Universidade de Lisboa
Faculdade de Ciências
Departamento de Biologia Animal

Phylogeography of the species *Psammodromus algirus*

Francisco Rente de Pina Martins

Mestrado em Biologia Evolutiva e do Desenvolvimento

2007

Universidade de Lisboa
Faculdade de Ciências
Departamento de Biologia Animal

Phylogeography of the species *Psammodromus algirus*

Francisco Rente de Pina Martins

Dissertação orientada por:
Professor Doutor Octávio Paulo

Mestrado em Biologia Evolutiva e do Desenvolvimento
2007

Index

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Appendix

Acknowledgements / Agradecimentos

Ao meu orientador, o Professor Octávio Paulo, por mais uma vez ter sido um excelente orientador e por todos os conselhos que me deu ao longo deste tempo.

Ao Mário Pulquério, pela amostragem que fez de *Psammodromus algirus*.

À Sofia Seabra, pelas revisões todas que fez a este trabalho.

Ás minhas colegas de laboratório, Vera, Margarida, Joana, Carla, Cândiada e outra vez Sofia, por todo o apoio prestado no trabalho laboratorial de todos os dias.

À Joana Morais e à turma de Filogenética do mestrado em BED 2007/2008 por terem sido fantásticos "Beta Testers" do *Concatenator*.

Ao Pedro e à Ana Rita pelo companheirismo e almoços quase diários.

Aos meus pais e ao meu irmão, pelo suporte familiar.

À Ana, por ser, como eu costumo dizer, a melhor mulher do Universo.

À FCT – "Padrões filogeográficos na bacia do Mediterrâneo" (POCTI/BSE/48365/2002), pelos fundos sem os quais este trabalho não teria sido possível.

Resumo

A presente tese de mestrado é composta por quatro capítulos, sendo o segundo e o terceiro independentes e dedicados a dois temas distintos e o primeiro e o último uma introdução geral e considerações finais respectivamente.

São propostos três objectivos para realizar nesta tese:

1. Inferir uma filogenia da espécie *Psammodromus algirus* com base numa amostragem ao longo de quase toda a sua área de distribuição (excepção feita para Algeria, Tunísia e Sul de França). Esta abordagem providenciará ainda dados básicos para o estudo populacional.
2. Contribuir para o debate científico relativa à filogeografia da espécie *P. algirus*, recorrendo ao melhor dos métodos utilizados em dois estudos anteriores dedicados à mesma problemática e adicionado-lhe uma componente ao nível populacional. Espera-se que estes novos resultados possam fazer avançar a resolução dos conflitos identificados.
3. Através da disponibilização do software *Concatenator*, facilitar a realização de análises filogenéticas, especialmente ao utilizadores que não estejam familiarizados com edição de ficheiros de texto UNICODE, através de uma interface gráfica simples do tipo “apontar e clicar”.

A espécie *Psammodromus algirus* tem sido estudada desde 1973, nessa altura principalmente ao nível da morfologia, e gradualmente a outros níveis, tal como da biologia, distribuição, taxonomia e ecologia. Mais tarde, em são realizados trabalhos com marcadores moleculares em que é incluída esta espécie mais mais recentes, são realizados dois trabalhos sobre a filogeografia de *P. algirus* com recurso a estes marcadores e que apresentam resultados discordantes.

Com vista à resolução deste conflito, foram experimentados diversos genes candidatos dos quais apenas quatro foram utilizados, apesar das diversas tentativas de amplificação e sequenciação efectuadas. Os quatro genes utilizados nas análises realizadas neste estudo são todos mitocondriais, e são o 12s rRNA, 16s rRNA, citocromo *b*, e NAD4.

Os resultados obtidos para os diversos conjuntos de dados analisados são apresentados sob a forma de árvores filogenéticas (apenas uma por conjunto de dados visto que os métodos utilizados apresentaram sempre resultados concordantes): 4 resultantes da análise de cada conjunto de sequências individualmente, 4 resultantes de dados concatenados dois a dois (2 árvores enraizadas e 2 desenraizadas) e 1 resultante da concatenação dos quatro conjuntos de dados.

É então analisado um quinto conjunto de dados, com sequências de apenas um gene (citocromo *b*), mas com um maior conjunto de amostras do que na abordagem filogenética. Com base nestes dados são efectuadas diversas análises, tais como mapas de "redes", gráficos de "mismatch", variância molecular (AMOVA), ou cálculo de diversidades nucleotídica e haplotípica.

Mostra-se então que as árvores apresentadas são de alguma forma contraditórias, e concluí-se que é devido a diferenças na relação do "ingroup" com o "outgroup", que difere de gene para gene. No entanto, é possível distinguir 6 "clades", Ibéria Este (IE), Ibéria Sul (IS), Ibéria Oeste (IW), Marrocos Interior (MI), Marrocos Este (ME) e Marrocos Oeste (MW), que são congruentes em todas as árvores. É ainda demonstrado nesta análise que os grupos IS e IW são bastante próximos e que existe uma maior clivagem entre IE e IS + IW do que entre os 3 "clades" Marroquinos.

Com base nestes resultados, são explicadas as diferenças encontradas entre os trabalhos anteriores sobre o mesmo tema. Uma vez que apenas com estes genes a análise filogenética clássica não permite resolver totalmente as relações entre os diversos grupos considerados recorre-se aos dados da abordagem populacional, principalmente valores de diversidades nucleotídica e haplotípica, para testar as hipóteses de explicação do padrão biogeográfico actual da espécie propostas nos trabalhos anteriores, e ainda um terceira hipótese proposta pela primeira vez neste trabalho. Os dados obtidos, principalmente a partir da abordagem populacional, dão um pouco mais de suporte à hipótese nova apresentada neste trabalho do que ás propostas em outros estudos, sem no entanto permitir colocar um ponto final na questão.

De qualquer das formas, uma vez que a hipótese apontada como mais provável exige dois momentos de atravessamento do mar Mediterrâneo por parte de uma espécie terrestre como é o caso de *P. algirus*, é de notar que esta barreira não será tão "impermeável" como se pensava outrora.

É importante ainda realçar que acrescentar um ou mais genes nucleares à análise, poderá com relativa facilidade alterar os resultados aqui obtidos.

Em relação ao terceiro objectivo desta tese, este é abordado no terceiro capítulo, que é um artigo publicado sobre o software desenvolvido neste âmbito. Este software foi escrito na linguagem “Perl”, hoje em dia bastante comum em bioinformática, recorrendo ao módulo “Perl/Tk” para a implementação de uma interface gráfica simples e elegante. Devido a este interface, é possível com este programa de uma forma simples e rápida efectuar conversões de formatos de ficheiros (FASTA para Nexus para FASTA) tendo em conta requerimentos específicos de outros programas populares em análise filogenética / filogeográfica (Ex. PAUP*, MrBayes, Network, TCS, etc...) e ainda efectuar a concatenação de até cinco matrizes de dados do tipo Nexus, mais uma vez com opções de preparação de ficheiros para outros programas populares neste tipo de análise (apenas PAUP* e MrBayes neste caso). Finalmente é dado um exemplo da utilização deste software em situações comuns neste tipo de análises. O programa encontra-se disponível para “download” em http://cobig2.fc.ul.pt na secção de “Downloads”. Estão disponíveis os ficheiros binários (executáveis) para Windows e o código fonte para Windows e sistemas UNIX (Linux e Mac OS X). A instalação em ambiente Windows e Linux resume-se à simples descompactação dos ficheiros e colocação numa pasta. Para Mac OS X este procedimento é mais complexo, mas está também disponível na mesma página um manual de instalação para este sistema.

Como considerações finais sobre os resultados obtidos neste trabalho, pode afirmar-se que:

Em relação ao estudo filogenético e filogeográfico são comentados os resultados pouco conclusivos, mas no entanto inovadores em relação aos trabalhos já existentes, não só pelo tipo de abordagem em termos de métodos, mas também pela distribuição dos locais amostrado como são também ainda expressas as expectativas futuras, em relação a outras hipóteses e aumento de amostragem e genes analisados (em relação à inclusão de um ou mais genes nucleares).

Relativamente à parte sobre software, biologia evolutiva e bioinformática é expressa a preocupação com a crescente dificuldade ao nível da análise de dados em biologia evolutiva, onde as ferramentas existentes hoje em dia não são “amigas do utilizador”. É necessário inverter esta situação, visto que hoje em dia é necessário saber diversas linguagens de (quase) programação para se poder efectuar uma análises filogenéticas

completas (principalmente com as mais recentes ferramentas de aferição de relógios moleculares). Porque biólogos não são programadores, espera-se que esta pequena aplicação desenvolvida nesta tese ajude a definir um padrão em termos de facilidade de utilização do software nesta área.

Palavras-chave: *Psammodromus algirus*, Filogenética, Filogeografia, mtDNA, Concatenação, Perl.

Abstract

In this work phylogenetic and phylogeoraphic analyses are conducted on the species *Psammodromus algirus* based on samples from the Iberian Peninsula and Morocco with the goal of resolving the controversy relative to this subject that exists in the literature. Four genes were used (12s, 16s, cytochrome *b* and NAD4) which were analysed one by one and concatenated in different combinations. The results differed from dataset to dataset if the trees were rooted; in the case of unrooted trees, the results were relatively congruent. The phylogenetic approach was thus not enough to resolve the addressed issues and reach satisfying conclusions regarding the species' present biogeographic patterns. In order to address this issue a population approach was made with a larger number of samples. With the combined results from both approaches it was possible to propose an explanation for this species' past migrations which is different from the ones presented in former publications.

Furthermore, software was developed in the context of this thesis, which is very useful in phylogenetic/phylogeographic analyses. The program's purpose is to make data matrix conversions (FASTA to Nexus to FASTA with several program requirements in mind) and concatenation (of up to five Nexus data files) an easy task, that anyone with minimum computer skills can easily use it. The chapter about this software is the content of a paper which is in press in at the time of writing this thesis.

Finally, comments are weaved on the outcome of this thesis, and brief remarks are made on the future of the two main components of this work.

Keywords: *Psammodromus algirus*, Phylogenetics, Phylogeography, mtDNA, Concatenated Data, Perl.

General Introduction

The Studied Species

A Brief Review

Psammodromus algirus was first described by Linnaeus in 1758 and later placed in the genus *Psammodromus* by Fitzinger in 1826 (Arnold, 1973).

The species *P. algirus* is exhaustively described in Arnold (1973), both in morphology and in internal function (skeletal and organ). This work considered 50 other reptile species, and finally suggested a full reclassification of the palaearctic lacertids, based on osteological and hemipenial characters, chromosomes, external morphology, coloring and distribution. Although this work was not centered on *P. algirus*, it is the first publishing where the species was throughly studied. It is also the first time that the genus *Psammodromus* was considered close to the genus Gallotia.

Between 1973 and 1987 several works were published including *P. algirus*. These papers assessed data like body temperatures (Busack, 1978) and ecological relations (Busack & Jaksic, 1982; Arnold, 1987).

The first work where molecular markers were applied to *P. algirus* is Busack & Maxson, (1987) where the taxonomic relations of lacertids from Arnold (1973) were confirmed using quantitative micro-complement fixation analysis of serum albumin.

No other studies using molecular markers on *P. algirus* were released until 1998. However, during this time, at least ten other works were published regarding the ecology (Carrascal & Diaz, 1989), biology (Veiga *et al.*, 1997), life history traits (Bauwens *et al.*, 1995) and thermal biology of the species (Belliure & Carrascal, 1996).

After this date, more then 40 other works were published including *P. algirus*, making this species well studied, on ecological, biological and behavioural levels.

Of particular importance to this thesis are the last few works, regarding the species' phylogeography, namely Carranza *et al.* (2006), Busack & Lawson (2006) and Busack *et al.* (2006). These works not only used molecular markers similar to what is done in this thesis, but they also reached different results. These will be compared later in this chapter.

In Busack *et al.* (2006) *P. algirus* was divided into two different species, *P. manuelae* and

P. jeanneae. However, since the species separation morphological characters are not very clear, and during the sampling for the present thesis all the individuals were considered the same species, only *P. algirus* will be considered here.

In Carranza *et al.* (2006) and Busack & Lawson (2006), the species was subjected to through phylogeographic and phylogenetic analysis, using several different approaches.

Distribution

This species is a typical inhabitant of very dense bushy habitats, although it sometimes occurs in more open areas. Often found in open or degraded woodland, in undergrowth in pine or eucalyptus forest, and among very dense spiny shrubs, dwarf oak, heather, gorse, brambles, and even prickly pear (Arnold, 2002). This sort of habitat (and thus, *P. algirus*) can be found throwout almost all of the Iberian Peninsula (except in the north Atlantic coast), in the southwest of France, and in the north of Tunisia and Algeria and in the north and center of Morocco (Miras *et al.*, 2006). The distribution, according to (Miras *et al.*, 2006) can be visualized in Figure 1.1. The species occurs mainly at altitudes below 400m, however, it can be found at up to 1600m in Portugal (Malkmus, 2004) and up to 2600m in the High Atlas in Morocco (Barbadillo *et al.*, 1999).

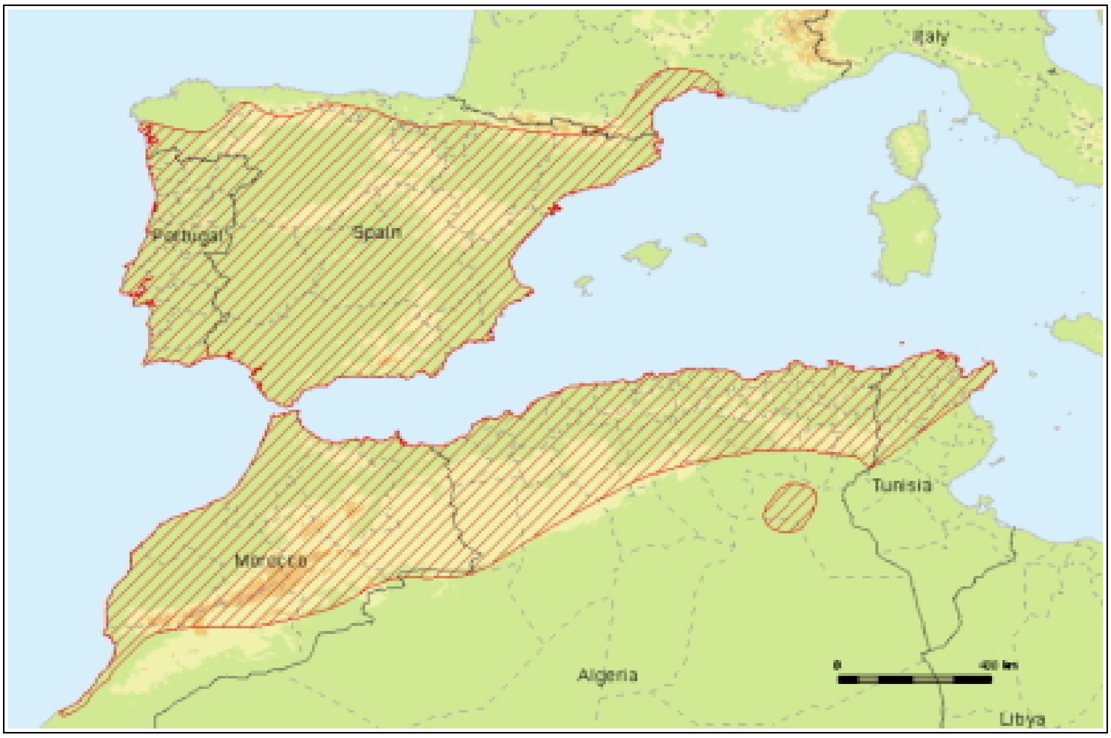


Figure 1.1. A map of the Iberian Peninsula and part the north of Africa. The areas dashed in red mark the distribution of *P. algirus*. Modified from Miras *et. al* (2006)

Biology

P. algirus adults usually measure up to 7.5 cm from snout to vent though sometimes they may reach 9 cm; their tail normally measures two or three times the body length. Relative size of males and females is variable from place to place. These thick-necked lizards have a thin, rather stiff, tail and no collar. The scales on their back and flanks are large, flat and pointed with a prominent keel, strongly overlapped, just like the belly scales. The species' color is fairly constant: usually metallic brownish with two conspicuous white or yellowish stripes on each side, the upper ones bordered above by dark dorsolateral stripes; the flanks are often dark and there may be vague dark stripes on back. Some animals, especially old males are almost uniform. Males often have one or more blue spots in the shoulder region. The underparts are slightly iridescent-whitish or even tinged green. The breeding males have orange, red or yellow sides of head and throat; the flanks and chest may also be yellow; the head coloring may occasionally show up on females. Infants look like adults, but the pale stripes are less obvious and tail is often of a more orange tone (Arnold, 2002).

The individuals of this species spend most of time around the base of plants hunting in leaf litter etc. but may climb in bushes and sometimes makes excursions across more open areas. It is the most abundant lizard species in many parts of Spain and Portugal but is not easily spotted, due to good camouflage that blend them with the landscape. They may sometimes squeak, especially when picked up but also at other times. Tends to be replaced by Spiny-footed Lizard (*Acanthodactylus erythrurus*) and Spanish Psammodromus (*Psammodromus hispanicus*) in more open areas (Arnold, 2002).

Their diet is composed mostly of arthropods, occasionally complemented with other small lizards and some vegetation (Arnold, 2002).

During mating, the male will hold the female's neck in his jaws which is different from most lacertids which hold the body flank. The eggs are laid from 2 to 4 weeks after mating. Females lay up to 2-3 clutches of 2-11 eggs every year. These range 12 – 14 mm x 7 – 9 mm in dimension, and hatch in 5 – 6 weeks producing infants of about about 2.5 – 3 cm from snout to vent. These mature in 1 or 2 years and may sometimes live up to 7 years.

These individuals have a pocket in the skin on each side of neck where chiggers (red larvae of trombiculid mites) often accumulate. These mites feed on the lizard's body fluids for some weeks and often attach around the eyes or ears, reducing efficiency of these organs, damaging delicate tissues. The pockets are believed to reduce this problem by luring mites into areas where they do less harm (Arnold, 2002).

The closest relatives

The genus includes four species: *P. algirus* (African and European), *P. blanci*, confined to the east of Morocco (Bons & Geniez, 1996), *P. microdactylus*, endemic to Morocco (Bons & Geniez, 1996) and *P. hispanicus* (Gasc *et al.*, 1997), endemic to the Iberian Peninsula.

