuming.

To address these shortcomings and to facilitate the integration of results from various selection assessments, we propose IMPACT_S, a free platform-independent user-friendly GUI software that integrates results of nucleotide and amino acid-level assessments by employing three widely used softwares: CODEML from Phylogenetic Analysis by Maximum Likelihood (PAML) package [43], DATAMONKEY, a web-server of the HYPHY [321] package

— www.datamonkey.org [19, 52], and Selection on Amino Acid Properties using phylogenetic trees (TREESAAP) [51]. Sites detected as positively and negatively selected can be automatically mapped onto the 3D structure of the protein. If experimentally determined protein structures are unavailable, IMPACT_S facilitates the prediction of models, through homology, using the built-in SWISS-MODEL [54–56] interface. Homology modeling is based on evolutionary relationships between target and template sequences. The template sequences result from homology searches of experimentally determined protein structures, through successive BLAST-p [322] searches. This method involves the construction of an atomic resolution model for the protein under study using an experimentally determined 3D-structure of a similar homologous protein [56]. The following criteria are automatically considered by SWISS-MODEL in order to obtain a single high quality template: *(i)* identification of related proteins with experimentally solved structures, *(ii)* mapping of corresponding residues of target and template structures, *(iii)* building of the 3D model on the basis of the alignment and finally, *(iv)* evaluation of the quality of the resulting model [56].

In the following sections, we introduce the foundations, the structure and the development of IMPACT_S and demonstrate the relevance and functionality of IMPACT_S using a case study on Cysteine-rich Secretory Proteins (CRiSPs), a class of reptilian and mammalian glycoproteins with extremely diversified functions. CRiSPs have been theorized to play important functions in the mammalian reproductive pathways, while in toxicoferan reptiles (venomous snakes and lizards) they have been hypothesized to participate in prey envenoming and capture [318].

5.3 Design and Implementation

IMPACT_S foundation

IMPACT_S integrates three broadly used bioinformatic softwares for assessing selection pressures shaping the evolution of protein-coding genes: *(i)* PAML and *(ii)* DATAMONKEY for assessing selection at the codon-level, and *(iii)* TREESAAP to detect selection at the protein-level.

CODEML from the PAML package [43], implements powerful site-specific [8, 45], branch-specific [6, 47], and branch-site specific models [45, 48] that detect the influence of natural

selection, easily and reliably. Site-specific models, which are the current focus of this work, comprise the alternative models — model 2a (M2a) and model 8 (M8) and the null models — model 1a (M1a) and model 7 (M7). The alternative models include the Bayes Empirical Bayes (BEB) [45] approach for identifying positively selected sites. The computed ω value and the sites detected by the alternative models as positively selected are only considered by the user if the likelihood ratio test (LRT) is significant [8, 45]. The LRT is conducted by comparing the null models with the alternative models (M1a *vs.* M2a and/or M7 *vs.* M8). Both M2a and M8 have one additional class or category when compared to their null counterparts M1a and M7, respectively [8]. In case the LRT is significant in both tests, with the alternative models as fitting the data better, then it is possible to assume stronger evidence for the presence of sites under positive selection [45]. However, PAML is incapable of identifying negatively selected sites, in contrast to certain models implemented in the DATAMONKEY web-server [19, 52]: Single-Likelihood Ancestor Counting (SLAC) [10], Fixed Effects Likelihood (FEL) [10], Random Effects Likelihood (REL) [10] and Fast Unconstrained Bayesian AppRoximation (FUBAR) [323] methods. The recently proposed Mixed Effects Model of Evolution (MEME) [199] is a state-of-the-art method for detecting sites that evolve under episodic selection pressures, which are often difficult to identify using traditional site-specific methods. This method allows the distribution of ω to vary not only across sites, but also from branch to branch at a site [199]. SLAC infers the number of non-synonymous and synonymous substitutions that have occurred at each site using Maximum likelihood (ML) reconstructions of ancestral sequences, while FEL estimates the ratio of non-synonymous to synonymous substitutions, not assuming *a priori* distribution of rates across sites substitution, on a site-by-site basis. REL involves fitting a distribution of substitution rates across sites and then inferring the rate at which individual sites evolve based on ML estimates. FUBAR on the other hand, which supersedes SLAC, FEL and REL, infers the rate at which individual sites evolve based on an approximate hierarchical Bayesian approach using a Markov Chain Monte Carlo sampler.

The choice of the method not only depends on the question being addressed, but also on the size of the dataset. When the positive selection on the evolution of protein-coding genes is sufficiently strong, the site-specific models of CODEML [20, 43, 45] can efficiently identify positively selected sites. However, due to the episodic or transient nature of natural selection on protein-coding genes, the precise identification of the regions that have undergone

adaptive evolution is often difficult [199]. Moreover, a strong influence of purifying selection on a majority of lineages can mask the signal of positive selection on others [199]. MEME was proposed to address these shortcomings and to reliably identify sites that are influenced by both episodic and pervasive influence of positive selection at the level of an individual site [199]. In order to identify sites influenced by positive and negative selection, the following methods can be chosen, considering the size of datasets to avoid biased results [10] SLAC for large datasets (over 40 sequences), REL for datasets of intermediate size (20–40 sequences), FEL for intermediate to large datasets (over 50 sequences) and FUBAR for analyzing very large datasets (the authors tested this model on a dataset of 3142 sequences, which is much faster when compared with the other methods [323]). While DATAMONKEY and CODEML assess the influence of natural selection at the codon-level, TREESAAP measures the selective influences on 31 biochemical and structural amino acid properties — such as hydrophobicity, polarity, solvent accessible reduction ratio and buriedness — during cladogenesis and performs goodness-of-fit and categorical statistical tests [51]. It enables the estimation of radicalness of a mutation at a particular site by revealing the influence of the novel amino acid, introduced by non-synonymous nucleotide substitution, on the structural and biochemical properties of the protein. The result files generated by TREESAAP are organized in two main subdirectories: Evpthwy and Substs. The former includes the sliding window analyses, while the latter contains results for particular sites [200, 201]. Both identify amino acid properties under selection, scored in every category, typically from 1 to 8. The assessment of biochemical and structural properties of the novel amino acids, introduced by non-synonymous substitutions, results in generation of numerous statistical files, especially when a large number of amino acid properties are evaluated. In order to detect sites that recurrently fall under the influence of natural selection or to detect sites with a defined number of selected properties, the user has to assess an array of statistical files. Thus, the interpretation of the result files generated by TREESAAP for downstream analyses can be an extremely time-consuming process.

Manually noting down log-likelihood (lnL) values, various Maximum likelihood parameter estimates and ω values from the result files of CODEML for every model analyzed is also tedious. These values are essential for conducting likelihood ratio tests and the identification of the appropriate regime of selection (for additional information regarding models and LRT tests, see [8]). This process is required to be repeated for each dataset being analyzed.

Similarly, tables generated by the DATAMONKEY web-server, which contain various site statistics, are required to be manually downloaded and processed to retrieve sites detected as positively and negatively selected at the desired significance level.

It has been suggested that the application of all the major selection assessment methods in HYPHY [321] (SLAC, FEL, REL, FUBAR and MEME), followed by the consensual identification of sites detected as positively selected among the methods, can minimize false positives, especially when analyzing smaller datasets [10]. This feature is indeed available on the DATAMONKEY web-server. However, the results of these analyses can only be integrated manually with the results of CODEML and TREESAAP selection assessments and the compilation of the sites detected in common by these approaches becomes a difficult task. Moreover, to our knowledge, there is no software to export these results for the downstream analyses, such as the mapping of the detected mutational sites onto the 3D structures of proteins.

Therefore, IMPACT_S has been designed to be a single platform, with which the user can employ various state-of-the-art selection assessments and identify the regime of natural selection on protein-coding genes. Furthermore, these sites could be automatically mapped onto the alignment and 3D structure (either in a homology model or a known crystallographic structure of the protein), making the entire process extremely rapid and efficient, even when a large number of datasets are analyzed.

IMPACT_S structure

IMPACT_S implements four main tabs: *(i)* 'PAML', *(ii)* 'Datamonkey', *(iii)* 'TreeSAAP' and *(iv)* 'Results & 3D'. The first three tabs enable the user to import the corresponding result files and automatically extract the sites detected as positively and negatively selected. IMPACT_S has been designed to organize these results in Comma Separated Values (CSV) format, which can be viewed and exported with the built-in CSV Viewer tool. Sites detected as positively and negatively selected by various selection assessment methods can be mapped individually for each method, or mapped in combination, as common sites, onto the 3D structure of proteins under the 'Results and 3D' tab. To map sites from two or more methods (considering the type of selection) in conjunction, as common sites, these must be firstly merged into a final Merged Results (MR) table. IMPACT_S provides several mapping schemes (please refer to Table 4 in the tutorial included in the package, for a list of all the

available schemes and descriptions), which are applied to the results and presented using the molecular visualization software, JMOL [324] (http://jmol.sourceforge.net/). The 3D structure can be obtained through homology modeling using the built-in Protein Modeler tool that acts as an interface to the SWISS-MODEL web-server [54–56] (http://swissmodel.expasy.org). Alternatively, these sites can be mapped onto the experimentally deduced crystal structures by importing PDB files from RCSB Protein Data Bank (RCSB PDB) [325, 326] (http://www.pdb.org) or the Swiss-Model Template Library (SMTL or ExPDB) [54–56] into IMPACT_S. Sites can also be marked onto the two-dimensional (2D) codon or protein sequence alignment using the built-in Alignment Filter tool. This useful feature enables the assessment of the influence of domain-specific accumulation of mutations and helps to generate publication quality figures (Figure 10 in [76]). This tool is available in every tab of IMPACT_S, except in the TreeSAAP Evpthwy, where the user can find the built-in GNUPLOT (http://www.gnuplot.info/) Options tool, which is specifically designed to plot graphs for the sliding window analysis using any amino acid property file. IMPACT_S also incorporates phylogenetic resources such as PHYML [41] and ARCHAEOPTERYX (http://www.phylosoft.org/archaeopteryx), the successor of ATV [327], and other built-in tools that deal with Multiple Sequence Alignment (MSA) and CSV files [328]. Throughout the process of dataset analyses the user may have different requirements and hence employ these tools to perform any alignment and/or phylogenetic tree alterations, as well as the generation of new phylogenetic trees, which are required for PAML and TREESAAP and optional in DATAMONKEY analyses.

IMPACT_S development

The IMPACT_S program was implemented in Java and has been tested on Linux, Windows and MacOS workstations. Java SwingWorker threads (from Java version 1.6.0) are used to run the entire integrated software, handle outputs and to allow GUI updates. Even though it is possible to process and filter PAML results using BioPerl libraries, it is a difficult task for users with no programming skills. Several tools found in IMPACT_S were improved from our previously developed software IMPACT [328], which include the MSA Editor/Viewer, the MSA Format Converter and the CSV Viewer. IMPACT_S uses the Internet to connect to the SWISS-MODEL [54–56] and RCSB PDB [325, 326] (http://www.pdb.org/) web-servers. The functionalities related to these web-servers are found in two distinct built-in

tools, the Protein Modeler tool which implements SWISS-MODEL functionalities and the PDB Downloader tool which includes the RSCB PDB and the SMTL.

IMPACT_S availability and requirements

IMPACT_S software was developed in Java with the platform-independent context in mind, providing the use of the software by a wider audience of users. The source code can be used by anyone with skills to manipulate and extend the program capabilities, or even used in other projects. It is an open source software distributed under GNU General Public License version 3.0 (GPLv3) and is freely available with an example dataset, a detailed tutorial and manual, to allow its user-friendly application, at http://impact-s.sourceforge.net. IMPACT_S is an integrated platform, enabling users to access results and knowledge from any stage of the user controlled workflow [329]. Being an integrated platform, users are not required to install independent programs, since they are already provided in the download package, but are allowed to replace (or include) the existing binaries. IMPACT_S requires the following:

Operating System: Linux/UNIX, Windows or MacOS;

Programming Language: Java;

Optional Programs Included: CODEML (PAML), TREESAAP, PHYML, ARCHAEOPTERYX, JMOL and GNUPLOT;

Other Requirements: Java version (minimum) 1.6.0.

5.4 Results

CRiSPs belong to a class of glycoproteins exclusive to vertebrates and have been implicated in a broad range of functions. Previously, we have demonstrated that toxicoferan-reptilian CRiSPs are significantly influenced by positive selection, and in snakes more than Anguimorpha and helodermatid lizards, while their mammalian homologues exhibit extreme coding sequence conservation (Appendices V: Table S1) [318]. By using a toxicoferan-reptilian CRiSPs dataset we demonstrate the applications of IMPACT_S and its useful designs and features. It should be noted that we have used the same dataset only to present all the tabs and associated methods. Nonetheless, in a real evolutionary study the researcher must choose the method that best fits the dataset.

Site-specific analyses were executed under the 'PAML' tab (Figure 5.1A) of IMPACT_S, and the value of ω was estimated using the following pairs of models: *(i)* M0 and M3; *(ii)* M1a and M2a and *(iii)* M7 and M8. IMPACT_S was then used to automatically extract all the relevant information from the generated result files, such as a table of ω, the ML parameter estimates associated with each model (Figure 5.1D) and lnL table for every model containing the corresponding number of parameters (np) (Figure 5.1E). The LRT was conducted in IMPACT_S to compare alternate models that allow $\omega > 1$ (M3, M2a and M8) with their null models that do not (M0, M1a and M7). The user can select from three significance thresholds, namely: p = 0.05 (by default selected), p = 0.01 and p = 0.001. The degrees of freedom (*df*) are calculated and may differ according to the pair of models selected due to the different np. For example, if the pairs (M1a *vs.* M2a) or (M7 *vs.* M8) are selected the *df* is 2. The (M0 *vs.* M3) pair is an exception, for which the *df* is 4. The sites detected as evolving under positive selection according to the BEB [45] approach (Figure 5.1B) were then extracted using IMPACT_S and automatically mapped onto the 3D structure of CRiSPs (PDB: 2GIZ-A; Figure 5.1C).

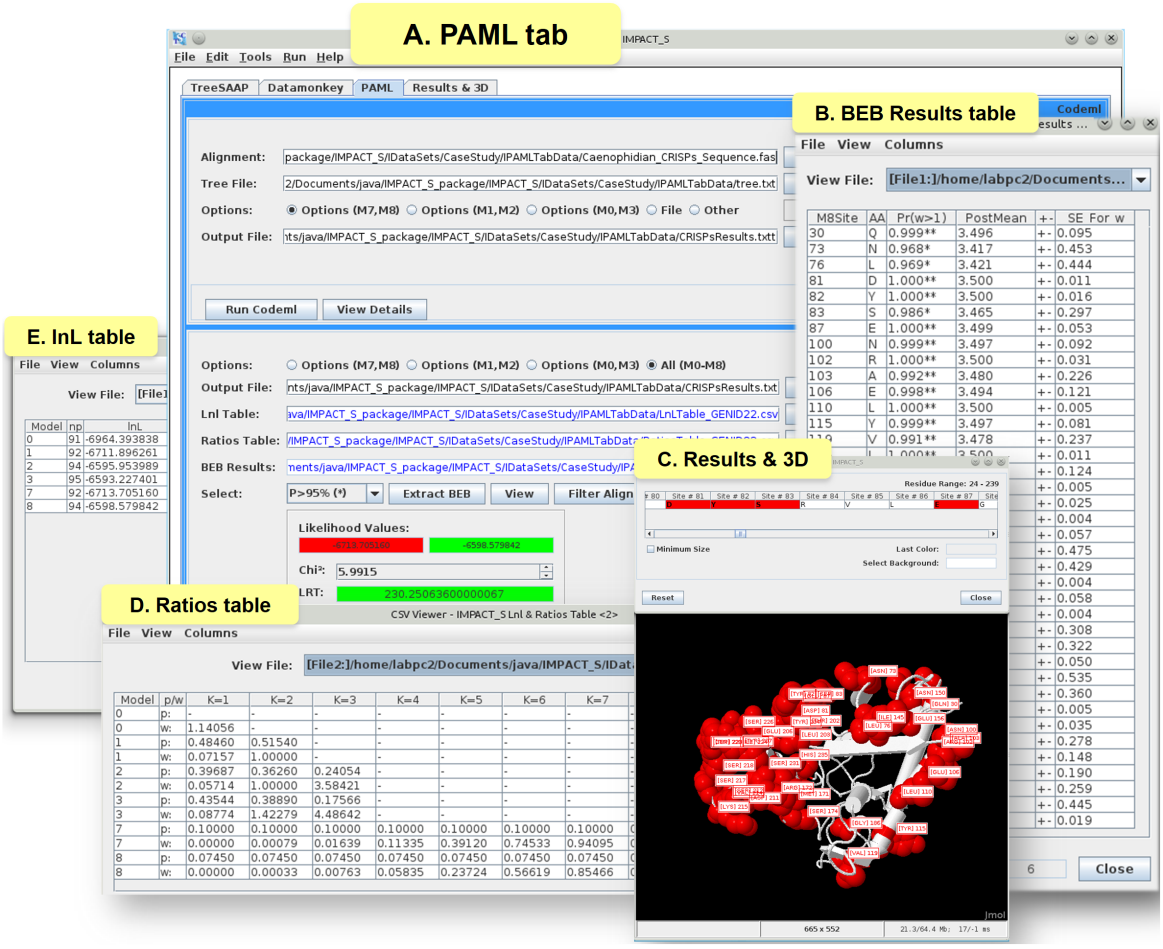
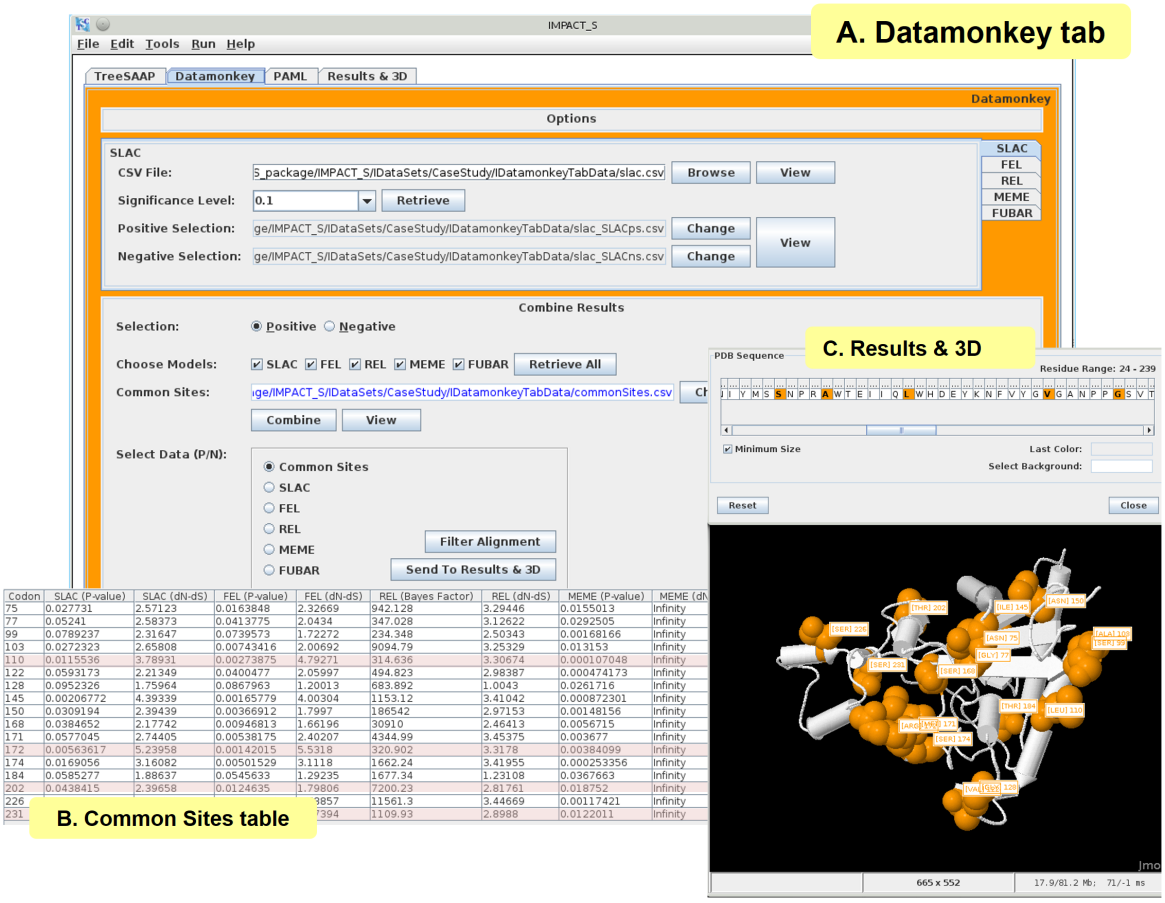


Figure 5.1: PAML tab and related results. (**A**) The 'PAML' tab from IMPACT_S showing the results of the LRT test (M7 *vs.* M8), (**B**) Automatically extracted and tabulated 'BEB results' table, (**C**) Mapping of the positively selected sites onto the 3D protein-structure of CRiSPs (2GIZ-A), (**D**) Tabulated results of ω estimation under the site-specific models (M0, M1, M3, M7 and M8) and (**E**) The 'lnL' table, showing all the log-likelihood values for these models.

All the models implemented in the DATAMONKEY web-server (SLAC, FEL, REL, MEME and FUBAR) were executed and the corresponding results were processed under the 'Datamonkey' tab of IMPACT_S (Figure 5.2A). Using the default significance cut-off, the results of positive and negative selection analyses were tabulated automatically for each method. The default significance cut-offs used in IMPACT_S are the same as on the DATAMONKEY web-server: p = 0.1 for SLAC, FEL and MEME; posterior probability = 0.9 for FUBAR and Bayes Factor = 50 for REL. The significance given by default can be changed by selecting from the available options or by typing the desired value. The positively selected sites detected using the default significance under these methods were tabulated, combined (Figure 5.2B: 'Common Sites' table) and then mapped onto the 3D structure of the protein (Figure 5.2C) in IMPACT_S.



Codon	SLAC (P-value)	SLAC (dN-dS)	FEL (P-value)	FEL (dN-dS)	REL (Bayes Factor)	REL (dN-dS)	MEME (P-value)	MEME (dN
75	0.027731	2.57123	0.0163848	2.32669	942.128	3.29446	0.0155013	Infinity
77	0.05241	2.58373	0.0413775	2.0434	347.028	3.12622	0.0292505	Infinity
99	0.0789237	2.31647	0.0739573	1.72272	234.348	2.50343	0.00168166	Infinity
103	0.0272323	2.65808	0.00743416	2.00692	9094.79	3.25329	0.013153	Infinity
110	0.0115536	3.78931	0.00273875	4.79271	314.636	3.30674	0.000107048	Infinity
122	0.0593173	2.21349	0.0400477	2.05997	494.823	2.98387	0.000474173	Infinity
128	0.0952326	1.75964	0.0867963	1.20013	683.892	1.0043	0.0261716	Infinity
145	0.00206772	4.39339	0.00165779	4.00304	1153.12	3.41042	0.000872301	Infinity
150	0.0309194	2.39439	0.00366912	1.7997	186542	2.97153	0.00148156	Infinity
168	0.0384652	2.17742	0.00946813	1.66196	30910	2.46413	0.0056715	Infinity
171	0.0577045	2.74405	0.00538175	2.40207	4344.99	3.45375	0.003677	Infinity
172	0.00563617	5.23958	0.00142015	5.5318	320.902	3.3178	0.00384099	Infinity
174	0.0169056	3.16082	0.00501529	3.1118	1662.24	3.41955	0.000253356	Infinity
184	0.0585277	1.88637	0.0545633	1.29235	1677.34	1.23108	0.0367663	Infinity
202	0.0438415	2.39658	0.0124635	1.79806	7200.23	2.81761	0.018752	Infinity
226				3857	11561.3	3.44669	0.00117421	Infinity
231				7394	1109.93	2.8988	0.0122011	Infinity

Figure 5.2: Datamonkey tab and related results. (**A**) 'Datamonkey' tab showing all the selected options, (**B**) 'Common Sites' table generated from the positive-selection analyses under various methods (SLAC, FEL, REL, MEME and FUBAR) — The sites 110, 172, 202 and 231 (highlighted in pink), were simultaneously detected by DATAMONKEY, PAML (CODEML) and TREESAAP as positively-selected (Appendices V: Table S1) [318], (**C**) 'Results & 3D' tab showing the mapping of sites found in the 'Common Sites' table.

Furthermore, we conducted a protein-level analyses under the 'TreeSAAP' tab to measure the influence of selection on 31 biochemical properties. Here, the data was processed from the 'Evpthwy' tab for the sliding window analyses and from the 'Substs' tab for mapping the substitutions on the 3D structure of the protein. In the Evpthwy directory under 'TreeSAAP' tab (Figure 5.3A), the 'Properties By Range (PBR)' table was generated (Figure 5.3B), which consists of the most significant properties under categories 7 and 8 (most radical changes). A graph was then plotted using the GNUPLOT Options dedicated tool for each significant property (Figure 5.3C).