There are two known subspecies of *P. algirus*: *P. a. nollii* from the algero-tunisian high plateau and Sahara, and *P. a. doriae* from the islet of Galitone north of Tunisia (Gasc *et al.*, 1997). The present work however does not contemplate any individuals from these subspecies.

Detailed Review on the Genetic Publications of the Species

This section will attempt to comprehensively describe the recent publications involving *P. algirus* and molecular markers.

The first work regarding molecular markers and *P. algirus* was Busack & Maxson (1987), which was essentially a taxonomic study. The used marker is the "quantitative micro-complement fixation analysis of serum albumin evolution", and it is indicated that *P. algirus* is quite separated from *Lacerta lepida* and *Lacerta monticola*, the central species in this study. The main conclusions were that the classification suggested in Arnold, 1973 is adequate and should be adopted.

The next work showed up only eleven years later. Harris *et al.*, 1998 made a review of the 1973 classification using more sophisticated molecular marker techniques and some changes to that classification were proposed, including the inclusion of the genus *Psammodromus* and *Gallotia* in the same group and considering them as a sister taxa to all other lacertids. This phylogenetic analysis used parts of three mitochondrial genes – 12s ribosomal RNA (12s), 16s ribosomal RNA (16s) and Cytochrome *b*.

Later, in Harris *et al.* (2001) a nuclear gene was included in order to complete the analysis made in Harris *et al.* (1998). Once again, the position of the genus *Psammodromus* is assessed as close to *Gallotia* and as a sister taxa to the remaining lacertids. Although it did not bring any new insights into the lacertid group resolution, this work resolved the relationships among other reptile groups.

After a five year period with no publications considering molecular approaches on *P. algirus*, in 2006 three publications arose: Carranza *et al.* (2006), Busack & Lawson (2006) and Busack *et al.* (2006).

Using partial sequences of the 12s, 16s and cytochrome *b*, Carranza *et al.* (2006) infer a phylogeography of the species based on two concatenated phylogenetic trees: one was constructed with the 3 mentioned gene sequences, using 10 samples of the genus *Gallotia*, 21 individuals of the species *P. algirus* and 3 more individuals of the genus *Psammodromus*, but not of the species *P. algirus* and another based on the 12s and 16s sequences using 66 individuals of *P. algirus* and 2 individuals of *Gallotia caesaris*.

The authors infer from these results that *P. algirus* existed initially in the east of Iberia and then separated into two clades (east and west) about 3.6 Million years ago. Later, this newly formed western clade separated into two sections – the Iberian section and the Maghreb section (in the north of Africa) at about 1.9 Million years ago. Finally, less then 1 Million years ago, the western Iberian clade divided again into a south and north clade.

Busack & Lawson (2006) use mitochondrial DNA (mtDNA) (6 complete sequences of cytochrome *b*, 11 complete NADH dehydrogenase subunit 2 (ND2) sequences and 13 partial NADH dehydrogenase subunit 4 (ND4) sequences) and allozymes to estimate a phylogeny of *P. algirus*. The mtDNA sequences were concatenated and outputted a phylogenetic tree that was considered for the analysis; they were further used to perform a Mantel test (only in Morocco, since the authors did not have enough Spanish samples to preform this test). The allozymes data was used essentially to perform Mantel tests as with the mtDNA.

After these results, the authors infer that *P. algirus* must have invaded the Iberian Peninsula from the north of Africa because the results indicate that the genetic differences per distance unit in Morocco are larger than in Spain (where the populations later subdivided into two isolated clades). They further estimate, based on other works regarding the genus *Podarcis*, that the last time that the European and African populations must have been in complete reproductive contact should have been at about 2.98 – 3.23 Million years ago. Likewise, the last time that the southern Spain populations must have been in complete reproductive contact must have been at about 1.40 – 1.54 Million years ago.

In Busack *et al.* (2006), the the species *P. algirus* is divided into 3 different species – *P. algirus* in the north of Africa, *P. manuelae* in the north of the Iberian Peninsula and *P. jeanneae* in the south of the Iberian Peninsula. These conclusions were drawn based on the results of Busack & Lawson (2006) and on morphological aspects exhaustively described in the paper.

The Selected Genes

Although the initial plan for the phylogeographic study included a nuclear gene, budget and time constraints impeded this plan to step forward. The study was thus completed using only mitochondrial DNA.

However, several attempts were made to amplify nuclear genes for this species. After a bibliographic search, several candidate genes were proposed and the following were thoroughly experimented:

- β fibrinogene intron 7 (FIB7) – Fibrinogene is a protein involved in blood clotting whereby clot formation involves the conversion of soluble fibrinogene into insoluble fibrin clot (Doolittle, 1984). This was the most promising nuclear gene found in the bibliography since it had been described as a very variable gene (Prychitko & Moore, 1997) and was successfully amplified in species close to *P. algirus* such as *Lacerta lepida* or *Lacerta schreiberi* (Godinho *et al.*, 2006). The primers for this sequence were described originally in Prychitko & Moore (1997). Despite all the efforts, however, no amplification of this intron was produced in the laboratory.
- Tropomyosin α subunit (TROP) – Tropomyosin is a myofibrillar protein involved in the regulation of contraction and relaxation of muscle fiber (Cummins & Perry, 1973). The primers for the amplification of the α subunit of this gene are described in Friesen *et al.* (1999). However, just like with the FIB7 gene, after many attempts no amplification of this gene could be obtained in the laboratory.
- C-MOS – This is a protooncogene that codes for the protein involved in the arrest of oocite maturation (Whiting *et al.*, 2003). It is not described as a very variable gene (Whiting *et al.*, 2003), however, it is adequate for phylogenetic analysis in species close to *P. algirus* (Harris *et al.*, 1998). The primers for this sequence were published in Saint *et al.* (1998). By the time erratic amplification for this gene in Iberian individuals were accomplished (no African individual was successfully amplified) the deadline for this work was close to due, so there was no opportunity to optimize the PCR protocols, and thus, this analysis was aborted.

- BOV-B lines – This gene has a quiet history in bibliography and was suggested for the first time in the context of phylogenetic analysis in Kordis & Gubensek (1997). It was later used in Piskurek *et al.* (2006) for a phylogenetic analysis in vipers. In the present thesis, this gene was sequenced initially for six individuals. The mean pairwise difference among these initially sampled individuals was of about 0.5%, and 3 haplotypes were found. Afterwards, the sequencing of twenty more samples was carried on. However, in these samples all of the variation ended (no new haplotypes and 18 of the 20 new samples were all the same haplotype), making this gene unusable for phylogenetic analysis in this species. Despite this lack of variability in *P. algirus*, it was successfully used to distinguish this species from *Lacerta lepida* and *Lacerta agilis*.

Finally, the used mitochondrial genes were:

- 12s rRNA – This is the gene for the small subunit ribosomal RNA in mitochondria (Palumbi, 2000). It is moderately conserved in squamates (Whiting *et al.*, 2003) (Palumbi, 2000), considering the overall mutation rate of more variable mitochondrial genes. It is thus, good to infer basal relations among closely related individuals in a phylogenetic tree. It is however, not so good to infer closer relationships among these individuals. Should the considered individuals belong to more separate groups, this gene will also resolve closer relations. The primers used to amplify this gene were described in Palumbi (2000).
- 16s rRNA – This is the gene for the large subunit ribosomal RNA in mtDNA (Palumbi, 2000). It is even more conserved than the 12s (Whiting *et al.*, 2003) (Palumbi, 2000), but despite this fact, its usage in phylogenetics is similar to that of 12s, due to being a larger fragment. The primers for this gene were described in Palumbi (2000).
- Cyt *b* – Cytochrome *b* is a protein in the electron transport chain (Palumbi, 2000). It is the only functional monomer in the mtDNA (Palumbi, 2000). It is also one of the most variables if not the most variable among lacertids in the mitochondrial genome (Whiting *et al.*, 2003). Should the individuals in a phylogenetic analysis be closely related, this gene will resolve most of the tree branches, both in the basal level and

in the end level. However, if the considered individuals are more diverged, polyctomies are likely to cause analysis problems. The primers used to amplify this gene are described in Paulo *et al.* (2001).

- NADH dehydrogenase subunit 4 (NAD4) – This gene is not as variable as the cytochrome *b* (this study), but is nevertheless a very variable gene in the Lacertid mitochondrial genome panorama. The phylogenetic analyses it is suited for are similar to those of the cytochrome *b*. This gene has been used in phylogenetic analysis in this species before (Busack & Lawson, 2006), but the primers used in this thesis were designed from gene bank sequences since the primers provided in Busack & Lawson (2006) did not amplify more than a fragment of 100 bp.

This set of genes was analysed both individually and concatenated with the others. This provided not only the individual gene “histories” but also a “big picture” approach of all of them together.

Bioinformatics and Software Development

Bioinformatics is an emerging area in biological sciences. More and more the use of computer science is being applied to other scientific areas and biology is no exception.

With the increase of DNA sequencing, bioinformatics tools were required not only to analyse these outputs, but also to assemble them in large databases and to group them in analysis specific clusters.

Phylogenetics (the main biological component of this thesis) is one of the main branches of bioinformatics and more and more often enters the laboratories and computers of molecular biologists.

It is in this context that software needs to be developed with the less informatics-wise informed user in mind. Since software is developed by computer science experts, the resulting programs are usually not so simple to use by biologists with scarce computer training. That is when command line driven programs stop fulfilling their objectives – users simply do not understand how to use them, because of their missing formation in computer science. To computer experts, this is completely trivial, but for the common user this might well be a potential nightmare and makes data analysis much more error prone.

That is the main reason why *Concatenator* was created: in order to simplify the task of file format transfers, and the managing of several of these into a single file. For most users the DNA sequences saves as a FASTA files are not a UNICODE text file with sequence names identifiers, followed by several lines of 60 characters ending in a “newline” character. For most users, DNA data files are sets of purines and pyrimidines identified by a given name, that can be aligned by mathematic algorithm at a click of a button. These regular users may well be excellent biologists, but they are not computer experts. *Concatenator* is for this sort of scientist. For those who dedicate to their area in exclusive. It is the author's sincere hope that the work and effort put into developing this tool will save great deals of time and reduce errors in future phylogenetic analysis, for all the molecular biologists who should find it useful.

Objectives

The objectives of this thesis are:

First: to outline a phylogeny of the species *P. algirus* based on a very wide sampling throughout the most of its distribution (only Algeria, Tunisia and southern France data are missing from this study). This approach will also provide the basic data for the population study.

Second: to contribute to the scientific debate about the phylogeography of the species *P. algirus*, using the best approaches considered in the two similar previous studies, Carranza *et al.* (2006) and Busack & Lawson (2006), adding them a new populational approach. This new data will contribute to the resolution of the conflicting results from other authors.

Third: with the release of the software *Concatenator*, to improve the usability and user friendliness of phylogenetic analysis, especially to users who are not familiarized with the UNICODE text file formats and who will benefit from having to skip this learning step in order to make their phylogenetic analysis, using a simple point and click graphical interface.

References

Arnold EN. 1973. Relationships of the Palaearctic lizards assigned to the genera Lacerta, Algyroides and Psammodromus (Reptilia: Lacertidae) . *Bulletin of the British Museum (Natural History)*. 25: pp. 289-366.

Arnold EN. 1987. Resource partitioning among lacertid lizards in southern Europe . *Journal of Zoology, London*. 1: pp. 739-782.

Arnold EN. 2002. *Reptiles and Amphibians of Europe*. Princeton University Press, Princeton, NJ. 288 pp.

Barbadillo L, Lacomba J, Pérez-Mellado V, Sancho V, López-Jurado L. 1999. Reptiles. *Anfibios y Reptiles de la Península Ibérica, Baleares y Canarias.* 3. pp. 306-308.

Bauwens D, Garland T, Castilla A and Vandamme R. 1995. Evolution of Sprint Speed in Lacertid Lizards - Morphological, Physiological, and Behavioral Covariation. *Evolution*. 49: pp. 848-863.

Belliure J and Carrascal L. 1996. Covariation of thermal biology and foraging mode in two Mediterranean lacertid lizards. *Ecology*. 77: pp. 1163-1173.

Bons J, Geniez P. 1996. Psammodromus algirus. *Amphibians & Reptiles of Morocco (Including western Sahara). Biogeographical Atlas*. pp.140-142.

Busack S. 1978. Body temperatures and live weights of five Spanish amphibians and reptiles. *Journal of Herpetology*. 12: pp. 256-258.

Busack S and Jaksic F. 1982. Ecological and historical correlates of Iberian herpetofaunal diversity: an analysis at regional and local levels. *Journal of Biogeography*. 9: pp. 289-302.

Busack S and Lawson R. 2006. Historical biogeography, mitochondrial DNA, and allozymes of Psammodromus algirus (Lacertidae): a preliminary hypothesis. *Amphibia-Reptilia*. 27: pp. 181-193.

Busack S and Maxson L. 1987. Molecular relationships among Iberian, Moroccan, and South African lacertid lizards (Reptilia: Lacertidae). *Amphibia-Reptilia*. 8: pp. 383-392.

Busack S, Salvador A and Lawson R. 2006. Two new species in the genus

Psammodromus (Reptilia : lacertidae) from the Iberian peninsula. *Annals of Carnegie Museum*. 75: pp. 1-10.

Carranza S, Harris D, Arnold E, Batista V and de la Vega J. 2006. Phylogeography of the lacertid lizard, Psammodromus algirus, in Iberia and across the Strait of Gibraltar. *Journal of Biogeography*. 33: pp. 1279-1288.

Carrascal L and Diaz J. 1989. Thermal ecology and spatio-temporal distribution of the mediterranean lizard psammodromus-algirus. *Holarctic Ecology*. 12: pp. 137-143.

Cummins P and Perry S. 1973. Subunits and biological-activity of polymorphic forms of tropomyosin. *Biochemical Journal*. 133: p. 765-777.

Doolittle R. 1984. Fibrinogen and fibrin. *Annual Review of Biochemistry*. 53: pp. 195-229.

Friesen V, Congdon B, Kidd M and Birt T. 1999. Polymerase chain reaction (PCR) primers for the amplification of five nuclear introns in vertebrates. *Molecular Ecology*. 8: pp. 2147-2149.

Gasc J, Cabela A, Crnobrnja-Isailovic J, Dolmen D, Grossenbacher K, Haffner P, Lescure J, Martens H, Martínez J, Maurin H, Oliveira M, Sofianidou T, Veith M, Zuiderwijk A. 1997. Reptilia. *Atlas of Amphibians and Reptiles in Europe.* pp. 302-303.

Godinho R, Mendonca B, Crespo E and Ferrand N. 2006. Genealogy of the nuclear beta-fibrinogen locus in a highly structured lizard species: comparison with mtDNA and evidence for intragenic recombination in the hybrid zone. *Heredity*. 96: pp. 454-463.

Harris D, Arnold E and Thomas R. 1998. Relationships of lacertid lizards (Reptilia : Lacertidae) estimated from mitochondrial DNA sequences and morphology. *Proceedings of the Royal Society of London Series B-Biological Sciences*. 265: pp. 1939-1948.

Harris D, Marshall J and Crandall K. 2001. Squamate relationships based on C-mos nuclear DNA sequences: increased taxon sampling improves bootstrap support. *Amphibia-Eeptilia*. 22: p. 242.

Kordis D and Gubensek F. 1997. Bov-B long interspersed repeated DNA (LINE) sequences are present in Vipera ammodytes phospholipase A(2) genes and in genomes of Viperidae snakes. *European Journal of Biochemistry*. 246: pp. 772-779.

Malkmus R. 2004. Reptiles. *Amphibians and Reptiles of Portugal, Madeira and the*

Azores-Archipelago. 3. pp. 287-290.

Miras J, Cheylan M, Nouira M, Joger U, Sá-Sousa P and Pérez-Mellado V. 2006 Psammodromus algirus. UICN 2007. 2007 IUCN Red List of Threatened Species. <www.iucnredlist.org>

Palumbi, S. 2000. Nucleic Acids II: The Polymerase Chain Reaction. *Molecular Systematics*. 7. pp. 205-247.

Paulo O, Dias C, Bruford M, Jordan W and Nichols R. 2001. The persistence of Pliocene populations through the Pleistocene climatic cycles: evidence from the phylogeography of an Iberian lizard. *Proceedings of the Royal Society of London Series b-Biological Sciences*. 268: pp. 1625-1630.

Piskurek O, Austin C and Okada N. 2006. Sauria SINEs: Novel short interspersed retroposable elements that are widespread in reptile genomes. *Journal of Molecular Evolution*. 62: pp. 630-644.

Prychitko T and Moore W. 1997. The utility of DNA sequences of an intron from the beta-fibrinogen gene in phylogenetic analysis of woodpeckers (Aves: Picidae). *Molecular Phylogenetics and Evolution*. 8: pp. 193-204.

Saint K, Austin C, Donnellan S and Hutchinson M. 1998. C-mos, a nuclear marker useful for squamate phylogenetic analysis. *Molecular Phylogenetics and Evolution*. 10: pp. 259-263.

Veiga J, Salvador A, Martín J, López P. 1997. Testosterone stress does not increase asymmetry of a hormonally mediated sexual ornament in a lizard. *Behavioral Ecology Sociobiology*. 41: pp. 171-176.

Whiting A, Bauer A and Sites J. 2003. Phylogenetic relationships and limb loss in sub-Saharan African scincine lizards (Squamata : Scincidae). *Molecular Phylogenetics and Evolution*. 29: pp. 582-598.

A new Perspective on the Evolutionary History of *Psammodromus algirus*

A new Perspective on the Evolutionary History of *Psammodromus algirus*

Abstract

A phylogeny of the species *Psammodromus algirus* was performed using four mitochondrial genes (12s rRNA, 16s rRNA, cytochrome *b* and NAD4). The phylogenetic analyses of the datasets one by one and on different combinations were not concordant on the rooted trees approach, but were concordant on the unrooted trees approach. These trees, were however, not able to fully resolve the existing controversy about the ancestral group. A population study was also used to address this issue, using more samples but with only one gene. Network analysis combined with other parameters enabled the testing of several hypotheses proposed by other authors for the species' phylogeography. Neither of these models was in full accordance with the obtained data and an additional hypotheses that is in much more concordance with the obtained data is proposed to explain today's *P. algirus* biogeographic patterns.

Keywords: Phylogeography; mtDNA; Lacertidae; Concatenated sequence analyses; west Mediterranean species.