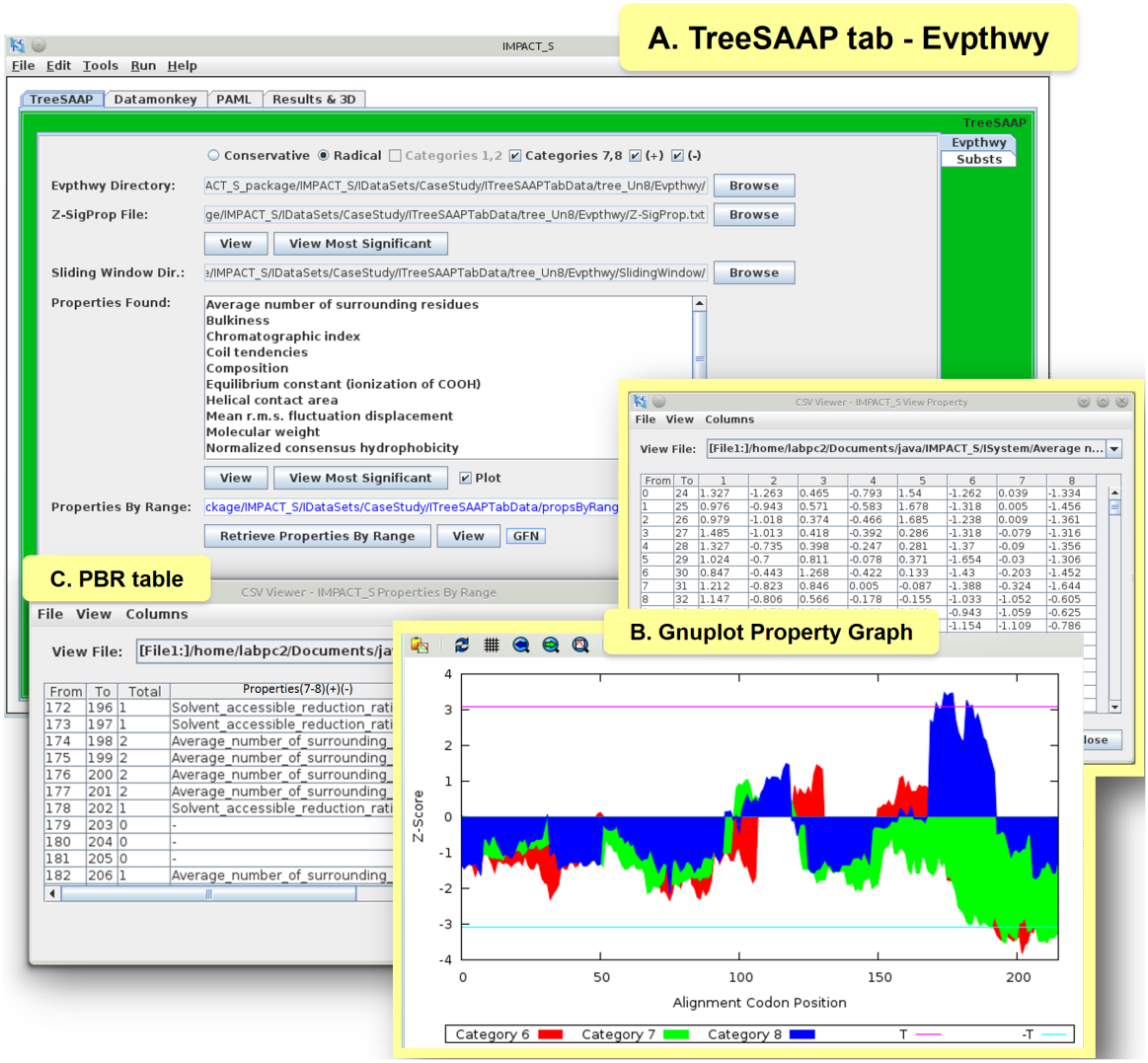


Figure 5.3: Evolutionary Pathway (Evpthwy) tab under TreeSAAP tab and related results. (**A**) 'TreeSAAP' tab showing the 'Evpthwy' tab with results of the CRiSP case study, (**B**) GNUPLOT graph for the property "Average number of surrounding residues" with Z-scores in y-axis and Codon positions in x-axis, (**C**) 'PBR' table showing all ranges ('From' – 'To') retrieved from the sliding window analysis with the 'Total' count of properties for each range and their names [Properties (7–8) (+) (-) — names of properties under significant positive (Z-score $\geq$ 3.09) and negative (Z-score $\leq$ 3.09) selection, found in the categories 7 and 8].

The 'Substs' tab (Figure 5.4A) was used to generate the 'Properties By Site (PBS)' table (Figure 5.4B) that consists of all the significant codons under categories 7 and 8. This tab provides the statistical files (Figure 5.4C) for counting properties. Using the 'Alignment Filter' tool, the sites reported in the PBS table as positively selected were highlighted in a new alignment view that keeps only these sites and their original positions in the provided alignment (Figure 5.4D). The information regarding the radical physicochemical amino acid changes varying across the reptilian-CRiSPs phylogenetic tree can also be obtained under

this tab (Appendices V: Table S2).

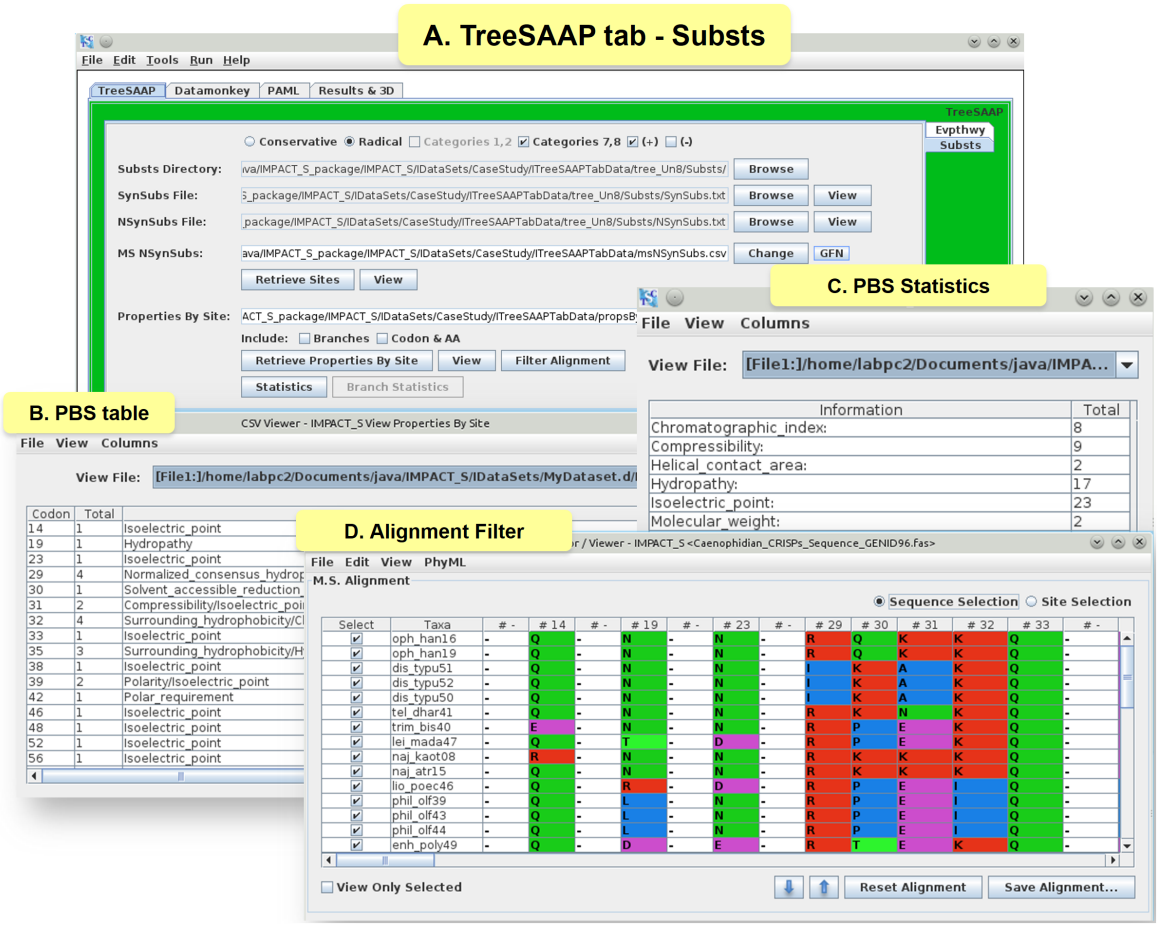


Figure 5.4: Substitutions (Substs) tab under TreeSAAP tab and related results. (**A**) 'TreeSAAP' tab showing the 'Substs' tab and the results of the CRiSPs case study, (**B**) 'PBS' table showing all the significant codons ('Codon' column) and the total count ('Total' column) of properties for each codon and their respective names ['Properties (7–8) (+)' — names of properties under significant (Z-score $\geq$ 3.09) positive selection], (**C**) 'PBS statistics', with respect to the PBS table, showing the number of times ('Total' column) that the same property ('Information' column) is selected across the data set provided and (**D**) Alignment Filter option with the final view of the alignment containing only the codons from the PBS table with their original positions.

In order to identify sites that were detected in common as positively selected by the aforementioned analyses (PAML, DATAMONKEY and TREESAAP), their resultant files were combined in IMPACT_S under the 'Results & 3D' tab (Figure 5.5A) as the MR table (Figure 5.5B; Appendices V: Table S3). The evolution of such sites are most likely to be influenced by natural selection, since they are detected in common by a diversity of selection assessment methodologies. Such sites are then automatically mapped onto the 3D structure (2GIZ-A; Figure 5.5C) using the 'Merged Results' Color Scheme option (Figure 5.5D).

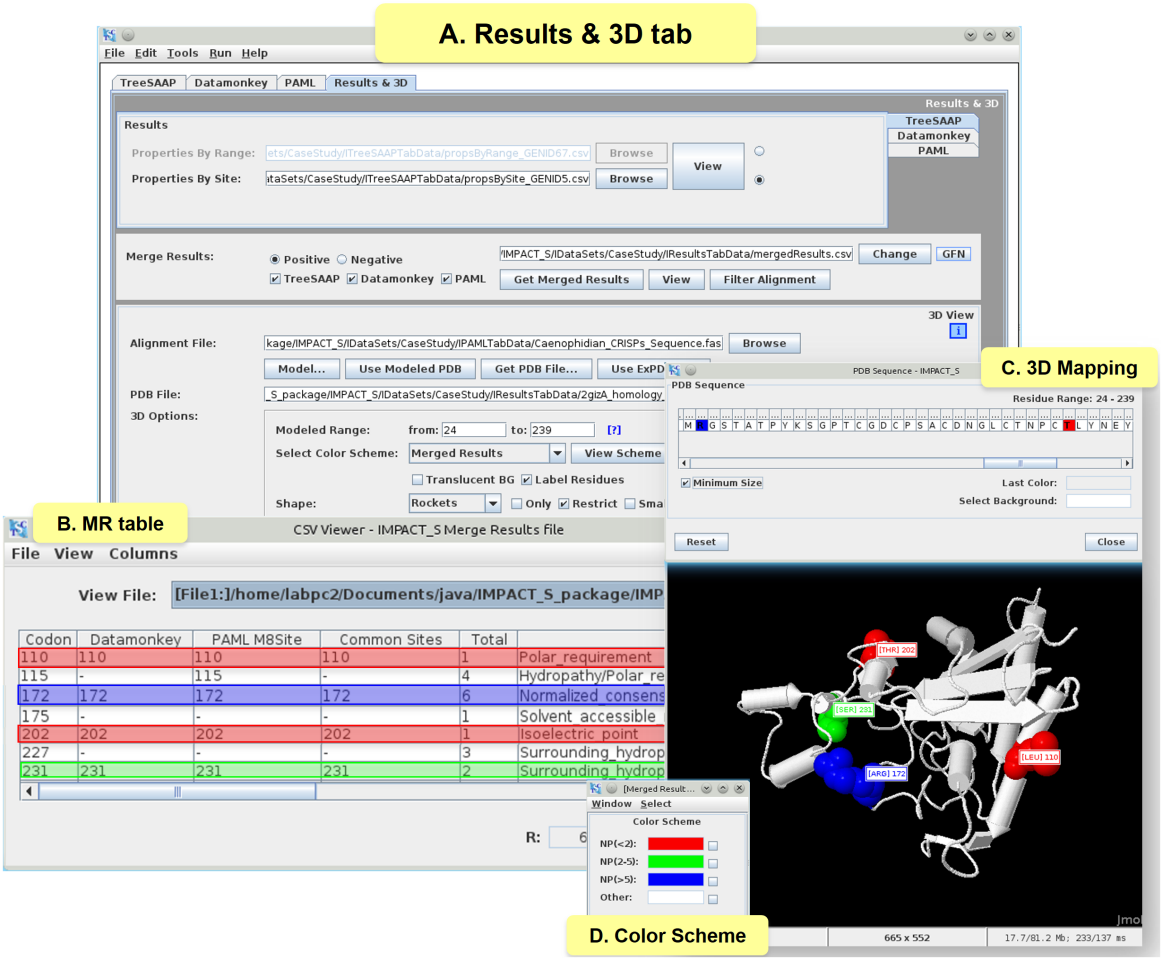


Figure 5.5: 3D mapping and related functionalities under Results & 3D tab. (**A**) 'Results & 3D' tab with the results of the CRiSPs case study, (**B**) 'MR table' (Appendices V: Table S3) presenting six columns: 'Codon' containing the site found as positive selected (TREESAAP — Z-score $\geq$ 3.09); 'Datamonkey' with the positive selected common sites positions from SLAC, FEL, REL, MEME and FUBAR; 'PAML M8Site' with the information from BEB results M8; 'Common Sites' — reporting the site position common to all the previous mentioned columns; 'Total' — count of the number of properties per site; 'Properties (7–8) (+)' with the names of properties found under categories 7 and 8 for Z-score $\geq$ 3.09, (only common sites are shown and highlighted) (**C**) 3D structure and 'PDB Sequence', mapping the sites positions from the 'Common Sites' column following the (**D**) Merged Results 'Color Scheme' according to a pre-defined number of properties: red — less than 2 (NP$<$2), green — between 2 and 5 (NP(2–5)) and blue — more than 5 (NP$>$5).

5.5 Conclusion

A diversity of selection assessment methodologies are employed to identify and corroborate sites that are influenced by various regimes of natural selection. Analyzing and integrating the result files from these methods is a tedious and time-consuming process. In this context, IMPACT_S offers a user-friendly interface to rapidly manage, filter, organize and interpret the results in different ways. Moreover, IMPACT_S provides the user the

freedom to choose different available methods, which should follow the respective authors' recommendations from programs integrated in this GUI. If properly used, this software will automatically map the positively and negatively selected sites onto the sequence alignment as well as the 3D protein structure. This enables the user to infer which sites are important, according to their specific location and significance for all types of selection. IMPACT_S is a new integrated multiprogram platform, which allows the user to combine the results from different bioinformatic programs. Merging PAML or DATAMONKEY with TREESAAP may provide additional information about amino acid properties, such as hydrophobicity, at the sites identified by the aforementioned codon methods. Overall, amino acid properties can provide crucial information that can be extremely relevant in protein-protein interactions [102] and thus a change in this biochemical property can compromise the folding of the protein interfering in its stability and function.

Part IV

GENERAL DISCUSSION AND CONCLUSION

Chapter 6

Discussion and Concluding Remarks

6.1 Discussion

The general discussion of the thesis has been divided into two major sections. The first section (**chapters 2-4**), contributes new insights into the adaptive evolution of mitochondrial and nuclear genomes of metazoans, especially in species adapted to an aquatic lifestyle (*i.e.,* ocean) and those that adapted to multiple environments (*i.e.,* terrestrial, aquatic and semi-aquatic). The second section (**chapter 5**), contributes with new bioinformatics software design to assist the analyses of large-scale genomic results.

The overall scope of **chapters 2-4** aligns closely with that of the "New Biology" (*i.e.,* foundation of genomics), where organisms' phenotypic variation is the result of their genetic makeup (*i.e.,* genes) and natural selection is the main "mechanism" that drives species evolution across environments [330]. Under this framework, the referred chapters contain the exploration of real genomic data aiming to examine patterns of adaptive selection on mitochondrial (critical for bioenergetics by aerobic respiration) and nuclear (essential for detoxification) protein-coding genes. These were tested by using different models of adaptive evolution (namely, random-site, branch, branch-site, clade models and different approaches accounting for changes in amino acid physicochemical properties).

In this regard, **chapters 2 and 3** examined the mitochondrial DNA (mtDNA) genome using the 13 protein-coding genes involved in the oxidative phosphorylation (OXPHOS) complex system: seven NADH dehydrogenases (ND1–ND6 and ND4L; Complex I), cytochrome b (CYTB; Complex III), three cytochrome c oxidases (COX1–COX3; Complex IV), and two ATPases (ATP6 and ATP8; Complex V). Here, the data set contains species from eight major taxonomic groups of the Class Cephalopoda (Oegopsida, Myopsida, Bathyteuthoidea, Sepiolidae, Sepiidae, Vampyromorpha, Octopoda and Nautiloidea). In **chapter 2**, analyses were performed considering 17 species, while in **chapter 3** further data analyses allowed to extend previous results to a total of 37 cephalopod species. The 13 mtDNA protein-coding genes were concatenated for phylogenetic reconstruction through different approaches. Notably, all methods provided a well-supported reconstruction of the Cephalopoda mitochondrial phylogeny, making clear the monophyly of Octopodiformes (*Vampyroteuthis infernalis* + Octopoda), which is consistent with the accepted Cephalopoda relationships based on multi-locus analyses [193] and mitochondrial data [184, 185, 331].

After determining the phylogeny, several major clades of species were chosen based on

important differences in terms of their ecology (*i.e.,* different ocean depths), physiology (*i.e.,* distinct metabolic rates), morphology (*i.e.,* presence or absence of cornea and number of arms) (only **chapter 2**) and gene duplication (only **chapter 3**). These traits were evaluated since they can influence the respiratory efficiency and may be subject to selective pressure. An important finding arising from these two chapters was the discovery of adaptive selection acting in the OXPHOS respiratory system of cephalopods, especially along Complex I (*nd5*), Complex III (*cytb*), Complex IV (*cox2-3*) and Complex V (*atp8*) in Octopodiformes (with eight arms), when analyzing ecophysio-morphological hypotheses (**chapter 2**). This was also the case for Complex IV (*cox2*) and Complex V (*atp6*) along Oegopsida and Bathyteuthoidea (OB; with ten arms) lineages when analyzing gene duplication hypothesis (**chapter 3**). Taken together, the discovery of positive Darwinian selection in cephalopods along with the information obtained from well-studied species revealed that positively selected sites were found to be lining along the proton translocation channels and mediating interactions between the OXPHOS subunits (**chapter 2**), or located in relevant regions predicted to influence enzymatic activity in other metazoans, and consequently ATP production (**chapter 3**). This would mean that the positively selected sites found in these studies may represent advantageous substitutions in cephalopods' energetic requirements. Possibly this has resulted from a more active mode of life (becoming agile and shell-less predators), which may have enhanced the locomotion performance with the development of a powerful jet propulsion system [163]. Over time this led to the ability to swim along the oceanic gradient and allowed a lifestyle transition from demersal to pelagic, which may have been associated with the OXPHOS complexes. Overall, positive selection may have precipitated substantial changes in Cephalopoda mitochondrial function, underlying adaptive improvements in their metabolic performance [78] in a lineage-specific manner.

Moreover, the physicochemical characterization of these sites found in specific lineages of cephalopods and others inferred through homology comparisons, conveyed possible implications for the effectiveness of molecular mechanisms of OXPHOS pathway, which could have determined the adaptation of these species to different environments (*e.g.,* varying gradients of oxygen, temperature, light, pressure, or even abundance of predators). In this regard, cephalopods from the OB clade revealed to be a valuable example because their ancestral underwent the most ancient transition from demersal to the contrasting pelagic lifestyle in Decapodiformes species [190, 237]. Therefore the positively selected sites

could benefit them by maintaining their bioenergetics efficiency across distinct ranges of the ocean column important to supply for example their mechanism of locomotion (swimming-jet propelled, to obtain food or escape from predators) and enhanced adaptation to changes in environmental temperature, pressure, and oxygen availability. In addition to that, the presence of mitochondrial duplicated genes just in this clade and the evidence of positive selection, suggest a genetic novelty occuring in this group of cephalopods, which likely assisted their high metabolic requirements and the possibility to handle the low levels of oxygen found in the deep ocean. These energetic adaptations may explain Oegopsida as the most species-rich clade within Decapodiformes. An additional finding worth highlighting, is the apparent absence of the transmembrane one (TM1 helix – **chapter 2**) of the ND5 subunit (Complex I) in cephalopods, an evidence previously reported only in arthropods and nematodes [101], which may constitute a lineage-specific bioenergetic adaptation.

Bioenergetics adaptations, as the ones previously mentioned (**chapters 2 and 3**), may play an important role on the longevity and environmental adaptation of metazoans, through a sustained production of energy for several body functions such as increasing locomotor activity (*e.g.,* swimming-jet propelled and flight). This requires additional ATP supplies to fuel the energetic expenditure, whereby is expected to increase the production of reactive oxygen species (ROS) during the normal process of OXPHOS [332]. Therefore, other potential effect of these innovations can be a balanced production of ROS through changes in the proton translocation pathway.

In line with an efficient controlling of ROS in the body of the organisms, assuring the molecular adaptation of the species to their local environment, **chapter 4** assesses several subfamilies of a family of detoxification genes, CYP2, from another successful group of metazoans (*i.e.,* vertebrates) found worldwide – the Class Aves. Uniqueness of this study is the overall genomic characterization of CYP2 subfamilies from all avian evolutionary groups: Palaeognathae, Galloanserae and Neoaves [180]. Overall, 12 CYP2 subfamilies were identified from 48 avian genomes revealing that some CYP2 genes, previously described as lineage-specific (*i.e.,* CYP2K and CYP2W), were indeed found in all the avian groups. Additionally, this study revealed the presence of CYP2F and CYP2G subfamilies in a reduced number of avian genomes. Using comparative analyses (**chapter 4**), the current knowledge of the substrate recognition sites of CYPs was updated, and several new regions (SRS0, SRS2_SRS3 and SRS3.1) were documented, highlighting their great spectrum of action

(*i.e.,* metabolism of numerous chemical compounds). Signatures of positive selection were identified in the six avian CYP2 subfamilies (CYP2C, CYP2D, CYP2H, CYP2J, CYP2K and CYP2AC), some of which are located in relevant SRS- and heme cofactors (HEM) that influence the diverse range of CYP2' substrates specificities and regulatory pathways (*i.e.,* structure and function). The array of CYP2 genes found in avian species, together with the detection of evidences of positive Darwinian selection revealed that the corresponding enzymes involved in the detoxification of dangerous endogenous and exogenous compounds, an in turn in the balancing of ROS in these species, are evolving differently according to distinct feeding habits, environments and migratory behaviors. The positively selected sites may indicate that the adaptation of detoxification enzymes requires different substrate affinities, thus potentially improving the electron-transfer chain systems. This could have contributed to the inactivation and excretion of noxious chemical compounds found in several habitats with distinct food resources, and facilitated the environmental adaptation and evolutionary success of the avian species.

In summary, the previously discussed work (**chapters 2-4**) indicates that variation in mitochondrial and nuclear (CYP2) encoded proteins has physiological implications relevant to the environmental adaptation of metazoans and provides further evidence for metabolic enzymes as targets of natural selection [333].

In the last decades, evidence of positive selection has been detected in several genes from diverse taxa (*e.g.,* reviewed in [78]). Thus, in the last section of this thesis (**chapter 5**) it was recognized the time-consuming task of manually analyzing the results for each of the adaptive selection analyses (as those performed in **chapters 2 to 4**). As such, **chapter 5** was devoted to the design of a bioinformatics software tool to assist researchers speeding up the processing of data in an easy and friendly manner, since handling enormous molecular data sets of great complexity (sequences) can be tedious and error-prone. In this context, it was developed and presented the IMPACT_S software, which allows the rapid and efficient filtering, combination and mapping of the selection results from three widely used programs for assessing the influence of selection (*i.e.,* PAML [20], DATAMONKEY [19] and TREESAAP [51]) on the three-dimensional (3D) protein structures; all this using a single easy-to-use graphical user interface.

6.2 Concluding Remarks and Future Directions

Overall, this study provides new insights to the environmental adaptation of poorly explored taxonomic groups of metazoans (*i.e.,* Classes Cephalopoda and Aves, which together cover all the environments on Earth), through the characterization of adaptive molecular evolutionary aspects of mitochondrial and detoxification genes (*i.e.,* modelling the evolution of a protein at the codon level). In this regard, the detection of patterns of positive Darwinian selection, analyzed in the context of a functional model (*i.e.,* protein 3D structure), uncovered potential implications in the physiological outputs of these metabolic enzymes, namely altered oxidative phosphorylation (bioenergetics) and variation in ROS production. Taken together, this study established a mechanistic link between the genetic variation and the environmental species' adaptation. In the future, these outcomes can be integrated with the increasing results of adaptive molecular adaptation from other taxonomic groups from similar environments and lifestyles, thus generating a global picture of important specializations that could be mediated by these genes. Moreover, these sites could be explored in an experimental framework to reveal the real consequences of the altered molecular interactions between the positively selected sites and the remaining sites of the considered pathway. Therefore, this thesis not only contributes to increase our understanding of protein-coding genes in an ecological context, but as well in the discovery of how relevant selected sites can be experimentally explored in a pharmacological context, thereby providing new tools that can help understanding several human health problems.

Finally, this study recognizes and reinforces the need for bioinformatics tools to help expedite the examination of results from the currently codon models of selection. The development of the software IMPACT_S contributes to rapid analyses, integration and even mapping of protein 3D structure, linking genotype with organismal phenotype through the identification of positively selected sites and their structural and functional implications. Softwares like this are particularly relevant in today's current large-scale data production (*e.g.,* next-generation sequencing), allowing the performance of analyses in timely manner of adaptive hypotheses that connect diverse metazoans (*i.e.,* large data sets), serving as guide to future experimental design in a cost-efficient approach.

Part V

REFERENCES

References

[1] Maloof AC, Porter SM, Moore JL, Dudás FÖ, Bowring SA, Higgins JA, et al. The earliest Cambrian record of animals and ocean geochemical change. Geological Society of America Bulletin. 2010;122(11-12):1731–1774.

[2] IUCN. The International Union for Conservation of Nature and Natural Resources Red List of Threatened Species; 2014. Available from: `http://www.iucnredlist.org`.

[3] Darwin C. On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life. J. Murray; 1859.

[4] Lemey P, Salemi M, Vandamme Am, Lehman JM, Heras J, Gómez-daglio L. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. vol. 16. Cambridge University Press; 2009.

[5] Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994;11(5):725–736.

[6] Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 1998;15(5):568–573.

[7] Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 2002;19(6):908–917.

[8] Wong WSW, Yang Z, Goldman N, Nielsen R. Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites 10.1534/genetics.104.031153. Genetics. 2004;168(2):1041–1051.

[9] Yang Z, Nielsen R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol. 2008;25(3):568–579.

[10] Kosakovsky Pond SL, Frost SDW. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. Molecular Biology and Evolution. 2005;22(5):1208–1222.

[11] Liò P, Goldman N. Models of molecular evolution and phylogeny. Genome research. 1998;8(12):1233–1244.

[12] Nei M. Selectionism and neutralism in molecular evolution. Molecular Biology and Evolution. 2005;22(12):2318–2342.

[13] Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, et al. What is a gene, post-ENCODE? History and updated definition. Genome Res. 2007;17(6):669–681.

[14] Telford MJ, Littlewood DTJ. Animal evolution: genomes, fossils, and trees. Oxford University Press; 2009.