Introduction

Understanding the pattern

This work is about the phylogenetic and phylogeographic patterns of the species *Psammodromus algirus*. The distribution of this species in Africa ranges from the western Morocco coast to the eastern coast of Tunisia and in Europe it occupies almost all of the Iberian Peninsula (it is only absent from the north Atlantic coast) and part of the south of France (Miras *et al.*, 2006). *P. algirus* is in this aspect, a typical west Mediterranean species (Arnold, 2002).

In the traditional view, today's biogeographic patterns for west Mediterranean species have been shaped by events of dispersal and vicariance. The Messinian Salinity Crisis (Hsu *et al.*, 1973) is considered to be one of the most important event in the region responsible for the species' distributions and genetic pattern we see today, but after that several other events seem to have shaped the current observed patterns (Harris *et al.*, 2004; Vasconselos *et al.*, 2006).

Despite being one of the most common reptiles found in it's distribution areas, the population structure of P. algirus was not very well known until recently with the works of Busack & Lawson, 2006 and Carranza *et al.* (2006). Until these works it was not even clear whether the colonization of the Iberian Peninsula by the species had been related to the closing of the Strait of Gibraltar.

More recently, several works using DNA markers have shown that some west Mediterranean terrestrial species have crossed the Mediterranean sea after the Messinian Salinity Crisis (Harris *et al.*, 2002; Carranza *et al.*, 2004) and *P. algirus* is thought to be no exception.

It has further been proposed in Busack *et al.* (2006), using both DNA markers and morphological data that the species *P. algirus* should be divided in two new species in the Iberian Peninsula, *P. jeanneae* and *P. manuelae*. We adopt only the designation of *P. algirus* since the sampling was made before this paper was published and the distinction of different species is outside the scope of this research.

The controversy

Recently, two papers that use DNA markers have shown different possibilities for explaining the species' present biogeographic patterns. However, these publications present contradicting results and propose different explanations for the mentioned patterns.

Busack & Lawson (2006) estimated that the last time that the populations of *P. algirus* from the north and south of the Strait of Gibraltar were in full reproductive contact was about 2.98-3.23 Million Years Ago (MYA). These estimates are made based on sequences of the full NADH dehydrogenase subunit 2 (11 samples), partial NADH dehydrogenase subunit 4 (13 samples), partial cytochrome *b* (6 samples) and allozymes data. The mutation rate of *Podarcis erhardi* according to Poulakakis *et al.* (2003) was used to calibrate the used molecular clock. Based on the same data, the authors further advance that the two southern Spanish populations must have been in last reproductive contact about 1.40-1.54 MYA. The direction of the migration is also assumed to have occurred from Morocco to Spain since the species has a more extensive and complex history in Morocco than in Spain.

Carranza *et al.* (2006), using partial sequences of the 12s rRNA, 16s rRNA and cytochrome *b* analysed in two separate datasets (one with the three genes' sequences but with only 21 samples of *P. algirus* and another only with 12s and 16s sequences, but using 66 *P. algirus* samples), suggested that the species is original from an eastern Iberian clade that originated about 3.6 MYA a western Iberian clade which later, about 1.9 MYA divided into an Iberian and a Maghreb sections. These conclusions were drawn based on a molecular clock calibrated by the differentiation between *Gallotia caesaris gomerae* and *Gallotia caesaris caesaris*.

The two scenarios proposed in the mentioned publications are displayed in **Fig. 2.1**.

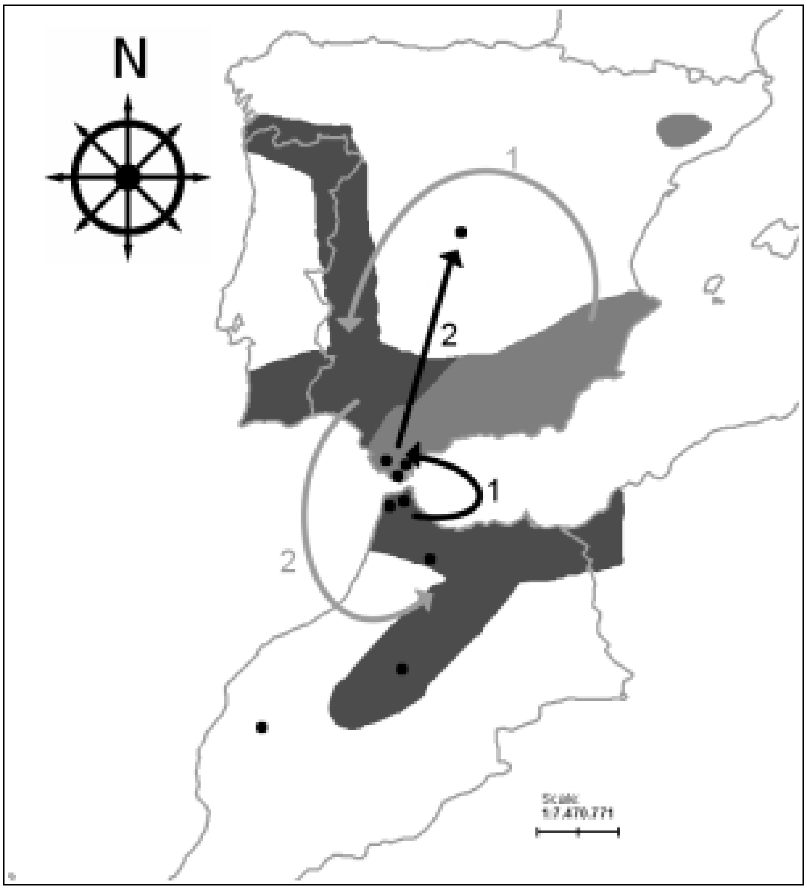


Fig. 2.1: The figure shows the models proposed in the papers of Busack & Lawson (2006) (Black dots and arrows) and Carranza *et al.* (2006) (Different shades of grey)for explaining the phylogeography of *P. algirus*. The dots and shaded zones represent sampling areas and the arrows represent the proposed migrations.

In this work, using four different DNA markers and a wide sampling, across most of the species' range, we try to clarify the phylogeography of *P. algirus* in the context of the western Mediterranean species dynamics.

Materials and Methods

Laboratory work and sample collection

Thirty six samples of *Psammodromus algirus* tails were collected under permit from the Iberian Peninsula and Morocco (**Fig. 2.2**); the tails were clipped, and the animals were immediately released. The outgroups for the phylogenetic analyses were selected from the sequences available in GenBank and are from the species *Lacerta lepida* and *Podarcis muralis*. Both strands of four fragments of mitochondrial DNA genes were amplified using the primers described in **Table 2.1**. DNA was extracted from tails using the *Jet Quick Tissue DNA Extraction Kit* (Genomed) according to the provided protocol. The amplification cycle consisted of 20 seconds at 94ºC, 30 seconds at 46ºC (51ºC for the NAD4 primers) and 42 seconds at 72ºC repeated 30 times. Conditions consisted of 1x PCR Buffer, 0.25mM $MgCl_2$, 0.20mM of each DNTP, 1mM of each primer, 1 unit of TAQ polymerase and approximately 2ng of genomic DNA on a 25µl reaction. PCR products were purified using a *SureClean Kit* (Bioline) with a few minor changes to protocol. Sequencing reactions were processed in the company *Macrogen Inc.* on ABI automatic sequencers. For the population approach, seventeen additional samples were collected, extracted, amplified and sequenced (only for the partial cytochrome *b* gene) for a total sum of 53 used samples.

Table 2.1: Primers for the sequences used in this work. Each pair is referenced and the size of the amplified fragment is indicated. Cyt *b* stands for cytochrome *b*, 12S stands for 12s ribosomal RNA, 16 stands for 16s ribosomal RNA and NAD4 stands for NADH dehydrogenase subunit 4.

Primer Name	Direction	Primer Sequence	Gene	Size	Reference
B1	F	CCA TCC AAC ATC TCA GCA TGA TGA AA	Cyt *b*	700 bp	Paulo *et al.* (2001)
B702	R	AAA TAG GAA GTA TCA CTC TGG TTT			
12 S L	F	TGA CTG CAG AGG GTG ACG GGC GGT GTG T	12S	450 bp	Palumbi (2000)
12 S H	R	CAA ACT GGA TTA GAT ACC CCA CTA T			
16 S L	F	CGC CTG TTT ATC AAA AAC AT	16S	600 bp	Paulo *et al.* (2002)
16 S H	R	CTC CGG TTT GAA CTC AGA TC			
NAD4 F	F	GGATCCATRGTACTAGCCGC	NAD4	650 bp	This study
NAD4 R	R	GTGAATGAGCTGGAAATTAGGC			

Data analyses

The sequences were verified and corrected using SEQUENCHER 4.0.5 (Gene Codes). DNA sequences were initially aligned using CLUSTAL X v1.83 (Thompson *et al.*, 1997) and then corrected manually using BioEdit (Hall, 1999). The two matrix concatenations were preformed using the software *Concatenator* (In press, Chapter 3 of this thesis), as were all the file conversions.

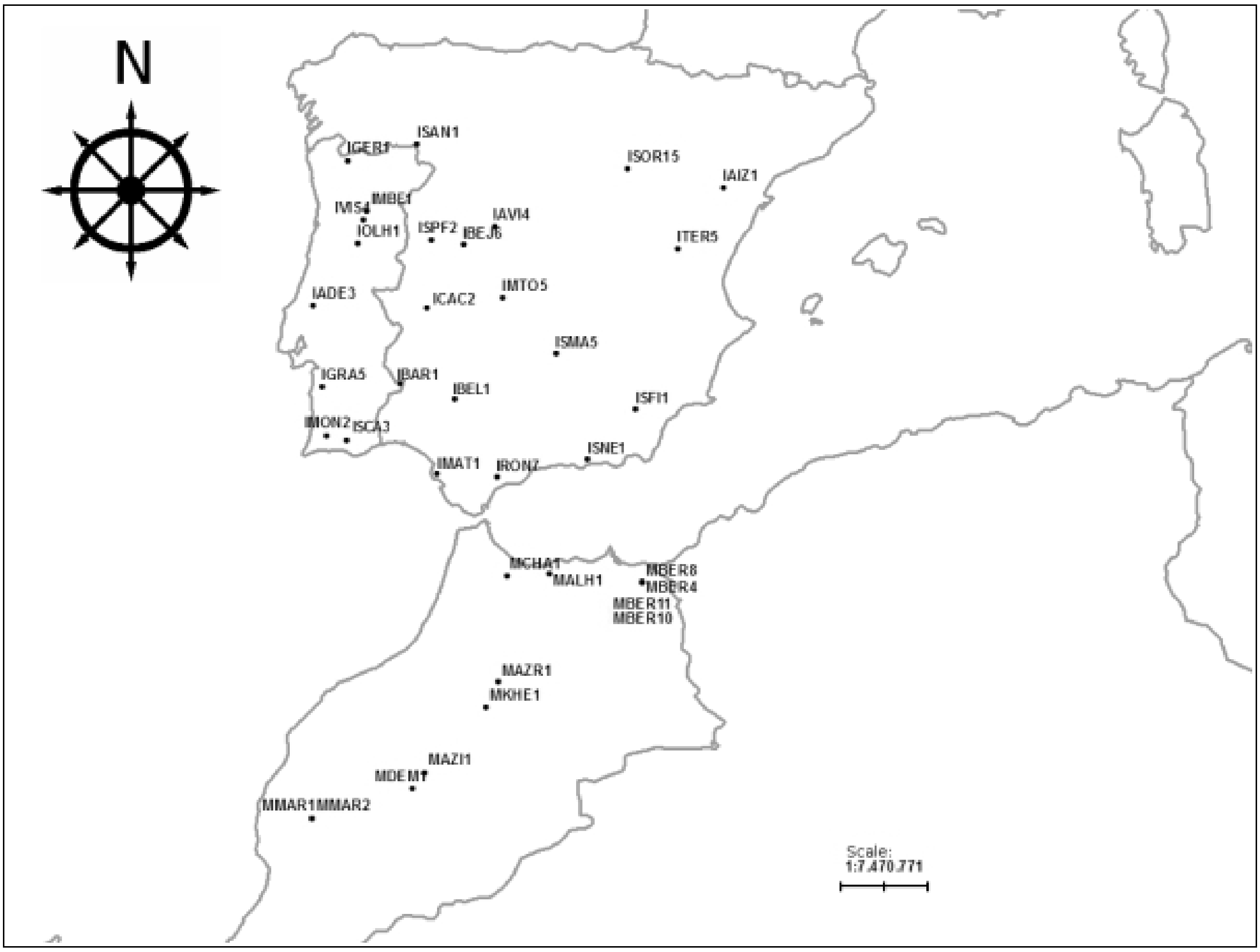


Fig. 2.2: Map of the sampled individuals. All samples starting with the letter "I" are Iberian, while all samples starting with the letter "M" are from Morocco. From north to south: ISAN1(Puebla de Sanabria) , IGER1 (Gerês), ISOR15 (Soria), IAIZ1 (Alcañiz), IMBE1 (Moimenta de Beira), IVIS4 (Viseu), IAVI4 (Ávila), ISPF2 (Sierra de Peña de Francia), IOLH1 (Oliveira do Hospital), IBEJ6 (Bejar), ITER5 (Teruel), IADE3 (Alcanede), IMTO5 (Montes de Toledo), ICAC2 (Cáceres), ISMA5 (Serra de Malcata), IGRA5 (Grândola), IBAR1 (Barrancos), IBEL1 (Belmez), ISFI1 (Sierra de los Fibrilares), IMON2 (Serra de Monchique), ISCA3 (Serra do Caldeirão), ISNE1 (Sierra Nevada), IMAT1 Matalascañhas), IRON7 (Ronda), MCHA1 (Chefchaouen), MALH1(Al-Hocemia), MBER4 (Berkane), MBER8 (Berkane), MBER10 (Berkane), MBER11 (Berkane), MAZR1 (Azrou), MKHE1 (Kenherifa), MAZI1 (Azilial), MDEM1 (Demnate), MMAR1 (Marrakech), MMAR2 (Marrakech). The samples exclusive of the population study are not indicated in the map, but are from roughly the same locations as other samples indicated: IVGU1 (Vale do Guadiana) ~ IBAR1, ISOR7 (Soria) ~ ISOR15, IEFS1 (Embalse de la Fuensanta) ~ ISOR15, IYES1 (Yeste) ~ ISOR15, ILGR1 (El Granado) ~ IMAT1. The Moroccan samples for the population study are from the same locations displayed in the map: MKHE2, MKHE3, MKHE4; MAZR2, MAZR3; MCHA3, MCHA4; MALH2, MALH3, MALH4, MALH5, MALH6.

Phylogenetic analyses were performed using PAUP* 4.0b10 (Swofford, 1993) and MrBayes 3.1.2 (Ronquist & Huelsenbeck, 2003). Modeltest 3.7 software (Posada & Crandall, 1998) associated with PAUP* 4.0 was used for choosing the most plausible evolutionary model for the different data sets according to Akaike Information Criterion (AIC). Both the Maximum Likelihood (ML) and the Neighbour-Joining (NJ) analysis were based on this model. MrModeltest 2.2 (Nylander, 2004) software was used in association with PAUP* 4.0b10 for selecting the most plausible evolutionary model for the different individual data sets (again, according to AIC); the concatenated data sets used separate evolutionary models with unlinked topology and unlinked parameters for the nucleotide substitution models across partitions. Bootstrapping consisted of 1000 replicates on every dataset. ML trees were obtained using 10 random sequence addition replicates. Parsimony analyses considered gaps as a fifth state. Bayesian analysis was conducted using an MCMC algorithm with 1.5 x 10^6 generations. The “burn in” was determined according to the plot of the average standard deviation of split frequencies; this value ranged from 500 to 1000, depending on the dataset. Two closely related species were used as outgroups on all analysis, *Lacerta lepida* and *Podarcis muralis* (data downloaded from GenBank). The incongruence length difference test (ILD), was implemented in PAUP* with all invariant characters removed. All datasets were analysed individually and concatenated with each other in every possible combination. However, only a few are mentioned in this work as many of these combinations did not yield any informative result.

The phylogenetic trees were drawn with the software *TreeView* (RodPage Software). They were further improved for easier visualization using a Perl script to scale the data and *The Gimp 2.4.0-rc3* (Spencer Kimball, Peter Mattis and the GIMP Development Team) to treat the image files. Branches with values of posterior probability or bootstrap below 0.50 or 50.0 respectively are collapsed, displaying polictomies.

The population analysis was conducted using the software Network 4.2.0.1 (Bandelt *et al.*, 1999), using a median joining approach, and the software Arlequin 3.1.1 for AMOVA analyses (Excoffier *et al.*, 1992) and to obtain F_{ST} values (100000 replicates on both cases). The division of individuals into samples for the AMOVA analyses were made according to the clades formed in the phylogenetic analysis divided in two “units” according to geographic locations. The groupings were made according to the results from the network analyses. The mismatch analysis and the calculation of nucleotidic diversity (π) and haplotipic diversity (H) were performed with DnaSP (Rozas *et al.*, 2003).

Results

Phylogenetic analysis:

The shown trees are the results of the combined 12s and cytochrome *b*, 16s and NAD4, and the concatenation of all four genes (plus the individual datasets). The summary of the tree data information is displayed in **Table 2.2**.

Table 2.2: Variability and phylogenetic model details for each gene and combination analysed: fragment size in base pairs, numbers of variable and parsimony informative sites, selected evolutionary model, shape parameter of the gamma distribution (Γ), proportion of invariable sites (I), individual substitution rates and number and length of maximum parsimony trees.

Fragment	Size (bp)	Var. Sites	Pars. Infrom. Sites	Evo. Model	Γ	I	A-C	G-A	A-T	C-G	C-T	Nº MP Trees (lenght)
12s rRNA	462	79	49	GTR + I	Equal	0,7268	16,39	78,48	26,08	1,00E-005	200,18	51 (107)
16s rRNA	547	117	66	GTR + I + G	0,6934	0,5518	05-01-00	16,65	10,21	1,00E-005	68,14	62 (149)
Cytochrome *b*	659	249	156	GTR + I + G	2,7656	0,5782	02-01-00	84,32	3,61	1,00E-005	45,03	18 (428)
NAD4	617	224	132	GTR + I + G	1,7222	0,5644	52218,9	1,60E+006	74438,80	1,00E-005	581512	38 (363)
12s + Cyt *b*	1121	328	205	TVM + I + G	1,1472	0,6120	07-01-00	94,49	8,35	1,00E-005	94,49	27 (566)
16s + NAD4	1164	341	198	TVM + I + G	0,4654	0,4369	06-01-00	68,58	10,64	1,00E-005	68,58	4 (539)
12s + 16s + Cytb + NAD4	2249	669	403	TVM + I + G	0,5766	0,5331	07-01-00	86,51	9,40	1,00E-005	86,51	3 (1133)

For each dataset, the tree topology was similar on all methods and thus, only one tree is displayed. These trees are shown in **Figs. 2.4 to 2.10**. The cytochrome *b*, NAD4 and concatenated trees have a better resolution (due to higher variability) than the 12s and 16s genes trees. These inferred phylogenetics patterns allow the division of the studied samples in 6 different geographic clades (**Fig. 2.3**): Iberia East (IE), Iberia South (IS), Iberia West (IW), Morocco Interior (MI), Morocco East (ME) and Morocco West (MW).