[15] Gregory TR. CHAPTER 1 - Genome Size Evolution in Animals. In: Gregory TR, editor. The Evolution of the Genome. Burlington: Academic Press; 2005. p. 3–87.

[16] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of the Cell, Fourth Edition. Garland Science; 2002.

[17] Barton NH. Mutation and the evolution of recombination. Philosophical transactions of the Royal Society B: Biological sciences. 2010;365(1544):1281–94.

[18] Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 1998;148(3):929–936.

[19] Pond SL, Frost SD. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics. 2005;21(10):2531–2533.

[20] Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences : CABIOS. 1997;13(5):555–556.

[21] Hogeweg P. The roots of bioinformatics in theoretical biology. PLoS Computational Biology. 2011;7(3):1–5.

[22] Apweiler R, Bateman A, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, et al. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Research. 2014;42(1):191–198.

[23] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research. 2003;31(1):365–370.

[24] Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, et al. The protein information resource. Nucleic Acids Research. 2003;31(1):345–347.

[25] Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL nucleotide sequence database. Nucleic Acids Research. 2005;33(Database iss.):29–33.

[26] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235–242.

[27] Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2012;40(Database issue):48–53.

[28] Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic Acids Res. 2014;42(Database issue):749–755.

[29] Mehmood MA, Sehar U, Ahmad N. Use of Bioinformatics Tools in Different Spheres of Life Sciences. Journal of Data Mining in Genomics & Proteomics. 2014;5(2):1.

[30] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

[31] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–1797.

[32] Pei J, Tang M, Grishin NV. PROMALS3D web server for accurate multiple protein sequence and structure alignments. Nucleic Acids Res. 2008;36(Web Server issue):30–34.

[33] Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Res. 2010;38(Web Server issue):23–28.

[34] Chang JM, Di Tommaso P, Lefort V, Gascuel O, Notredame C. TCS: a web server for multiple sequence alignment evaluation and phylogenetic reconstruction. Nucleic Acids Research. 2015;43(W1):3–6.

[35] Jordan G, Goldman N. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Molecular biology and evolution. 2012;29(4):1125–1139.

[36] Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. Molecular biology and evolution. 2012;29(1):1–5.

[37] Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9(8):772.

[38] Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27(8):1164.

[39] Xia X, Xie Z. DAMBE: software package for data analysis in molecular biology and evolution. J Hered. 2001;92(4):371–373.

[40] Xia X. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol. 2013;30(7):1720–1728.

[41] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307–321.

[42] Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61(3):539–542.

[43] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–1591.

[44] Yang Z. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J Mol Evol. 2000;51(5):423–432.

[45] Yang Z, Wong WS, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 2005;22(4):1107–1118.

[46] Swanson WJ, Nielsen R, Yang Q. Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol. 2003;20(1):18–20.

[47] Yang Z, Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 1998;46(4):409–418.

[48] Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005;22(12):2472–2479.

[49] Bielawski JP, Yang Z. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J Mol Evol. 2004;59(1):121–132.

[50] Weadick CJ, Chang BS. An improved likelihood ratio test for detecting site-specific functional divergence among clades of protein-coding genes. Mol Biol Evol. 2012;29(5):1297–1300.

[51] Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA. TreeSAAP: selection on amino acid properties using phylogenetic trees. Bioinformatics. 2003;19(5):671–672.

[52] Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics. 2010;26(19):2455–2457.

[53] Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5(4):725–738.

[54] Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 2003;31(13):3381–3385.

[55] Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics. 2006;22(2):195–201.

[56] Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. Protein structure homology modeling using SWISS-MODEL workspace. Nature protocols. 2009;4(1):1–13.

[57] Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014;42(Web Server issue):252–258.

[58] DeLano WL. The PyMOL Molecular Graphics System. Schrödinger, LLC.; 2002. Available from: `http://www.pymol.org`.

[59] Cannarozzi GM, Schneider A. Codon Evolution: Mechanisms and Models. 1st ed. Oxford University Press; 2012.

[60] Schierwater B, Stadler P, DeSalle R, Podsiadlowski L. Mitogenomics and metazoan evolution. Mol Phylogenet Evol. 2013;69(2):311–312.

[61] Barber GP, Raney BJ, Clawson H, Learned K, Karolchik D. UCSC California Sea Hare Genome Browser Gateway; 2008. Available from: `http://genome-euro.ucsc.edu/cgi-bin/hgGateway?hgsid=174282363&clade=other&org=Sea+hare&db=0&redirect=manual&source=genome.cse.ucsc.edu`.

[62] Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, et al. The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Research. 2012;40(D1):918–923.

[63] Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Briefings in Bioinformatics. 2013;14(2):144–161.

[64] Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, et al. The Genome Portal of the Department of Energy Joint Genome Institute. Nucleic Acids Research. 2012;40(D1):26–32.

[65] NCBI. NCBI Genome database; 2012. Available from: `http://www.ncbi.nlm.nih.gov/genome/?term=Biomphalaria+glabrata`.

[66] Jex AR, Hall RS, Littlewood DT, Gasser RB. An integrated pipeline for next-generation sequencing and annotation of mitochondrial genomes. Nucleic Acids Res. 2010;38(2):522–533.

[67] Maddison DR, Schulz KS. The Tree of Life Web Project; 2007. Available from: http://tolweb.org.

[68] D'Onorio de Meo P, D'Antonio M, Griggio F, Lupi R, Borsani M, Pavesi G, et al. MitoZoa 2.0: a database resource and search tools for comparative and evolutionary analyses of mitochondrial genomes in Metazoa. Nucleic Acids Res. 2012;40(Database issue):1168–1172.

[69] Boore JL. Animal mitochondrial genomes. Nucleic Acids Res. 1999;27(8):1767–1780.

[70] Kolesnikov AA, Gerasimov ES. Diversity of mitochondrial genome organization. Biochemistry Mosc. 2012;77(13):1424–1435.

[71] Gregory TR. Animal Genome Size Database; 2016. Available from: http://www.genomesize.com.

[72] Porcelli D, Butlin RK, Gaston KJ, Joly D, Snook RR. The environmental genomics of metazoan thermal adaptation. Heredity (Edinb). 2015;114(5):502–514.

[73] Blier PU, Dufresne F, Burton RS. Natural selection and the evolution of mtDNA-encoded peptides: evidence for intergenomic co-adaptation. Trends Genet. 2001;17(7):400–406.

[74] Scott GR, Schulte PM, Egginton S, Scott AL, Richards JG, Milsom WK. Molecular evolution of cytochrome C oxidase underlies high-altitude adaptation in the bar-headed goose. Mol Biol Evol. 2011;28(1):351–363.

[75] Fontanillas P, DÉPraz A, Giorgi MS, Perrin N. Nonshivering thermogenesis capacity associated to mitochondrial DNA haplotypes and gender in the greater white-toothed shrew, Crocidura russula. Molecular Ecology. 2005;14(2):661–670.

[76] da Fonseca RR, Johnson WE, O'Brien SJ, Ramos MJ, Antunes A. The adaptive evolution of the mammalian mitochondrial genome. BMC Genomics. 2008;9:119.

[77] Foote AD, Morin PA, Durban JW, Pitman RL, Wade P, Willerslev E, et al. Positive selection on the killer whale mitogenome. Biol Lett. 2011;7(1):116–118.

[78] Garvin MR, Bielawski JP, Gharrett AJ. Positive Darwinian Selection in the Piston That Powers Proton Pumps in Complex I of the Mitochondria of Pacific Salmon. Plos One. 2011;6(9).

[79] Silva G, Lima FP, Martel P, Castilho R. Thermal adaptation and clinal mitochondrial DNA variation of European anchovy. Proc Biol Sci. 2014;281(1792).

[80] Teacher AG, Andre C, Merila J, Wheat CW. Whole mitochondrial genome scan for population structure and selection in the Atlantic herring. BMC Evol Biol. 2012;12:248.

[81] Duchene S, Frey A, Alfaro-Nunez A, Dutton PH, Thomas PGM, Morin PA. Marine turtle mitogenome phylogenetics and evolution. Mol Phylogenet Evol. 2012;65(1):241–250.

[82] Dowling DK, Friberg U, Lindell J. Evolutionary implications of non-neutral mitochondrial genetic variation. Trends Ecol Evol. 2008;23(10):546–554.

[83] Nelson DR. The cytochrome p450 homepage. Hum Genomics. 2009;4(1):59–65.

[84] Palmer G, Reedijk J. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature of electron-transfer proteins. Recommendations 1989. Biochim Biophys Acta. 1991;1060(3):599–611.

[85] Scott JG, Wen Z. Cytochromes P450 of insects: the tip of the iceberg. Pest Manag Sci. 2001;57(10):958–967.

[86] Nelson DR, Kamataki T, Waxman DJ, Guengerich FP, Estabrook RW, Feyereisen R, et al. The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. DNA Cell Biol. 1993;12(1):1–51.

[87] Nebert DW. Suggestions for the nomenclature of human alleles: relevance to ecogenetics, pharmacogenetics and molecular epidemiology. Pharmacogenetics. 2000;10(4):279–290.

[88] Nebert DW, Russell DW. Clinical importance of the cytochromes P450. Lancet. 2002;360(9340):1155–1162.

[89] Nebert DW, Wikvall K, Miller WL. Human cytochromes P450 in health and disease. Philos Trans R Soc Lond B Biol Sci. 2013;368(1612):20120431.

[90] Kirischian N, McArthur AG, Jesuthasan C, Krattenmacher B, Wilson JY. Phylogenetic and functional analysis of the vertebrate cytochrome p450 2 family. J Mol Evol. 2011;72(1):56–71.

[91] Korsloot A, van Gestel CA, Van Straalen NM. Environmental stress and cellular response in arthropods. CRC Press; 2004.

[92] Bertin G, Averbeck D. Cadmium: cellular effects, modifications of biomolecules, modulation of DNA repair and genotoxic consequences (a review). Biochimie. 2006;88(11):1549–1559.

[93] Ephrussi B, Slonimski PP. Yeast Mitochondria: Subcellular Units involved in the Synthesis of Respiratory Enzymes in Yeast. Nature. 1955;176(4495):1207–1208.

[94] Margulis L. Origin of Eukaryotic Cells: Evidence and Research Implications for a Theory of the Origin and Evolution of Microbial, Plant, and Animal Cells on the Precambrian Earth. Yale University Press; 1970.

[95] Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR. Mitochondrial origins. Proc Natl Acad Sci U S A. 1985;82(13):4443–4447.

[96] Castellana S, Vicario S, Saccone C. Evolutionary patterns of the mitochondrial genome in Metazoa: Exploring the role of mutation and selection in mitochondrial protein-coding genes. Genome Biology and Evolution. 2011;3(1):1067–1079.

[97] Butler JM. Chapter 2 - Basics of DNA Biology and Genetics. In: Butler JM, editor. Fundamentals of Forensic DNA Typing. San Diego: Academic Press; 2010. p. 19–41.

[98] Danovaro R, Dell'Anno A, Pusceddu A, Gambi C, Heiner I, Møbjerg Kristensen R. The first metazoa living in permanently anoxic conditions. BMC Biology. 2010;8(1):30.

[99] Saraste M. Oxidative Phosphorylation at the fin de siècle. Science. 1999;283(5407):1488–1493.

[100] Mimaki M, Wang X, McKenzie M, Thorburn DR, Ryan MT. Understanding mitochondrial complex I assembly in health and disease. Biochimica Et Biophysica Acta-Bioenergetics. 2012;1817(6):851–862.

[101] Efremov RG, Sazanov LA. Structure of the membrane domain of respiratory complex I. Nature. 2011;476(7361):414–420.

[102] Nabholz B, Ellegren H, Wolf JB. High levels of gene expression explain the strong evolutionary constraint of mitochondrial protein-coding genes. Mol Biol Evol. 2013;30(2):272–284.

[103] Brandt U. Energy converting NADH:quinone oxidoreductase (complex I). Annual Review of Biochemistry. 2006;75(Complex I):69–92.

[104] Haltia T, Saraste M, Wikström M. Subunit III of cytochrome c oxidase is not involved in proton translocation: a site-directed mutagenesis study. The EMBO journal. 1991;10(8):2015–2021.

[105] Belevich I, Verkhovsky MI, Wikström M. Proton-coupled electron transfer drives the proton pump of cytochrome c oxidase. Nature. 2006;440(7085):829–832.

[106] Smet J, Seneca S, De Paepe B, Meulemans A, Verhelst H, Leroy J, et al. Subcomplexes of mitochondrial complex V reveal mutations in mitochondrial DNA. Electrophoresis. 2009;30(20):3565–3572.

[107] Osigus HJ, Eitel M, Bernt M, Donath A, Schierwater B. Mitogenomics at the base of Metazoa. Molecular Phylogenetics and Evolution. 2013;69(2):339–351.

[108] Burger G, Gray MW, Lang BF. Mitochondrial genomes: Anything goes. Trends in Genetics. 2003;19(12):709–716.

[109] Breton S, Milani L, Ghiselli F, Guerra D, Stewart DT, Passamonti M. A resourceful genome: updating the functional repertoire and evolutionary role of animal mitochondrial DNAs. Trends Genet. 2014;30(12):555–564.

[110] Attardi G. Animal mitochondrial DNA: an extreme example of genetic economy. Int Rev Cytol. 1985;93:93–145.

[111] Greaves LC, Reeve AK, Taylor RW, Turnbull DM. Mitochondrial DNA and disease. J Pathol. 2012;226(2):274–286.