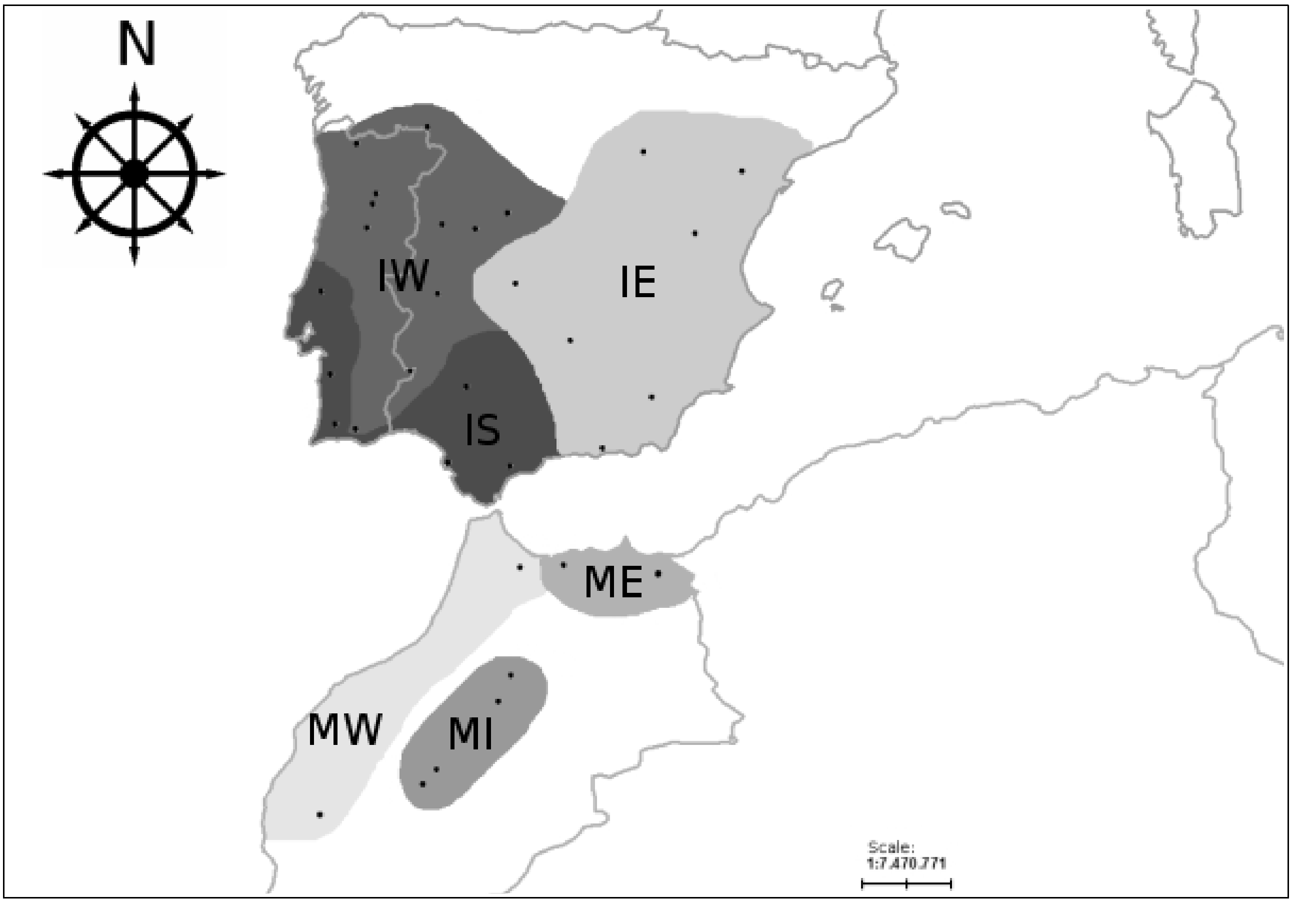


Fig. 2.3: Map of the zones sampled for this work divided by clades according to the phylogenetic analyses. IE, IS, IW, MI, ME and MW stand for Iberia East, Iberia South, Iberia West, Morocco Interior, Morocco East and Morocco West respectively. The samples in each group are: **IE** – IAIZ1, ISOR15, ITER5, ISFI1, ISMA1, IMTO5, ISNE1; **IS** – IRON7, IMAT1, IBEL1, IMON2, IGRA5, IADE3; **IW** – ISCA3, IBAR1, ICAC2, IAVI4, IBEJ6, ISFP2, IOLH1, IVIS4, IMBE1, IGER1, ISAN1; **MI** – MDEM1, MAZI1, MKHE1, MAZR1; **ME** – MALH1, MBER10, MBER8, MBER11, MBER4; **MW** – MCHA1, MMAR1, MMAR2.

The topology of the trees differs among datasets, even tough the formed clades/groups are always very similar. The main differences are on the placement of the basal group and on the groups that actually form in each dataset.

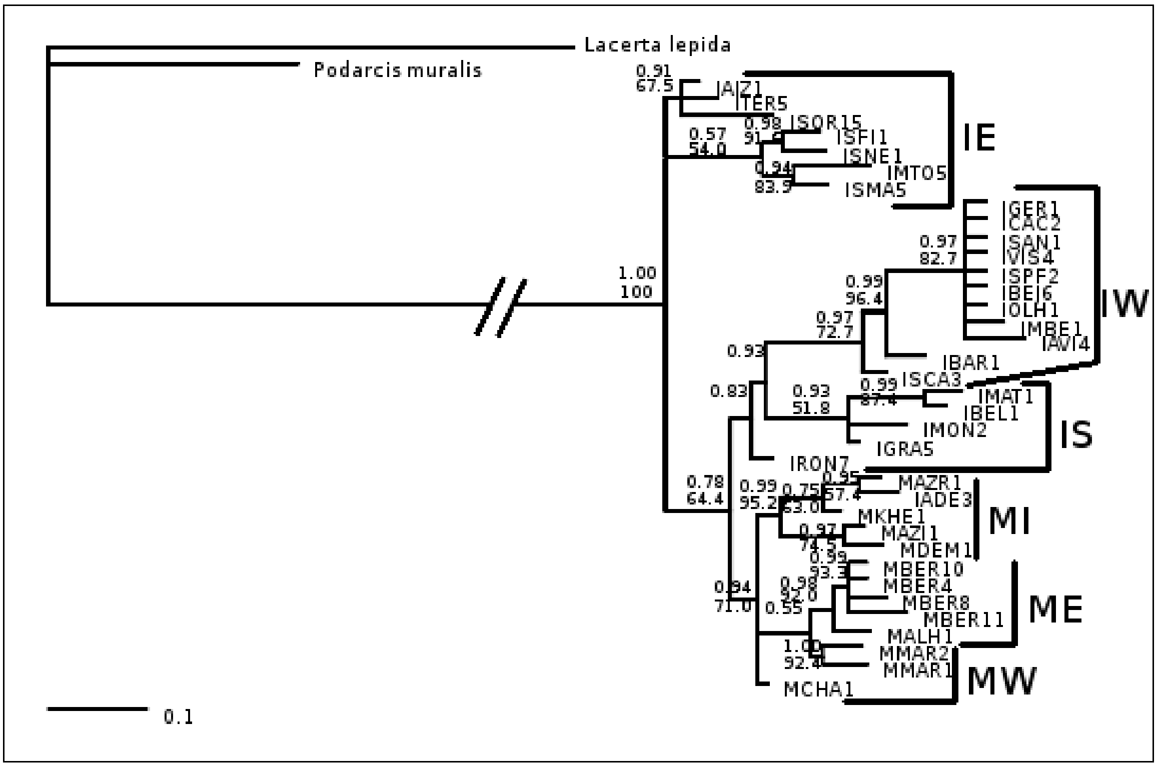


Fig. 2.4: Phylogenetic tree of the cytochrome *b* sequences dataset. The values above the branches are posterior probabilities and ML bootstraps (above and below respectively); When only one value is displayed the bootstrap support was lower than 50. Right of the tree, the formed groups/clades are indicated.

Cytochrome *b* (Fig. 2.4): The trees obtained for this dataset are well resolved (with high posterior probabilities/bootstraps in general) with the IE group in the tree base. In the crown of the tree, two groups can be found, one composed of the African clades and one composed of the IS clade basal to the IW clade. This tree resolves all of the groups considered in this work due to highly variable sequences. It is important to notice that the IE group is displayed as a polytomy, rather than a single clade/group as happens with the other groups. In the African groups, the MW group is not as supported as the other clades or even as in the other genes' sequences. The homoplasy index for this tree was 0.3270.

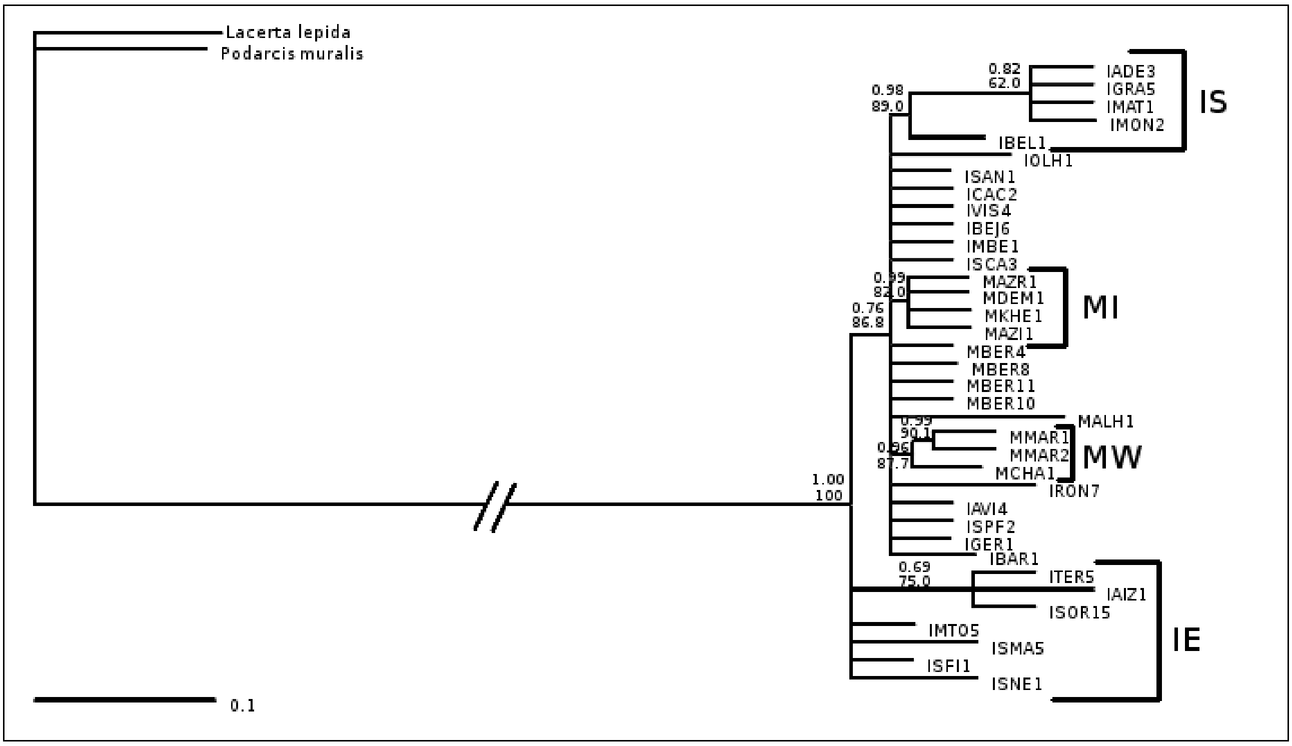


Fig. 2.5: Phylogenetic tree of the 12s rRNA sequences dataset. The values above the branches are posterior probabilities and ML bootstraps (above and below respectively). Right of the tree, the formed groups/clades are indicated.

12s ribosomal RNA (12s) (Fig.2.5): The trees obtained for this dataset place the IE group as basal and are unable to differentiate the remaining groups except for the IS, MI and MW clades in the tree crown. Nevertheless, the bootstrap/posterior probability support is high for these relationships. Due to being a relatively conserved gene and it's small size, these trees are not able to resolve many of the existing relations among the expected groups. It is important to note that in this tree too, the IE group is polytomic, although, in this dataset most relationships between the existing groups are not resolved. The homoplasy index for this tree was 0.1441.

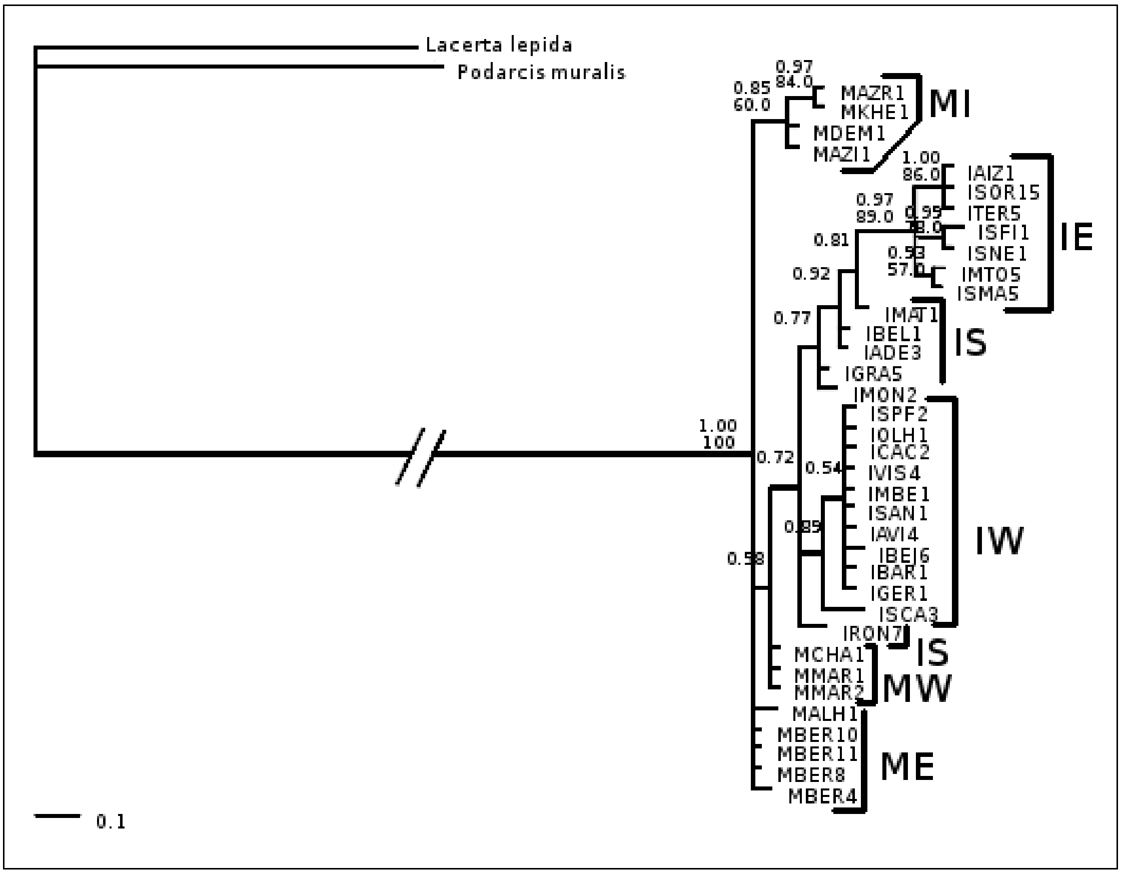


Fig. 2.6: Phylogenetic tree of the 16s rRNA sequences dataset. The values above the branches are posterior probabilities and ML bootstraps (above and below respectively); When only one value is displayed the bootstrap support was lower than 50. Right of the tree, the formed groups/clades are indicated.

16s ribosomal RNA (16s) (Fig. 2.6): The results obtained for this dataset place the Moroccan clade/groups as basal and further discern three Iberian groups. The IW clade is displayed very evidently and the IS group shows up as basal to IE clade. The posterior probabilities from the bayesian analysis give a good support to this tree's groups/branches, but the bootstrap values are considerably lower, most times below 50. Nevertheless, the tree topologies are concordant, regardless of the method used to obtain them. Despite being a somewhat conserved dataset, the 16s resolves most of the relationships between groups. Still, the ML tree is not very well supported by the bootstraps. It is important to note that the ME and MW groups are polytomic on this tree. Another important detail to highlight is that in the base of the Iberian groups, is the MW group, which is a geographically coastal group. The homoplasy index for this tree was 0.1258.