[112] Moritz C, Dowling TE, Brown WM. Evolution of animal mitochondrial DNA: Relevance for population biology and systematics. Annu Rev Ecol Syst. 1987;18:269–292.

[113] Ladoukakis ED, Zouros E. Direct evidence for homologous recombination in mussel (Mytilus galloprovincialis) mitochondrial DNA. Molecular biology and evolution. 2001;18(7):1168–1175.

[114] Passamonti M, Boore JL, Scali V. Molecular evolution and recombination in gender-associated mitochondrial DNAs of the manila clam Tapes philippinarum. Genetics. 2003;164(2):603–611.

[115] Trinei M, Berniakovich I, Pelicci PG, Giorgio M. Mitochondrial DNA copy number is regulated by cellular proliferation: A role for Ras and p66Shc. Biochimica et Biophysica Acta - Bioenergetics. 2006;1757(5–6):624–630.

[116] Doucet-Beaupre H, Breton S, Chapman EG, Blier PU, Bogan AE, Stewart DT, et al. Mitochondrial phylogenomics of the Bivalvia (Mollusca): searching for the origin and mitogenomic correlates of doubly uniparental inheritance of mtDNA. BMC Evol Biol. 2010;10:50.

[117] Hoffmann RJ, Boore JL, Brown WM. A Novel Mitochondrial Genome Organization for the Blue Mussel, Mytilus-Edulis. Genetics. 1992;131(2):397–412.

[118] Rigaa A, Monnerot M, Sellos D. Molecular cloning and complete nucleotide sequence of the repeated unit and flanking gene of the scallop Pecten maximus

mitochondrial DNA: putative replication origin features. Journal of molecular evolution. 1995;41(2):189–195.

[119] Barr CM, Neiman M, Taylor DR. Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. New Phytologist. 2005;168(1):39–50.

[120] Saccone C, Lanave C, De Grassi A. Metazoan OXPHOS gene families: evolutionary forces at the level of mitochondrial and nuclear genomes. Biochim Biophys Acta. 2006;1757(9–10):1171–1178.

[121] Meiklejohn CD, Montooth KL, Rand DM. Positive and negative selection on the mitochondrial genome. Trends in Genetics. 2007;23(6):259–263.

[122] Bazin E, Glemin S, Galtier N. Population size does not influence mitochondrial genetic diversity in animals. Science. 2006 Apr;312(5773):570–572.

[123] Finch TM, Zhao N, Korkin D, Frederick KH, Eggert LS. Evidence of positive selection in mitochondrial complexes I and V of the African elephant. PLOS ONE. 2014;9(4).

[124] Consuegra S, John E, Verspoor E, de Leaniz CG. Patterns of natural selection acting on the mitochondrial genome of a locally adapted fish species. Genetics Selection Evolution. 2015;47(1):58.

[125] Lott MT, Leipzig JN, Derbeneva O, Michael Xie H, Chalkia D, Sarmady M, et al. MtDNA variation and analysis using Mitomap and Mitomaster. Current Protocols in Bioinformatics. 2013;44(SUPPL.44):1–26.

[126] Arends A, McNab BK. The comparative energetics of 'caviomorph' rodents. Comp Biochem Physiol A Mol Integr Physiol. 2001;130(1):105–122.

[127] Bonaccorso FJ, McNab BK. Standard energetics of leaf-nosed bats (Hipposideridae): its relationship to intermittent- and protracted-foraging tactics in bats and birds. J Comp Physiol B. 2003;173(1):43–53.

[128] McNab BK. Energy constraints on carnivore diet. Nature. 2000;407(6804):584.

[129] Rand DM, Haney RA, Fry AJ. Cytonuclear coevolution: The genomics of cooperation. Trends in Ecology and Evolution. 2004;19(12):645–653.

[130] Brown TA. Genomes. 2nd. UK: Wiley-Liss Manchester; 2002.

[131] Ponting CP. The functional repertoires of metazoan genomes. Nature Reviews Genetics. 2008;9(9):689–698.

[132] Griffiths AJF. Introduction to genetic analysis. 8th ed. New York : W. H. Freeman ; Basingstoke : Palgrave; 2005.

[133] West SC. Molecular views of recombination proteins and their control. Nat Rev Mol Cell Biol. 2003;4(6):435–445.

[134] Neale MJ, Keeney S. Clarifying the mechanics of DNA strand exchange in meiotic recombination. Nature. 2006;442(7099):153–158.

[135] Vawter L, Brown WM. Nuclear and mitochondrial DNA comparisons reveal extreme rate variation in the molecular clock. Science. 1986;234(4773):194–196.

[136] Gu X, Wang Y, Gu J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nature genetics. 2002;31(2):205–9.

[137] Dehal P, Boore JL. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. PLOS Biology. 2005;3(10).

[138] Rogers J. Retroposons defined. Nature. 1983;301(5900):460.

[139] Brown TA. Genomes 3. 3rd ed. Taylor & Francis group, an informa business. Garland Science Pub.; 2007.

[140] Saccone C, Barome PO, D'Erchia AM, D'Errico I, Pesole G, Sbisa E, et al. Molecular strategies in Metazoan genomic evolution. Gene. 2002;300(1-2):195–201.

[141] D'Errico I, Gadaleta G, Saccone C. Pseudogenes in metazoa: origin and features. Briefings in functional genomics & proteomics. 2004;3(2):157–167.

[142] Ohno S. Evolution by Gene Duplication. Springer; 2014.

[143] Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999;151(4):1531–1545.

[144] Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290(5494):1151–1155.

[145] Ohno S. So much "junk" DNA in our genome. Brookhaven Symp Biol. 1972;23:366–370.

[146] Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annual review of genetics. 2005;39:309–338.

[147] Emes RD, Goodstadt L, Winter EE, Ponting CP. Comparison of the genomes of human and mouse lays the foundation of genome zoology. Human molecular genetics. 2003;12(7):701–709.

[148] Emes RD, Beatson SA, Ponting CP, Goodstadt L. Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents. Genome research. 2004;14(4):591–602.

[149] Emes RD, Riley MC, Laukaitis CM, Goodstadt L, Karn RC, Ponting CP. Comparative evolutionary genomics of androgen-binding protein genes. Genome Res. 2004;14(8):1516–1529.

[150] Yadetie F, Butcher S, Førde HE, Campsteijn C, Bouquet JM, Karlsen OA, et al. Conservation and divergence of chemical defense system in the tunicate Oikopleura dioica revealed by genome wide response to two xenobiotics. BMC Genomics. 2012;13:55.

[151] Lowry JA GAVMTNSPLJ Pearce RE, GL K. Lead and its Effects on Cytochromes P450. Journal of Drug Metabolism & Toxicology. 2012;3(6):4–7.

[152] Zhou S. Cytochrome P450 2D6: Structure, Function, Regulation and Polymorphism. CRC Press; 2016.

[153] Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM, Nebert DW. Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics. 2004;14(1):1–18.

[154] Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th ed. W.H. Freeman; 2002.

[155] Kroemer HK, Eichelbaum M. It's the genes, stupid". Molecular bases and clinical consequences of genetic cytochrome P450 2D6 polymorphism. Life Sciences. 1995;56(26):2285–2298.

[156] Kawashima A, Satta Y. Substrate-dependent evolution of cytochrome P450: rapid turnover of the detoxification-type and conservation of the biosynthesis-type. PLoS One. 2014;9(6):16.

[157] Gotoh O. Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. J Biol Chem. 1992;267(1):83–90.

[158] Johnson EF. The 2002 Bernard B. Brodie award lecture: Deciphering substrate recognition by drug-metabolizing cytochromes P450. Drug Metabolism and Disposition. 2003;31(12):1532–1540.

[159] Grespi BJ, Summers K. Positive selection in the evolution of cancer. Biological Reviews. 2006;81(3):407–424.

[160] Negishi M, Iwasaki M, Juvonen RO, Sueyoshi T, Darden TA, Pedersen LG. Structural flexibility and functional versatility of cytochrome {P450} and rapid evolution. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 1996;350(1):43–50.

[161] da Fonseca RR, Antunes A, Melo A, Ramos MJ. Structural divergence and adaptive evolution in mammalian cytochromes P450 2C. Gene. 2007;387(1-2):58–66.

[162] Young RE, Vecchione M, Mangold KM. Cephalopoda Cuvier 1797. Octopods, squids, nautiluses, etc.. Version 27 February 2016 (under construction) . in The Tree of Life Web Project, http://tolweb.org/; 2016. Available from: `http://tolweb.org/Cephalopoda/19386/2016.02.27`.

[163] Kroger B, Vinther J, Fuchs D. Cephalopod origin and evolution: A congruent picture emerging from fossils, development and molecules: Extant cephalopods are younger than previously realised and were under major selection to become agile, shell-less predators. Bioessays. 2011;33(8):602–613.

[164] Albertin CB, Bonnaud L, Brown CT, Crookes-Goodson WJ, da Fonseca RR, Di Cristo C, et al. Cephalopod genomics: A plan of strategies and organization. Stand Genomic Sci. 2012;7(1):175–188.

[165] Hunsicker ME, Essington TE, Watson R, Sumaila UR. The contribution of cephalopods to global marine fisheries: can we have our squid and eat them too? Fish and Fisheries. 2010;11(4):421–438.

[166] Brandt A, Gooday AJ, Brandao SN, Brix S, Brokeland W, Cedhagen T, et al. First insights into the biodiversity and biogeography of the Southern Ocean deep sea. Nature. 2007;447(7142):307–311.

[167] Gage JD, Tyler PA. Deep-Sea Biology: A Natural History of Organisms at the Deep-Sea Floor. Cambridge University Press; 1991.

[168] Levin LA. Oxygen minimum zone benthos: Adaptation and community response to hypoxia. Oceanography and Marine Biology. 2003;41:1–45.

[169] Seibel, Thuesen, Childress. Flight of the vampire: ontogenetic gait-transition in vampyroteuthis infernalis (Cephalopoda: vampyromorpha). The Journal of experimental biology. 1998;201:2413–2424.

[170] Kawashima Y, Nishihara H, Akasaki T, Nikaido M, Tsuchiya K, Segawa S, et al. The complete mitochondrial genomes of deep-sea squid (Bathyteuthis abyssicola), bob-tail squid (Semirossia patagonica) and four giant cuttlefish (Sepia apama, S. latimanus, S. lycidas and S. pharaonis), and their application to the phylogenetic analysis o. Mol Phylogenet Evol. 2013;69(3):980–993.

[171] Gissi C, Iannelli F, Pesole G. Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity (Edinb). 2008;101(4):301–320.

[172] Ostrom JH. Archaeopteryx and the origin of birds. Biological Journal of the Linnean Society. 1976;8(2):91–182.

[173] Xu X, You H, Du K, Han F. An Archaeopteryx-like theropod from China and the origin of Avialae. Nature. 2011;475(7357):465–470.

[174] Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014;346(6215):1311–1320.

[175] Ellegren H. The Evolutionary Genomics of Birds. Annual Review of Ecology, Evolution, and Systematics. 2013;44(1):239–259.

[176] Gill FB. Ornithology. 3rd ed. W. H. Freeman; 2007.

[177] Holmes DJ, Ottinger MA. Aging in birds. Age. 2005;27(4):3–5.

[178] Lupi R, de Meo PD, Picardi E, D'Antonio M, Paoletti D, Castrignano T, et al. MitoZoa: a curated mitochondrial genome database of metazoans for comparative genomics studies. Mitochondrion. 2010;10(2):192–199.

[179] Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank; 2016. Available from: `http://www.ncbi.nlm.nih.gov/GenBank/`.

[180] Zhang G, Li B, Li C, Gilbert MTP, Jarvis E, The Avian Genome C, et al.. The avian phylogenomic project data. GigaScience Database; 2014.

[181] Janssens TKS, Roelofs D, Van Straalen NM. Molecular mechanisms of heavy metal tolerance and evolution in invertebrates. Insect Science. 2009;16(1):3–18.

[182] Nateewathana A, Munprasit A, Dithachey P. Systematics and Distribution of Oceanic Cephalopods in the South China Sea, Area III: Western Philippines. Proceedings of the SEAFDEC Seminar on Fishery Resources in the South China Sea, Area III: Western Philippines (SEAFDEC). 1998;p. 76–100.

[183] MarineBio. Vampire Squids, Vampyroteuthis infernalis - MarineBio.org.; 2012.

[184] Allcock AL, Cooke IR, Strugnell JM. What can the mitochondrial genome reveal about higher-level phylogeny of the molluscan class Cephalopoda? Zoological Journal of the Linnean Society. 2011;161(3):573–586.

[185] Akasaki T, Nikaido M, Tsuchiya K, Segawa S, Hasegawa M, Okada N. Extensive mitochondrial gene arrangements in coleoid Cephalopoda and their phylogenetic implications. Mol Phylogenet Evol. 2006;38(3):648–658.

[186] Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010;27(2):221–224.

[187] Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647–1649.

[188] Xia X, Xie Z, Salemi M, Chen L, Wang Y. An index of substitution saturation and its application. Mol Phylogenet Evol. 2003;26(1):1–7.

[189] Rambaut A, Suchard MA, Xie W, Drummond AJ. Tracer v1.6.; 2013. Available from: `http://tree.bio.ed.ac.uk/software/tracer`.