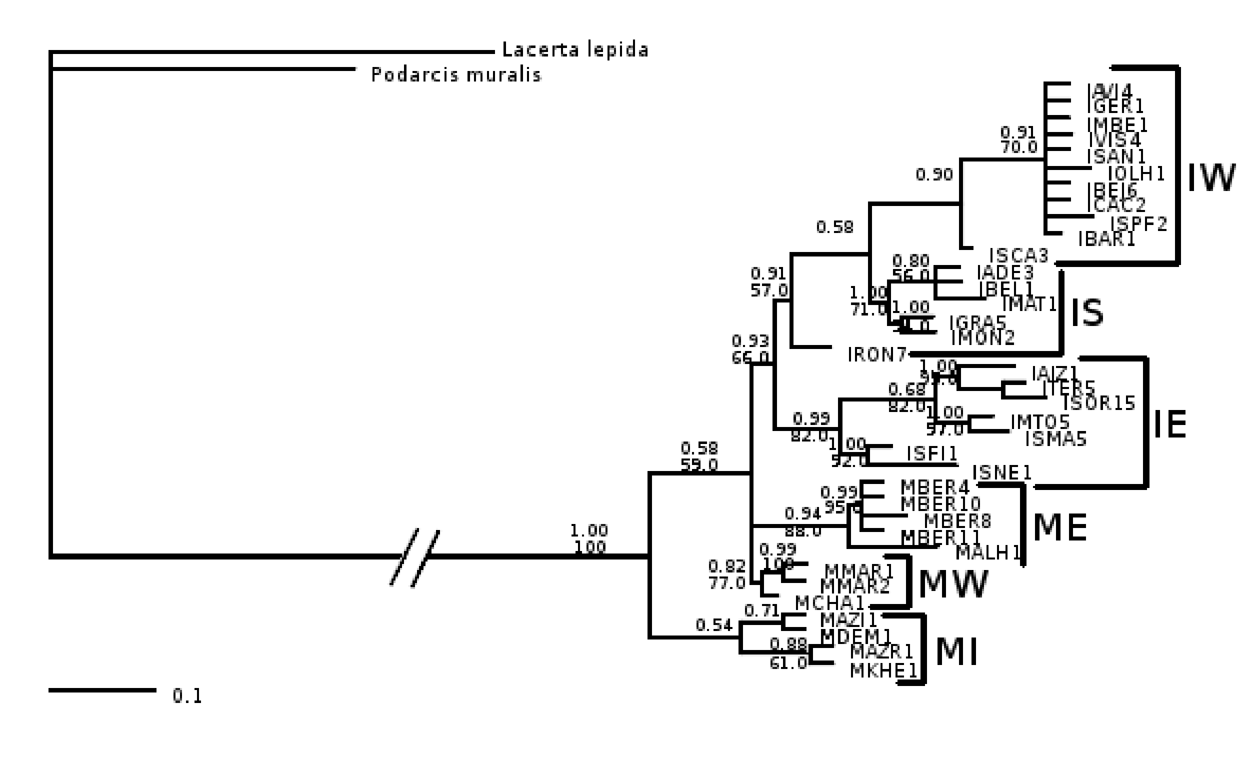


Fig. 2.7: Phylogenetic tree of the NAD4 sequences dataset. The values above the branches are posterior probabilities and ML bootstraps (above and below respectively); When only one value is displayed the bootstrap support was lower than 50. Right of the tree, the formed groups/clades are indicated.

NADH dehydrogenase subunit 4 (NAD4) (Fig.2.7): The obtained trees for this dataset are well resolved in both methods and place the MI clade at the base of the tree; three groups derive from this clade – one is the ME clade, the second is the MW clade and the third is composed of the IE clade, and the IS group basal to the IW clade. Once again, due to a high variability of the sequences the resulting tree is well resolved and all the relationships between the considered groups are displayed. Similar to what happened with the 16s, but in a smaller scale, the ML bootstrap support tends to be lower than the posterior probabilities. The homoplasy index for this tree was 0.3202.

For the three concatenated datasets, the results are shown in **Figs. 2.8 – 2.10**, for the 12s and cytochrome *b*, 16s and NAD4 and the total evidence dataset, respectively.

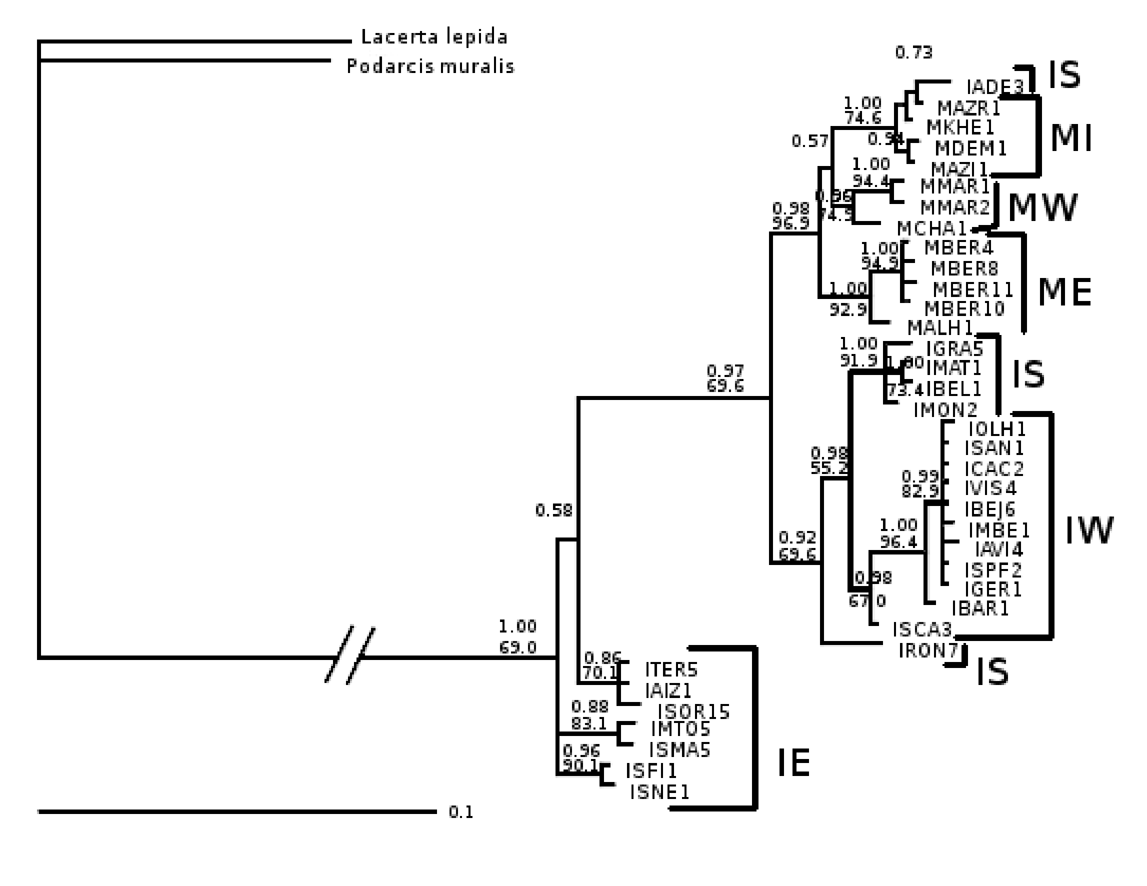


Fig. 2.8: Phylogenetic tree of the concatenated 12s rRNA and cytochrome *b* sequences dataset. The values above the branches are posterior probabilities and ML bootstraps (above and below respectively); When only one value is displayed the bootstrap support was lower than 50. Right of the tree, the formed groups/clades are indicated.

When the 12s and cytochrome *b* datasets are combined **(Fig. 2.8)**, the result is a well resolved tree with the IE clade in the base of the tree and the African clades and the IS and IW groups as derived. Most relationships in this tree are very well supported by the posterior probabilities/bootstrap values. Like in the individual datasets, the IE group does not resolve here as a single clade, though the existing polytomies from the individual datasets are almost completely resolved. Although the ILD partition test result was significant (p<0.01), the datasets were concatenated and analysed, since both datasets presented a similar phylogenetic signal and it is generally accepted that the mitochondria is behaves like a single molecule, and suffers no recombination, thus transferring both these genes as a single unit. The homoplasy index for this tree was 0.3291.

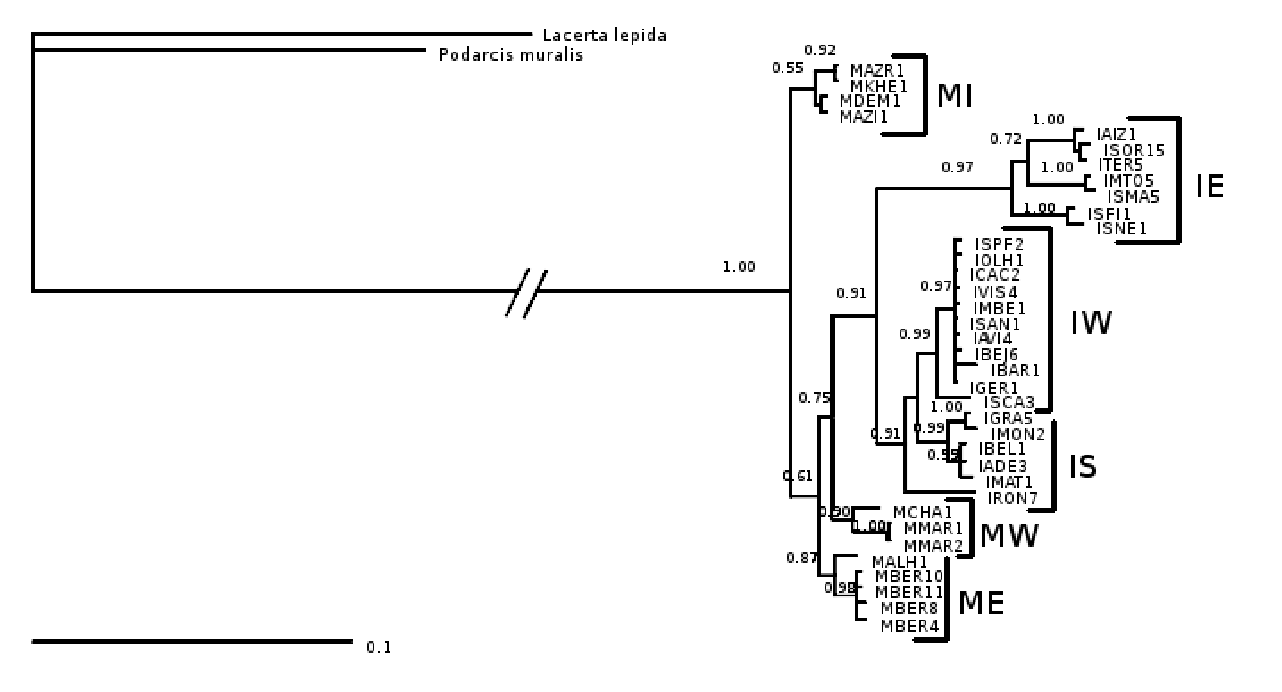


Fig. 2.9: Phylogenetic tree of the concatenated 16s rRNA and NAD4 sequences dataset. The values above the branches are posterior probabilities and ML bootstraps (above and below respectively); When only one value is displayed the bootstrap support was lower than 50. Right of the tree, the formed groups/clades are indicated.

When the 16s and NAD4 datasets are combined **(Fig. 2.9)**, the result is also a well resolved tree where the three African clades are at the base of the tree and shows the Iberian groups as two clades, IE as a crown clade and IS as a group basal to the IW clade, derived from the Moroccan populations. The posterior probability/bootstrap support for this tree is very high on most branches and the polytomies displayed in the two individual datasets are resolved with the concatenation. The ILD partition test results were not significant ($p<0.9$), meaning that these datasets are similar enough to be concatenated. The homoplasy index for this tree was 0.2812.

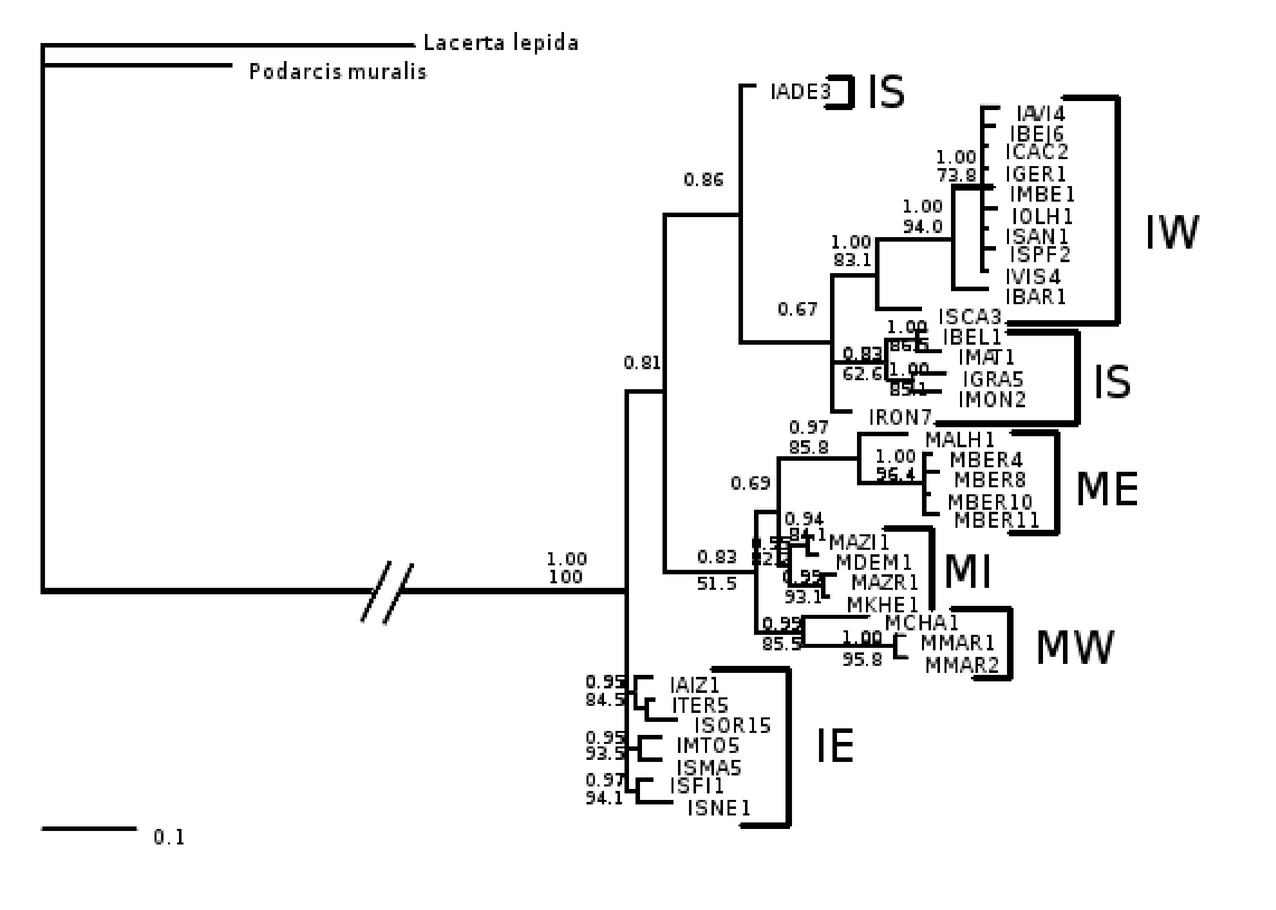


Fig. 2.10: Phylogenetic tree of the four concatenated genes' sequences dataset. The values above the branches are posterior probabilities and ML bootstraps (above and below respectively); When only one value is displayed the bootstrap support was lower than 50. Right of the tree, the formed groups/clades are indicated.

When all four datasets are concatenated **(Fig. 2.10)**, the IE group (once again, polytomic) is placed on the tree base with relatively good posterior probability/bootstrap support for each subgroup. The tree further distinguishes two crown groups, one composed of the three Moroccan clades, and one composed of the IS group, basal to the IW clade. The polytomies may eventually be explained by contradictory phylogenetic signal from the different datasets, which also explains the ILD partition test results (p<0.01). Regardless of this, the dataset was analysed when concatenated since the tree is composed of only mitochondrial gene sequences. The homoplasy index for this tree was 0.3320.

The sample IADE3 displayed a great affinity with the MI clade for the cytochrome *b* sequences. Although this sample was amplified and sequenced before any African sample had ever reached the laboratory, it was sequenced twice, to make sure there was no cross contamination of any sort and the results were the same. This could be indicative of recombination, but since it is outside the scope of this paper, it will not be addressed any longer. The effect it had on phylogenetic analysis was not important, and the trees produced without it were not significantly different from those where it was included, and so it was left in the analysis. In the population analyses it was always included in the MI clade, despite having been captured in Iberia.

Two other phylogenetic trees were obtained in order to more easily visualise the differences between the two different sets of results. These datasets are identical to the combined 12s + cytochrome *b*, and 16s + NAD4 but the outgroups were removed. The results are displayed in **Figs. 2.11 & 2.12**.

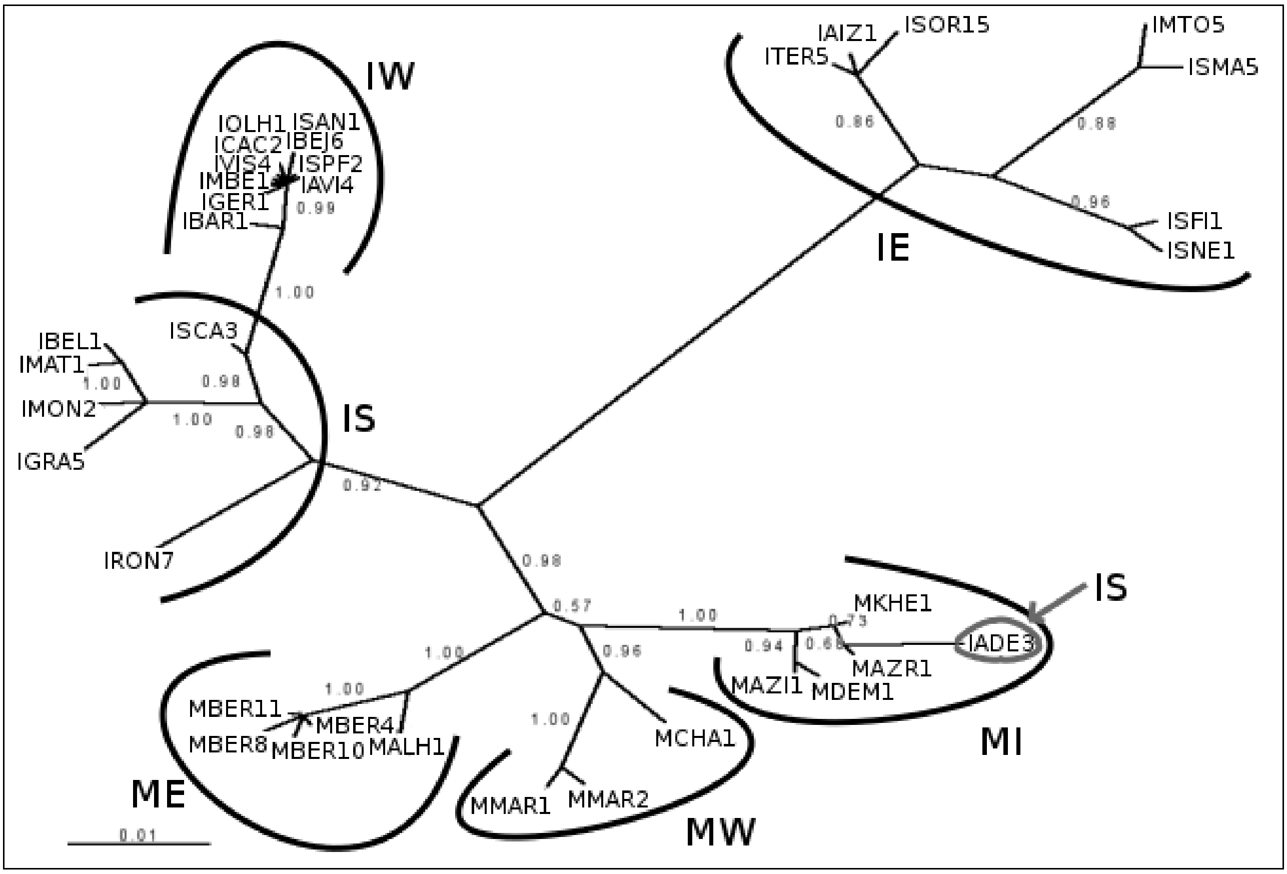


Fig. 2.11: An unrooted bayesian inference phylogenetic tree of the 12s rRNA and cytochrome *b* concatenated datasets. The values close to the branches are posterior probabilities. The clade/group that each set of samples forms is indicated with the larger font. **IE** – Iberia East, **IS** – Iberia South, **IW** – Iberia West, **MI** – Morocco Interior, **ME** – Morocco East, **MW** – Morocco West.

The unrooted 12s and cytochrome *b* **(Fig. 2.11)** and 16s and NAD4 **(Fig. 2.12)** trees are very similar to the rooted ones. The same clades/groups are formed and the distances between groups are similar as expected.

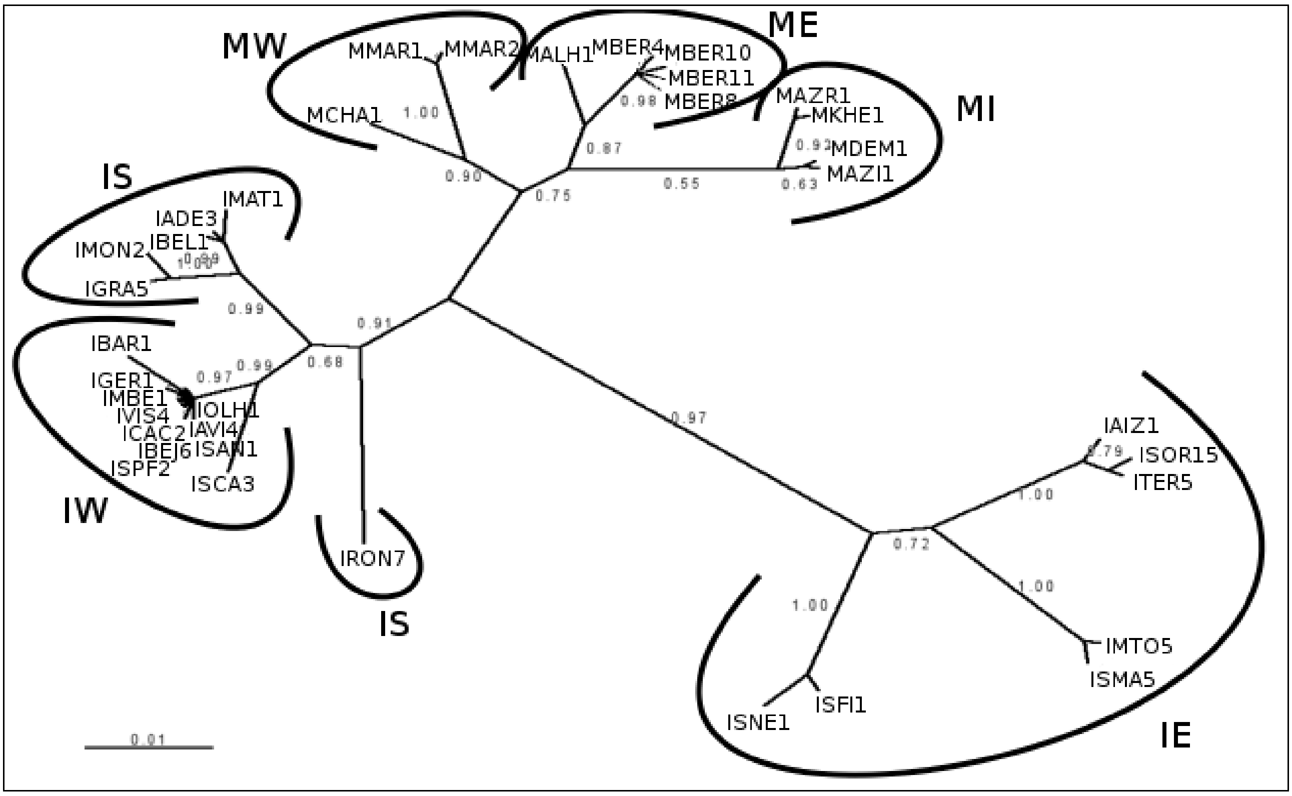


Fig. 2.11: An unrooted bayesian inference phylogenetic tree of the 16s rRNA and NAD4 concatenated datasets. The values close to the branches are posterior probabilities. The clade/group that each set of samples forms is indicated with the larger font. **IE** – Iberia East, **IS** – Iberia South, **IW** – Iberia West, **MI** – Morocco Interior, **ME** – Morocco East, **MW** – Morocco West.

This set of results clearly indicate that the IE group is the more differentiated group of the ones detected independently of his position in relation to the outgroups

Population analysis:

The results of the network analyses are shown in **Figs. 2.13 & 2.14**. These are congruent with the species' geographical distribution, except for the distance from the eastern Iberian clade to the others which is much more separated genetically than it is geographically. This population analysis using the cytochrome *b* sequences with 53 samples is displays more complex links than the analysis of all four concatenated genes with only 36 samples.

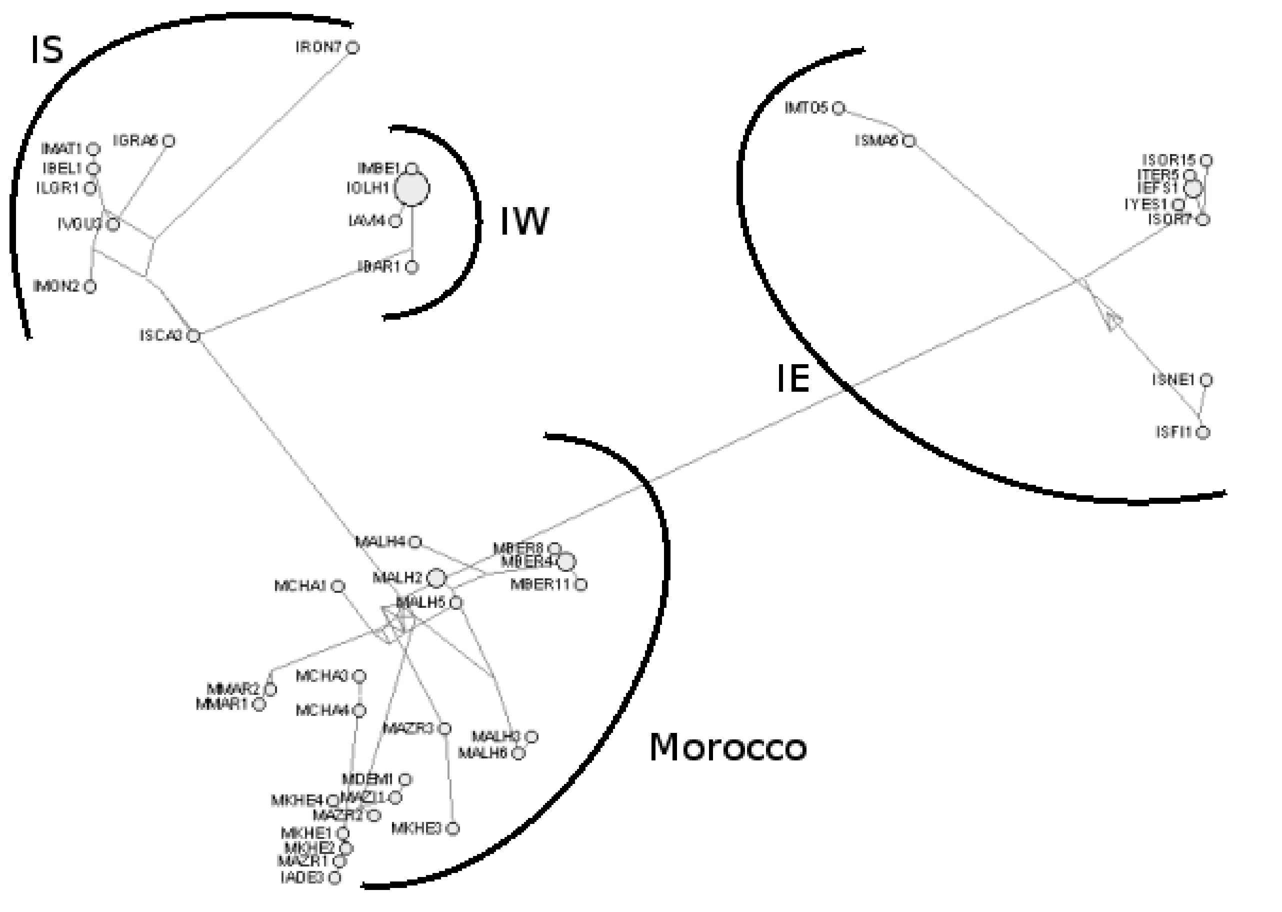


Fig. 2.13: A median-joining network of the cytochrome *b* dataset with 53 samples. The formed relations respect the geographic positions of the samples, except for the distance from the IE clade to the others, which is longer genetically relatively to the geographic distance. Larger circles represent several individuals sharing the same haplotype (the larger the circle, the more individuals share that haplotype).

The four performed AMOVAs outputted in Arlequin the results shown in **Table 2.3**. The groupings are the only difference in the analyses inputs. These results are congruent with the phylogenetic analysis when the 6 clades are separated (70% of variation is explained among groups). All results are significant (p < 0.04624±0.00186). The 6 clades grouping clearly makes the best separation between groups followed by IE Vs. IS + IW Vs. Morocco which means that this grouping is not as good as the former to explain variance, but it is still quite good (explaining 54.12% of the variance).

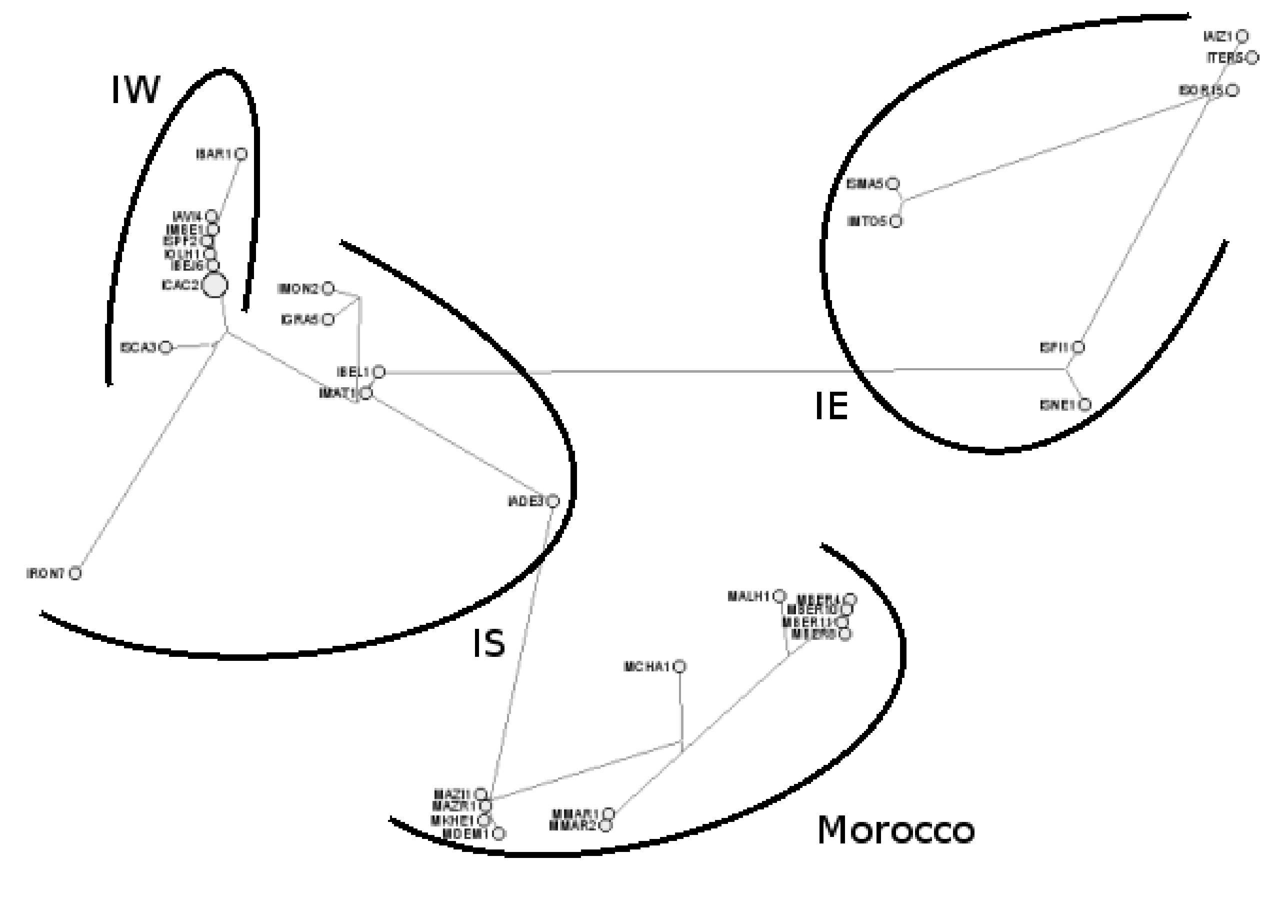


Fig. 2.14: A median-joining network of the four genes concatenated dataset with 36 samples. The formed relations respect the geographic positions of the samples, except for the distance from the IE clade to the others, which is longer genetically relatively to the geographic distance. The resolution is low due to the large dataset used. Larger circles represent individuals sharing the same haplotype (the larger the circle, the more individuals share that haplotype).

The obtained nucleotide diversity (π) and haplotype diversity (H) values are shown in **Table 2.4**. The values of π are relatively high and the values of H are very high. Since the number of samples was low for mismatch analyses on each group, these analyses were performed including multiple groups. The graphs (**Figs. 2.15-2.17**) display multi-modal profiles in all cases: the mismatch peaks are either separated by a slope (IE and IS+IW) or are roughly constant (Moroccan clades).

Table 2.3: AMOVA results with the percent variation obtained with four different groupings. All results are significant (p < 0.04624±0.00186).

Source of Variation	IE Vs. IS+IW Vs. Morocco	Iberia Vs. Morocco	IE Vs. IS+IW +Morocco	6 clades
Among groups	54,12	24,29	45,81	70,00
Among populations within groups	24,67	53,54	36,75	9,36
Within populations	21,21	22,17	17,44	20,64

Table 2.4: Observed π and H in each of the considered clades and in several combinations.

Dataset	Number of Samples	Observed	
		π	H
IE	10	0,02441	0,97778
IS	7	0,01413	1,00000
IW	11	0,00558	0,61818
MI	10	0,01550	1,00000
MW	5	0,02449	1,00000
ME	10	0,01678	0,95556
IS+IW	18	0,02171	0,86275
Iberia	28	0,05162	0,94180
Morocco	25	0,02814	0,99333
M+IS+IW	43	0,04206	0,97453
All	53	0,05429	0,98258

Table 2.5 shows the pairwise F_{ST} values. All values are significant (p<0.01564 ± 0.0012). The higher values are found between IW and all the groups except IS and the lowest is found between MI and MW.

Table 2.5: Pairwise F_{ST} values between each of the considered clades.

	IE	IS	IW	MI	MW	ME
IE						
IS	0.73704					
IW	0.82839	0.75411				
MI	0.76010	0.74500	0.82633			
MW	0.68837	0.66420	0.80683	0.21198		
ME	0.74369	0.70969	0.82860	0.57617	0.42429	

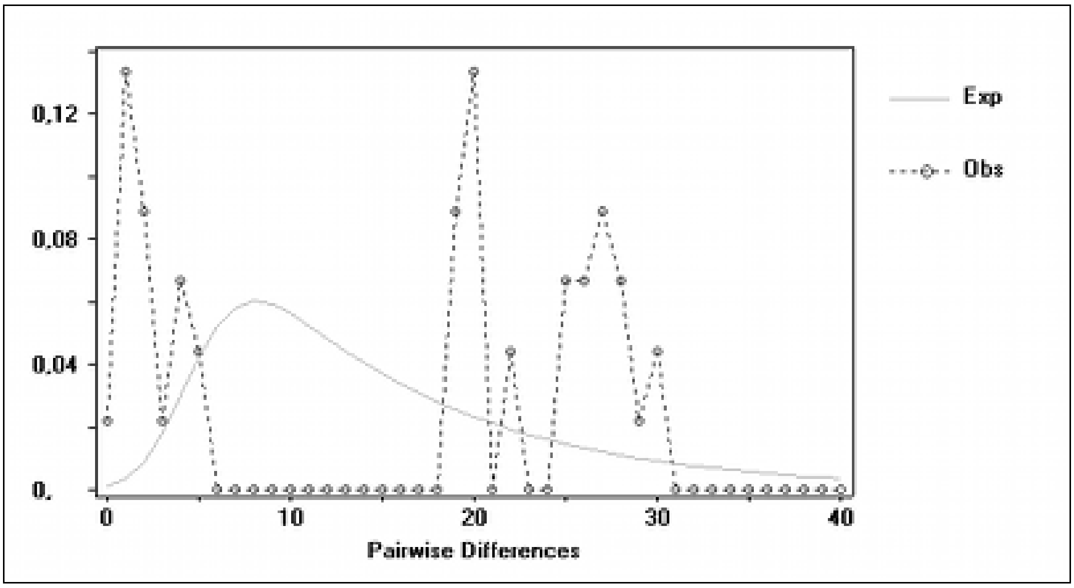


Fig. 2.15: Mismatch distribution of the IE group. The dashed line represents the observed mismatch values and the whole line represents the expected line under a recent growth model.