[190] Lindgren AR, Pankey MS, Hochberg FG, Oakley TH. A multi-gene phylogeny of Cephalopoda supports convergent morphological evolution in association with multiple habitat shifts in the marine environment. BMC Evol Biol. 2012;12:129.

[191] Phillips MJ, Delsuc F, Penny D. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol. 2004;21(7):1455–1458.

[192] Meiklejohn KA, Danielson MJ, Faircloth BC, Glenn TC, Braun EL, Kimball RT. Incongruence among different mitochondrial regions: a case study using complete mitogenomes. Mol Phylogenet Evol. 2014;78:314–323.

[193] Strugnell J, Norman M, Jackson J, Drummond AJ, Cooper A. Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) using a multigene approach; the effect of data partitioning on resolving phylogenies in a Bayesian framework. Mol Phylogenet Evol. 2005;37(2):426–441.

[194] Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 2000;155(1):431–449.

[195] Anisimova M, Bielawski JP, Yang Z. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol. 2001;18(8):1585–1592.

[196] Anisimova M, Bielawski JP, Yang Z. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol. 2002;19(6):950–958.

[197] Suzuki Y, Nei M. Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Mol Biol Evol. 2001;18(12):2179–2185.

[198] Suzuki Y, Nei M. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. Mol Biol Evol. 2002;19(11):1865–1869.

[199] Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012;8(7):1002764.

[200] McClellan DA, McCracken KG. Estimating the influence of selection on the variable amino acid sites of the cytochrome B protein functional domains. Mol Biol Evol. 2001;18(6):917–925.

[201] McClellan DA, Palfreyman EJ, Smith MJ, Moss JL, Christensen RG, Sailsbery JK. Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. Mol Biol Evol. 2005;22(3):437–455.

[202] Maldonado E, Sunagar K, Almeida D, Vasconcelos V, Antunes A. IMPACT_S: Integrated Multiprogram Platform to Analyze and Combine Tests of Selection. PLoS One. 2014;9(10):96243.

[203] Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305(3):567–580.

[204] Alnajjar KS, Hosler J, Prochaska L. Role of the N-terminus of subunit III in proton uptake in cytochrome c oxidase of Rhodobacter sphaeroides. Biochemistry. 2014;53(3):496–504.

[205] Brown MD, Voljavec AS, Lott MT, Torroni A, Yang CC, Wallace DC. Mitochondrial DNA complex I and III mutations associated with Leber's hereditary optic neuropathy. Genetics. 1992;130(1):163–173.

[206] Carelli V, Ghelli A, Bucchi L, Montagna P, De Negri A, Leuzzi V, et al. Biochemical features of mtDNA 14484 (ND6/M64V) point mutation associated with Leber's hereditary optic neuropathy. Ann Neurol. 1999;45(3):320–328.

[207] Chinnery PF, Brown DT, Andrews RM, Singh-Kler R, Riordan-Eva P, Lindley J, et al. The mitochondrial ND6 gene is a hot spot for mutations that cause Leber's hereditary optic neuropathy. Brain. 2001;124(Pt 1):209–218.

[208] De Vries DD, Went LN, Bruyn GW, Scholte HR, Hofstra RM, Bolhuis PA, et al. Genetic and biochemical impairment of mitochondrial complex I activity in a family with Leber hereditary optic neuropathy and hereditary spastic dystonia. Am J Hum Genet. 1996;58(4):703–711.

[209] Horvath R, Scharfe C, Hoeltzenbein M, Do BH, Schroder C, Warzok R, et al. Childhood onset mitochondrial myopathy and lactic acidosis caused by a stop mutation in the mitochondrial cytochrome c oxidase III gene. J Med Genet. 2002;39(11):812–816.

[210] Huang LS, Cobessi D, Tung EY, Berry EA. Binding of the respiratory chain inhibitor antimycin to the mitochondrial bc1 complex: a new crystal structure reveals an altered intramolecular hydrogen-bonding pattern. J Mol Biol. 2005;351(3):573–597.

[211] Kao MC, Di Bernardo S, Nakamaru-Ogiso E, Miyoshi H, Matsuno-Yagi A, Yagi T. Characterization of the membrane domain subunit NuoJ (ND6) of the NADH-quinone oxidoreductase from Escherichia coli by chromosomal DNA manipulation. Biochemistry. 2005;44(9):3562–3571.

[212] Kirby DM, Kahler SG, Freckmann ML, Reddihough D, Thorburn DR. Leigh disease caused by the mitochondrial DNA G14459A mutation in unrelated families. Ann Neurol. 2000;48(1):102–104.

[213] Marechal A, Meunier B, Lee D, Orengo C, Rich PR. Yeast cytochrome c oxidase: a model system to study mitochondrial forms of the haem-copper oxidase superfamily. Biochim Biophys Acta. 2012;1817(4):620–628.

[214] Mayorov V, Biousse V, Newman NJ, Brown MD. The role of the ND5 gene in LHON: characterization of a new, heteroplasmic LHON mutation. Ann Neurol. 2005;58(5):807–811.

[215] Meunier B, Fisher N, Ransac S, Mazat JP, Brasseur G. Respiratory complex $\{$III$\}$ dysfunction in humans and the use of yeast as a model organism to study mitochondrial myopathy and associated diseases. Biochimica et Biophysica Acta (BBA) - Bioenergetics. 2013;1827(11-12):1346–1361.

[216] Michel J, DeLeon-Rangel J, Zhu S, Van Ree K, Vik SB. Mutagenesis of the L, M, and N subunits of Complex I from Escherichia coli indicates a common role in function. PLoS One. 2011;6(2):17420.

[217] Nakamaru-Ogiso E, Kao MC, Chen H, Sinha SC, Yagi T, Ohnishi T. The membrane subunit NuoL(ND5) is involved in the indirect proton pumping mechanism of Escherichia coli complex I. J Biol Chem. 2010;285(50):39070–39078.

[218] Patsi J, Kervinen M, Finel M, Hassinen IE. Leber hereditary optic neuropathy mutations in the ND6 subunit of mitochondrial complex I affect ubiquinone reduction kinetics in a bacterial model of the enzyme. Biochem J. 2008;409(1):129–137.

[219] Rahman S, Taanman JW, Cooper JM, Nelson I, Hargreaves I, Meunier B, et al. A missense mutation of cytochrome oxidase subunit II causes defective assembly and myopathy. Am J Hum Genet. 1999;65(4):1030–1039.

[220] Valentino ML, Avoni P, Barboni P, Pallotti F, Rengo C, Torroni A, et al. Mitochondrial DNA nucleotide changes C14482G and C14482A in the ND6 gene are pathogenic for Leber's hereditary optic neuropathy. Ann Neurol. 2002;51(6):774–778.

[221] Wenz T, Covian R, Hellwig P, Macmillan F, Meunier B, Trumpower BL, et al. Mutational analysis of cytochrome b at the ubiquinol oxidation site of yeast complex III. J Biol Chem. 2007;282(6):3977–3988.

[222] Wissinger B, Besch D, Baumann B, Fauser S, Christ-Adler M, Jurklies B, et al. Mutation analysis of the ND6 gene in patients with Lebers hereditary optic neuropathy. Biochem Biophys Res Commun. 1997;234(2):511–515.

[223] Xia D, Yu CA, Kim H, Xia JZ, Kachurin AM, Zhang L, et al. Crystal structure of the cytochrome bc1 complex from bovine heart mitochondria. Science. 1997;277(5322):60–66.

[224] Efremov RG, Baradaran R, Sazanov LA. The architecture of respiratory complex I. Nature. 2010;465(7297):441–445.

[225] Melo-Ferreira J, Vilela J, Fonseca MM, da Fonseca RR, Boursot P, Alves PC. The elusive nature of adaptive mitochondrial DNA evolution of an arctic lineage prone to frequent introgression. Genome Biol Evol. 2014;6(4):886–896.

[226] Mathiesen C, Hagerhall C. Transmembrane topology of the NuoL, M and N subunits of NADH:quinone oxidoreductase and their homologues among membrane-bound hydrogenases and bona fide antiporters. Biochim Biophys Acta. 2002;1556(2-3):121–132.

[227] Drummond DA, Wilke CO. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet. 2009;10(10):715–724.

[228] Stefani M, Dobson CM. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. J Mol Med (Berl). 2003;81(11):678–699.

[229] Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, et al. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature. 2015;524(7564):220–224.

[230] Boyle P, Rodhouse P. Cephalopods: Ecology and Fisheries. Blackwell Publishing Ltd; 2007.

[231] Nesis KN. Distribution of recent Cephalopoda and implications for Plio-Pleistocene events. Berliner Paläobiologische Abhandlungen. 2003;3:199–224.

[232] Winkelmann I, Campos PF, Strugnell J, Cherel Y, Smith PJ, Kubodera T, et al. Mitochondrial genome diversity and population structure of the giant squid Architeuthis: genetics sheds new light on one of the most enigmatic marine species. Proc Biol Sci. 2013;280(1759):20130273.

[233] Cheng R, Zheng X, Ma Y, Li Q. The complete mitochondrial genomes of two octopods Cistopus chinensis and Cistopus taiwanicus: revealing the phylogenetic position of the genus Cistopus within the order Octopoda. PLoS One. 2013;8(12):84216.

[234] Dai L, Zheng X, Kong L, Li Q. DNA barcoding analysis of Coleoidea (Mollusca: Cephalopoda) from Chinese waters. Molecular ecology resources. 2012;12(3):437–447.

[235] Zheng X, Yang J, Lin X, Wang R. Phylogenetic relationships among the decabrachia cephalopods inferred from mitochondrial DNA sequences. Journal of Shellfish Research. 2004;23(3):881–887.

[236] Almeida D, Maldonado E, Vasconcelos V, Antunes A. Adaptation of the Mitochondrial Genome in Cephalopods: Enhancing Proton Translocation Channels and the Subunit Interactions. PLoS One. 2015;10(8):0135405.

[237] Kröger B, Servais T, Zhang Y. The origin and initial rise of pelagic cephalopods in the Ordovician. PLoS ONE. 2009;4(9):7262.

[238] Yokobori S, Fukuda N, Nakamura M, Aoyama T, Oshima T. Long-term conservation of six duplicated structural genes in cephalopod mitochondrial genomes. Mol Biol Evol. 2004;21(11):2034–2046.

[239] Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res. 2011;39(Database issue):289–294.

[240] Altenhoff AM, Gil M, Gonnet GH, Dessimoz C. Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS One. 2013;8(1):53786.

[241] Fernández R, Laumer CE, Vahtera V, Libro S, Kaluziak S, Sharma PP, et al. Evaluating topological conflict in centipede phylogeny using transcriptomic data sets. Molecular Biology and Evolution. 2014;31(6):1500–1513.

[242] Wernersson R, Pedersen AG. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 2003;31(13):3537–3539.

[243] Pattengale ND, Alipour M, Bininda-Emonds OR, Moret BM, Stamatakis A. How many bootstrap replicates are necessary? J Comput Biol. 2010;17(3):337–354.

[244] Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 2002;51(3):492–508.

[245] Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 2001;17(12):1246–1247.

[246] Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol. 1989;29(2):170–179.

[247] Weadick CJ, Chang BS. Complex patterns of divergence among green-sensitive (RH2a) African cichlid opsins revealed by Clade model analyses. BMC Evol Biol. 2012;12:206.

[248] Fletcher W, Yang Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol. 2010;27(10):2257–2267.

[249] Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol. 2009;1:114–118.

[250] Schott RK, Refvik SP, Hauser FE, Lopez-Fernandez H, Chang BS. Divergent positive selection in rhodopsin from lake and riverine cichlid fishes. Mol Biol Evol. 2014;31(5):1149–1165.

[251] Lin YG, Weadick CJ, Santini F, Chang BS. Molecular evolutionary analysis of vertebrate transducins: a role for amino acid variation in photoreceptor deactivation. J Mol Evol. 2013;77(5-6):231–245.

[252] Sanderson MJ. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol. 2002;19(1):101–109.

[253] Bollback JP. SIMMAP: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics. 2006;7:88.

[254] Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution. 2012;3(2):217–223.

[255] Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, et al. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics. 2014;30(15):2216–2218.

[256] Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.; 2015.

[257] Conant GC, Wagner GP, Stadler PF. Modeling amino acid substitution patterns in orthologous and paralogous genes. Mol Phylogenet Evol. 2007;42(2):298–307.

[258] Atchley WR, Zhao J, Fernandes AD, Druke T. Solving the protein sequence metric problem. Proc Natl Acad Sci U S A. 2005;102(18):6395–6400.

[259] Xu T, Pagadala V, Mueller DM. Understanding structure, function, and mutations in the mitochondrial ATP synthase. Microb Cell. 2015;2(4):105–125.

[260] Zhang F, Broughton RE. Heterogeneous natural selection on oxidative phosphorylation genes among fishes with extreme high and low aerobic performance. BMC Evolutionary Biology. 2015;15(1):1–15.

[261] Garvin MR, Bielawski JP, Sazanov LA, Gharrett AJ. Review and meta-analysis of natural selection in mitochondrial complex I in metazoans. J Zool Syst Evol Res. 2015;53.

[262] Allegretti M, Klusch N, Mills DJ, Vonck J, Kuhlbrandt W, Davies KM. Horizontal membrane-intrinsic alpha-helices in the stator a-subunit of an F-type ATP synthase. Nature. 2015;521(7551):237–240.