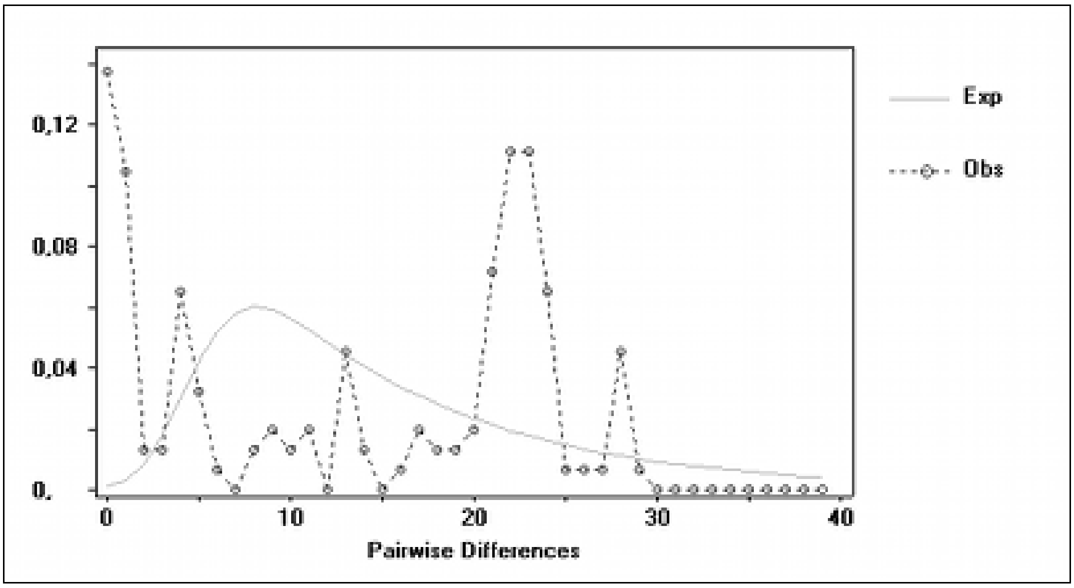


Fig. 2.16: Mismatch distribution of the IS + IW group. The dashed line represents the observed mismatch values and the whole line represents the expected line under a recent growth model.

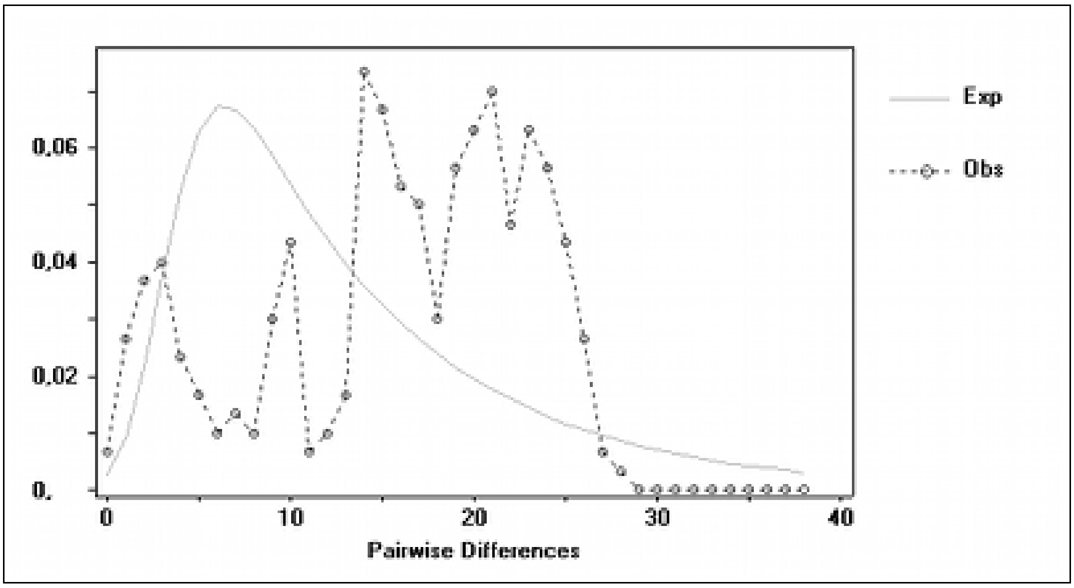


Fig. 2.17: Mismatch distribution of the Moroccan group (all three clades). The dashed line represents the observed mismatch values and the whole line represents the expected line under a recent growth model.

Discussion

The phylogenetic analyses

Regardless of the methods used (parsimony, neighbour-joining, maximum likelihood or bayesian) the topologies of the found trees are always very similar. This congruence among the analyses is not very common and it indeed reinforces the validity of the obtained results.

The data is congruent when it comes to defining groups. The six considered groups/clades are very well supported on most trees, regardless of the sequence set that originated them but in some trees not all groups are defined, however the formed groups are identical across datasets. This seems to be indicative of a well established separation between these groups.

The main issue on this data, was defining the ancestral clade and thus, a putative origin of the species' phylogeographic patterns. If two of the genes seem to indicate an Iberian origin for the species, the other two suggest an African origin. When concatenated, the agreeing datasets provide the same results as the individual ones. When all four datasets are analysed as a single unit, the obtained tree indicates the IE group as ancestral. However, this result is not final, since the ILD partition test indicated that these sequences may be providing distinct phylogenetic signal; furthermore, it may be due to the fact that gene sequences with the phylogenetic signal indicating an Iberian origin have 7 more parsimonious informative sites than the one that indicates an African origin.

Despite all the differences found in the rooted trees, when the unrooted trees are analysed, the differences between the datasets are minor, both in the relationships among groups and in the support that each branch has. This is due to the outgroup being closer to the Moroccan clades in the 16s and NAD4 datasets and closer to the IE group in the 12s and cytochrome *b* datasets. These discrepancies may be originated by homoplasies. Nevertheless, these trees only marginally support one of the theories regarding the ancestral clade.

The classical phylogenetic analysis was performed using four different genes and yielded two different results regarding the ancestral clade and unrooted trees were added to try to understand the controverse results. However it the number of used genes were to be increased, several more hypotheses of evolutionary histories with different kinds of supports could have been generated. Could this issue be solved with the use of a different outgroup? And what if nuclear genes had been included in the analyses? Would we be looking at even another scenario?

The different perspectives revisited

After looking at our own data, we seem to have found the origin of the differences found between Carranza *et al.* (2006) and Busack & Lawson (2006). By using different genes' sequences, the authors obtained different topologies due to different relations between the analysed genes and the outgroup which caused different types of tree rootings. In Busack & Lawson (2006), despite the use of two genes that produce different results, it is likely that the lack of cytochrome *b* samples (only 6) caused them to miss this point, even if it is mentioned in their work that this gene's sequences pointed their trees in a different direction.

In Carranza *et al.* (2006) the tree that is mainly used to propose a biogeographic model is based on two genes (12s and 16s) that our work has shown to produce contrasting results about the basal group. This reason, coupled with relatively high conservation of the datasets, may be why regardless of the large amount of samples used their tree is not very well supported by bootstrap values.

In order to better understand the differences obtained by these authors, we have proposed hypotheses similar to the ones proposed in the mentioned papers but based on our datasets. However, since our main phylogenetic dataset only provides marginal support when resolving the ancestral clade, a populational approach with additional samples was used, increasing their number especially in the Moroccan clades. With both the phylogenetic and the phylogeographic approach, we try to clarify the biogeographic pattern of this species.

The Population Approach

Since the differentiation among the Moroccan clades is not as marked as in the Iberian clades, the analyses in the population approach (except the AMOVA) consider all of the Moroccan clades as a single group. The same is also applied to the IS and IW groups which were considered as a single group in many analyses.

The two network analyses are relatively congruent and the differences in complexity can be explained by the number of bases and the number of included samples. The great increase in the number of bases in the concatenated genes analysis makes the Median--Joining algorithm lose resolution, producing an “all straight lines” output, where the increase in the sample size will make the connections between individuals (especially among the most differentiated individuals) more difficult to resolve, resulting in complex median points arrangement in a “net” shape.

The percentage of variance explained by the groupings in the AMOVA analyses indicates that the six clades division is indeed where the great cleavages exist in our dataset. This grouping alone explains 70.0% of the found variance. The other groupings explain only low to moderate amounts of variation (24.29% to 54.12%).

Regarding the values of nucleotide diversity (π) and haplotipic diversity (H), it is important to mention that the H values are much more sensitive to small sample sizes than π. When both of these values are mentioned later in the analyses, we attribute a much greater weight to π than to H.

The results displayed in **Table 2.5** show that the F_{ST} results are congruent with the phylogenetic analyses when considering which groups are closer. The Moroccan clades are displayed as the closest to each other and IS is shown to be closer to the Moroccan group than to IE. the IW clade is also closer to IS than to any of the other clades. The IS group is also roughly equidistant to IE and the Moroccan clades.

Despite the relatively low number of samples for the mismatch analyses, they were still performed and are somewhat informative. In the IE group, this might be indicative of two expansions (an older one and a more recent one), but in other analyses such as in the IS +IW grouped the bimodal profile might represent an old expansion (IS expanding) and a recent expansion (IW expanding) or simply the differences between two established populations. In the Moroccan groups we see a many peaks profile (although always lower than in the other graphs) that must represent the differences between the three clades (regardless of expansions). The AMOVA has shown that this is a good way to join the groups and any of the isolated groups provided too little information to be analysed by itself.

The proposed hypotheses

Based on the obtained results, we assesed the hypotheses proposed by other authors (and our own) to explain *P. algirus*' current phylogeography.

Hypothesis 1:

This hypothesis is according to the conclusions of Busack & Lawson (2006). The model proposes an African origin followed by a crossing of the Mediterranean sea into the Iberian Peninsula. The newly established Iberian population would then expand separately to the east and to the west. The model is displayed in **Fig. 2.18**.

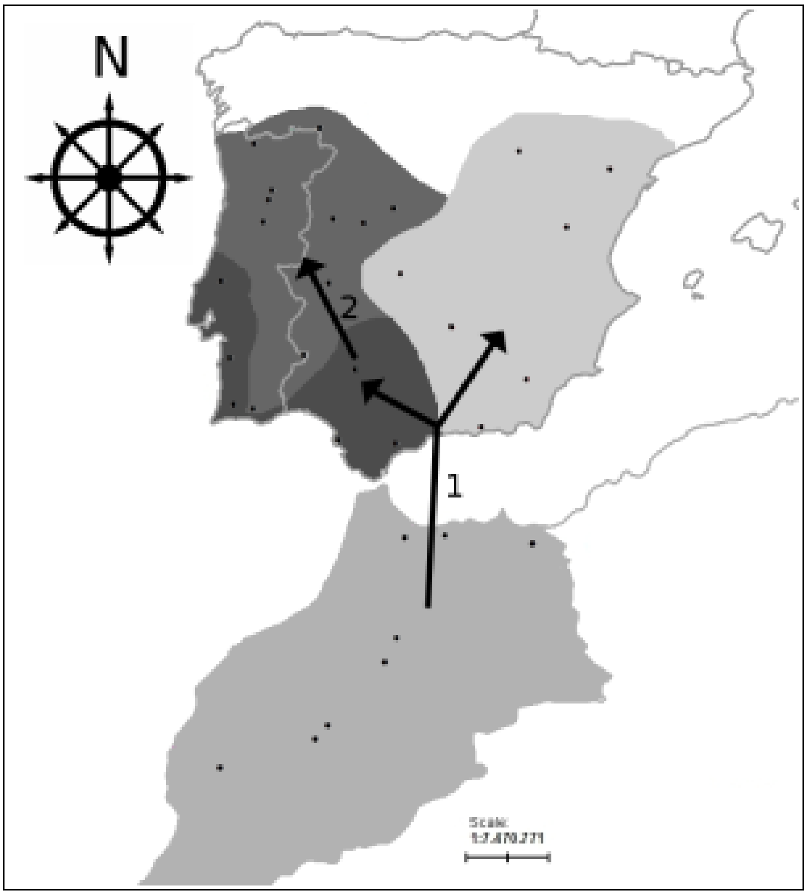


Fig. 2.18: Map of the Iberian Peninsula with *P. algirus* movimentations according to Hypothesis 1. The numbers next to the arrows are the order of the movements.

This model is only only supported by some results of the phylogenetic analyses, but not by infered trees of the concatenated four genes dataset, the Moroccan clades should be basal in relation to the Iberian one, which is not verified. Nevertheless, it can be further tested under our populational data. According to this model we expected the data to show the highest values of π and H in the African clades, medium values for all Iberian groups together and the lowest of these values for each of the Iberian groups. However, what we observe are higher values of π in Iberia than in Morocco (and similar values of H), and the lowest of these in each of the Iberian groups. Most of the expectations are not verified in this model which means that the support from our data to this hypothesis is limited.

Hypothesis 2:

The model proposed by Carranza *et al.* (2006) suggests an Iberian origin, which according to our results this means an origin in the IE group, followed by an expansion to the southwest of Iberia and then, from this area, across the Mediterranean sea into Africa, forming the Moroccan clades. Later the IS clade would expand to the north forming the IW clade. The model is shown in **Fig. 2.19**.

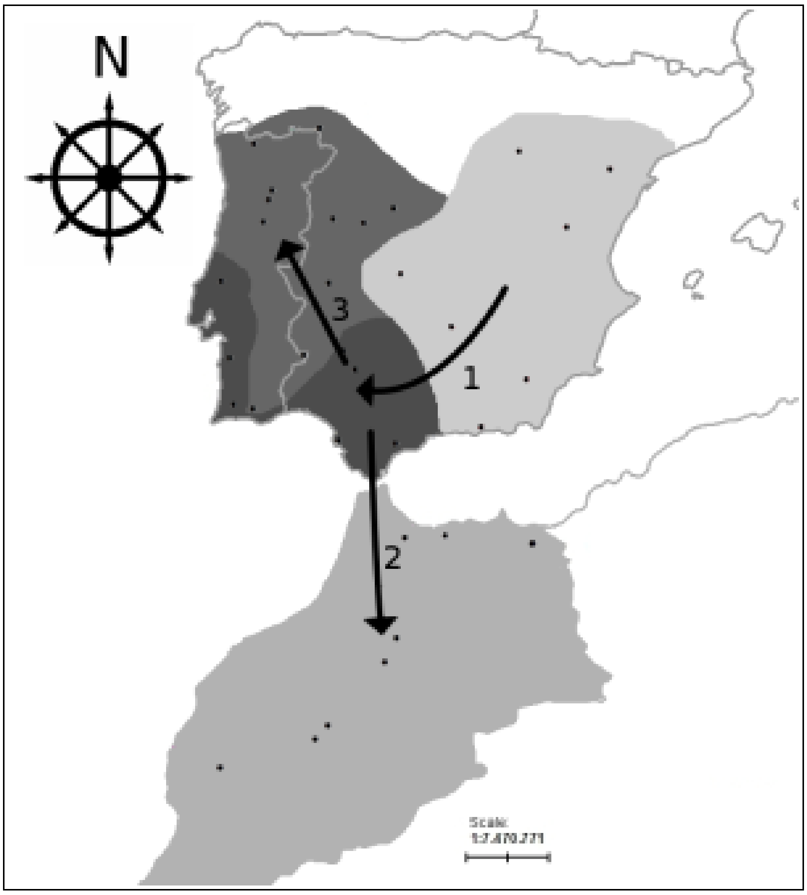


Figure 2.19: Map of the Iberian Peninsula with *P. algirus* movimentations according to Hypothesis 2. The numbers next to the arrows are the order of the movements.

This model is in agreement with our phylogenetic analyses. Not only the distance from the ancestral clade is congruent, as the two derived clades are roughly equidistant from their origin. Under it, we would expect to find the highest values of π and H in Iberia (all three clades together), followed by IE, followed by IS + IW and finally, the Moroccan clades. In fact, the highest π is found on all the Iberian groups together (although H is not, but it is due to a low value in the IW clade), but the Moroccan clades display a higher value of π and H (except for H in IS which is 1, but has only 7 samples, all them unique haplotypes) than any of the Iberian groups. This too is incompatible with the proposed hypothesis, unless we consider the lower values of π and H as effects of glaciations on the Iberian Peninsula. This hypothesis is thus, only marginally supported by our population study.

Hypothesis 3:

This model is based on the data obtained on this study and represents the hypothesis that seems more reasonable according both our phylogenetic and population study. Like the former, this hypothesis assumes an Iberian origin for the species (also in the IE group/clade). However, in our model, we propose an initial expansion from this clade to the south, across the Mediterranean sea, establishing the Moroccan clades; later, individuals from these populations crossed the Mediterranean sea heading north and established the IS clade, with a later expansion, once again northwards forming the IW clade. The model is shown in **Fig. 2.20**.

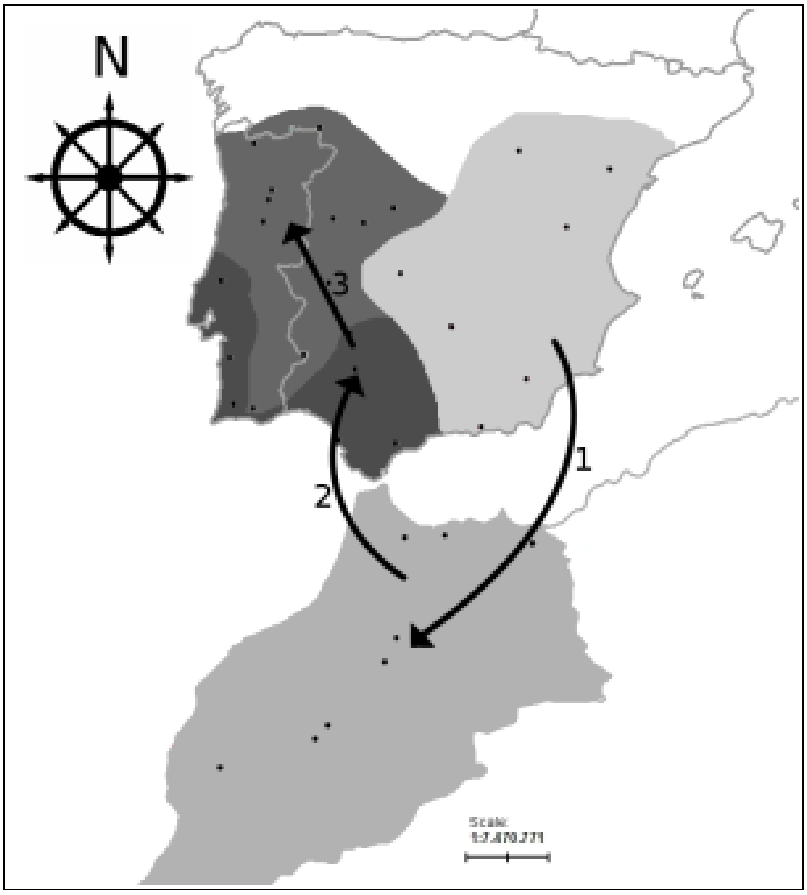


Fig. 2.20: Map of the Iberian Peninsula with *P. algirus* movimentations according to Hypothesis 3. The numbers next to the arrows are the order of the movements.

This hypothesis is also supported by the phylogenetic analyses. The IE group/clade is the most distant from the rest of the groups and Morocco and IS+IW are roughly equidistant from the ancestral clade. According to this hypothesis we would expect to find the highest π and H values in the conjunction of all Iberian clades, followed by the IE clade, then the

Moroccan clades, then the set of IS + IW, then IS and finally IW. When looking at the H values, they are not verified in some cases for the same reasons than in the above models – low sample size in one case, and high amount of samples sharing the same haplotype in the other. Regarding the π values, all of these expectations are verified, except the order of IE – Morocco, which is reversed. However, this may well be due to the fact that the IE has only 10 samples and the Morocco group is composed of 25. Thus, this hypothesis cannot be rejected by the data. Although it is less parsimonious than the second hypothesis, because it requires two crossings of the Mediterranean, the data seem to be much more supportive of this model than of any of the others.

Final Remarks

The data presented here shows that conventional phylogenetic analysis, based only in four is not sufficient to discern the origin of *P. algirus* as was presented in other studies (Busack & Lawson, 2006; Carranza *et al.*, 2006). The use of a larger set of gene sequences not only explained the differences found between the two mentioned studies, as it raised new questions regarding the phylogeography of the species. In an attempt to resolve the conflicting data (or to rephrase it, the unresolving data) a population approach was required using an increased sampling. This has enabled the testing of a set of hypotheses that could explain the species' phylogeography. Of these hypotheses, the data give more support to one that implies two migration events across the Mediterranean sea. However, the data does not support this hypothesis to it's full extent, though it is likely that it is due to differences in the number of samples on the groups/clades involved in the non matching expectations. A larger sampling in the IE and Moroccan clades/groups could resolve this issue either by giving support to this hypotheses or by pointing in a different direction. The use of nuclear genes could also eventually provide a better insight on this issue.

It is also important to note that, when the sampling in Morocco was increased from 12 to 25 samples the network complexity increased much more in this location then in any of the Iberian clades. If the rate of increase in complexity per additional sample was kept should we continue to increase the sampling, it is possible that the data could start pointing towards an African origin rather than an Iberian one.

The data marginally favours an hypothesis that requires two crossings of the Mediterranean sea. This is not a parsimonious explanation for *P. algirus*' biogeographic patterns, but it is in fact the most supported we have found. The idea of a single crossing of the Mediterranean sea after the Messinian Salinity crisis was already present in the works of Busack & Lawson (2006) and Carranza *et al.* (2006), but in here we present crossings in two distinct moments. This reinforces the already present idea that the Mediterranean sea is not so impermeable to the crossing of terrestrial species such as *P. algirus* as it was once thought.

References

Arnold E. 2002. *Reptiles and Amphibians of Europe*. Princeton University Press, Princeton, NJ. 288 pp.

Bandelt H, Forster P and Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution. 16: p. 48.

Busack S and Lawson R. 2006. Historical biogeography, mitochondrial DNA, and allozymes of Psammodromus algirus (Lacertidae): a preliminary hypothesis. Amphibia-Reptilia. 27: pp. 181-193.

Busack S, Salvador A and Lawson R. 2006. Two new species in the genus Psammodromus (Reptilia : lacertidae) from the Iberian peninsula. Annals of Carnegie Museum. 75: pp. 1-10.

Carranza S, Arnold E, Wade E and Fahd S. 2004. Phylogeography of the false smooth snakes, *Macroprotodon* (Serpentes, Colubridae): Mitochondrial DNA sequences show European populations arrived recently from northwest Africa. Molecular Phylogenetics and Evolution. 3: pp. 523-532.

Carranza S, Harris D, Arnold E, Batista V and de la Vega J. 2006. Phylogeography of the lacertid lizard, *Psammodromus algirus*, in Iberia and across the Strait of Gibraltar. Journal of Biogeography. 33: pp. 1279-1288.

Excoffier L, Smouse P and Quattro J. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes - application to human mitochondrial-DNA restriction data. Genetics. 131: p. 491.

Hall T. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp.. 41: pp. 95-98.

Harris D, Batista V, Carretero M. 2004. Assesment of genetic diversity within *Acantodactylus erythrurus* (Reptilia: Lacertidia) in Morocco and the Iberian Peninsula using mitochondrial DNA sequence data. Amphibia-Reptilia. 25: pp. 227-232.

Harris D, Carranza S, Arnold E, Pinho C and Ferrand N. 2002. Complex biogeographical distribution of genetic variation within podarcis wall lizzards across the Strait of Gibraltar. Journal of Biogeography. 29: pp. 1257-1262.

Hsu K, Ryan W and Cita M. 1973. Late miocene desiccation of mediterranean. Nature. 242: p. 244.

Miras J, Cheylan M, Nouira M, Joger U, Sá-Sousa P and Pérez-Mellado V. 2006 Psammodromus algirus. UICN 2007. 2007 IUCN Red List of Threatened Species. <www.iucnredlist.org>

Nylander JAA, Ronquist F, Huelsenbeck JPP, Nieves-Aldrey JL. 2004 Bayesian phylogenetic analysis of combined data. Systemaic Biology. 53: pp. 47-67

Posada D and Crandall K. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. 14: p. 818.

Poulakakis N, Lymberakis P, Antoniou A, Chalkia D, Zouros E, Mylonas M and Valakos E. 2003. Molecular phylogeny and biogeography of the wall-lizard *Podarcis erhardii* (Squamata : Lacertidae). Molecular Phylogenetics and Evolution. 28: p. 46.

Ronquist F and Huelsenbeck J. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19: p. 1574.

Rozas J, Sanchez-DelBarrio J, Messeguer X and Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 19: pp. 2496-2497.

Swofford D. 1993. PAUP - A computer-program for phylogenetic inference using maximum parsimony. Journal of General Physiology. 102: p. a9.

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

Vasconselos R, Carretero M and Harris D. 2006. Phylogeography of the genus Blanus (Worm Lizzards) in Iberia and Morocco based on mitochondrial and nuclear markers - preliminary analysis. *Amphibia-Reptilia*. 27: pp. 339-346.

Software Development

Chapter 3

Developed Software

Article. Pina-Martins, F.; Paulo, O.S. 2008. Concatenator: Data Matrices Handling Made Easy. Molecular Ecology Notes. In Press.

Concatenator: Sequence Data Matrices Handling Made Easy

F. Pina-Martins[1] and O. S. Paulo[1]

[1]Centro de Biologia Ambiental, Departmento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal.

Keywords: Sequence data; concatenation; Nexus; FASTA; software; perl.

Corresponding author:

Name: Francisco Rente de Pina Martins;

Address: Centro de Biologia Ambiental, Departmento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal;

E-mail: f.pinamartins@gmail.com

Abstract

Concatenator is a simple and user friendly software that implements two very useful functions for phylogenetics data analysis. It concatenates Nexus files of several fragments in a single NEXUS file ready to be use in phylogenetics softwares, such as PAUP and MrBayes and it converts FASTA sequence data files to NEXUS and vice-versa. Additionally, concatenated files can be prepared for partition tests in PAUP. It is freely available in the downloads section of http://cobig2.fc.ul.pt/.

The Program

Sequence data files can be organized in many different formats. Different sequence analysis software require differently formatted input files. The FASTA format has become very popular due to its simplicity and the capacity to quickly compare sequences (Pearson & Lipman 1988); these characteristics made this format one of the NCBI default outputs. The Nexus format became popular due to its modular format which is at the same time flexible and standardized (Maddison *et. al* 1997).

The existence of different file formats for the same data types require investigators to know how to handle them since they are not shared by some of the most common phylogenetic analysis software.

Concatenator's main purpose is to turn data matrix handling into a simple task, by allowing intuitive format conversions and concatenations of data matrices.

Concatenator is written in Perl using the Perl/TK module in order to give it a GUI for simplifying usage. The software was compiled using PAR module. It is available in the Win32 version and source code at the author's group website.

The only requirements are either a system with a Perl interpreter and the Tk module installed (source code version) or a system running Microsoft Windows XP (not tested on other versions). The software was developed with a very specific aim – the simple handling of data matrices from one program to another and the concatenation of several of these data matrices. All the functions that the program performs can be accomplished manually provided the user has some knowledge about the involved file formats; however, even in such case this process is very error prone due to complex data organization such as in the interleave Nexus format.

The user interface is very simple (**Fig. 3.1**) and consists of a window that accommodates essentially the input and output entry boxes; these files can be selected thru the File menu, a browse button located on the right of every entry box or by entering the path and filename directly on the entry box.

Concatenator can be used to accomplish 2 essential tasks chosen from the welcome window buttons or from it's "File" menu.

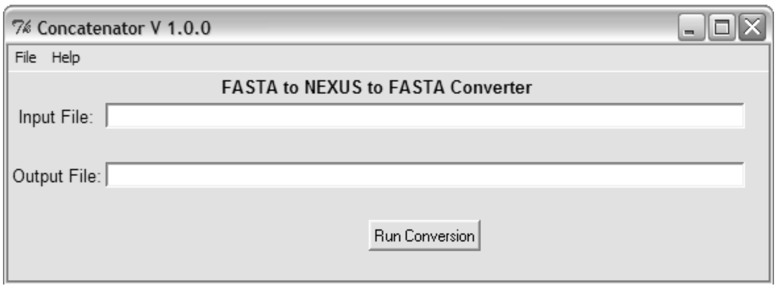
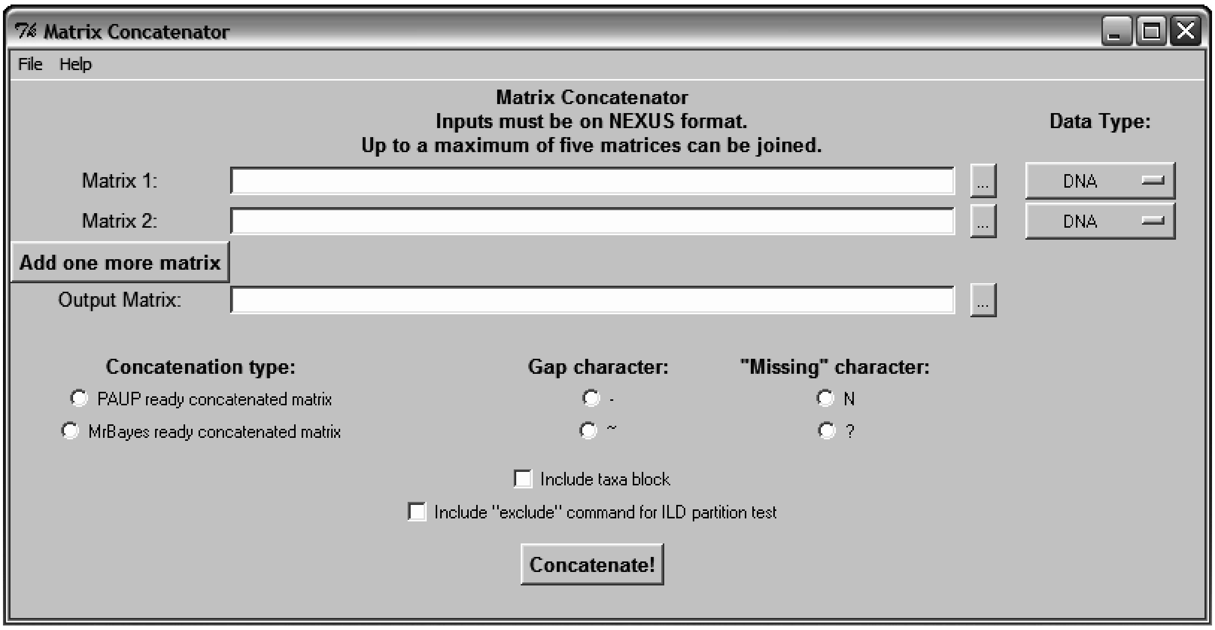


Fig. 3.1. *Concatenator interface on Win XP when converting matrices from Nexus to FASTA. Should the input be a FASTA file, the interface changes to accommodate the extra options to outputting a Nexus file.*

Fasta-Nexus-Fasta Converter – It converts files from FASTA to Nexus format and from Nexus to FASTA format. When converting from Nexus to FASTA, there are no options available to chose from, however, when converting from FASTA to Nexus the user can choose whether to include a Taxa block, a leave or interleave organization, the type of data, the character for missing data and the gap character. File comments are ignored when converting.

Matrix Concatenator – This function takes 2 to 5 Nexus formatted matrices and concatenates them into a single file. Two output formats are possible, one formatted to be used with PAUP* (Swofford 2003), and the other prepared to input to MrBayes (Ronquist & Huelsenbeck 2003). Several parameters are customizable such as the inputs' data type, the gap character, the missing character, whether or not to include the Taxa block and a pre input for performing a “partition test” in PAUP* excluding constant characters (**Fig. 3.2**).

Each function has a help file. The whole program is simple to use, but the help files are nevertheless as descriptive as possible.



Matrix Concatenator
Inputs must be on NEXUS format.
Up to a maximum of five matrices can be joined.

Fig. 3.2: Concatentor interface on Win XP when concatenating two data matricres.

Example Usage

The user downloads two arrays of sequences (e.g. two different genes from the same species) from the NCBI database using a program such as BioEdit (Hall 1999). After a proper alignment session, the program outputs two FASTA files – one for each gene.

The user then wants to analyze these files using PAUP*, MrBayes, TCS (Clement *et. al* 2000) or Network (Bandelt *et al.* 1999). Concatenator is useful in this step, because it provides a simple way to convert these FASTA files into Nexus files, ready to use in the analysis programs.

If after this first analysis the user decides to analyze both genes as a single block, Concatenator can join the two Nexus files in a single data matrix, ready to input on software such MrBayes or PAUP*; the built in function for data partitioning will automatically add the required commands for partitioning data required for a “Partition Test” in PAUP*.

References

Bandelt HJ, Forster P, Röhl A (1999) *Median-joining networks for inferring intraspecific phylogenies.* In: Molecular Biology and Evolution 16:37-48.

Clement M, Posada D, Crandall K (2000) *TCS: a computer program to estimate gene genealogies.* In: Molecular Ecology 9(10): 1657-1660.

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

Maddison DR, Swofford DL, Maddison WP (1997) *NEXUS: An Extensible File Format for Systematic Information.* In: Systematic Biology 46 4: 590-621.

Ronquist F, Huelsenbeck JP (2003) *MrBayes 3: Bayesian phylogenetic inference under mixed models.* In: Bioinformatics 19:1572-1574.

Swofford DL (2003) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.* Sinauer Associates, Sunderland, Massachusetts.

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. In: Biochemistry 85: 2444-2448.

Final Considerations

The *Psammodromus algirus* case study

This study aims to address the controversy of the explanation of the evolutionary history of the species *P. algirus*. However, it ended up rising more questions than provided answers. No conclusions on the evolutionary history of *P. algirus* could be drawn from the exhaustive phylogenetic analysis of the four datasets and their combinations. Nevertheless, a new hypothesis is proposed for the explanation of *P. algirus* present phylogeographic patterns, based on a large number of samples. In this case the determination of the migrations order was not only based on a phylogenetic tree, but on the values of nucleotidic diversity and haplotipic diversity coupled with the information provided in the phylogenetic analyses of four datasets from different genes.

Just as it was shown in chapter 2 of this thesis, concatenated datasets can in fact be a great aid in resolving phylogenetic relations, since the resolution of the trees can be greatly improved by adding “size” to the dataset. Some of the relations evident in concatenated trees could not be seen in any of the individual datasets.

In order to fully support the proposed hypotheses, more sampling in a few key areas, such as the Iberia east zone or some previously unsampled areas in Morocco would be required.

Additionally, completing these analyses with information from nuclear genes, could also provide the information that is missing to reach a conclusive outcome and would probably lead to a different phylogenetic history for the species.

Evolutionary Biology and Bioinformatics

Evolutionary genetics do rely a lot on bioinformatics. It is a reality and the association will only become stronger in the future.

In the second chapter of this thesis 9 different software packages were used to conduct the data analysis. More and more, evolutionary biologists are required to learn new sets of skills in the area of informatics. And it's not just learning to use a spreadsheet, or a word processor.

Biologists are required to know how to modify and even create UNICODE datafiles which have to be formatted according to the analysis program input type. This is about learning new languages – those that the programs we use speak.

Who in evolutionary biology has not lost hours of valuable research time figuring out data inputs alone? The field of bioinformatics is very advanced, but it lack the user friendliness that is required for the less computer instructed audience – biologists included. It should not be required of biologists to know the syntax of several programming languages to perform their research.

The recent software BEAST is a dire example of this. The input file has to be loaded in XML format, and so far, no program can replace the manual edition of the input datafile (BEAUti tries to accomplish this task, but it is very limited according to the program's manual).

That is why software such as *Concatenator* is useful. The user only has to click in order to get the desired results. No coding, no text editing, no learning curve. In a future version, this small program is intended to be able to convert between all the file formats used in molecular phylogenetics. Due to it's modular architecture, this change will take some time, but not that much effort. It is the author's hope that the amount of people benefiting from this work increases and that it helps set a new standard for the next generation of data analysis software.

Appendix

Informatic Support

In the provided CD's inside the folder “Concatenator”, the program can be found in three versions:

- Windows Binaries – Ready to use on a system running Microsoft Windows XP (it was not tested on earlier versions of this OS, but should run fine on them; it will not run on Windows Vista).
- Windows Source Code – This version is not compiled and will run on any machine with Perl and the Perl/Tk module. It is optimized for running on windows though.
- Mac OS X and Linux Source Code – Another pre compiled version, but optimized for running on Unix based systems, such as the above mentioned. Will have issues running on a Microsoft Windows system.