[263] Overholtzer MH, Yakowec PS, Cameron V. The effect of amino acid substitutions in the conserved aromatic region of subunit II of cytochrome c oxidase in Saccharomyces cerevisiae. J Biol Chem. 1996;271(13):7719–7724.

[264] Rawson PD, Burton RS. Molecular evolution at the cytochrome oxidase subunit 2 gene among divergent populations of the intertidal copepod, Tigriopus californicus. Journal of Molecular Evolution. 2006;62(6):753–764.

[265] Zhou A, Rohou A, Schep DG, Bason JV, Montgomery MG, Walker JE, et al. Structure and conformational states of the bovine mitochondrial ATP synthase by cryo-EM. Elife. 2015;4:10180.

[266] Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28(10):2731–2739.

[267] Portner HO, Zielinski S. Environmental constraints and the physiology of performance in squid. South African Journal of Marine Science-Suid-Afrikaanse Tydskrif Vir Seewetenskap. 1998;20mm:207–221.

[268] Allcock AL, Lindgren A, Strugnell JM. The contribution of molecular data to our understanding of cephalopod evolution and systematics: a review. Journal of Natural History. 2015;49(21-24):1373–1421.

[269] Schirtzinger EE, Tavares ES, Gonzales LA, Eberhard JR, Miyaki CY, Sanchez JJ, et al. Multiple independent origins of mitochondrial control region duplications in the order Psittaciformes. Mol Phylogenet Evol. 2012;64(2):342–356.

[270] Bonnaud L, Pichon D, Boucher-Rodoni R. Molecular approach of decabrachia phylogeny: is Idiosepius definitely not a sepiolid. Phuket marine biological Centre Research Bulletin. 2005;66(January):203–212.

[271] Young RE, Vecchione M. Morphological observations on a hatchling and a paralarva of the vampire squid, Vampyroteuthis infernalis Chun (Mollusca : Cephalopoda). Proceedings of the Biological Society of Washington. 1999;112(4):661–666.

[272] Carlini DB, Reece KS, Graves JE. Actin gene family evolution and the phylogeny of coleoid cephalopods (Mollusca: Cephalopoda). Mol Biol Evol. 2000;17(9):1353–1370.

[273] Strugnell J, Nishiguchi MK. Molecular phylogeny of coleoid cephalopods (Mollusca : Cephalopoda) inferred from three mitochondrial and six nuclear loci: A comparison of alignment, implied alignment and analysis methods. Journal of Molluscan Studies. 2007;73:399–410.

[274] Naef A. Die Cephalopoden, Systematik. Fauna e Flora del Golfo di Napoli, Monograph. 1921;35(I-1):1–863.

[275] Kucharczyk R, Ezkurdia N, Couplan E, Procaccio V, Ackerman SH, Blondel M, et al. Consequences of the pathogenic T9176C mutation of human mitochondrial DNA on yeast mitochondrial ATP synthase. Biochim Biophys Acta. 2010;1797(6-7):1105–1112.

[276] Jacobsen MW, Pujolar JM, Hansen MM. Relationship between amino acid changes in mitochondrial ATP6 and life-history variation in anguillid eels. Biology Letters. 2015;11(3).

[277] Holm L, Saraste M, Wikstrom M. Structural Models of the Redox Centers in Cytochrome-Oxidase. Embo Journal. 1987;6(9):2819–2823.

[278] Hendrickx ME, Urbano B, Zamorano P. Distribution of pelagic squids Abraliopsis Joubin, 1896 (Enoploteuthidae) and Pterygioteuthis P. Fischer, 1896 (Pyroteuthidae)

(Cephalopoda, Decapodiformes, Oegopsida) in the Mexican Pacific. Zookeys. 2015;537:51–64.

[279] Zhao J, Benlekbir S, Rubinstein JL. Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase. Nature. 2015;521(7551):241–245.

[280] Guengerich FP. Cytochrome p450 and chemical toxicology. Chem Res Toxicol. 2008;21(1):70–83.

[281] Konstandi M, Johnson EO, Lang MA. Consequences of psychophysiological stress on cytochrome P450-catalyzed drug metabolism. Neurosci Biobehav Rev. 2014;45C:149–167.

[282] Pochapsky TC, Kazanis S, Dang M. Conformational plasticity and structure/function relationships in cytochromes P450. Antioxid Redox Signal. 2010;13(8):1273–1296.

[283] Nebert DW, Gonzalez FJ. P450 genes: structure, evolution, and regulation. Annu Rev Biochem. 1987;56:945–993.

[284] Jirimutu, Wang Z, Ding G, Chen G, Sun Y, Sun Z, et al. Genome sequences of wild and domestic bactrian camels. Nat Commun. 2012;3:1202.

[285] Sullivan RJ, Hagen EH, Hammerstein P. Revealing the paradox of drug reward in human evolution. Proc Biol Sci. 2008;275(1640):1231–1241.

[286] Saxena A, Tripathi KP, Roy S, Khan F, Sharma A. Pharmacovigilance: effects of herbal components on human drugs interactions involving cytochrome P450. Bioinformation. 2008;3(5):198–204.

[287] Danielson PB. The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. Curr Drug Metab. 2002;3(6):561–597.

[288] Nelson DR. Comparison of P450s from human and fugu: 420 million years of vertebrate P450 evolution. Arch Biochem Biophys. 2003;409(1):18–24.

[289] Nelson DR. Cytochrome P450 nomenclature. Methods Mol Biol. 1998;107:15–24.

[290] Thomas JH. Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet. 2007;3(5):67.

[291] Watanabe KP, Kawai YK, Ikenaka Y, Kawata M, Ikushiro S, Sakaki T, et al. Avian cytochrome P450 (CYP) 1-3 family genes: isoforms, evolutionary relationships, and mRNA expression in chicken liver. PLoS One. 2013;8(9):75689.

[292] Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 2014;346(6215):1320–1331.

[293] Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg Le A, et al. Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. PLoS Biol. 2010;8(9).

[294] Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, et al. The genome of a songbird. Nature. 2010;464(7289):757–762.

[295] Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics. 2010;26(19):2462–2463.

[296] Csuros M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics. 2010;26(15):1910–1912.

[297] Farris JS. Phylogenetic Analysis Under Dollo's Law. Systematic Biology. 1977;26(1):77–88.

[298] Antunes A, Ramos MJ. Gathering computational genomics and proteomics to unravel adaptive evolution. Evol Bioinform Online. 2007;3:207–209.

[299] Poulos TL, Finzel BC, Howard AJ. High-resolution crystal structure of cytochrome P450cam. J Mol Biol. 1987;195(3):687–700.

[300] Watson GE, Watson GE. In: Penguins: Spheniscidae. American Geophysical Union; 2013. p. 63–84.

[301] Sheng J, Guo J, Hua Z, Caggana M, Ding X. Characterization of human CYP2G genes: widespread loss-of-function mutations and genetic polymorphism. Pharmacogenetics. 2000;10(8):667–678.

[302] Richardson TH, Johnson EF. Alterations of the regiospecificity of progesterone metabolism by the mutagenesis of two key amino acid residues in rabbit cytochrome P450 2C3v. J Biol Chem. 1994;269(39):23937–23943.

[303] Baader M, Gnerre C, Stegeman JJ, Meyer UA. Transcriptional activation of cytochrome P450 CYP2C45 by drugs is mediated by the chicken xenobiotic receptor (CXR) interacting with a phenobarbital response enhancer unit. J Biol Chem. 2002;277(18):15647–15653.

[304] Yasukochi Y, Satta Y. Evolution of the CYP2D gene cluster in humans and four non-human primates. Genes Genet Syst. 2011;86(2):109–116.

[305] Yasukochi Y, Satta Y. Molecular evolution of the CYP2D subfamily in primates: purifying selection on substrate recognition sites without the frequent or long-tract gene conversion. Genome Biol Evol. 2015;7(4):1053–1067.

[306] Kanetoshi A, Ward AM, May BK, Rifkind AB. Immunochemical identity of the 2,3,7,8-tetrachlorodibenzo-p-dioxin- and beta-naphthoflavone-induced cytochrome P-450 arachidonic acid epoxygenases in chick embryo liver: distinction from the omega-hydroxylase and the phenobarbital-induced epoxygenase. Mol Pharmacol. 1992;42(6):1020–1026.

[307] Hu SX. Effect of age on hepatic cytochrome P450 of Ross 708 broiler chickens. Poult Sci. 2013;92(5):1283–1292.

[308] Ourlin JC, Baader M, Fraser D, Halpert JR, Meyer UA. Cloning and functional expression of a first inducible avian cytochrome P450 of the CYP3A subfamily (CYP3A37). Arch Biochem Biophys. 2000;373(2):375–384.

[309] Yu Z, Huse LM, Adler P, Graham L, Ma J, Zeldin DC, et al. Increased CYP2J expression and epoxyeicosatrienoic acid formation in spontaneously hypertensive rat kidney. Mol Pharmacol. 2000;57(5):1011–1020.

[310] Rainio MJ, Kanerva M, Wahlberg N, Nikinmaa M, Eeva T. Variation of basal EROD activities in ten passerine bird species–relationships with diet and migration status. PLoS One. 2012;7(3):33926.

[311] Sunagar K, Jackson TN, Undheim EA, Ali SA, Antunes A, Fry BG. Three-Fingered RAVERs: Rapid Accumulation of Variations in Exposed Residues of Snake Venom Toxins. Toxins (Basel). 2013;5(11):2172–2208.

[312] Brust A, Sunagar K, Undheim EA, Vetter I, Yang DC, Casewell NR, et al. Differential evolution and neofunctionalization of snake venom metalloprotease domains. Mol Cell Proteomics. 2013;12(3):651–663.

[313] Kini RM, Chan YM. Accelerated evolution and molecular surface of venom phospholipase A2 enzymes. J Mol Evol. 1999;48(2):125–132.

[314] Kozminsky-Atias A, Zilberberg N. Molding the business end of neurotoxins by diversifying evolution. FASEB J. 2012;26(2):576–586.

[315] Low DH, Sunagar K, Undheim EA, Ali SA, Alagon AC, Ruder T, et al. Dracula's children: molecular evolution of vampire bat venom. J Proteomics. 2013;89:95–111.

[316] Ruder T, Sunagar K, Undheim EA, Ali SA, Wai TC, Low DH, et al. Molecular phylogeny and evolution of the proteins encoded by coleoid (cuttlefish, octopus, and squid) posterior venom glands. J Mol Evol. 2013;76(4):192–204.

[317] Sunagar K, Fry BG, Jackson TN, Casewell NR, Undheim EA, Vidal N, et al. Molecular evolution of vertebrate neurotrophins: co-option of the highly conserved nerve growth factor gene into the advanced snake venom arsenalf. PLoS One. 2013;8(11):81827.

[318] Sunagar K, Johnson WE, O'Brien SJ, Vasconcelos V, Antunes A. Evolution of CRISPs associated with toxicoferan-reptilian venom and mammalian reproduction. Mol Biol Evol. 2012;29(7):1807–1822.

[319] Tian C, Yuan Y, Zhu S. Positively selected sites of scorpion depressant toxins: possible roles in toxin functional divergence. Toxicon. 2008;51(4):555–562.

[320] Zhu S, Bosmans F, Tytgat J. Adaptive evolution of scorpion sodium channel toxins. J Mol Evol. 2004;58(2):145–153.

[321] Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21(5):676–679.

[322] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–3402.

[323] Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, et al. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol. 2013;30(5):1196–1205.

[324] Herráez A. Biomolecules in the computer: Jmol to the rescue. Biochemistry and Molecular Biology Education. 2006;34(4):255–261.

[325] Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, et al. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res. 2011;39(Database issue):392–401.

[326] Rose PW, Bluhm WF, Beran B, Bi C, Dimitropoulos D, Goodsell DS, et al. The RCSB Protein Data Bank: site functionality and bioinformatics use cases. NCI Nature Pathway Interaction Database. 2011;.

[327] Zmasek CM, Eddy SR. ATV: display and manipulation of annotated phylogenetic trees. Bioinformatics. 2001;17(4):383–384.

[328] Maldonado E, Dutheil JY, da Fonseca RR, Vasconcelos V, Antunes A. IMPACT: integrated multiprogram platform for analyses in ConTest. J Hered. 2011;102(3):366–369.

[329] Ghosh S, Matsuoka Y, Asai Y, Hsin KY, Kitano H. Software for systems biology: from tools to integrated platforms. Nat Rev Genet. 2011;12(12):821–832.

[330] Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013;47:97–120.

[331] Yokobori S, Lindsay DJ, Yoshida M, Tsuchiya K, Yamagishi A, Maruyama T, et al. Mitochondrial genome structure and evolution in the living fossil vampire squid, Vampyroteuthis infernalis, and extant cephalopods. Mol Phylogenet Evol. 2007;44(2):898–910.

[332] Ballard JW, Melvin RG. Linking the mitochondrial genotype to the organismal phenotype. Mol Ecol. 2010;19(8):1523–1539.

[333] Marden JH. Nature's inordinate fondness for metabolic enzymes: why metabolic enzyme loci are so frequently targets of selection. Mol Ecol. 2013;22(23):5743–5764.

Part VI

APPENDIX

Appendix A

Supplementary Materials

A.1 Appendices II – Chapter 2

https://goo.gl/eQptuN



A.2 Appendices III – Chapter 3

https://goo.gl/Gvh1fe



A.3 Appendices IV – Chapter 4

https://goo.gl/67jpqG



A.4 Appendices V – Chapter 5

https://goo.gl/I83gNo




U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO