Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.
The double-edged complexity of molecular sequence evolution

Thesis submitted for the degree of
Doctor of Philosophy

April 2008

Valentín Ruano Rubio

Smurfit Institute of Genetics
Trinity College Dublin
DECLARATION

This work has not been submitted as an exercise for a degree at this or any other University. It is result of the author's entirely own work unless otherwise stated in relevant parts of the text. The author agrees that the Library, Trinity College Dublin, may lend or copy the thesis upon request. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgment.

Signature ............................................................................................................

Valentín Ruano Rubio

May 2, 2008
Traditionally, phylogenetic studies based on sequence data assume simple models of evolution. This is due to computational and time constraints and to our incomplete understanding of the actual complexity of the evolutionary process. The importance of producing tools and protocols to manage this complexity is twofold: to improve the efficacy of phylogenetic reconstruction methods, and to detect peculiarities of the evolutionary process that may lead to biologically interesting conclusions and predictions.

In the first part of this doctoral work, I focused on the detection and measurement of phylogenetic artifacts due to non-stationary evolutionary patterns that render common models used in phylogenetic inference simplistic. Substitution rate shifts are possible sources of phylogenetic artifacts and they have been subjected to recent controversy regarding relative efficiency of phylogenetic reconstruction methods. First, I conducted simulation studies to determine the likelihood for substitution rate shifts to produce artifactual phylogenies. I contributed to a previous work on this subject by studying more realistic scenarios of subfunctionalisation. This can occur either after gene duplications in multi-gene family phylogenies or following drastic changes in the environment or molecular context in single gene families. Results confirm that bias due to substitution rate shifts should result in attraction of lineages evolving under convergent model parameters. Nevertheless due to the complex multifactorial nature of these artifacts, divergent taxa may occasionally attract each other. I also have developed a semi-graphical test based on triangular geometry to determine the presence of bias in quartets of well-resolved taxon groups.

Later, I investigated the effect of substitution rate shifts and compositional heterogeneity in a real dataset: the SecA gene phylogeny in chloroplasts. Early evolutionary studies on this gene suggested the existence of two separate homologous SecA groups: one including green plants and another red and brown algae. In turn, these observations were used as supporting evidence for a polyphyletic origin for chloroplasts. Now there is strong evidence that supports a single primary endosymbiosis theory. A later study strongly suggested that compositional bias was the cause of the incogruence of the SecA phylogeny. Nevertheless, results indicate that this is in fact not the case and that the most likely explanation is the presence of substitution rate shifts. Here I show that substitution rate shift structure present in the data can explain the apparent polyphyly obtained using distance based methods as was done in the original study. To this end I applied the quartet based method above and fitted models that account for substitution rate shifts using maximum-likelihood. I also used other existent tools to assess the presence of compositional bias and covarion structure and proposed a simple parsimony approach to detect compositional shifts. Additionally, I briefly explored the possibility of using functional divergence to reconstruct trees.
In the second part of this work I used the presence of non-stationary evolution patterns to elucidate important biological aspects of the evolution of gene duplicates. I focused on the analysis of a single multi-gene family, chaperonin homologs in archaeabacteria, and the joint analysis of multiple gene paralogies in yeast.

Chaperonin homologs are found in all three domains of life and are essential due to their primary role in the proper folding of a wide range of proteins. Surprisingly this gene family has evolved very differently across domains. In bacteria the chaperonin system consists of homooligomeric complexes where all subunits are encoded by a single gene. In contrast, eukaryotic chaperonin complexes are heterooligomers encoded by eight genes resulting from a succession of ancient duplications at the base of the eukaryotic tree. Interestingly, archaeal species present a mixed picture where several lineages have independently acquired additional gene copies resulting in gene families of one to five members. In this work, I conducted an in silico approach to test a theory of neutral fixation of hetero-oligomerism proposed by an earlier study. This neutral hypothesis states that hetero-oligomerism in archaeal CCT was fixed without functional gain or loss as opposed to a neo- or subfunctionalisation (with changes in substrate specificity). I performed various evolutionary analyses to test this hypothesis. These include detecting asymmetrical evolution across paralogs, accelerated fixation of amino acid replacements, adaptive evolution and epistasis. The results indicate that this neutral theory explains well the evolution of chaperonin in archael lineages with two paralogous genes. Nevertheless, lineages with more subunits show evidence of functional specialisation of subunits as is observed in eukaryotes.

Finally, I investigated the presence of concerted evolution in a large set of simultaneous gene paralogies. These resulted from a single whole-genome duplication event that took place around a 100Myr ago in the yeast *Saccharomyces* lineage. On the one hand, many evolutionary analysis methods assume that paralogues evolve independently at the sequence level after duplication. Therefore it is important to detect and report the presence of concerted evolution in order to handle convergently evolving copies appropriately. In this study I engineered a simple and effective novel method to detect convergent evolution using pairwise distances. I also made important remarks about possible methodological pitfalls that may result in an increased number of false positive detections. On the other hand, concerted evolution is an interesting aspect of multi-gene family evolution. Therefore I explored possible links between the incidence of concerted evolution and biological aspect of the data. I found interesting connections between the degree of concerted evolution, patterns of gene loss across species, the yeast species analysed and the functional role of genes. In this study I used synteny and viability information available and ontology enrichment analyses.
ACKNOWLEDGMENTS

My most sincere thanks go to my supervisor Dr. Mario A. Fares for giving me the opportunity to work in his laboratory and all his invaluable advice, encouragement, trust and time dedicated to supervise my work.

I would also like to thank current and past members of Mario A. Fares' laboratory for their support and having a good time together: Damien Tully, Frances Smyth, David McNally, Elena Ruiz Zamora, Christina Toft, Aisling Ní Ruairc, Dr. Simon Travers, Dr. Francisco Codoñer, Dr. Laura Madrigal, Dr. Jennifer Commins, Tom Williams and Xiaowei Jiang.

Especial thanks to Dr. Jennifer Commins and Tom Williams for proof-reading this thesis and also thesis examiners, Dr. Eladio Barrio and Prof. Kenneth H. Wolfe, and Dr. Simon Travers, Dr. Davide Pisani, Prof. Dan Bradley, and other anonymous referees that reviewed publications integrated in this thesis for their critical review and constructive suggestions.

Also I would like to thank all programmers around the world that have unselfishly dedicated part of their lifes to develope Free¹ and Open Source software used in this study.

Contents in chapter 2 and 4 are adapted from previously published work (Syst. Biol., 2007, 56:68-82) and (Mol. Biol. EvoL, 2007, 24:1384-96) respectively. I am first author in both publications.

This work has been supported by Science Foundation Ireland and National Development Plan.

¹ "... free as in 'free speech', not as in 'free beer'." Free Software Fundation.
Dedico este trabajo a mi familia y amigos
Contents

1 Introduction .................................................. 1
  1.1 Stationary molecular evolution ......................... 1
    1.1.1 Briefly: molecular evolution and some key concepts .. 1
      1.1.1.1 Definition and genesis .......................... 1
      1.1.1.2 Molecular sequence data: nucleotides, amino acids and codons . 2
      1.1.1.3 Mutagenesis and selection ..................... 2
      1.1.1.4 Gain of biological functions through gene duplication . 4
    1.1.2 Phylogenetics ......................................... 5
      1.1.2.1 Definition and graphical representations ....... 5
      1.1.2.2 Phylogeny reconstruction in general .......... 7
      1.1.2.3 Optimal tree search criteria .................. 7
      1.1.2.4 Fast tree search algorithms and heuristics .... 11
      1.1.2.5 Assessing the goodness-of-fit of phylogenetic trees .... 12
    1.1.3 Mainstream (quasi-)stationary models of molecular evolution ........ 13
      1.1.3.1 Why independent, identical and stationarily evolving sites? .... 13
      1.1.3.2 Evolutionary models based on time homogeneous Markovian pro-
                  cesses ........................................... 14
      1.1.3.3 Stationary residue frequencies .................. 15
      1.1.3.4 DNA and protein models ........................ 15
      1.1.3.5 Codon-based models ............................ 16
      1.1.3.6 Heterogeneous substitution rates across sites ........ 18
      1.1.3.7 Heterogeneous substitution rates across lineages .... 19
  1.2 Non-stationary and dependent evolution: an inconvenient truth? .......... 19
    1.2.1 The real complexity of molecular sequence evolution .......... 19
      1.2.1.1 Heterogeneous and non-stationary residue frequencies ...... 19
      1.2.1.2 Changes in evolutionary pressure in coding nucleotide sequences 20
      1.2.1.3 Non-independent evolution: epitasis and coevolution ....... 21
CONTENTS

6 General discussion 139

A Genetic code 145
  A.1 The twenty amino acids ..................................................... 145
  A.2 Standard Genetic Code ....................................................... 146

B Evolutionary models and distances formulae 147
  B.1 DNA evolution models ....................................................... 147
  B.2 Codon-based models ........................................................ 148
  B.3 Model-based pairwise distances measures .......................... 148
  B.4 LogDet correction ............................................................ 149

C Covarion bias test equation demonstration 151

D Investigating the origin of the green plastid SecA \( \beta \) homolog 155

E Hetero-oligomerism in Archaeal CCT: Supplementary material 157
  E.1 Full phylogeny of non-redundant archaeal CCT homologs ............. 157
  E.2 Fit comparison for diverse codon-based models ........................ 159

F Least-squares branch length estimation ignoring dubious pairwise distances 161
  F.1 Method explained ............................................................ 161
  F.2 Real control dataset results using GENECOV ........................ 164
  F.3 Numeric results among yeast WGD paralogs .......................... 166
  F.4 GO term enrichment analysis ............................................. 168

Bibliography 175
# List of Figures

1.1 Phylogeny examples ................................................................. 6  
1.2 Example unrooted quartet tree ..................................................... 10  
1.3 Gu (1999) divergence models ...................................................... 30  

2.1 Simple bipartition mixture model that produces biased phylogenies .......... 40  
2.2 Covarion-like model used for the simulation study .............................. 41  
2.3 Triangular layout used for representing bias visually ............................ 43  
2.4 Covarion-like “bad” bias RELL 3-D plot towards the wrong topology .......... 48  
2.5 Covarion-like “bad” bias effect based on different model parameters .......... 49  
2.6 Covarion-like “bad” bias effect based on asymmetric tree heights .......... 50  
2.7 Covarion-like “good” bias RELL 3-D plot towards the correct topology .......... 51  
2.8 Covarion-like ugly bias RELL 3-D plot towards the wrong topology .......... 52  
2.9 Covarion-like “ugly” bias effect based on asymmetric tree heights .......... 54  
2.10 Example of bias visualisation plots ............................................. 55  
2.11 Bias test sensitivity graphs ....................................................... 57  

3.1 Example of type-I divergence model four taxon groups ........................ 65  
3.2 Consensus maximum likelihood tree of SecA homologs ........................ 69  
3.3 Possible origins of the green plastid β homologous group ....................... 71  
3.4 Principal component analysis of compositional bias among SecA homologous groups 72  
3.5 SecA Phylogeny corrected for compositional bias .................................. 72  
3.6 Parsimony tree of compositional shifts between major SecA groups .......... 74  
3.7 Heterotachous bias analysis among SecA groups ................................ 76  
3.8 Tree-like representation of divergence type-I between SecA groups .......... 78  

4.1 Consensus phylogeny of the archaeal CCT ..................................... 88  
4.2 Three-dimensional localisation of accelerated and positive selected replacements in CCT ............................................................. 96  
4.3 Three-dimensional localisation of coevolution in CCT ........................... 100
LIST OF FIGURES

5.1 Distribution of post-WGD paralogous pairs across species and gene loss patterns 109
5.2 Unrooted quartet necessary for detecting concerted evolution using distances 112
5.3 Rooted linearised quartet tree assumed to calculate the duration of concerted evolution 115
5.4 Unrooted ML tree of yeast species 117
5.5 Empirical cumulative p-value distribution using the control dataset 124
5.6 Summary of numerical results of concerted evolution detection methods applied on nucleotide alignments 126
5.7 Concerted evolution in WGD paralogs across species and gene loss patterns 130
5.8 Box-and-whisker diagram of concert duration in WGD paralogs 131
5.9 Amino acidic pairwise distance profiles between paralogs in yeast species 135

D.1 Extended SecA phylogeny incorporating some bacterial taxon groups close related to Cyanobacteria 156

E.1 Full archaeal CCT phylogeny 158

F.1 Illustrative unrooted tree including post-WGD duplicates and pre-WGD homologous sequences 163
List of Tables

3.1 Divergence type-II analysis between SecA groups ............................................. 74
3.2 Changes in likelihood of the SecA phylogeny when we allow for covarion like evolution .......................................................... 75
3.3 Differences in $\Gamma$ shape parameter $\alpha$ across SecA groups .......................... 77
3.4 Pairwise $\theta$ value matrix between different SecA subgroups ....................... 77
4.1 Detection of accelerated rates of amino acid substitution in CCT ................ 94
4.2 Sites under accelerated evolution in CCT ...................................................... 95
4.3 Sites under adaptive evolution in CCT ......................................................... 98
5.1 Performance summary of Sawyer's approach applied to the empirical control dataset 122
5.2 Summary of the effect of codon usage bias in the number of positive detections of concerted evolution ........................................... 129
5.3 Comparison of $\tau$ distributions across gene loss patterns ......................... 131
5.4 List of GO terms that are significantly over- or under-represented amongst paralogs that have evolved in concert. ....................... 133
6.1 Domains of life and CCT hetero-oligomerism. .............................................. 142
A.1 Amino acid table and Dayhoff recoding groups. ......................................... 145
A.2 Standard genetic code .................................................................................. 146
B.1 Popular DNA models $Q$ matrices ................................................................ 147
E.1 Likelihood ratio test for the comparison between codon-based models ........ 159
F.1 Summary of the performance of Sawyer's approach on the control dataset ..... 165
F.2 Summary of numerical results of concerted evolution detection methods applied on nucleotide alignments .................................... 167
F.3 Full interspecies GO term enrichment analysis ........................................... 168
F.3 Full interspecies GO term enrichment analysis ........................................... 169
F.3 Full interspecies GO term enrichment analysis ........................................... 170
F.4 *Saccharomyces cerevisiae* GO term enrichment analysis .............................................. 170
F.4 *Saccharomyces cerevisiae* GO term enrichment analysis .............................................. 171
F.4 *Saccharomyces cerevisiae* GO term enrichment analysis .............................................. 172
F.5 *Saccharomyces castellii* GO term enrichment analysis .............................................. 172
F.5 *Saccharomyces castellii* GO term enrichment analysis .............................................. 173
F.6 *Candida glabrata* GO term enrichment analysis .................................................. 174
Chapter 1

Introduction

Preface. The introductory section 1.1 gives some theoretical background on molecular evolution, mainstream evolutionary models and phylogenetic methods relevant for this study. In section 1.2 I comment on the real complexity of molecular sequence evolution and its dualistic implications: challenges and opportunities. Finally, in section 1.3 I connect the aspects of molecular evolution discussed in sections 1.1 and 1.2 with the objectives and achievements of the work carried out in the course of this PhD tenure.

1.1 Stationary molecular evolution

1.1.1 Briefly: molecular evolution and some key concepts

1.1.1.1 Definition and genesis

Molecular evolution is an inter-disciplinary field devoted to the study and understanding of the ancestral relationships among species and changes at the molecular level. In the mid-60s, the advent of protein and DNA sequencing allowed its emergence as a result of the intermingling between evolutionary and molecular biology. The fact that molecular data is an abundant source of non-subjective traits common to all forms of life gives molecular evolution an edge in the pursuit of crucial biological endeavours: for example the resolution of the Tree of Life, characterisation of population dynamics, and investigation of forces that dictate genome evolution or the acquisition of new biological functions through gene duplication.

From the very beginning, applied mathematics has had a vital role in the development of molecular evolution. In their pioneer work, Pauling and Zuckerkandl (1962) proposed the pairwise number of molecular differences as an estimator of divergence time between species under the molecular clock hypothesis (MCH). Likewise, Edwards and Cavalli-Sforza (1964) proposed several numerical approaches to reconstruct population phylogenies.
From that point on, new methods or variations on existing ones have continuously emerged in the literature, not only devoted to phylogenetic reconstruction but also to study important aspects of the evolution of molecules such as coevolution, detection of adaptive constraints, modelling of interaction networks and so forth. In this introduction I focus on those methods and evolutionary aspects most relevant to this work.

1.1.1.2 Molecular sequence data: nucleotides, amino acids and codons

Regarding sequence based analyses, molecular evolutionary studies focus on two major types of organic molecules: nucleotide (DNA and RNA) and amino acid (protein) polymers. Four possible bases constitute the alphabet of nucleotide polymers \{A,C,G,T/U\}. In protein-coding nucleic open reading frames (ORF), consecutive tri-nucleotides, or codons, indicate the primary structure of the resulting protein: sorted sequence of amino acids along the protein's main-chain. In the standard genetic code, each of the 61 sense codons is translated into one of the 20 fundamental amino acids. Any of the remaining 3 stop codons mark the end of protein translation (Appendix A.2).

The genetic code is said to be “degenerate” or “redundant” as there are several alternative encoding codons for most amino acids. Accordingly, for every codon each of its three nucleotide positions is:

- **non-degenerate** if any base change at that position induces an amino acid replacement,
- **two-fold degenerate** if only two of the three possible changes induce an amino acid replacement, or
- **four-fold degenerate** if no base change would induce an amino acid replacement.

In turn, base changes or mutations that induce a nucleotide change but do not result in an amino acid replacement are called synonymous or “silent” mutations. In contrast those mutations that result in a change at the codon and amino acid levels are called non-synonymous or replacement mutations. Since most four-fold degenerated sites are found at third codon positions, most mutations at these positions are silent. A reduced number of mutations at first codon positions are also synonymous whereas all sense substitutions at second positions result in an amino acid replacement.

1.1.1.3 Mutagenesis and selection

Mutations can result from copying errors in DNA replication prior to cell division or natural mutagenic mechanisms such as cross-over during meiosis or somatic hypermutation in adaptive immunity. Mutations can also occur due to DNA damage caused by stress or mutagens (e.g.
carcinogenic chemicals or radiation). Mutations are the source of *genetic variation* in populations and therefore provide the raw material for evolution.

Depending on the effect that mutations may have on the organism's fitness, these can be subjected to *positive selection* (the mutation induces an increment of fitness, i.e. advantageous), *neutral evolution* (no fitness change, i.e. neutral) or *purifying selection* (decrement of fitness, i.e. deleterious). However whether a mutation is retained in the population or not also depends on several other factors such as population size, diploidicity, dominance relationship between gene alleles and so forth. Here I discuss briefly only those details of population genetics relevant to this work. A good introductory text on this subject can be found in Graur and Li (1999), and Nei and Kumar (2000). Here I briefly summarise details relevant to this work.

For example, if we consider a simple diploid model population of constant size $N$, the probability to fix a neutral mutation, denoted $P_N$, is inversely proportional to the population size (Kimura, 1962):

$$ P_N = \frac{1}{2N} $$

That is, the larger the population the more difficult to fix a new change. In contrast, advantageous and deleterious mutations are respectively more and less probable to be maintained in the population based on their relative effect on the fitness. If we quantify this effect with a parameter $s$, and the number of individuals that participate in producing the next generation or *effective population size* as $N_e$, the fixation probability can be roughly calculated as:

$$ P_\pm(s) = \frac{1 - e^{-(2N_e s/N)}}{1 - e^{-4N_e s}} $$

Thus in sufficiently extensive populations advantageous mutations are more probably maintained due to positive selection, whereas deleterious mutations are efficiently eliminated:

$$ P_+(s > 0) \gg P_N \gg P_-(s < 0) $$

Nevertheless, neutral and slightly deleterious substitutions can generate the *genetic variation* or *polymorphism* observed because (a) populations may be small or may have been small at certain periods in its recent history (population bottlenecks), and (b) deleterious mutations are much more probable to occur due to the random nature of mutagenesis. Proponents of the *neutral theory of evolution* (Kimura, 1983) believe that most of this variation is dynamic and results from neutral or slightly deleterious mutations due to genetic drift. In contrast, the *selectionist* theory states that most genetic variation is result of positive selection. In practice
both forces are not mutually exclusive and can jointly explain the genetic polymorphism and adaptation observed (Kimura, 1986).

1.1.1.4 Gain of biological functions through gene duplication

Pioneering work in the 50s to the 70s (e.g. Stephens, 1951; Horowitz, 1965; Nei, 1969; Ohno, 1970; Orgel, 1977) set up the foundations for a now well-established view that gene duplication is a main source of molecular functional novelty in genome evolution. This hypothesis was widely popularised by Ohno's textbook *Evolution by gene duplication* where he proposed the neofunctionalisation model of divergent evolution of new pairs of identical paralogous genes resulting of a single gene duplication event (Ohno, 1970).

Under the neofunctionalisation model one of the two copies would stochastically fix non-synonymous changes under neutral evolution; without affecting the fitness of the specimen. In a "business as usual" fashion, the conserved copy would carry out the ancestral locus functions. In most cases these changes should result in pseudogenisation, also called nonfunctionalisation, and subsequent gene loss of the fast-evolving copy (e.g. by insertion of stop codons, frame shifts or silencing changes in associated regulatory elements). Conversely, in a small percentage of cases some of these substitutions would result in new specific interactions and advantageous molecular functions. These could be later optimised by positive evolution.

Hughes (1994; 1999) proposed an alternative subfunctionalisation model (term not introduced until later by Force et al., 1999) in where both copies could initially accumulate a (similar) number of neutral and positively selected substitutions resulting in a repartition and specialisation across copies of the functions carried out by the ancestral locus. Under this model gain of functions precedes gene duplication. Later, Force and colleagues (Force et al., 1999; Lynch and Force, 2000) focused their attention on substitutions in regulatory elements rather than in the coding sequence. They formulated the duplication-degeneration-complementation (DDC) model. Under the DDC model, neutrally selected substitutions in regulatory elements could well result in the complementary expression profiles of both copies. In this case both paralogs would maintain identical ancestral molecular functions but across different biological conditions or locations (organ, tissue or cell compartment).

More recent work has proposed a hybrid neo-subfunctionalisation model where new loci first subfunctionalise to later neofunctionalise (He and Zhang, 2005; Rastogi and Liberles, 2005). Then subfunctionalisation would become a mechanism of new gene preservation for a later creation of new molecular functions. This is in agreement with Hughes' early suggestion that creation of new functions precedes gene duplication rather than be an immediate consequence (Hughes, 1994).

Additionally, on occasions copies evolve convergently or in concert, that is, sequences of inter-
genomic paralogous genes are more similar (even identical) than expected under independent evolution. This can result from sequence homogenizing mechanisms, such as gene conversion or recombination (Drouin et al., 1999; Kellis et al., 2004), or simply due to selective constraints (Nei et al., 2000; Wagner, 2002), for example due to selection for "more of the same product" (Sugino and Innan, 2006) or genetic robustness (Gu et al., 2003; Gu, 2003).

1.1.2 Phylogenetics

1.1.2.1 Definition and graphical representations

In origin, phylogenetics is a key subfield in evolutionary biology committed to unveiling the divergence events that explain the concurrent existence of collections of extant organisms (species or individuals) assuming that all of them derive from a unique ancestor. Nonetheless phylogenetics is also used in the investigation of gene duplications that generated multiple gene families.

The resulting reconstructed evolutionary history or phylogeny is best depicted as a directed graph with no cycles. This is commonly refer as rooted tree (Figure 1.1a) where existent (or extinct) taxa are leaf nodes (those without leaving edges). In contrast, internal nodes represent the most recent common ancestors for all their descendants also known as a clade (set of leaf nodes reachable from that point by following directed edges). Accordingly, they also represent a divergence (split, speciation or duplication) event. Thus, the root of the tree represents the original ancestor of all extant taxa. The edges between nodes are commonly referred as tree branches whereas the edge path leading from the root or some internal node to a particular extant taxa represent its evolutionary history or lineage.

In some studies the location of the tree root is not exactly known, relevant or applicable. In this case the tree is said to be unrooted, and is represented with an undirected cycle-free graph (Figure 1.1b). In an unrooted tree, the extant taxa or leaves are those nodes with a single incident edge. Internal nodes, although they still represent divergence events, cannot be considered as the most recent common ancestor of any subset of taxa unless there is some notion of the whereabouts of the actual root on the tree.

A phylogeny is said to be resolved when all internal nodes are bifurcated: in rooted trees each internal node has always two leaving edges or three incident edges in unrooted phylogenies. In contrast, non-resolved trees have one or more internal multifurcated nodes representing either a genuine (yet uncommon) multiple simultaneous split event or simply lack of resolution to determine the order of bifurcated events that gave rise to lineages bellow that point.

In many occasions, the single common ancestor assumption is not suitable, for example when compelling yet contradictory evidence support different incompatible locations for the tree root. Also the existence of biological processes that make possible the non-vertical inheritance of genetic material such as lateral gene transfer (LGT) or recombination cast doubt on whether the
search for a unique origin makes sense at all (Doolittle and Bapteste, 2007). These scenarios are better represented using potentially independent tree topologies for each trait (e.g. a phenotype, a gene, a molecular regions or even individual molecular sites) or phylogenetic networks (PNs). PNs are graphs with no directed cycles where each node, either internal or leaf, may have several parent nodes (Figure 1.1c). If the location of the root or roots is known the PN becomes a directed acyclic graph (DAG). Thus PN generalise rooted and unrooted phylogenetic trees. PN may not only represent a reticulated evolutionary history but may also model uncertainty or lack of resolution when we assume that there is indeed a unique true underlying phylogenetic tree.

As regarding edges or branches of trees and networks, depending on the nature of the analyses these may or may not be labelled with additional properties of the evolutionary process between adjacent nodes. For example, in studies where we are only interested in the parental relationship among certain species or whether a group of taxa is mono-, poly- or paraphyletic, only the topology (and root location) is of any interest to us. In this case, branches do not hold any additional information apart from support values that indicate the degree of confidence in their existence. In contrast, if we want to find out the historical order of divergence events or which taxon cluster evolved faster than others (asymmetric evolution), branches must be also decorated...
with their length (time, substitution rate or the product of both as discussed below).

Here I will deal no further with PNs and the rest of the text focuses on single cycle-free tree phylogenies. A good introduction to phylogenetic networks can be found in Huber and Moulton (2005).

1.1.2.2 Phylogeny reconstruction in general

Phylogeny reconstruction, as any other kind of inference, is based on two major elements: data and assumptions. These assumptions inspire the development of one or more methods (algorithms and criteria) that, applied to the data, yield a phylogenetic tree and additional affixed metrics (typically branch lengths and supports). Although assumptions may vary substantially from one method to another, they are all based on the idea that common traits may well reflect a conserved ancestral characteristic thus pinpointing kinship between taxa. Nevertheless, most methods also take into account the possibility that convergent evolution may explain some of these apparent homologies; they are not truly the result of common ancestry but coincidental homoplasies.

The data used for phylogenetic inference is diverse in nature: phenotypic traits (morphology, physiology, behaviour, habitat and so forth) or genotypic characters (aligned molecular sequences, restriction sites, allele frequencies and so on). This information is best represented as a table or data matrix where each row represents a taxon (species, specimen, gene copy etc.), each column represents a different trait, and each cell indicates the observed value of the column trait for that row taxon. Unknown or non-existent entries can be noted as such on the data matrix permitting methods to handle them appropriately (e.g. gap and unknown characters in molecular sequences).

Before the development of molecular sequencing, early systematic studies would rely more frequently in phenotypic data. Although this kind of data is still in used, specially in phenetics, nowadays most phylogenetic analyses are based on genetic data, or more precisely DNA and protein multi-sequence alignments. Below I briefly introduce major categories of phylogenetic reconstruction methods relevant to this work and some general aspects of the search and the validation of results.

1.1.2.3 Optimal tree search criteria

Maximum parsimony and compatibility methods. In science and philosophy, the principle of parsimony states the preference for the simplest alternative explanation or hypothesis. In phylogenetics, maximum-parsimony (MP) methods look for the tree topology that minimises the overall number of substitutions, steps, required to explain the differences observed between taxa. Thus MP criterion considers that the true topology is the most economical in terms of
changes required with or without character weighting schemes (Camin and Sokal, 1965; Eck and Dayhoff, 1966; Kluge and Farris, 1969; Farris, 1969; Fitch, 1971; Le Quesne, 1974; Farris, 1978; Golboossik, 1997; Salisbury, 1999). MP methods do not assume (at least not explicitly\(^1\)) a probabilistic model of residue substitution and therefore are regarded as non-parametric. Accordingly, they can simultaneously analyse mixtures of diverse data types including those without a sound probabilistic model (e.g. phenotypic traits and some genotypic traits such as genomic rearrangements, repeats, insertions and deletions).

Another non-parametric approach consists in looking for those tree topologies with the maximum number of compatible characters (Le Quesne, 1969; Estabrook et al., 1976; Meacham, 1981). A character is said to be compatible with a tree topology if its evolution along the tree can be explained without homoplasies. Although not identical, compatibility methods are similar to parsimony since trees that maximise the number of compatible characters require a minimal number of substitutions for each of those characters. Accordingly compatibility may exhibit roughly the same advantages and disadvantages as MP methods.

**Maximum-likelihood approach.** In contrast to MP, maximum-likelihood (ML) explicitly relies on a probabilistic state substitution model (discussed below in section 1.1.3). It consists in the search for the combination of model parameter and topology (the hypothesis) that maximises the likelihood of the data.

The Bayes theorem formula indicates how to calculate the posterior probability of each hypothesis based on the likelihood of the data and the prior probability distribution of plausible hypotheses:

\[
\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{normalizing constant}}
\]

\[
\Pr(T, \Theta | \text{Data}) = \frac{\Pr(T, \Theta | \text{Data})}{\Pr(\text{Data})}
\]

\[
T \equiv \text{tree topology and branch lengths}
\]

\[
\Theta \equiv \text{substitution process parameters}
\]

If we assume a uniform prior, the maximisation of the posterior probability and the likelihood become the same problem. R. A. Fisher introduced maximum-likelihood in early 20th century (reviewed in Aldrich, 1997). It was first applied to molecular sequences by Neyman (1971) and Kashyap and Subas (1974) (cf. Felsenstein, 2004). Nowadays, likelihood approaches are quite popular due to their mathematical soundness and robustness to violations of the probabilistic model assumed.

---

\(^1\)MP objectors argue that in fact the parsimony criterion, as is applied in phylogenetics, does assume that multiple substitution are rare if not absent (slow-evolving sites) and that the data roughly evolves under a molecular clock.
Maximum-likelihood searching relies on common optimisation algorithms such as expectation maximisation (EM) and computational procedures such as hill-climbing, simulated-annealing or genetic programming. ML optimal phylogeny reconstruction presents two main technical issues: how to handle potential existence of local maxima that may prevent convergence to the targeted global maximum and how to efficiently explore a vast parameter space. These not only affect maximum-likelihood but other non-greedy approaches like MP or some distance methods. I discuss them separately below in section 1.1.2.4.

Bayesian approach. In contrast to ML methods, full Bayesian approaches explicitly specify a prior probability distribution for model parameters in order to compute their posterior probability. The application of Bayesian methods in phylogenetics suffered from a major technical setback: the computational cost of calculating the normalizing constant (eq. 1.3) by numerical integration of a vast multi-dimensional parameter space. The development of an effective solution to circumvent this issue eventually permitted the use of Bayesian methods in the field of phylogenetics (Rannala and Yang, 1996; Mau et al., 1999; Larget and Simon, 1999; Huelsenbeck et al., 2001): Markov chain Monte Carlo (MCMC) methods.

MCMC-based approaches compute the posterior distribution by means of time-spaced and repetitive sampling from one or several running Markov chains (see section 1.1.3 for further details on Markovian processes). Although convergence to the desired posterior probability distribution is guaranteed, it is unknown a priori how long the chains must run for. Additionally, MCMC methods allow for different sampling strategies, next chain state proposal functions and prior distribution choice.

In contrast with other methods, Bayesian analyses do not result in a single best solution but a set of observations drawn from the posterior probability distribution of the model parameters. Then, we can use this sample to estimate statistical properties of interest, for example the mean and variance of branch lengths, residue substitution rates or the probability of the monophyly of a group of taxa (percentage of times this group appears in the output sample).

Distance methods. A metric or measure computable between taxa that complies with the four-point condition (FPC) (Buneman, 1971) is a tree-additive distance. If we consider any possible combination of four taxa \( \{a, b, c, d\} \) as illustrated in Figure 1.2, under the FPC:

\[
\begin{align*}
    d_{ab} + d_{cd} &\leq \max \{d_{ac} + d_{bd}, d_{ad} + d_{bc}\} \\
    d_{xy} &\equiv \text{distance between taxa } x \text{ and } y
\end{align*}
\]  

(1.4)

Tree-additive distances, as their own name implies, can be used to reconstruct the phylogeny. These distances are computed based on the content of the data-matrix and collected in a square
symmetric distance matrix.

The simplest measure, $p$-distance ($p$), is the percentage of characters that differ between two taxa (e.g. polymorphisms between DNA or protein sequences). Multiple substitutions may render $p$-distances as underestimates of the actual number of changes between distant taxa thus losing the tree-additive property. This can be corrected by more robust measures based in substitution models as further described in Appendix B.3.

There are several methods to reconstruct the phylogeny from a pairwise distance matrix based on either a specific optimal criterion, such as minimum evolution (ME) and least squares (LS), or a particular clustering algorithm like, for instance, UPGMA$^2$ and neighbour-joining (NJ). Least squares consists of fitting observed distances to those induced by the candidate phylogeny (Fitch and Margoliash, 1967). The optimal hypothesis $T$ (topology and branch lengths) must minimise the sum of squared differences:

$$T_{WLS} = \arg\min_T \left\{ \sum_{ij} w_{ij} (d_{ij} - d'_{ij})^2 \right\}$$

where $d_{ij}$ is the observed pairwise distance and $d'_{ij}$ is the sum of branch lengths ($l_k$) along the path that connects taxa $i$ and $j$ on the tree. Since different distance estimates have different variance, weights $w_{ij}$ should be used to correct for levels of informativeness among pairwise distances: weighted least squares (WLS). The LS criterion directly selects the best fitting branch lengths but not the topology. Thus the tree space is explored using the same techniques as in ML and MP approaches described below in section 1.1.2.4. This is also the case for ME (Kidd and Sgaramella-Zonta, 1971; Rzhetsky and Nei, 1992), a variant of LS that consists of selecting the topology that involves the minimum amount of evolution; it minimises the tree length (sum

---

$^2$acronym for Unweighted Pair Group Method with Arithmetic mean.
of branch lengths):

\[ T_{ME} = \arccosT \left( \sum_k l_k \right) \]  

(1.6)

For each topology branch lengths are constrained to the LS or WLS estimates in order to avoid the optimal but trivial uninformative solution where all branches are equal to zero.

Clustering algorithms like UPGMA and NJ are extremely fast as they do not search across the tree topology space; these greedy algorithms steadily reconstructs the tree, one internal node at each step. UPGMA is only applicable with ultrametric trees (Sokal and Michener, 1958). At each pass it merges the taxon pair with minimal pairwise distance. Then the immediate common ancestor node replaces the pair for the remainder of the run. This is repeated until only one taxon remains (the root).

NJ, although similar to UPGMA, is able to cope with more general non-ultrametric trees (Saitou and Nei, 1987). The main difference between UPGMA and NJ, resides in the criterion to choose the next taxon pair to merge. Although it is difficult to summarise it in a few words, roughly speaking, NJ merges at each step the pair with the last common ancestor furthest from a “synthetic” outgroup taxon. The distance of this outgroup from the pair is calculated as the average pairwise distance to all other taxa. Despite its simplicity, NJ and variants have proved to be quite efficient and perhaps the only real choice for analyses that involve great amounts of data and taxa.

1.1.2.4 Fast tree search algorithms and heuristics

As mentioned above, non-greedy phylogeny reconstruction approaches need to explore a high-dimensional parameter space for the optimal tree topology. Unfortunately, the number of possible topologies increases factorially with the number of taxa \( N \). If we only consider fully resolved unrooted topologies, we can generate a total of:

\[ \frac{(2N - 4)!}{2^{N-2} (N - 2)!} \]  

(1.7)

This means that for only 10 taxa we have over 2 million possible trees and for 20 taxa more than \( 2^{20} \) possibilities. These figures are even greater if we are interested in rooted or unresolved trees.

Consequently, an exact and exhaustive search for the best tree topology becomes impractical very quickly as we add more taxa. Moreover, the expectancy is that most tree topologies are rather far from the optimal solution and just a very small proportion of the topology parameter space is worth “visiting”. Thus phylogeny reconstruction methods employ quicker heuristic based searches to speed up the process. Typically, the more optimistic the heuristic the quicker and less exact the search. There is always going to be a compromise between speed and efficacy considering the amount of data to analyse, computational resources and time.
1 Introduction

In a typical heuristic search, the start-point topology is selected based on more approximate and quick clustering algorithms such as NJ. The optimisation criterion is evaluated for that tree and its “neighbourhood” is explored for a better topology. The tree neighbourhood is the set of other adjacent topologies reachable from that tree by simple branch and subtree rearrangements. For example nearest-neighbour interchange (NNI) consists of swapping around the four substrees delimited by each internal branch (further explained in Felsenstein, 2004). This procedure generates $2(N - 3)$ neighbours at a time. More ambitious approaches expand the size of the neighbourhood increasing search efficacy but also computational time. The search would proceed in this “crawling” fashion until a maximum is found. Repetitive searches using several start-points may ameliorate the negative effect of possible local maxima.

Several studies have proposed other alternative strategies to speed up tree searching. Strimmer and von Haeseler (1996) contributed an approximate likelihood method where they use ML marginal estimators of taxon quartets to work out the whole tree. There are only three possible unrooted topologies for a quartet making an exhaustive search rather quick. The end result is the full phylogeny that best agrees with the quartet topologies. Guindon and Gascuel (2003) proposed another very quick and popular approach implemented in their PhyML program. In their algorithm, tree topology and branch lengths change jointly during the search. Also, algorithm parallelisation speeds up phylogenetic reconstruction when copious computational resources are available (Stamatakis et al., 2005; Keane et al., 2007).

1.1.2.5 Assessing the goodness-of-fit of phylogenetic trees

Among all approaches mentioned above in section 1.1.2.3, Bayesian reconstruction is the only one that generates a multiple tree sample that can directly be used to determine the support or Bayesian posterior probability (BPP) of a clade $X$:

$$\text{BPP}_X = \frac{\text{number occurrences of } X \text{ in sample}}{\text{total size of sample}}$$  \hspace{1cm} (1.8)

In contrast, MP, ML and distance-based methods return a single “best” estimate where only a zero length edge hints the lack of reliability on the existence of that branch. Consequently many theoretical phylogenetic studies have focused on the development of procedures to assess the goodness-of-fit of phylogenies. Here I comment on just a few of the most popular ones including those used in this work. More elaborate but less popular set of solutions consists in defining a tree distance measure and build confidence regions (introduced in Holmes, 2005).

Non-parametric bootstrapping. This is a general procedure to find out properties of sample-based statistics. In phylogenetics, bootstrapping is used to assess the variability or robustness of the topology, individual clades and the variance of some other model parameters such as
branch lengths (Felsenstein, 1985). It consists of repeated phylogenetic reconstructions based on pseudoreplicates of the data-matrix where traits (e.g. molecular sites) are sampled at random with replacement. Then the credibility of a clade is assessed as is done with BPPs (eq. 1.8).

Despite its popularity, non-parametric bootstrapping presents two major inconveniences. On the one hand, bootstrapping is not scalable in computationally intensive methods such as ML where a single reconstruction may already take a significant amount of time. Running analyses in parallel could reduce the total time if many processors are available (Stamatakis et al., 2005). On the other hand, bootstrap support values are not good estimates of the posterior probability of a clade as they tend to be too conservative (Hillis and Bull, 1993).

Bootstrapped site likelihoods. This class of methods uses resampling of site-wise likelihood values of different tree hypotheses. Thus it is only applicable in probabilistic frameworks like in the ML approach. These methods are quite popular as they avoid the computational burden of performing costly reconstruction repeats. For example the RELL test (Kishino and Hasegawa, 1989) is typically used to compare two trees. It counts how many times across replicates each tree has a greater likelihood. Under the null hypothesis 50% is expected. Similarly the SH test performs a correction for multiple tree comparisons (Shimodaira and Hasegawa, 1999).

A short-coming of these approaches is that only a concrete and reduced set of trees can be compared at a time; if we would like to assess the confidence in the existence of a clade, in theory, this would involve comparing all possible trees with or without that clade (eq. 1.7).

1.1.3 Mainstream (quasi-)stationary models of molecular evolution

1.1.3.1 Why independent, identical and stationarily evolving sites?

Stationary evolutionary models are those where the residue substitution process remains invariable through time. This is roughly equivalent to say that evolutionary constraints remain unchanged overtime. This is in general quite improbable if we take into account dynamic factors that certainly influence the evolution of molecular sites, domains, genes or entire genome regions (e.g. epistatic interactions, changes in molecular interaction networks, mutation fixation rates, population size and so forth).

In addition to stationarity, molecular sites are traditionally assumed to evolve independently from each other under an identical\(^3\) model. Therefore any kind of site-specific metric is supposed to be independent and identically distributed (i.i.d.) across sites. This assumption drastically simplifies analyses based on those metrics. For example, ML and Bayesian tree reconstruction methods evaluate the likelihood independently one site at a time speeding up calculations (see

\(^3\)Notice that here "identical" is not synonymous of "homogeneous" (see section 1.1.3.6). Nevertheless the inverse implication is true: a non-identical process across sites is heterogeneous across sites.
1 Introduction

Therefore these models are taken on board due to their mathematical and computational simplicity and the limiting knowledge of the actual complexity of the evolutionary process. We do so hoping that departures from the model assumed at different sites cancel each other out resulting in random statistical noise. Several studies have shown that these models are quite robust to the violation of their assumptions.

In the remaining of section 1.1.3 I give a brief insight into some general aspects of stationary evolutionary models. While base, amino acid, residue or codon are names linked to specific bio-sequence data types, here I refer to them in general terms as character states or simply states. I denote the set of possible states as alphabet or $A$ in equations throughout the text.

1.1.3.2 Evolutionary models based on time homogeneous Markovian processes

Most probabilistic models of molecular evolution are based on Markovian processes where the probability of the state of a site at time $t$ only depends on its state at time $t-1$ (eq. 1.9). That is, the history of transitions at that site does not have any influence. This is what is called the memoryless Markov property. As the number of possible states in biological sequences is finite (4 for DNA/RNA, 20 for amino acids and 61+3 for codons) we can represent the transition probability distribution as a stochastic matrix, $P$, where each element $p_{ij}$ represents the probability that site in state $i$ at an arbitrary instant $t$ is at state $j$ after a unit of time:

$$p_{ij} = \Pr(x^{t+1} = j | x^t = i)$$

Hence, the stochastic probability matrix for an arbitrary time interval $\Delta t$ can be calculated via the exponentiation of $P^{\Delta t}$.

The average number of transitions per time unit is commonly named substitution rate $r$. The actual number of substitutions, $s$, that take place during time interval $\Delta t$ at rate $r$ follows a Poisson distribution:

$$\Pr(s = S) = \frac{(r\Delta t)^S e^{-r\Delta t}}{S!}$$

In practice, unless we are able to employ palaeontological evidence to date evolutionary events, we cannot separate time from substitution rate: slow-evolving and early divergent pairs of sequences look similar to fast-evolving and recently divergent pairs. Consequently we normally work instead with the product of these two quantities, $u$:

$$u = r\Delta t$$
1.1 Stationary molecular evolution

Typically, evolutionary models are specified in terms of the *instantaneous transition rate matrix*, $Q$, rather than the probability matrix $P$ in order to ease literature notation and computational formulation. This $Q$-matrix is in fact conveniently equal to the subtraction $R - I$ where $I$ is identity matrix and $R$ denotes the transition probability between states when a substitution occurs (including the possibility of self-substitution, hence no effective state change). The non-diagonal entries in the $Q$-matrix, $q_{ij}$, indicate the relative instantaneous transition rate from state $i$ to $j$ whereas diagonal entries are adjusted so that each row adds up to zero:

$$
Q = \begin{pmatrix}
-\sum_{j \neq 0} q_{0j} & q_{01} & \cdots & q_{0|\mathcal{A}|} \\
q_{10} & -\sum_{j \neq 1} q_{1j} & \cdots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
q_{|\mathcal{A}|0} & \cdots & \cdots & -\sum_{j \neq |\mathcal{A}|} q_{|\mathcal{A}|j}
\end{pmatrix}
$$

(1.12)

Then, based on equation (1.10), $P^n$ can be calculated from $Q$ as:

$$
P^n = \sum_{s=0}^{\infty} \frac{(R - I)^s u^s}{s!} = \sum_{s=0}^{\infty} \frac{(Q u)^s}{s!} \approx e^{Qu}
$$

(1.13)

In practice the exponent in equation (1.13) is computed using matrix diagonalisation based on a single value decomposition (SVD) of $Q$ saving computational time with respect to the matrix exponentiation $P^n$.

1.1.3.3 Stationary residue frequencies

Typical evolutionary models as described above have an associate constant stationary residue frequency, $\Pi = \{\pi_i | i \in \mathcal{A}\}$. This frequency vector is equal to the equilibrium frequency of the Markovian process:

$$
\Pi = \Pi P^\infty
$$

(1.14)

Therefore, under these models the state frequencies is supposed to be constant across all extant taxa and ancestral (unseen) sequences. Although this assumption may hold well for close related organism this is generally not true and a major source of artifacts in phylogenetic studies (see sections 1.2.1.1 and 1.2.2.3).

1.1.3.4 DNA and protein models

Mainstream evolutionary models differ in the set of contributed parameters to tune the $Q$ matrix entries in order to better fit the data. The simplest and oldest DNA model, Jukes and
Cantor (JC), has no parameters and assumes that all possible mutations are equally probable. Consequently, all base frequencies are also equal (0.25) (Jukes and Cantor, 1969). Later work has progressively added more biological realism. For example, Kimura (1980) included a parameter to account for different rates for transition and transversions\(^4\), \(\kappa\). Hasegawa and coworkers merged Kimura's model with Felsenstein's unequal base frequencies model (Felsenstein, 1981) in a more general and popular HKY model (Hasegawa et al., 1985). At the other end of the spectrum we found the generalised time reversible (GTR) model, the most flexible of all reversible\(^5\) models (Tavare, 1986). \(Q\) matrices for these models can be found in the Appendix (Table B.1).

In contrast, protein sequence analyses rely on a precalculated empirical transition matrix based on a large training dataset. This is due to a larger number of possible residues (20) with more complex biochemical properties. On one hand, a simple Poisson matrix (analogous to JC in DNA) is only accurate when handling close related sequences. On the other hand, the corresponding GTR model \(Q\) matrix would have a total of 190 parameters. Therefore, the estimation of the \(Q\) matrix entries by means of diverse optimisation techniques is computationally intensive and is only applicable to large datasets.

Dayhoff and coworkers (Dayhoff et al., 1972, 1978) published the first empirical protein evolutionary models named Dayhoff or PAM set of matrices. Different powers of that matrix can be used for database searches, such as BLAST, and alignment reconstruction between homologs with different time divergence. Several other studies generated alternative protein empirical models aiming to increase the accuracy for general applications. For example BLOSUM matrices used local ungapped alignments using homologs with certain identity thresholds (Henikoff and Henikoff, 1992). Jones and coworkers used a more complete and modern protein database than the one used by Dayhoff (Jones et al., 1992). Whelan and Goldman (2001) focused on globular proteins using a maximum-likelihood approach. Some designed empirical protein models to deal with more specific and peculiar datasets such as CpREV for plastid encoded proteins (Adachi et al., 2000) or RtREV for retrotranscriptase analysis (Dimmic et al., 2002).

1.1.3.5 Codon-based models

Codon-based evolutionary models treat codons as the evolutionary unit, filling the gap between nucleotide and protein based models; they take into account the biochemical parameters that affect the mutational process at the nucleotide level (e.g. base frequencies, transition versus transversion ratio) in light of the selective pressures that govern the evolution of the resulting functional proteins.

\(^4\)Transversions are mutations between a purine and a pyrimidines (A\(\Rightarrow\)T or G\(\Rightarrow\)C) whereas transitions are changes within purines or within pyridimines (A\(\Rightarrow\)G or T\(\Rightarrow\)C).

\(^5\)A model is said to be reversible when the likelihood of the data does not depend on the position of the root on the tree.
Traditionally, the effect of selection is normalised based on the assumption that synonymous mutations have no effect on the fitness of the carrier because they do not result in an amino acid replacement. Moreover, if we assume that mutagenesis is uniform along the sequence, the synonymous substitution ratio should also be homogeneous across sites. In contrast, non-synonymous mutations may well have an effect of the fitness since they induce a protein amino acid replacement. Therefore, they may be subject to purifying or positive selection.

If we consider the average number of synonymous substitution, $d_S$, and the average number of non-synonymous substitutions, $d_N$, per site, the selective pressure can be quantified as their ratio (Kimura, 1983):

$$\omega = \frac{d_N}{d_S}$$  \hspace{1cm} (1.15)

If $d_S$ is a proxy to the substitution rate under neutral evolution:

- $\omega = 1$ indicates that the gene, region or site is under neutral selection since amino acid replacements do not have any effect; they are fixed neutrally.
- $\omega < 1$ indicates that the gene, region or site is under purifying selection since amino acid replacements are discarded more often than neutral mutations.
- $\omega > 1$ indicates that the gene, region or site is under positive selection since amino acid replacements are fixed more often than neutral mutations.

Several studies have suggested different methods to count the number of synonymous, non-synonymous substitutions and, therefore, the $\omega$ ratio between pairs of taxa with increasing accuracy (Perler et al., 1980; Miyata and Yasunaga, 1980; Nei and Gojobori, 1986; Li et al., 1985; Li, 1993; Pamilo and Bianchi, 1993; Comeron, 1995; Ina, 1995; Moriyama and Powell, 1997; Yang and Nielsen, 2000). A clear limitation of these methods is that they are quite susceptible to saturation and to selection of synonymous sites (underestimation of $d_S$). Additionally, a pairwise $\omega$ cannot be used to tell which lineage of the two is actually under the estimated selective constraint (perhaps both).

First stationary codon-based probabilistic models of evolution were time and space homogeneous: a single immutable $\omega$ value for each gene throughout all lineages and sites (Goldman and Yang, 1994; Muse and Gaut, 1994; Pedersen et al., 1998). Accordingly, these models only allow to detect positive selection if it has been strong and persistent through time (affect the whole sequence or is very intense in a molecular region). Later, a stationary site-heterogeneous model was proposed to determine evolutionary constraints at particular regions or sites although selective pressure is still assumed to be stationary (Nielsen and Yang, 1998; Yang et al., 2000; Suzuki, 2004a).
1.1.3.6 Heterogeneous substitution rates across sites

In protein-coding sequences some sites may evolve faster than others due to contrasting evolutionary constraints. Moreover, this is also true if we compare overall substitution rates of whole molecular regions or genes. For example, this could be due to heterogeneous effect in the three-dimensional conformation of the protein or coevolution interactions between molecular regions (discussed in section 1.2.1.3). Accordingly, evolutionary analyses often allow molecular sites to evolve at different substitution rates. In this case, the probability transition matrix at each site is computed using a different exponent in equation 1.13 although the $Q$ matrix remains invariable across sites:

$$P^{u_i} = e^{u_i Q},$$

where $u_i = r_i \Delta t$ and $r_i$ is the substitution rate at site $i$.

Allowing each site to evolve at its own site-specific rate (SSR) would substantially increase the number of model parameters with its inconveniences: introducing computational complexity while probably maintaining a model with low explanatory power. The most popular solution consists in modelling rate-across-site (RAS) heterogeneity using a reduced or fixed collection of possible substitution rates categories, $C = \{1, 2, \ldots, |C|\}$. The total number of categories is significantly smaller than the number of molecular sites under analysis ($|C| < N$). The discrete $\Gamma$ distribution is the most commonly chosen (Yang, 1994). In this case, a single distribution shape parameter, $\alpha$, is enough to tune the relative substitution rate of each equiprobable category.

The resulting discrete distribution approaches the continuous $\Gamma$ while keeping the average rate equal to 1 (tied scale parameter $\beta = \alpha^{-1}$) so that the average rate, $\bar{u} = \sum_i u_i$, can be directly interpreted as the average number of substitutions per site. Alternative RAS distributions include the log-normal (Olsen, 1987), general discrete distribution (Susko et al., 2003) or a combination of a number of mixed $\Gamma$ distributions (Mayrose et al., 2005).

Additionally some models assume that a proportion of sites, $I$, are invariant, that is, not allowed to accept changes as they are subject to very strict functional constraints. The discrete rate distribution above such as $\Gamma$ and invariants are not mutually exclusive approaches. For example, in literature “WAG” makes reference to the across-site homogeneous WAG model of protein evolution whereas “WAG+$\Gamma$,” “WAG+$\Gamma$” and “WAG+$\Gamma$+$\Gamma$” denote across-site heterogeneous WAG with invariants, with $\Gamma$ distributed site substitution rates and with $\Gamma$ distributed site substitution rates plus invariants respectively. Whatever the RAS heterogeneous model of choice, a site is assumed to have evolved under the same relative substitution rate (constant relative rate $v_i = u_i/\bar{u}$) throughout its evolutionary history. As I discuss later in section 1.2.1.4...
1.2 Non-stationary and dependent evolution: an inconvenient truth?

1.2.1 The real complexity of molecular sequence evolution

As mentioned in section 1.1.3.1 there are many aspects of molecular evolution that render stationary models simplistic. This simplicity facilitates the interpretation of the results while keeping calculations computationally affordable. Nevertheless departures from assumptions may lead to spurious results. Next I discuss some of the main cases of model departures described in the literature.

1.2.1.1 Heterogeneous and non-stationary residue frequencies

Even before the first publications on the main numerical phylogenetic methods based on molecular data, Sueoka already noticed differences in concentration of nucleotide base pairs and amino acids across phyla (Sueoka, 1961, 1962). Several research groups have proposed diverse explanatory theories with different degree of success. Foerstner and colleges have recently reviewed them (Foerstner et al., 2005). These include for example bacterial genome size (Heddi et al., 1998; Moran, 2002; Rocha and Danchin, 2002), metabolic requirements (Rocha and Danchin, 2002), lack of DNA repair mechanisms (Glass et al., 2000), ideal growth temperatures (Musto et al., 2004) or anaerobic lifestyle (Naya et al., 2002). In contrast, using environmental samples Foerstner and colleagues showed correlation between base-pairs concentrations and habitats regardless
of the taxonomical composition of the sample (Foerstner et al., 2005).

Compositional heterogeneity is also found across regions within the same genome (Nekrutenko and Li, 2000; Bernaola-Galván et al., 2004). Interestingly, compositional variation distribution can be itself heterogeneous with constant GC concentration regions of variable length called _isochores_ (Macaya et al., 1976; Thiery et al., 1976; Bernardi et al., 1985). Isochores have been extensively studied in vertebrate genomes and specially warm-blooded animals (Bernardi, 2000; Eyre-Walker and Hurst, 2001). GC-rich isochores seem to have a denser concentration of ORFs. Nevertheless, it remains unclear what is the actual cause for the existence of isochores especially when there is evidence indicating that they are no longer being maintained (Duret et al., 2002; Arndt et al., 2003; Belle et al., 2004). Some studies argued that heterogeneity may result from mutational biases (Wolfe et al., 1989; Filipski, 1987; Fryxell and Zuckerkandl, 2000), natural selection (D’Onofrio et al., 1991, 1999; Bernardi, 2007), or _biased gene conversion_ (BGC) (Eyre-Walker, 1993; Duret et al., 2006).

Differential composition is not only observable at the nucleotide level. Protein and codon sequences also show differential composition between species and genes. Grantham and coworkers were the first to notice that certain synonymous codons were used more frequently than others and already connected this fact to a potential mechanism of expression regulation (Grantham et al., 1980, 1981). Later work showed that effectively these differences affect translational efficiency as highly expressed genes show a “preference” for codons linked to more abundant tRNAs at least in unicellular organisms (Ikemura, 1985; Kanaya et al., 1999; Rocha, 2004). In contrast, on multicellular organism and specially in vertebrates, GC content bias seem to have a greater influence in the synonymous codon and amino acid preferences (Kanaya et al., 2001). Also translation robustness seems to be an important factor in the selection for particular codons. This is specially so in conserved protein regions (and corresponding codon sites) where mistranslation may bear a more severe effect (Akashi, 1994; Hartl et al., 1994; Archetti, 2004). Some other factors that may influence codon usage include DNA methylation (Tazi and Bird, 1990; Kanaya et al., 2001), mRNA stability (Andersson and Kurland, 1990) or horizontal gene transfer (Lawrence and Ochman, 1997).

### 1.2.1.2 Changes in evolutionary pressure in coding nucleotide sequences

Stationary codon-based models assume that selective pressures are invariable (section 1.1.3.5). Nevertheless we expect the presence of short episodic changes in constraints along the evolutionary history of the molecule. Stationary models do not handle this scenario properly. For example, episodic positive selection at a particular region would be averaged out through time (lineages) and space (full molecular sequence). Also occasional changes in other population parameters such as population size, growth rate or migration between sub-population invariably
have an impact (section 1.1.1.3).

In addition, most codon-based probabilistic models assume that synonymous substitutions are fixed under neutrality and changes are uniformly distributed across the sequence. Nevertheless there is an overwhelming body of evidence against this premise (Chamary et al., 2006; Schattner and Diekhans, 2006; Sauna et al., 2007; Parmley and Hurst, 2007). To avoid flawed detection of positive selection based on purifying selection occurring at synonymous substitutions, Pond and Muse proposed a model where synonymous substitution are heterogeneous across sites (Pond and Muse, 2005). Recently, Maryrose and colleagues have contributed an extended model adding the possibility of non-independent evolution between adjacent codons (Mayrose et al., 2007). They argue that this add-on not only makes the model more realistic but also increases robustness of site-specific \( \omega \) estimates.

Additionally, some are exploring the informativeness of the contrasting nature of different amino acid replacements (Wong et al., 2006; Suzuki, 2007). Some amino acid changes are more chemically radical than others thus they could well be subject of different selective pressures. Non-synonymous substitutions, \( d_N \), can be classified into mild and radical substitutions; \( d_C \) and \( d_R \) respectively. The selection for conserved or radical changes can be quantified using ratios between these figures and the number of synonymous substitution \( d_S \):

\[
\omega_{RC} = \frac{d_R}{d_C}, \quad \omega_R = \frac{d_R}{d_S}, \quad \omega_C = \frac{d_C}{d_S}
\]  

\[1.17\]

### 1.2.1.3 Non-independent evolution: epistasis and coevolution

Since they form part of the same biological system, processes that shape the evolution of different inter-specific molecules must be entangled to a certain extent. Moreover, coevolution should be more prevalent between intra-molecular site pairs. Intuitively, coevolution depends on the "distance" between both elements within the system: whether they interact directly or only through mediators (common interacting partners).

**Epistatic** interactions between intra- and inter-gene substitutions may well be subject to selection. For example, if one substitution leads to a possible improvement of a preexisting- or a novel advantageous biological function, a second mutation in a coevolving site can be fixed through adaptive selection. Additionally, **compensatory mutations** (conditional advantageous mutations) may salvage slightly deleterious mutations making the conjugate selectively neutral.

Evolution of paired RNA sites in stem regions is a good example of coevolution (Wheeler and Honeycutt, 1988; Dixon and Hillis, 1993; Muse, 1995; Chen et al., 1999). Maintenance of these structures is crucial for the function of catalytic RNA or expression regulation of mRNA transcripts. Stem regions result from affinity hybridisation between inverse repeats. A change in a nucleotide makes an inverse mutation in the complementary site conditionally advantageous.
The same principle is applicable to residues at interacting motifs or surfaces between or within proteins in order to maintain their functional three-dimensional conformation (Kondrashov et al., 2002; Ortlund et al., 2007). Moreover, interdependences do not only involve direct contacts as replacements at distant sites in the structure can be mutually compensating (Martin et al., 2005; Fares and Travers, 2006).

1.2.1.4 Non-stationary rates over time across sites

As mentioned in section 1.1.3.6, different sites are normally subject of alternative evolutionary constraints resulting in contrasting substitution rates across sites; homologous sites tend to be consistently slower or faster across homologous sequences. Nevertheless, episodic changes in constraints due to diverse circumstances discussed above render this assumption over-simplistic. Moreover, literature in the field suggests that this is the norm, not only in multi-gene families that include paralogies, but also among strictly orthologous sequences.

Alternating changes in substitution rates were first hypothesised by Fitch and Markowitz (1970) who coined the term concomitantly variable codons or, in short, covarions to refer to this phenomenon in coding nucleotide sequences. They proposed a model where substitution rates at codons can vary independently in an auto-correlated fashion. Recently, Lopez and colleagues (Lopez et al., 2002) introduced a more general concept of heterotachy ("different speeds" in Greek) that encompasses any variability of substitution rates across sites overtime without a concrete model. Thus covarions are a particular case of heterotachy. Conversely, traditional models with constant relative site substitution rates are particular cases of homotachy.

Despite Fitch and Markowitz early work, covarions have received little attention in the literature until recently. Tuffley and Steel (1998) brought them back to focus and set-up the mathematical framework for their analysis. They proved that typical RAS models may not generate tree additive distances on data evolving under a covarion model, therefore possibly resulting in phylogenetic inconsistency. Some more recent studies have proposed generalisations of this model for nucleotide and protein analyses (Galtier, 2001; Huelsenbeck, 2002; Galtier and Jean-Marie, 2004; Wang et al., 2007).

Based on Tuffley and Steel (1998) formulae, time heterogeneous character evolution is modelled using two independent time homogeneous Markovian processes. One process controls how sites pass between a change accepting "on" state and an invariant "off" state. A molecular site is only allowed to change when it remains in the "on" state. This Markovian process has two parameters that determine transition ratios between both states: $s_{on\rightarrow off}$ ($s_{10}$) and $s_{off\rightarrow on}$ ($s_{01}$). The instantaneous substitution rate matrix is then:
The second Markovian process models a regular substitution process much like the one used by stationary models (section 1.1.3). Tuffley and Steel noticed that both processes can be conveniently merged into a single time homogeneous Markovian process. The rate-matrix for the joint process, $R_{cov}$, with dimensions $2|A|$ computes as:

$$
R_{cov} = \begin{pmatrix}
R - s_{01}I & s_{01}I \\
s_{10}I & R - s_{10}I
\end{pmatrix}
$$

where $R$ is the transition rate matrix for the second process and $I$ the identity matrix. In this model states have two components: the visible character state (base, amino acid or codon), and the hidden "on" or "off" status. Thus, equation (1.13) can be used to compute the likelihood as it is usually done with regular homotachous models.

Heterotachy is not limited to covarion models as formulated above. Any change in substitution rates per site constitutes heterotachy. Other instances of heterotachous evolution include subfunctionalisation models discussed below in section 1.2.3.2 and some mixed models described in section 1.2.2.4.

### 1.2.2 A certain issue for systematic studies

#### 1.2.2.1 Definition of inconsistency, bias and identifiability

A statistic or estimator is said to be asymptotically consistent if it converges to the real value when more data (independent observations) is added to the analysis. The estimator’s efficiency is quantified in terms of how much data is needed to approach the right answer to a certain accuracy level (inversely proportional to the stochastic error, $e^{-1}$). Accordingly an inconsistent estimator is the one that does not converge to the right value as more data becomes available.

An estimator is biased when the real value does not fall among estimates obtained from independent repetitions (i.e. it is not the estimator’s expectation). Consequently, an inconsistent estimator is always biased. Nevertheless, an estimator can be biased but asymptotically consistent. For example the variance of a sample:

$$
\hat{\sigma}^2 = \sum_i \frac{(x_i - \bar{x})^2}{n}
$$

is a underestimation of the real population’s variance, $\sigma^2$, if the mean $\bar{x}$ is estimated from the population. In this case $n - 1$ would be the actual number of degrees of freedom. Nevertheless $\hat{\sigma}^2$ converges to $\sigma^2$ with a great sample size.
Finally, model or function parameters are said to be identifiable if each possible combination of parameter values generates a different distinguishable result. Thus, inversely, if the identifiable condition is met, there is hope that we can unequivocally estimate the true single set of generator parameter values for any dataset.

A central objective of theoretical work in phylogenetics is to propose criteria and models that approach the real evolutionary process with a sufficient degree of accuracy so that parameter estimators are consistent, unbiased and, if possible, identifiable.

1.2.2.2 Phylogenetic inconsistency and long-branch attraction

Accurate estimation of the topology discrete parameter is paramount for systematic biology studies and also for any other down-stream phylogeny dependent evolutionary analysis. Consequently, the literature in the field of phylogenetics has extensively addressed inconsistency of tree topology estimators.

Felsenstein (1978) and Cavender (1978) were the first to present mathematical arguments demonstrating the inconsistency of MP methods. Felsenstein (1978) showed that under certain circumstances (including non-molecular clock like evolution) the most parsimonious tree converges to a topology where faster evolving lineages are clustered together disregarding actual phylogenetic relationships. This phenomenon is popularly known as long-branch attraction (LBA). Later work found additional cases where parsimony is inconsistent (Hendy and Penny, 1989; Zharkikh and Li, 1993).

ML and Bayesian methods are regarded as more robust than MP and distance methods and can cope with violations of assumptions. Nevertheless they only guarantee consistency when the model is correct. Therefore, in general, they are also susceptible to systematic errors. As a result, whatever the method used, LBA is pointed out as the cause nearly every time putatively non-closely related and relatively longer lineages cluster on the outcome tree. Many studies preventively repeat phylogenetic analyses using several alternative methods in order to work upon a robust consensus topology (e.g. Bellinvia, 2004; Perkins et al., 2007).

Felsenstein’s germinal work marks the beginning of a long lasting controversial debate between parsimony advocates and objectors where the latter are typically ML or Bayesian method proponents. These studies have rarely challenged the possibility of inconsistency of parsimony or probabilistic methods. Instead, they have discussed the preference for a method based in pragmatic (topology reconstruction efficacy) and philosophical grounds (Siddall, 1998; Swofford et al., 2001; Sober, 2004; Kolaczkowski and Thornton, 2004; Philippe et al., 2005; Thornton and Kolaczkowski, 2005; Steel, 2005).

Consequently, there is a great interest in knowing in what circumstances each method may fail or at least perform better than the others (Kolaczkowski and Thornton, 2004; Mossel and
1.2 Non-stationary and dependent evolution: an inconvenient truth?

Vigoda, 2005; Rodríguez-Ezpeleta et al., 2007; Kubatko and Degnan, 2007). Some have pointed out that saturation is in general problematic in phylogenetic analyses and a possible source of bias (Olsen, 1987; Brinkmann and Philippe, 1999). Thus, discarding fast-evolving characters or lineages whenever possible is considered a good practice by many (Brinkmann and Philippe, 1999; Burleigh and Mathews, 2004; Brinkmann et al., 2005; Rodríguez-Ezpeleta et al., 2007).

Additionally, in an attempt to smooth over discrepancies between data and model, thus to avoid phylogenetic artifacts, some studies have designed tests to determine what genes are most appropriate for phylogenetic studies (Brinkmann et al., 2005; Roure et al., 2007) or what probabilistic models are most appropriate for the analysis (Posada and Crandall, 1998; Bollback, 2002; Ané et al., 2005; Keane et al., 2006).

1.2.2.3 Artifacts caused by compositional bias

Stationary models of evolution assume that overall state frequencies do not change and are identical for all extant taxa. As a result, sequences that coincidentally have more similar composition come across as being closer related as this is the only plausible explanation under the stationary model. Moreover, some factors that cause compositional convergence, also intuitively may seem to indicate true phylogenetic kinship; thus potentially resulting in error reinforcement. As discussed in section 1.2.1.1, these factors include similar or identical habitat (Foerstner et al., 2005), anti-codon tRNA pool (Rocha, 2004) and distribution of isochores (Bernardi, 2000).

Some studies in late 80s were the first to report phylogenetic artifacts due to base-pair compositional bias (Saccone et al., 1989; Weisburg et al., 1989). This phenomenon has been extensively studied since then (e.g. Loomis and Smith, 1990; Lockhart et al., 1992; Hasegawa and Hashimoto, 1993; Steel et al., 1993; Collins et al., 1994; Chang and Campbell, 2000; Tarrio et al., 2000, 2001). Many have proposed different methods to avoid these artifacts. Some rightfully suggested that bias could be ameliorated, without explicit correcting methods, using protein data or discarding fast changing third codon positions (Loomis and Smith, 1990; Lockhart et al., 1992). Nevertheless nucleotide and codon bias can affect also amino acids frequencies (Foster et al., 1997; Foster and Hickey, 1999; Singer and Hickey, 2000).

Phillips and Penny (2003) reported that recoding reduces the effect of composition bias in DNA phylogenetic analyses. Typically, recoding of nucleotide data consists in grouping purines (AG) and pyrimidines (TC) in two states single-letter coded R and Y respectively; consequently this method is called RY-recoding. Analogously, in protein analyses the 20 amino acids can be reduced to a smaller number of classes or “bins” based in similar biochemical properties. Some have used preestablished bins such the Dayhoff 6 or Dayhoff 4 recoding® (Embley et al., 2003; Hrdy et al., 2004; Rodríguez-Ezpeleta et al., 2007). Others have proposed methods to estimate

®recoding groups can be found in Table A.1 in the Appendix.
optimal groups dynamically based on an input data alignment and or a substitution rate matrix (Cannata et al., 2002; Kosiol et al., 2004; Susko and Roger, 2007).

Methods that explicitly account for heterogeneous composition include distance-based corrections and full probabilistic model approaches. Independently, Lake (1994) and Steel and coworkers (Steel, 1994; Lockhart et al., 1994) developed very similar pairwise distance formulae to correct for heterogeneous composition across lineages. These are called paralinear and LogDet distances respectively. The LogDet distance between two sequence, \( x \) and \( y \), is proportional to the determinant of the divergence matrix \( J_{xy} \) whose entries represent the joint occurrence of residues in the same site for both sequences. In order to ensure that the LogDet distances equals zero for cases with no state differences, the distance is rescaled using the determinant of the diagonal frequency matrix product \( F_x F_y \) of both sequences. For DNA analysis:

\[
d_{xy} = -\frac{1}{4} \left( \ln |J_{xy}| - \frac{1}{2} \ln |F_x F_y| \right)
\]  

Waddell and colleagues extended this approach in order to accommodate heterogeneous rate across sites by discarding a number of putatively invariant sites. This modification is called Constant Site Removal (CSR) LogDet (Waddell, 1995; Penny et al., 1999). Later, Tamura and Kumar (2002) provided corrections for heterogeneous composition for several model based pairwise distance formulae. They also included a correction for the LogDet distance itself so that it approximates the average number of substitutions per site even if stationary frequencies are not equal (further discussed in Appendix B.4).

Main probabilistic models that account for heterogeneous composition across lineages are devised for nucleotide sequence analyses as it is easier to handle 4 states as compared to 20 or 61 in amino acid and codon sequence data. Galtier and Gouy (1995; 1998) proposed an alternative model in where GC-content can vary between tree branches adding an extra parameter per branch \((2N - 3 \text{ in unrooted trees with } N \text{ taxa})\). However, this model does not account for difference frequencies within purines or pyrimidines that may result from DNA skew due to mutational bias (Lobry, 1996). Yang and Roberts (1995) proposed more general models where each tree branch has its own specific frequencies for all four nucleotides. Nevertheless, this means a notable increase in the number of parameters that does not scale well for large dataset analyses \((6N - 9 \text{ extra parameters in unrooted trees with } N \text{ taxa})\).

Later, Foster (2004) proposed a model that groups branches in a reduced number of classes \(C\) each with its own base frequencies. This binning reduces the number of additional parameters in those datasets where the data-fit improvement does not justify an elevated number of classes \(C \ll 2N - 3\). Recently, Blanquart and Lartillot (2006) tackled the problem of over-parameterisation in a different way. Their model assumes that changes in composition take place at instantaneous events exponentially spaced in time along lineages. They removed the need of
prespecifying the number of classes and allowed for compositional shifts to occur along branches and not necessarily at internal nodes of the tree.

### 1.2.2.4 Artifacts caused by data evolving under mixture models

As discussed in section 1.1.3.6, RAS models may successfully handle substitution rate heterogeneity due to prevailing evolutionary constraints. Nevertheless, the expectancy is that in occasions these constraints may relax or increase. Episodic changes should be specially frequent after gene duplications (Gu, 1999, 2001; Yang and Nielsen, 2002; Abhiman et al., 2006). Therefore, phylogeny reconstruction that assumes a stationary model may potentially be inconsistent. Intuitively taxa with convergent evolutionary constraints would tend to cluster in the outcome phylogeny.

Multiple simulation studies have already reported inconsistency of ML and other methods in this scenario. For example Chang (1996) and, more recently, Kolaczkowski and Thornton (2004) used a simple mixture model where sites could alternatively evolve following two distinct negatively correlated sets of branch lengths (Figure 2.1 on page 40). Takashita and coworkers (Takahata et al., 2005) found a empirical example based on a multi-gene cytoskeletal (actin and tubulin) phylogenetic analysis in protist.

Mixture models can be classified in two major categories depending on how the data is partitioned. On the one hand, if we have a priori non-sequential information that clearly tells apart characters in distinct evolutionary classes we can analyse each class separately and merge results. For instance, in multi-locus analyses we can reconstruct independent trees for each gene and in single gene analyses we can analyse different site categories separately (e.g. intramembrane and cytoplasmic domains, or core and surface sites). On the other hand, when there are not clear data partitions, we may assume that each character has a non-null probability of evolving under each proposed process.

In probabilistic approaches such as ML and Bayesian tree reconstruction the likelihood is calculated by integration over data characters and possible partitions. In its most general form it is computed as:

\[
L_{k_{\text{total}}} = \sum_{k} \prod_{i} w_{ik} L_k(X_i | \Theta_k, T_k)
\]

\(X_i\) \(\equiv\) pattern at site \(i\)

\(\Theta_k\) \(\equiv\) model parameters for partition \(k\)

\(T_k\) \(\equiv\) topology and branch lengths for partition \(k\)

\(w_{ik}\) \(\equiv\) weight or probability of partition \(k\) at site \(i\)

Weights variables, \(w_{ik}\), become additional parameters that characterise the mixture.
In the \textit{a priori} fixed data partition approach, all weights are set to zero except those that correspond to the partition each site belongs to which are set to 1. For example, Pupko and colleagues proposed an "unlinked" likelihood model where several partitions (normally one per gene) are bound to have the same topology but with completely independent sets of branch lengths (Pupko et al., 2002). In many non-fixed partition approaches, the probability of the partition is site-independent thus the number of weights to estimate is reduced to the number of partitions minus 1 where:

\begin{equation}
\forall k : w_{ik} = p_k
\end{equation}

\begin{equation}
p_k \equiv \text{probability of partition } k
\end{equation}

Typical RAS heterogeneous models are a special case where model parameters are the same for all partitions and tree branch length sets are simply scaled version of the same set.

Mixture models can also be defined in terms of alternative \(Q\) matrices and compositional heterogeneity across sites. Likewise, partitioning may also rely on \textit{a priori} information (Goldman et al., 1996, 1998; Koshi and Goldstein, 1998, 2001; Crooks and Brenner, 2005; Gowri-Shankar and Rattray, 2006) or rely on site-independent probabilities (Pagel and Meade, 2004; Lartillot and Philippe, 2004). Recently, Lartillot and colleagues have shown that this kind of mixture model can resolve artifacts in real data caused by traditional homogeneous models (Lartillot et al., 2007).

\subsection*{1.2.2.5 Artifacts caused by heterotachy}

Several studies indicate that heterotachous evolution may also have a negative impact in phylogenetic reconstruction analyses of real data-sets using stationary models. For example, Inagaki and colleagues (Inagaki et al., 2004) showed that in the elongation factor-\(\alpha\) (EF-\(\alpha\)) phylogeny, microsporidian (a fungal phylum) homologs appear basal in the Eukaryota group rather than among other fungi. This is putatively due to a covarion shift artifact. This seems to be also the case for the RNA polymerase-C (rpoC) gene in chloroplast (Lockhart et al., 2006) or the puromycin-sensitive aminopeptidase (psa) complex subunits in dinoflagellates (Shalchian-Tabrizi et al., 2006).

Nevertheless, covarion shifts may well favour the recovery of the correct topology as we intuitively expect that convergent substitution rate changes correlate with phylogenetic relationships. Lockhart and coworkers (Lockhart et al., 1998) reported this phenomenon in the elongation factor Tu (tufA) phylogeny.

Currently there is only a few covarion-like models available in phylogenetic reconstruction computer programs. These include the original model proposed by Tuffley and Steel (1998) and a few more recent generalisations. Galtier (2001) and Huelsenbeck (2002) proposed two...
different ways to mix Tuffley and Steel's covarion model with the RAS heterogeneous Γ(α) model. Recently, Wang and colleagues produced a model implemented for protein analysis that includes those two as special cases (Wang et al., 2007). All these models assume that substitution rate shifts occur at a constant pace across sites and lineages.

1.2.3 The bright side of complex evolution

1.2.3.1 The interesting stuff is in the “nasty” details

The non-stationary complexity of molecular evolutionary processes hampers analyses based on sequence data. This is specially problematic for systematic studies where resolution of the Tree of Life is the central problem. Nevertheless, these peculiarities are key aspects that may well close the wide gap between unidimensional sequence data to biological function and complex system dynamics. In this section I introduce some of this peculiarities including those most relevant to the scientific work carried out and conveyed in this thesis.

1.2.3.2 Inferring the fate of duplicated genes using protein sequences

As discussed in section 1.1.1.4, paralogous copies that result of a gene duplication event are unlikely to evolve under a stationary model due to changes in evolutionary constraints. When subfunctionalisation is not only due to changes in associated regulatory elements (under a strict DDC model), at least one copy's coding sequence should have accepted non-synonymous mutations at a greater rate than usual.

Under a neo- or subfunctionalisation model, these changes can occur not only in fast-evolving sites but also in conserved regions, thus resulting in heterotachous evolutionary patterns. If incidentally some of these changes “bump into” a novel function then this could be refined through positive selection (additional optimal substitutions) and later stabilised by purifying selection.

Molecular clock hypothesis (MCH) tests could detect accelerated evolution and determine whether it affects both or only one copy (asymmetric evolution). Unfortunately, even strict orthologous evolution may not follow a MCH due to lineage specific evolutionary parameters (Ayala, 1999). Some have proposed relaxation of the molecular clock that may overcome this problem (Sanderson, 1997; Thorne et al., 1998; Drummond et al., 2006). Nevertheless, if one suspected of what lineages may be under different substitution rates, one could use specific local clock approaches (Kishino and Hasegawa, 1990; Takezaki et al., 1995; Uyenoyama, 1995; Douzery et al., 2003). For example, the two-cluster test (Takezaki et al., 1995) assesses rate constancy between two paralogs looking at the difference in the average distances for each paralogous group extant sequences to the most recent common ancestral node. Some have used this test to study divergent evolution between fucosyltransferases in vertebrates (Petit et al.,
Gu and colleagues have focused on the development of methods that model subfunctionalisation in particular molecular sites (Gu, 1999, 2001; Gaucher et al., 2001, 2002; Gu, 2006). Some others have developed similar methods (Knudsen and Miyamoto, 2001; Gaucher et al., 2001; Dermitzakis and Clark, 2001; Pupko and Galtier, 2002; Gaucher et al., 2002; Blouin et al., 2003). The compendium of these approaches provides statistical tests to determine the presence of subfunctionalisation, to estimate the percentage of sites involved and to build a site-specific profile that highlights such sites in primary and ternary structure.

Gu (1999) classified sequence or functional divergence in two classes: type-I and type-II divergence (Figure 1.3). Type-I divergent sites exhibit different site-specific relative substitution rates between two or more (paralogous) taxon groups: a substitution rate shift (thus heterotachy). This phenomenon is characteristic of sites that are functionally constrained in one group but have become less or completely unconstrained in others (e.g. Fares and Wolfe, 2003; Knudsen et al., 2003; Suárez-Castillo and García-Arrarás, 2007). In contrast, type-II divergence consists in a compositional shift between groups. That is, different (paralogous) taxon groups present a chemically distinct composition at that particular site. Compositional shift may mark a change in the specificity of the gene such as different interacting partners or substrate (e.g. Knudsen et al., 2003; Lynch et al., 2006; Zheng et al., 2007).
Some others have proposed that changes in the shape parameter of RAS distributions are indicators of subfunctionalisation (Siltberg and Liberles, 2002; Abhiman et al., 2006). Stronger heterogeneity in subtrees in comparison with the whole tree may well indicate the presence of rate substitution shifts due to subfunctionalisation. If we assume a discrete \( \Gamma(\alpha) \) RAS distribution, as described in section 1.1.3.6, then a global shape parameter \( \alpha \) greater than its counterpart for each individual copy group, \( \alpha_1 \) and \( \alpha_2 \), would report subfunctionalisation.

1.2.3.3 Detection of lineage and site specific selective constraints

In orthologous homologies, stationary site-heterogeneous models (see 1.1.3.6) can satisfactorily determine site-specific substitution rates (Mayrose et al., 2005). Similarly, site-heterogeneous codon-based models and the \( \omega \) ratio criterion (described on page 17) estimate selective constraints at each molecular site. For example, stationary models are suitable to detect non-directional selection constraint such as relentless positive selected changes in viral envelope proteins to elude adaptive immune response mechanisms (e.g. Nielsen and Yang, 1998; Yang, 2000) and the “chasing” adaptation of host immunological molecules (e.g. Swanson et al., 2001; Sawyer et al., 2005).

Unfortunately, stationary codon-based models are not appropriate to search for punctual changes in constraints such as directed selection; the signal may be lost in a predominant purifying selection background due to a global \( \omega \) averaged across sites or branches. Extending the repertoire of codon-based models introduced in section 1.1.3.5, Yang and Nielsen (Yang and Nielsen, 1998; Yang, 1998) proposed a time heterogeneous model where different branches of the phylogenetic tree can be subject to divergent selective constraints. Under this model each branch can have its own \( \omega \) ratio. Nevertheless, using a priori information, branches may be grouped in a discrete number of classes with common \( \omega \). The same authors formulated a branch-site model where some sites are allowed to have a divergent \( \omega \) ratio at selected branches on the phylogeny (Yang and Nielsen, 2002; Zhang et al., 2005).

Later work by Forsberg and Christiansen (2003), and Bielawski and Yang (2004) proposed a model where selection pressure may change in some sites across clades (full subtrees). This reassembles Gu (2001) substitution rate shift model for protein analyses described above but detecting selective shifts instead. More recently, Guindon and colleagues elaborated a covarion-like approach in order to handle non-stationary \( \omega \) values at sites and branches simultaneously without falling in over-parameterisation (Guindon et al., 2004). In their model, at each site the \( \omega \) value changes along the tree following a Markovian process similar to the one described by Galtier (2001).

Alternatively, some studies put forward counting methods able to determine site and branch specific \( \omega_{ib} \) values without the computational burden of ML and Bayesian inference. Counting
1 Introduction

methods reconstruct ancestral states using parsimony (Suzuki and Gojobori, 1999; Fares et al., 2002; Berglund et al., 2005) or distance-based methods (Suzuki, 2004a). Then, each site and branch specific $\omega_{ij}$ values are quantified using synonymous and non-synonymous changes at that site $(i)$ between sequences at both ends of that branch $(j)$.

However, when estimating selection constraints in specific positions, high variability of single codon distance estimates can lead to false positives due to stochastic error. In order to avoid this issue, Fares and colleagues (Fares et al., 2002) proposed a method based in sliding a window of optimal size when looking for selective constraints. In the same line, Suzuki (2004b) and Berglund and colleagues (Berglund et al., 2005) considered the effect of structural proximity between residues by sliding a tree-dimensional window (or "bubble") instead.

These models and methods have successfully detected functional and evolutionary shifts in a uncountable number of occasions. For example, Levasseur and colleagues (Levasseur et al., 2006) used them to link the occurrence of adaptive evolution to environmental and functional changes in fungal lipase- and feruloyl esterases (ester link cleaving enzymes). In this case positive selection was only evident using the branch-site model (Yang and Nielsen, 2002) at branches right after gene duplication. Semple and colleagues used a combination of probabilistic and counting methods to determine the selective pressures that have shaped the evolution of $\beta$-defensins in mammalian genomes (Semple et al., 2005).

1.2.3.4 Convergent evolution of homologous molecules

Typically, after gene duplication we assume that paralogous molecules follow an independent substitution process. Thus, only a relatively small number of convergent substitutions may occur in both lineages resulting in homoplasies. Nevertheless the observed level of sequence identity is often superior to the expected at random. Although this could result from using an over-simplistic evolutionary model (e.g. ignoring the existence of invariant sites), occasionally convergence may be due to biologically relevant evolutionary processes that jointly shape the evolution of paralogous sequences. For example, sequence homogenisation through concerted evolution among intra-specific paralogs is not an uncommon phenomenon.

The term concerted evolution was coined by Zimmer and colleagues (1980) to identify the tendency of multi-gene families inter-specific homologs to evolve "in unison". They pointed out intra-genomic recombination mechanisms, such as gene conversion and unequal cross-over, as the direct cause of concerted evolution. Later, Dover and colleagues encompassed these mechanisms together with population genetic processes into a unified concept that they termed "molecular drive" (Dover et al., 1982a,b). Nevertheless, concerted evolution like patterns may arise from other aspects of multi-gene family evolution (Nei et al., 2000). Selective pressures may also couple the evolution of paralogous genes. Thus concerted evolution could indicate the preservation of
1.2 NON-STATIONARY AND DEPENDENT EVOLUTION: AN INCONVENIENT TRUTH?

Genetic redundancy in order to increase gene dosage (Sugino and Innan, 2006). Additionally, the reciprocal lost of alternative redundant genes may be responsible of reproductive isolation and consequent speciation (Scannell et al., 2006).

Between orthologous sequences, species barriers do make more difficult that homologs could have an influence over each other's evolution. Nevertheless since homologous sites and regions, at least initially, carry out identical biological functions their evolution should show certain parallelism. For example homologous amino acid sites tend to have similar composition due to purifying selection on the conservation of their chemical properties and structural context. Nonetheless it has been shown that directional selection may be responsible for sequence and function convergence emphasising the importance of some key amino acid replacement in shaping molecular function. For instance, Stewart and coworkers noticed convergent amino acid changes at key sites of the lysozyme protein sequence (Stewart and Wilson, 1987; Stewart et al., 1987). These substitutions parallel the recurrent appearance of the fore-gut fermentation in mammals and birds. Later evolutionary studies confirmed this observation using statistical tests (Zhang and Kumar, 1997) and connected it to directional positive selection (Yang, 1998; Creevey and McInerney, 2002; Yang and Nielsen, 2002).

There are several methods to detect convergent evolution, either between paralogous or homologous sequences. These are rarely based in probabilistic sequence evolution models due to the complexity of the nature of converging forces. Exceptionaly, Sugino and Innan have recently proposed a model to determine the duration of concerted evolution of simultaneously duplicated genes (Sugino and Innan, 2006). Drouin and coworkers described a protocol for the characterisation of gene conversion including a review of many methods contributed to detect gene conversion and recombination (Drouin et al., 1999). A very reduced number of tests is available for detecting convergent evolution (Goldman, 1993; Zhang and Kumar, 1997; Kitazoe et al., 2005).

1.2.3.5 Functional and interaction prediction through coevolution

As discussed in section 1.2.1.3, molecules seldom evolve totally independent from each other. Moreover, intra-molecular sites should be subject to coevolution quite often due to their epistatic effect on the molecular ternary structure. Nevertheless coevolution of this sort has a limited effect on the efficiency of phylogenetic reconstruction methods when a sufficiently great number of sites is used in analyses (Dixon and Hillis, 1993; Huelsenbeck and Nielsen, 1999). Consequently, coevolutionary sequence patterns, rather than being an inconvenient, are a good source of information to determine interesting facts about the integrative evolution of molecules within biological system. Codoner and Fares (2008) have recently reviewed methods to detect sequence coevolution. The rest of this section synthesises those aspects most relevant to this
work covered in that review.

Researchers have used coevolution as a proxy to detect functional dependences between residues (Fares and Travers, 2006; Travers and Fares, 2007; Travers et al., 2007), to predict molecular interactions (Goh et al., 2000; Ramani and Marcotte, 2003; Ju et al., 2003; Pazos et al., 2005; Kim and Subramaniam, 2006; McNally and Fares, 2007) and protein function (Zheng et al., 2002) among other interesting aspects of molecules.

Many of those studies also contributed new methodologies to tell apart real coevolutionary signal from independent evolutionary patterns. For example, some researchers used reconstructed phylogenies to detect coevolution between proteins (Fryxell, 1996; Goh et al., 2000; Ramani and Marcotte, 2003; Pazos et al., 2005; Kim and Subramaniam, 2006; Zheng et al., 2002). These approaches search for correlation signatures in topology, phylogenetic profiles, branching history or pairwise distances between candidate coevolutionary molecules or domains.

Other methods scan for coevolution between particular sites using non-parametric approaches such as the mutual information content (MIC) method (e.g. Korber et al., 1993; Tillier and Lui, 2003; Martin et al., 2005). The MIC quantifies mutual uncertainty between two variables. If single variable uncertainty or entropy is computed as:

$$H(X) = - \sum_x \Pr(x) \ln \Pr(x)$$

(1.24)

the mutual information (MI) is an index that measures the co-occurrence of variable values:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y)$$

(1.25)

$$H(X,Y) = - \sum_{x,y} \Pr(x,y) \ln \Pr(x,y)$$

where $MI$ is close to zero only under independence. Using MIC we can determine whether two molecular regions or sites (variables) seem to covary compositionally. Alternatively, some other studies have proposed parametric methods to identify compositional correlations (Pollock et al., 1999; Dimmic et al., 2005; Fares and Travers, 2006). For example, Pollock and colleagues proposed a Markov process that explains the joined evolution of pairs of sites (Pollock et al., 1999).

In contrast to composition analysis, some approaches look for covariation of substitution rates. That is, sites that tend to fix changes simultaneously are supposed to coevolve (Pazos et al., 1997; Pazos and Valencia, 2002; Göbel et al., 1994; Fares and Travers, 2006). For example, Fares and Travers (2006) proposed a method that looks for coevolutionary occurrence of replacements using an amino acid similarity scoring matrix. These methods seem more appropriate to
find coevolution that goes beyond direct interactions between residues detecting functional and structural coevolution (Fares and Travers, 2006; Travers and Fares, 2007; Travers et al., 2007).

1.3 Summary, goals and achievements

Under the theory of evolution (Darwin, 1859) homologous molecular sequences cannot be considered independent observations; they represent the outcome of a mutational or substitution process and several duplication or speciation events that trace back to a common unique origin. Failure to acknowledge this attribute of the data may well lead to artifactual conclusions; apparent correlation between diverse features within the data may just be due to overlooked relatedness between sequences. Moreover the evolutionary history of molecules also contributes valuable information to understand their function in the present and to make predictions as to the effect of sequences alterations through genetic engineering. Consequently, the reconstruction of the evolutionary process is a step required previous to any other in silico of multi-taxon sequence studies. A precise knowledge of such a process has a positive impact in the accuracy of results attained by down-stream analyses. This task is the central aim of phylogenetics.

Traditionally, phylogeny inference relies on simple stationary probabilistic models of evolution or other ad hoc criteria. This is due to bound computational and time constraints, and our sketchy understanding of the actual complexity of the evolutionary process. However, it is important to produce tools to “manage” this complexity in order to (a) improve the efficacy of phylogenetic reconstruction methods and (b) detect peculiarities of the evolutionary process that may lead to interesting conclusions and predictions. Thus, the development of more realistic evolutionary models presents an ever-lasting challenge in the advance of phylogenetics.

In this work I conduct research on both aspects of non-stationary evolution. In the first part of this thesis, Chapters 2 and 3, I show possible phylogenetic artifacts resulting from heterotachous evolution when the reconstruction method assumes an ordinary homotachous model. This is of especial interest in analyses involving gene family phylogenies either for the study of the evolutionary history of the family itself or in building the associated species tree. In Chapter 2, I include a simulation study under different heterotachous evolutionary scenarios and propose a statistical tool to determine the presence of phylogenetic bias in subtree quartet phylogenies. Later, in Chapter 3, I apply this test to a real data set of the SecA gene family. This phylogeny was used as an argument supporting a paraphyletic origin of green and red-brown chloroplast as opposed to a now well established single primary plastid endosymbiosis view.

In the second part of this work, Chapters 4 and 5, I use non-stationary and non-independent evolutionary traits to obtain biologically meaningful conclusions about the evolution of two different datasets. In Chapter 4, I indagate in the process that have repetitively given rise to hetero-oligorism of the chaperonin containing T-complex peptide (CCT) in Archaeobacteria
in contrast to strict homo-oligomeric and fully subfunctionalised hetero-oligomeric eubacterial and eukaryotic homologs respectively. I do so through evolutionary constraint analysis looking for adaptive evolution, acceleration and coevolutionary traits. In Chapter 5, I propose a distance-based method for detection of convergent evolution in multi-gene families. I apply it to a well characterised collection of yeast duplicates resulting from a single whole genome duplication event (WGD). I report interesting links between the occurrence of concerted evolution to evolutionary and functional properties of these genes.
Chapter 2

Artifactual Phylogenies Caused by Heterotachy


**Abstract** Despite the advances in understanding molecular evolution, current phylogenetic methods barely take account of a fraction of the complexity of evolution. We are chiefly constrained by our incomplete knowledge of molecular evolutionary processes and the limits of computational power. These limitations lead to the establishment of either biologically simplistic models. Such models may assign high confidence to an incorrect tree (inconsistency). For example, assuming stationary rate-across-site (RAS) in the presence of heterotachy may lead to systematic errors in tree inference. In this work we show possible misleading effects in tree inference when the assumption of constant within-site substitution rates across lineages is violated using maximum likelihood and simulations. More precisely, we show that different degrees of heterotachy may hamper phylogenetic inference when the model assumed is stationary. Finally, we propose a geometry-based approach to visualise and to test for the possible existence of bias due to heterotachy.

2.1 Introduction

There are several factors that may hamper the inference of reliable phylogenies, including insufficient data and oversimplified models. At best, these model deficiencies may lead to random noise that does not result in erroneous tree topologies. However, correlated parameters unaccounted for in the tree inference model can result in systematic biases towards highly supported
erroneous topologies, branch lengths, and other parameter values.

Typically, probabilistic models proposed for the molecular evolution at each site are based on a discrete Markov process where each state represents a different character value (a nucleotide or amino acid) and transitions between states represent a site change fixation. Early methods assumed an identical and independent evolution process for all sites. Uzzell and Corbin (1971) accounted for the evolutionary heterogeneity among sites using rate-across-site (RAS) models. In these RAS models, site-specific relative substitution rates are tied under a parametric sampling distribution, usually a gamma $\Gamma(\alpha, \beta)$ distribution. Despite their usefulness to approximate empirical data, RAS models are simplistic, as they do not take into account changes in the functional, environmental, and selective constraints acting on amino acid sites of a protein over time.

Covarion-like models were among the first to account for variations in evolutionary constraints among lineages. Fitch and Markowitz (1970) coined the term covarion, but covarion-like models were not given much attention until recently (e.g. Tuffley and Steel, 1998; Lopez et al., 1999; Galtier, 2001; Penny et al., 2001; Huelsenbeck, 2002; Kolaczkowski and Thornton, 2004; Philippe et al., 2005; Ané et al., 2005). Under these models, two separate Markov processes describe molecular evolution. The first is the typical transition between character states (amino acids for protein, nucleotides for DNA), whereas the second allows sites to switch between on-off states. Sites in the off state will remain constant, whereas those in the on state accept changes following the former process. This covarion model can explain an overall RAS-like heterogeneity, as every site's relative substitution rate would depend on the time interval it has spent in the on state; the longer the time, the higher the substitution rate. It also justifies the presence of extensive heterotachy, or changing substitution rates, for a particular site across lineages that cannot be well explained by RAS models (Lockhart et al., 1998; Lopez et al., 2001; Philippe et al., 2005; Ané et al., 2005).

Extensions to this initial covarion model may convey more accurately the complexity present in real data with a limited set of parameters. For example, Galtier (2001) considered a discrete gamma $\Gamma(\alpha)$ model that adds a fixed instantaneous rate of change, rate-of-rates ($\upsilon$), between substitution rate categories for each lineage on every site. This extended covarion model does not allow, however, for rate-of-rates change between lineages that may be responsible of producing phylogenetic biases.

Currently, most phylogenetic studies use RAS models on simplicity and software availability grounds. Simultaneously, some have proposed methods to determine significant within-site changes in evolutionary substitution rates across lineages (Gu, 1999; Mooers and Holmes, 2000; Gu, 2001; Knudsen and Miyamoto, 2001; Susko et al., 2002). Other authors have carried out work on topologically misleading effects with special focus on shared invariant sites (Lockhart et al., 1998; Susko et al., 2004), on highly divergent or even random sites (Brinkmann and
Philippe, 1999; Pisani, 2004; Susko et al., 2005), or without restriction to any of those two extremes (Chang, 1996; Huelsenbeck, 1998; Inagaki et al., 2004; Kolaczkowski and Thornton, 2004).

Deviations from the RAS assumptions may lead to systematic errors. Chang (1996) showed that taxa tend to cluster together in a phylogeny when they display similarities in the distribution of substitution rates among sites. Susko and colleagues reported that distance correction to account for heterogeneity in distance-based methods might become a one-way trip into long-branch attraction artifacts (Susko et al., 2004). Their results pinpoint a tendency towards long-branch attraction or repulsion, even using maximum likelihood (ML) methods, when the model and parameter values induce systematic misestimation of genetic distances. A covarion-like evolving dataset, as those described by Tuffley and Steel (1998) or Galtier (2001), does not necessarily lead to an incorrect phylogeny. In fact, true sister groups may well present a greater correlation in their relative substitution rates and hence tend to cluster together in the phylogeny. Heterogeneity in the rate-of-rates parameter values among lineages may however lead to mistaken phylogenetic groupings, as results in this study suggest.

Here, we first investigated possible sources of bias from assuming RAS models in applying the maximum likelihood criterion on data that evolved following a covarion-like model of evolution. More precisely, we focused on evolutionary scenarios where lineages differ in the assignation of substitution rate category per site. Then we proposed a tool to visualise and test the presence of such biases.

2.2 Methods

2.2.1 Simulation Model

To reproduce bias caused by changes in evolutionary constraints, we used a similar approach as that used in previous work (Chang, 1996; Kolaczkowski and Thornton, 2004; Spencer et al., 2005; Philippe et al., 2005). However, we introduced important differences regarding the phylogeny used in simulations. The simulation guide tree consisted in a quartet of subtrees rather than a quartet of taxa and we defined a set of coefficients to control the conservation of relative rate categories within-sites at the deepest divergence events (internal nodes) of the tree. We also allowed for different RAS $\Gamma(\alpha)$ distribution across subtrees; divergent shape parameter, $\alpha$, values. In this study we did not consider the effect of data-matrix bipartitions based on different sets of branch lengths as done in other studies (Fig. 2.1); however, heterotachy may result from random independent changes in relative substitution rate category per site and from differences in substitution rate distributions across lineages.

We focused on unrooted phylogenies presenting four well-resolved monophyletic groups (Fig.
Figure 2.1: A simple bipartition mixture model that produces biased phylogenies. A percentage, \( p \), of sites evolves under a set of branch lengths, whereas the rest does so under a complementary set of branch lengths. Non-sister group relative substitution rates are positively correlated, whereas real phylogenetic groups are negatively correlated. This conflicts with an RAS evolutionary process where relative branch lengths must be kept constant across all sites.

2.2. Each group is composed of a number of taxa and evolves under a standard RAS model. The shape of each subtree is controlled by three parameters (Fig. 2.2c): the total height of the tree, \( h \), the number of divergence events along every lineage, \( d \) (thus the total number of leaves is \( 2^d \)), and the relative branch length ratio, \( r \), that modulates the time intervals between divergence events. The motivation behind considering subtrees rather than the typical unrooted four-taxa phylogeny was twofold: first, to generate enough data (extant taxa) to infer back the substitution rate for each site and subtree and with that its relative substitution rate in that clade, and second, to record the influence of taxa sampling on the bias strength.

As we anticipated above, we modelled substitution rate heterogeneity across sites using a discrete \( \Gamma(\alpha) \) distribution with \( C = 16 \) rate categories. Starting from a standard RAS model, where each site belongs to the same rate category throughout the tree, we added the possibility of instantaneous substitution rate category changes per site on the two most internal nodes or divergence events, that led to the four well-supported subtrees (labelled \( a \) and \( b \) in Figure 2.2b). On both nodes we defined three transition probability matrices \( P_{\alpha \alpha}, P_{\alpha \beta}, \) and \( P_{\alpha \gamma} \), one per each incident branch. Each matrix element \( p_{ij} \) indicates the probability for a site currently evolving under substitution rate category \( i \) on that node to change to category \( j \) for the corresponding incident branch and the subtree beyond (except for the subtree quartet’s internal branch). Since we assumed a discrete \( \Gamma \) distribution for substitution rates across sites, all matrices must yield equally probable stationary category frequency vectors \( F = \{C^{-1}\}_C \). We further constrained the family of valid matrices in order to be able to parameterise each with a single tuning value, \( \theta \) or substitution rate category conservation coefficient (Fig. 2.2a, Fig. 2.2b). \( \theta \) indicates the
2.2 Methods

Figure 2.2: Covarion-like model used for the simulation study. (a) Site relative substitution rate categories may change across the two innermost nodes of the tree. Conservation coefficients $\theta_{na}$, $\theta_{n\beta}$ and $\theta_{n\gamma}$ represent the substitution rate category transition probabilities between the node (n) and each branch and beyond. (b) The resulting overall model is a quartet, with four resolved RAS processes and six conservation coefficients. We collapsed coefficients at the inner branch into one, $\theta_{ab}$, being $\theta_{a1}$ and $\theta_{b1}$ equal to $\sqrt{\theta_{ab}}$ for simplicity. (c) Each subtree shape is determined by three parameters: $h$ indicates the total height of the tree, $r$ the ratio between branch length after and before each divergence event, and $d$ the number of divergence events from root to leaves. This figure illustrates a subtree with $d = 3$, therefore with 8 taxa. The branch length nuisance variable $l$ is set so that the sum $l + lr + lr^2 + \ldots + lr^d$ matches the total height $h$.

Proportion of sites whose substitution rate category remains unchanged between incident branch and node, whereas, $1 - \theta$, indicates the percentage of sites that is susceptible to change to another substitution rate category. In the latter case, the new category is selected at random from all $C$ possible categories with replacement. Thus the resulting transition matrix is:

$$P_{ij} = \begin{bmatrix} \theta''_{xy} & \theta'_{xy} & \ldots & \theta'_{xy} \\ \theta'_{xy} & \theta''_{xy} & \ldots & \theta'_{xy} \\ \vdots & \vdots & \ddots & \vdots \\ \theta'_{xy} & \ldots & \theta'_{xy} & \theta''_{xy} \end{bmatrix}$$

(2.1)

where:

$$\theta''_{xy} = C^{-1}(1 - \theta_{xy}), \forall xy : \theta_{xy} \in [0, 1]$$

(2.2)

$$\theta'_{xy} = \theta_{xy} + \theta''_{xy}$$

Under these restrictions, reversion and transitivity are trivial because $P_{xy}$ is symmetric:

$$P_{xy} = P_{yx}, \theta_{xy} = \theta_{yx}$$

(2.3)

$$P_{xz} = P_{xy}P_{yz}, \theta_{xz} = \theta_{xy}\theta_{yz}$$
where \( z \) is reachable from \( x \) through node \( y \). We based our simulations on a JTT (Jones et al., 1992) Markov process. Nonetheless, qualitative conclusions are extensible to other evolutionary models.

### 2.2.2 Simulation case studies

Due to computational limitations we only investigated covarion bias effects on three different parametric scenarios: (a) presence of changes in substitution rate categories that are equally frequent in non-sister subtrees, (b) presence of changes in rate categories in the inner-most branch connecting phylogenetic pairs of subtrees; and (c) existence of different site substitution rate sampling distributions. Each parameter value combination was repeated 50 to 100 times to reduce stochasticity. Unless otherwise indicated, alignments were 1000 amino acids long. We utilised the program CODEML from the PAML package version 3.14 (Yang, 1994) for evaluating the phylogenetic support based on RELL bootstrap proportions (Kishino and Hasegawa, 1989). For each dataset we compared the support obtained by the three possible unrooted trees resulting from rearranging the four subtrees; we assume that all four subtrees are monophyletic and their topologies are known. Parameter values were specific for each simulation batch. RELL support values plotted in diagrams are the mean among repetitions. Notice that this mean is not equivalent the percentage of cases where that associated tree topology is the one that maximises the likelihood, which is the statistic typically reported in this kind of studies. We used the R statistical package to produce all diagrams, the locfit module to fit surfaces in three-dimensional RELL support plots, and the geneplotter module for the triangular plots in Figure 2.10.

### 2.2.3 Geometrical bias visualisation

Intuitively, sites that may introduce bias towards a wrong hypothesis (topology) are those for which two selected non-sister subtree pairs have similar relative substitution rates and different across pairs (Fig. 2.1). Therefore, assuming that we have enough data to accurately estimate substitution rates per site per subtree, we can classify sites according to the hypothesis they favour due to substitution rate shift bias. We expect that a proportion of sites supports arbitrary topologies due to stochasticity alone. However, their overall effect must cancel out and result in an unbiased sample.

There are three possible topologies for a fully resolved subtree unrooted quarted phylogeny. A simple way to represent graphically these hypotheses is using the equidistant vertices of a regular triangle (Fig. 2.3). Others have used triangle geometry previously in a phylogenetic context although with a different purpose (Strimmer and von Haeseler, 1997; Kim, 2000). Within this layout, impartial sites (those not favouring a particular hypothesis) would be distributed around to the middle point, the barycentre, whereas biased or partial sites would accumulate towards
2.2 Methods

Figure 2.3: Triangular layout used for representing bias visually. The three equidistant vertices represent the three possible unrooted topologies. The distribution of sites that show the same relative substitution rates for all subtrees should be centred on the barycentre of the triangle, whereas different degrees of site substitution rate correlation between subtrees should result in greater concentration of points located closer to one or two of the topologies represented by the triangle vertices.

Peripheral areas close to a vertex or edge depending on which hypothesis or hypotheses they do favour.

In order to compare substitution rates per site between different subtrees, we used empirical Bayesian estimates as a proxy (Mayrose et al., 2004). We calculated these for each subtree independently. For this calculation in particular we assumed a Poisson model (Jukes and Cantor, 1969) rather than the actual JTT rate-matrix because using an empirical matrix for low evolving sites would yield rates highly dependent on site composition. Here it is not as critical to obtain accurate estimators as to produce a reasonable order that is not sensitive to convergent site composition. For each subtree we ranked sites by their substitution rate estimator, resolving ties by randomisation. Thus, we defined $R_x(i)$ as the rank for site $i$ on the subtree $x$ taking values between 0 and 1 in steps of $1/($number of sites$)$.

2.2.4 Plotting individual sites

Under the null hypothesis (RAS model) all subtrees should consistently have equal relative substitution rates at each site across lineages; that is, a fast-evolving site on a lineage has to be proportionally fast evolving in all other lineages. Consequently, independent ranking based on relative substitution rates carried out in different subtrees should yield a similar order (positively correlated across trees) depending on the variance of the substitution rate estimator used. Nonetheless, in non-RAS processes, as the covarion-like process described above, within-site substitution rate changes across groups break to certain degree this correspondence rendering site rankings less correlated. Subtrees with a more coincidental distribution of substitution rate

---

$^1$When subtree height differences are pronounced, Bayesian estimates on data simulated based under empirical models tend to converge for slow evolving taxon groups.
categories across sites should yield more strongly correlated site rankings.

Based on previous observations, lineages with more similar evolutionary processes should tend to attract each other. In this context, we may say that a site "prefers" a hypothesis \(((x, y), v, w)\) if for that site, ranks on subtrees \(x, y\) and \(v, w\) are closer within pairs and different between pairs. A way to reflect this in quantitative terms is that a site will prefer a particular hypothesis if \(|R_x(i) - R_y(i)|\) and \(|R_v(i) - R_w(i)|\) are small (close to zero) and the average \(AVG(|R_x(i) - R_v(i)|, |R_x(i) - R_w(i)|, |R_y(i) - R_v(i)|, |R_y(i) - R_w(i)|)\) is big (close to one).

Instead of determining individual support indices for each site and hypothesis, we calculated coefficients that give to each hypothesis a weight proportional to the ranking similarity of their two hypothetically sister subtree pairs. Following this approach we computed three weight values, with only two degrees of freedom as indicated below.

Let us consider the three hypotheses:

\[
H_0 \equiv ((a_1, a_2), b_1, b_2) \\
H_1 \equiv ((a_1, b_1), a_2, b_2) \\
H_2 \equiv ((a_1, b_2), a_2, b_1)
\]

From each site \(i\) we calculate the value \(F\) that quantifies the amount of relative substitution rate change between two subtrees \((x\) and \(y)\) as:

\[
F(x, y, i) = 1 - |R_x(i) - R_y(i)|
\]  

Small differences in substitution rates (rate conservation or convergence) will yield values close to 1, whereas extreme differences should yield \(F\) values closer to 0. \(F\) values would be directly comparable only if they are estimated from subtrees with equivalent shapes (topology and branch lengths) as distributions of rate estimators used to rank sites depend on subtree model parameters (topology, branch lengths and RAS shape parameter \(a\)). Therefore, in order to normalise ranking differences between all subtrees, we performed parametric simulations (using 1000 replicates) for each substitution rate category based on RAS model assumptions. In these simulations, we used the parameter estimators obtained by ML analysis performed in PAML. Then we substituted the \(F\) values at each site by the complement of their corresponding cumulative probability \(F'\) assuming that its substitution rate category is the one with maximum posterior probability. This will yield uniformly distributed values from 0 to 1.

\[
F'(x, y, i) = \Pr \left( F'_{xy} > F(x, y, i) \right)
\]
Then we evaluated substitution rate changes on a site $i$ for each hypothesis with a second function, $G$, defined as the product of the indices for its two individual phylogenetic subtree pairs:

$$G_{H_0}(i) = F'(a_1, a_2, i)F'(b_1, b_2, i)$$
$$G_{H_1}(i) = F'(a_1, b_1, i)F'(a_2, b_2, i)$$
$$G_{H_2}(i) = F'(a_1, b_2, i)F'(a_2, b_1, i)$$

(2.7)

The resulting hypothesis indices also vary from 1 (total conservation of the substitution rate categories per site within pairs) to 0 (extreme change of the substitution rate categories per site at least within one pair). Then, we defined the weight, $w$, on site $i$ associated to a particular hypothesis $j$ as:

$$w_{H_j}(i) = \frac{G_{H_j}(i)}{\sum_k G_{H_k}(k)}$$

(2.8)

Finally, we used these weights for combining arithmetically the three triangle vertices (hypotheses) in order to plot a single point that represents the site within the triangle. The bidimensional location $(p_i)$ of that point can be calculated as:

$$p_i = \frac{1}{3} (w_{H_0}(i)h_0 + w_{H_1}(i)h_1 + w_{H_2}(i)h_2)$$

(2.9)

Here $h_j$ stands for the two-dimensional coordinates of the vertex that represents hypothesis $H_j$. Notice that we can implement many alternative $G$ or $F$ functions generating different weight coefficients and plots with different statistical properties. Nonetheless, any alternative must be designed as to allow impartial sites to be distribute around the centre of the triangle, whereas partial sites must cluster towards the corners (one hypothesis wins) or the edges (two draw and one loses).

### 2.2.5 Bias test and measurement

We can evaluate the assumptions in the RAS models by testing the goodness-of-fit of the expected to the observed distribution densities of the points in the triangle. Even though this may spot nonstandard samples, it would not automatically indicate which hypothesis benefits from that deviation. Besides, some samples may still be balanced and yet have significantly different plot densities. Alternatively, we could take a look at the mass centre of the plot, that is, the arithmetic mean of all site points.

In the ideal case of equivalent subtrees with the same shape (topology and branch lengths), and assuming a genuine RAS evolution model, the mass centre must be placed on the triangle.
barycentre equidistant from each tree hypothesis (topology) regardless of the presence or lack of any phylogenetic signal. Asymmetric heavy tails must pull this mean towards the winning alternative hypothesis (or further from the losing one). In consequence, one could use this to infer whether the sample is biased with respect to one or more hypotheses. Provided that the distribution of the mean plot point position approaches a bivariate normal distribution, we can readily obtain a simple formula to calculate a confidence area for the sample mass centre as developed in the Appendix C:

\[
\Pr(D > d) = \exp\left(-d^2 \left(2\sigma_x^2\right)^{-1}\right)
\]

where \(D\) designates the Euclidean distance from the sample mean to the barycentre, \(\sigma_x\) is the standard deviation of the sample’s projection on any axis of choice with origin in the barycentre, and \(L\) is the sample size (data-matrix length).

Nonetheless, distinct branch lengths and different number of taxa may affect the distribution of dots on the plot and change the mass centre. This is due to the fact that the \(G\) values for each hypothesis, as described in Methods, are not independent from each other because they combine shared subtree estimators. In practice, however, the test starts to become liberal just with a great amount of data (10,000 sites) and severe tree shape height asymmetry (e.g., when trees present subtrees over 8 times taller, or with 4 times greater number of taxa, than others). A general solution to this issue is to further simulate the expected site plot and mass centre, given the ML estimators of the full phylogeny, and to use the distance to this alternative centre. Even though Equation (2.10) is still a good approximation under such circumstances, sample mean comparison tests, such as the Wilcoxon test, are probably more appropriate.

Another aspect to bear in mind is that since we sort ranking ties by randomisation, each execution can potentially return a different p-value. In order to stabilise the outcome, the test must be repeated a number of times and an averaged p-value must be selected. However, under this procedure the test becomes somewhat conservative. In this work we proceeded without repeats, as results for single datasets are not critical in simulations.

As an alternative to the null hypothesis presented above, datasets can show different degrees of relative substitution rate correlation among lineages that may cause systematic distortions in the resulting topology and branch estimators. Interpretation of the results of such a test is independent of the amount of actual phylogenetic signal (inner branch length). Therefore, results can only point to the possibility of having inferred an artifactual phylogeny by ML under a RAS model.

The simulation and test programs were implemented in the Java programming language utilising functionality from the Java standard library, PAL library (Drummond and Strimmer, 1996).
2.3 Results and Discussion

We investigated the effect of inducing different levels of site substitution rate correlation between subtrees using the simulation model based on four monophyletic groups of taxa $a_1$, $a_2$, $b_1$, and $b_2$ as described in Methods and Figure 2.2. We intuitively expected that lineages with a more similar evolutionary process would attract each other due to a greater similarity of distribution of substitution rates across groups at each site, or alternatively that groups with more divergent evolutionary processes would repel accordingly. Higher pairwise site substitution rate category conservation coefficients $\theta$ (product of individual $\theta$s along connecting nodes) should cluster lineages with increasing confidence as more data becomes available.

2.3.1 The “bad” bias against phylogenetic signal

In the first simulation, we fixed $\theta_{ab}$ (Fig. 1b) on the inner branch connecting the $a$ and $b$ groups to 1 so that there is no intrinsic substitution rate category shuffle along this branch. $\theta_{aa_1} = \theta_{bb_1} = \theta_1$ and $\theta_{aa_2} = \theta_{bb_2} = \theta_2$ were assigned arbitrary values between 0 and 1 in steps of 0.1 (Fig. 2.4a). Initially, for each of these cases we fixed all subtrees total heights, $h$, to a common average of one change per site, resulting in a symmetric and molecular clock-compatible phylogeny. The shape parameter value, $\alpha$, was also set to 1 (i.e., a moderate to high degree of substitution rate heterogeneity among sites). Subtrees were composed of four taxa; the result of two rounds of duplication or speciation events on all lineages ($d = 2$). Each branch was twice as short as the one that leads to the parent node ($r = 2$). The inner branch was set to 0.04 changes per site. This last value was chosen so that our simulated datasets were forced towards the inconsistency parametric edge. Consequently, relatively small changes in parameter values can make the results wander from yielding the right topology to the wrong one with increasing probability as the size of the dataset increases. Under these conditions, the bootstrap support for the right topology following an RAS model of evolution was 80% out of 100 replicates.

A possible biological interpretation of such an arrangement is that smaller $\theta$s indicate a greater dissimilarity between evolutionary constraints on different lineages. For instance, paralogs within a gene family with distinct functions would show a similar degree of disparity in within-site evolutionary rates. If $\theta_1 = \theta_2$, there is the same degree of covariance between all subtrees. Consequently, we expect to obtain a roughly random type of error with no systematic bias. However, if two lineages have a higher degree of substitution rate category conservation; for example, when the two paralogous groups of sequences are functionally redundant, there will
Figure 2.4: Plot of the mean RELL support value for the wrong topology \(((a_1, b_1), a_2, b_2)\). (a) Simulated process: the conservation coefficient on the inner branch was fixed to 1 and the same values were assigned to non-sister subtrees: \(\theta_1\) for \(a_1\) and \(b_1\), and \(\theta_2\) for \(a_2\) and \(b_2\), respectively. Tree heights are set to 1 substitution per site, whereas the innermost branch length is set to 0.04. There are a total of four taxa per each group \((d = 2)\) and each branch is twice as long as the next branch towards the tree leaves \((r = 2)\). Rates are drawn from a gamma distribution \(\Gamma(\alpha)\) for all subtrees. (a) A plot of the mean RELL support value (z-axis) and the conservation coefficients \(\theta_1\) (x-axis) and \(\theta_2\) (y-axis). For this particular set of parameters, at very low conservation distance, \(|\theta_1 - \theta_2|\), the mean RELL value approaches a constant low support value (expected under a RAS model), whereas it increases considerably when the difference is extreme (= 1).

be a higher correlation of site rates and a stronger attraction under non-covarion-aware models between these two lineages. The opposite can be also true for the two lineages with lower site substitution rate category conservation. Thus we would expect to see wrong topology resolution as the difference between \(\theta_1\) and \(\theta_2\) increases. In fact, our simulations, based on a symmetrical phylogeny where all subtrees have the same height, show that such an increase leads from reasonable support for the true topology towards false support for the hypothesis in which lineages with greater site substitution rate conservation appear together (Fig. 2.4b).

Further, we took a particular combination of \(\theta\) values \((\theta_1 = 1, \theta_2 = 0.2)\) and explored the effect of changing values for key characteristics of the input phylogeny at different datasets sizes (Fig. 2.5). As expected, the increase of the amount of phylogenetic signal (inner branch length) neutralises systematic error effects. On the other hand, the increase of subtree height does reinforce systematic error because of the loss of phylogenetic signal in the innermost branch. Covarion-like bias relies on the existence of heterogeneity of substitution rates among sites because no site rate categories would be considered otherwise. Accordingly, a lower shape parameter value (more severe heterogeneity) favours resolution for the wrong topology, whereas a higher value does just the opposite. Besides, proper taxon sampling may reduce the effect of bias. In fact, adding more taxa reduces the bias strength. Short deep and long-tip branches seem to be less problematic than long deep and short-tip branches are. Thus, we can conclude that the characteristics of the evolutionary process at the subtrees have a remarkable effect on
2.3 Results and Discussion

Figure 2.5: Effect that changes in simulation parameters have on the RELL support for the three possible topologies. We investigate the marginal effect of five different parameters (one per each diagram column), fixing all other parameters to the values used in the surface plot in Figure 2.4: inner branch length \( \{0.01, 0.02, 0.04, 0.08, 0.16\} \), shape parameter \( (\alpha) \) \( \{0.05, 0.25, 0.5, 1, 1.5, 2\} \), subtree common height \( (h) \) \( \{0.25, 0.5, 0.75, 1, 1.5, 2\} \), branch length factor \( (r) \) \( \{0.25, 0.5, 1, 2, 4, 8\} \), and number of taxa per subtree \( (d^2) \) \( \{1, 2, 4, 8, 16\} \). We tested these changes on a particular combination of conservation coefficients \( \theta_1 = 1 \) and \( \theta_2 = 0.2 \) situated at an average height on the plotted slope in Figure 2.4. These values are close to the inconsistency critical region. For each combination of values, we took the mean RELL support of 50 replicates for all three possible topologies. We performed the experiment for a number of different alignment lengths, although here we show results just for 1000, 3000, and 10,000 sites. There are no data for 10,000 sites and 16 taxa per subtree because it was computationally prohibitive.

phylogeny recovery because long deep branches seem to have a greater impact, something that is in agreement with previous studies (Kim, 1996, 1998; Bremer et al., 1999).

Finally, within-site rate heterogeneity among lineages can also alter the outcome significantly. Among all alternative combinations studied, only the ones that elongate or shorten lineages together with the same \( \theta \) values have a pronounced effect in comparison with the results obtained with the symmetric settings (Fig. 2.6). In fact, bias increases when lineages with high \( \theta \) values are accelerated, whereas it decreases when lineages with low \( \theta \) values accumulate a greater number of changes. We also studied the effect that different numbers of taxa among subtrees has on producing such phylogeny biases. Results here are qualitatively parallel to the ones obtained by tree height asymmetry, with accelerated lineages and more conserved lineages having equivalent effects to poorly sampled and well-sampled lineages, respectively. This, together with the results presented in Figure 2.5, suggests that long deep branches (caused by rate acceleration,
Figure 2.6: Effect of subtree height asymmetry on the covarian bias strength. Here we investigate the effect of five distinct relative subtree height rearrangements. We simulated datasets of variable data lengths, although here we show results just for 1000 and 10,000 sites alignments. As in Figure 2.5, groups $a_1$ and $b_1$ share conservation coefficient $\theta_1 = 1$, whereas $a_2$ and $b_2$ share a different and much lower value $\theta_2 = 0.2$. In order to create variability in subtree heights, we study five different assortments where two or three subtrees share the same height ($h$) and the remaining share the inverse ($h^{-1}$). $h = 1$ means the original symmetric tree that becomes more asymmetric as $h$ approaches 0 or infinite.

poor sampling or high tree branch length ratio values) may have a major role in catalysing phylogenetic bias.

2.3.2 The “good” bias that enhances true phylogenetic signal

For the second simulation scenario, we allowed $\theta_{ab}$ to vary from 0 to 1 in steps of 0.1, while we fixed the $\theta$ for each external branch to the same value $\theta_{12}$ taking the same set of values as formerly (Fig. 2.7a). The tree shape is the same as for "the Bad" case, although here we set up the internal branch to 0.01 to hamper the recovery of the correct topology (48% bootstrap out of 100 repeats). Intuitively, simplistic assumptions may actually assist in obtaining the right topology, as closely related taxa should share evolutionary traits not accounted for by the model more often. Accordingly, decreasing $\theta_{ab}$ (no substitution rate category conservation across the inner branch) and increasing subtree $\theta_{s}$ (conservation between sister groups) yields high support.
2.3 Results and Discussion

(a) (b)

**Figure 2.7:** Plot of the mean RELL support value for the right topology (\((a_1, a_2), b_1, b_2\)) with limited phylogenetic signal (inner branch length = 0.01). (a) In this simulation, the inner branch had a variable conservation \(\theta_{ab}\), taking values from 0 to 1 in steps of 0.1, whereas the branches leading to each one of the subtrees share the same conservation coefficient \(\theta_{12}\) that takes the same set of values. Under this procedure, the coefficient between sister subtrees is \(\theta_{12}\) whereas between non-sister subtrees is \(\theta_{ab}\theta_{12}\). (b) Plot that illustrates the effect of the difference between conservation coefficients on the mean RELL value.

for the true topology (Fig. 2.7b) despite the limited phylogenetic signal available.

Although this may seem to be preferable for obvious reasons, we may well want to have an objective measure of the evidence towards each hypothesis. This same reasoning was used to debunk the arguments concerning the superiority of parsimony over likelihood in the so-called Farris zone (Siddall, 1998; Swofford et al., 2001), which is part of the long-standing literature battle regarding long-branch attraction effects and the performance of maximum-likelihood and parsimony (e.g. Sullivan and Swofford, 2001; Kolaczkowski and Thornton, 2004; Gadagkar and Kumar, 2005; Gaucher and Miyamoto, 2005; Brinkmann et al., 2005).

2.3.3 The "ugly" bias caused by heterogeneous RAS distributions

In the third simulated scenario, we explored the effect of distinct shape parameter values under different degrees of site substitution rate category conservation. Two pairs of phylogenetic groups evolved under a \(\Gamma\) RAS model, with the first pair of subtrees \(a_1\) and \(b_1\) using a shape parameter, \(\alpha_1\), and the other two subtrees \(a_2\) and \(a_2\) using a different shape parameter, \(\alpha_2\). For the inner branch, we used one of the pair values, \(\alpha_{ab} = \alpha_1\) (Fig. 2.8b) and an interpolated version of both values, \(\alpha_b = \exp[\ln(\alpha_1) + \ln(\alpha_2)]/2\) (Fig. 2.8c, Fig. 2.8d, Fig. 2.8e, and Fig. 2.8f). This means that even if a site belongs to the same rate category throughout the phylogeny, its relative rate will change among lineages depending on the difference in site rate distributions. Once more, we intuitively expected to reproduce systematic error that would join lineages with convergent relative rate among site distributions \(\alpha_1 \neq \alpha_2\). Topology and branch lengths parameters are set.
up as in the good case including weak phylogenetic signal, so that we can see support oscillations in both directions (pro and against the wrong rate distribution convergent topology). We also simulated different shape parameters under different conservation levels between 1 and 0 in steps of 0.25 but maintained them equal among all subtrees (Fig. 2.8a) so that:

\[
\begin{align*}
\theta_{ab} & = 1 \\
\theta_1 & = \theta_2 = \theta_{12} \in [0, 1] \\
\theta_{xy} & = \theta_{12}^2, \forall xy \in \{a_1, a_2, b_1, b_2\}^2
\end{align*}
\] (2.11)

The outcome is quite surprising. When \( \theta \)s are high (close to 1), for the most part, different shape parameter values drag together those subtree pairs with equal rate distribution shapes (Fig. 2.8b). Nonetheless, greater difference in shape parameters does not necessarily produce greater attraction between convergent subtrees. Moreover, there are zones where support falls below the level achieved when there are no differences between substitution rate distributions (\( \alpha_1 = \alpha_2 \)). Consequently, repulsion takes place between lineages with the same shape parameter. In this particular simulation batch, this phenomenon occurred when medium/high heterogeneity (around \( \alpha = 1 \)) and high heterogeneity (when \( \alpha \) approaches zero) are combined. In fact, these sinuous support surface changes as the relative subtree heights vary to break the initial clock like overall structure (Fig. 2.9) or when the inner branch shape parameter is set up differently (Fig. 2.8b versus 2.8c).

This may be due to a delicate equilibrium between counterpoised attraction and repulsion effects between all lineages because a single shape parameter is used to maximise the global likelihood, despite that each combination of two subtrees would have a different marginally

---

Figure 2.8 (following page): Plot of the mean RELL support value for the wrong topology ((\( a_1, b_1 \), \( a_2, b_2 \)) with different RAS sampling distributions. (a) The simulated process: initially, rate category conservation is full along the inner branch (\( \theta_{ab} = 1 \)) and equals \( \theta_{12} \) for all branches leading to subtrees, so that the pairwise conservation coefficient is equal for all pairs of groups. RAS distribution shape parameters change across subtrees, \( \alpha_1 \) for \( a_1 \) and \( b_1 \), and \( \alpha_3 \) for \( a_2 \) and \( b_2 \). We assigned two different values to the shape parameter along the inner branch: we took one of the group values \( \alpha_{ab} = \alpha_1 \) (b) or alternatively used an interpolated value \( \alpha_{ab} = \exp(\ln(\alpha_1) + \ln(\alpha_2))/2 \) (c, d, e, and f). We simulated 100 repeats at shape parameter values \( \alpha_1 \) and \( \alpha_2 \in \{\exp(x) \mid x \in \{-5, -4, \ldots, 5\}\} \). (b and c) The effect of different combinations of shape parameters in the mean RELL support value where all \( \theta \)s are set to 1. (d, e, and f) Changes in RELL support surface in (c) as we decreased the coefficient \( \theta_{12} \) to 0.75, 0.5, and 0.25, respectively. Bias fades away as more sites are given different random substitution rate categories per subtree. Convergent RAS distributions have a weakening effect in this case.

52
2.3 Results and Discussion
Figure 2.9: Effect of subtree height differences in the RELL support surface for the wrong topology \(((a_1, b_1), a_2, b_2)\). (a) Tree with the wrong topology. (b) RELL support surface where the mean support RELL value is plotted against the natural logarithm of the shape parameter for the substitution rates in long branches \((a_1)\) and short branches \((a_2)\). We used the same parameterisation as for Figure 2.8c but changed subtree heights to 2 for clades \(a_1\) and \(b_1\), and 0.5 changes for clades \(a_2\) and \(b_2\).

best-fitting pairwise shape parameter. Susko et al. (2004) concluded that the bias strength and the inconsistency zone would depend on the relative branch lengths.

 Nonetheless, one may question the biochemical relevance of such a particular scenario; changes on the overall distribution of site substitution rates under the RAS model or its tuning parameters must be accompanied by arbitrary changes in site rate categories, resulting in decreasing \(\theta_s\). We simulated the same shape parameter subspace at different overall rate category conservation coefficients (Fig. 2.8c, Fig. 2.8d, Fig. 2.8e, and Fig. 2.8f). The outcome revealed that repulsion or attraction between subtrees caused by convergence in the sampling distribution of relative rates between lineages strongly depends on the conservation of relative substitution rates within-sites. This puts forward the conclusion that site rate category dependency between lineages is required for attraction or repulsion over possible influences of relative rate probability distribution differences or similarities. Therefore, different RAS distributions cannot be used alone as an argument to postulate possible systematic bias in a dataset or to indicate what lineages should attract.

2.3.4 Visualisation and test

Using the visualisation method described in Methods, we plotted datasets simulated under different model conditions including the first and second parameter subspaces previously investigated. A regular RAS process \((\theta = 1\) for all branches\) generates a balanced density surface (Fig. 2.10a) around the centre of the triangle, although segments from the centre to the vertices do accu-
2.3 Results and Discussion

Covarion unbiased datasets (θ values are 0 for all the subtree-leading branches and θ = 1 for the innermost branch) produce a visibly different point distribution with similar amount of points cumulating towards each hypothesis vertex (Fig. 2.10b). In contrast, covarion-biased datasets generated by positively correlated conservation coefficients between pairs of subtrees yield unbalanced plots (Fig. 2.10c, Fig. 2.10d and Fig. 2.10f). Plot differences between biased and unbiased alignments are evident to the naked eye. The overall plot distribution may change if subtrees have clearly distinct shapes (number of taxa, branch length, and topology) as illustrated in Figure 2.10e and Figure 2.10f. Nonetheless, even in these circumstances, the averaged mean point position remains close to the centre of the triangle in genuine RAS datasets.

These features match the resulting real or artificial tree support obtained by ML tree reconstruction. Accordingly, application of the proposed test on these simulated datasets allows the detection of deviations with increasing sensitivity as the bias strength grows (Fig. 2.11). Sensitivity here is measured as the percentage of positives with a significance level of 0.05 in 1000 amino acid long alignments.

Regarding "the Ugly" case, the subtree pairwise substitution rate difference ranking described above also detects parametric areas where systematic error occurs (Fig. 2.11d). Nonetheless, it seems not to be able to discern between attraction and repulsion, supporting the hypothesis that joins subtrees with similar site relative rate distributions. The fact that the parametric area of attraction and repulsion changes as different modelling elements are altered suggests that this will not be easily resolvable with an ad hoc approach. Here, the use of the right or fairly well

Figure 2.10 (following page): Example of bias visualisation plots based on the layout described in Figure 2.3. These are the product of 10000 amino acid long alignments after smoothing the resulting scattered site plot. Darker areas indicate a greater concentration of site points. The circle indicates the centre of the triangle, whereas the x points to the mass centre of the cloud. (a) Plot resulting from a RAS substitution process. In this figure we used the parameterisation of "the Bad" case but fixing all θ6 values to 1. (b) "Unbiased" covarion process: as in (a) but setting all conservation coefficients θ to zero so that substitution rates are selected at random for each subtree. (c) A covarion evolving dataset in which non-sister lineages attract: "the Bad" case where θ1 = 1 and θ2 = 0. (d) Covarion process in which actual sister lineages attract: "the Good" case where θab = 1 and θ12 = 0. (e) Represents an RAS process with a combination of different number of taxa and subtree heights. Subtrees a1 and b1 are 2 substitutions tall and have 8 taxa each, whereas the height of subtrees a2 and b2 is set to 0.25 substitutions and contain just 2 taxa each. (f) "The Bad" case applied to (e) setting where θ1 = 1 and θ2 = 0. The distribution of dots clearly changes if subtrees are asymmetric. Biased datasets differ in that points migrate to the favoured hypothesis vertex. Consequently, the sample mass centre clearly breaks off from the centre of the triangle (c, d, and f).
2 Artifactual Phylogenies Caused by Heterotachy

(a) (b) (c) (d) (e) (f)
approximated model of evolution is required. This issue will require further investigation; for the moment, we recommend to perform simulations when there is an important difference between shape distributions in different regions of the phylogeny.

Figure 2.11: Sensitivity (solid line) of the proposed test at different values for model parameters that cause bias, plotted against the average RELL value of the favoured hypothesis (dashed line). (a) "The Bad" case, distantly related lineages attract as difference between conservation coefficient \(|\theta_1 - \theta_2|\) increases from 0 to 1 (Fig. 2.4). \(\theta_1\) takes values from 0.5 to 1 in steps of 0.05. \(\theta_2\) is set to \(1 - \theta_1\). (b) On "the Good" case when closely related lineages attract, the inner branch conservation coefficient \(\theta_{ab}\) varies from 0 to 1 in steps of 0.1. Other coefficients on branches leading to the four subtrees are set to 1 (Fig. 2.7). (c) Again on the second simulation batch, but in this case as lineage conservation coefficients vary from 0 to 1 and the inner branch coefficient is set to 1. (d) In "the Ugly" case, setting \(\alpha_{ab} = \alpha_1 = 1\), all \(\theta=1\) and giving \(\alpha_2\) values in \(\{\exp(x)\mid x \in \{-2, -1.5, -1, \ldots, 3\}\}\). Notice that in this particular case the bias is pro the wrong topology in when \(\alpha_1 < \alpha_2\) (\(\ln(\alpha_2) > 0\)) and pro the right topology when \(\alpha_1 > \alpha_2\). In all cases we repeated the experiment 100 times.

2.4 Remarks

It is important to find solutions to the limitations of current phylogenetic methods and evolutionary models. Incorrect phylogenetic resolution may invalidate any result obtained by downstream analyses. A first step towards this aim is to characterise peculiarities of evolutionary processes...
that may cause bias and provide assessing tools to localise their occurrence in real datasets. Moreover, differential variation of substitution rates across sites or composition is an indicator of gene subfunctionalisation (Gu, 1999). Intuitively homologous sequences with more similar distribution of substitution rate across sites probably share more functionality too. Therefore these computational tools may also be used in the characterisation of gene function.

In this work, we illustrated possible misleading effects of some violations of RAS model assumptions regarding evolutionary rates among sites and lineages. Here we extend the observation done by previous work on this aspect of molecular evolution (Chang, 1996; Kolaczkowski and Thornton, 2004), with a more general and relaxed heterogeneous model of evolution. Bias is caused by different degrees of site rate drift among lineages rather than specially tailored conflictive site substitution rate convergence. This phenomenon may also be reproducible when assuming a process accounting for continuous changes in site rates categories (Tuffley and Steel, 1998; Galtier, 2001) that allows for different rate-of-rates among lineages. We investigated different overlapping effects of some model parameters on the intensity of the systematic error caused by covarion-like evolution. It is important to note the importance that taxa sampling may have in reducing the resulting systematic error. Also, we tentatively proposed a tool to visualise and test for the presence of bias on four well-supported subtree phylogenies based on simple geometric concepts. In other words, we test the suitability of using RAS models for resolving the deepest branches of the phylogenetic tree. Reliability is especially dubious at this depth, as any change fixed later (rest of the tree height) constitutes noise regarding the resolution of these relationships. The closest procedure to our approach has been to check, normally a posteriori, on changes in the topology support values after selective resampling of data (Lockhart et al., 1998; Inagaki et al., 2004; Pisani, 2004; Brinkmann et al., 2005).

Although here we have focused only on quartet analysis, it may be extendable to a greater number of monophyletic groups by means of quartet puzzling-like techniques (Strimmer and von Haeseler, 1996) or porting the geometrical framework described herein to a higher dimensional space. In order to estimate discrepancies in site substitution rates among lineages, we needed to have a reasonable number of taxa per each group. Therefore this method cannot be used accurately with poorly taxed groups. Unfortunately, this is not an uncommon situation especially in studies performed on poorly sequenced taxonomy groups. Perhaps this may be avoided using the other dimension of the data, increasing the number of sites compared each time using a sliding window.

We may devise some correcting protocols and alternative methods inspired by the simulation model and test proposed. Certainly, in molecular datasets, the accuracy of ML and Bayesian analyses may increase by including more sets of parameters in existing models such as the conservation coefficient per branch or subtree and by accounting for within-site rate changes. In this context, one could use likelihood-ratio tests or alternative model sorting criteria to determine
whether the enhanced model explains the data significantly better and whether the simplified model is susceptible to bias. However, that may unbearably increase the computational cost and risk of overfitting. Model-based distance methods such as neighbour-joining could benefit from the fact that each pair distance does not need to be estimated using the same global model. A plausible alternative is then to generate a mixed distance matrix, where many inter-subtree or inter-taxa models approximate better the evolutionary path between pairs of monophyletic groups. Another possibility is to perform \textit{a priori} site-weighted bootstrapping or removal in order to improve data fitness for the model of choice. Although this last option may not seem to be the most attractive, as we may need to discard a sizable portion of the data, it is in fact constantly being done; e.g., leaving out the third codon position in coding DNA analyses. 

In any case, it is important to bear in mind that there is more than a single source of systematic error; for example, compositional heterogeneity (Mooers and Holmes, 2000; Foster, 2004). Moreover, bias may reinforce the right topology, as close phylogenetic relatives should intuitively tend to share traits that may not be accounted for by mainstream models. Therefore, thoughtful phylogeny curation is necessary in order to accurately assess the cause and effect of such model singularities as we show later in Chapter 3.

In summary, diagnostic analysis approaches, like the one proposed in this work, may complement existing methods such as the maximum-likelihood and Bayesian approaches and provide a visual statistical framework to take into account hidden parameters (covarion evolution, compositional bias, and so forth) that, when ignored, may hamper accurate phylogenetic inference.

### 2.5 Acknowledgments

We would like to thank Davide Pisani for his valuable comments on the manuscript. We are also thankful to Professor Dan Bradley and Simon A. Travers for revising the grammar and spelling of the manuscript. Also we would like to express our gratitude to the editor and to both anonymous reviewers whose comments and suggestions were vital for the development of the final version of the manuscript. Computational resources were provided by Science Foundation Ireland (SFI) under the basic research grant program awarded to M.A.F. We would also like to express our gratitude to the computational resources hosted by the Bioinformatics Laboratory and Computer Science Department at the National University of Ireland, Maynooth.
Chapter 3

Reassessing the evolutionary story of the SecA gene

The materials in this chapter have been recently submitted for review to a peer-reviewed journal.

Abstract The ATPase SecA gene is a key element of the machinery that permits the export of oligo-peptides across the enveloping membrane in free-living bacteria and plastids alike. Early phylogenetic analyses on this gene provided molecular evidence supporting the existence of two distinct polyphyletic groups among chloroplast SecA homologs. These groups would have resulted from multiple primary endosymbiosis events. A later study challenged this claim suggesting that the apparent polyphyly was due to compositional bias. Here we reassessed the phylogeny of the SecA gene in light of more complete data and tools to handle compositional bias and heterogeneous evolution. Our results based on protein sequences indicate that non-orthologous evolutionary patterns such as covarion shifts seem to be the reason for overconfidence in the polyphyly of chloroplast SecA. We further discuss the importance of detecting and discarding functionally divergent genes from systematic analyses, and the prospect of using heterotachous evolutionary patterns as a means to phylogenetic reconstruction.

3.1 Background

The SecA gene encodes a peripheral membrane protein involved in the SEC-mediated export of proteins in free living bacteria and chloroplasts. Based on the model postulated by Vrontou and Economou (2004), a chaperone, such as SecB, collects the neo-nascent unfolded client protein chain and targets it to SecA. Then, SecA uses the energy produced from ATP hydrolysis to mechanically "push" the client precursor through the channel constituted by other inter-membrane subunits of the translocase complex (SecY, E and G).
Early phylogenetic analysis of SecA homologs by Valentin (1997) gave support to the existence of two distinct types of polyphyletic SecA groups in chloroplasts. He based his conclusions on bootstrap support obtained using distances and the neighbour-joining (NJ) method on protein sequences. This result added evidence in support of a polyphyletic origin of plastids at a time when this was still an open debate (reviewed in Stiller et al., 2003). In contrast, more recent and exhaustive phylogenetic studies provided compelling support for a now well-established single primary endosymbiosis theory (Rodríguez-Ezpeleta et al., 2005). Accordingly, single- or a few-gene phylogenies that recover polyphyletic plastid lineages should arise as a result of recombination, lateral gene transfer (LGT), old paralogy followed by reciprocal gene loss (e.g. Delwiche and Palmer, 1996) or mathematical artifacts (Barbrook et al., 1998; Lockhart et al., 2006).

Specifically in the SecA gene case, Barbrook and colleagues (Barbrook et al., 1998) concluded that compositional bias was the most plausible explanation for the strong support obtained by Valentin. In their phylogenetic analyses using DNA sequences they never found strong support for a polyphyletic plastid SecA. They, however, found significant support for the monophyly of red and brown algae homologs compatible with both polyphyletic and monophyletic hypotheses. They noticed that red and brown algae homologs are AT-rich as compared with other taxa in the dataset. Accordingly, they reconstructed the phylogeny using heterogeneous composition corrected distances (see LogDet below). This reduced significantly the confidence in an otherwise well-supported red and brown algae monophyletic group. Then, they extrapolated these findings in DNA to protein analysis suggesting that heterogeneous amino acid preference conditioned by heterogeneous nucleotide composition could offer an alternative explanation to Valentin’s findings. Additionally, Barbrook and colleagues reported the presence of covarion structure in the dataset that could also explain part of the support observed, at least, for the monophyletic red and brown algae group.

Concomitantly variable codons (covarions), changes in proportion of invariant sites and other related concepts are all forms of heterotachy or heterotachous evolution (Lopez et al., 2002). A molecule is said to evolve heterotachously if it undergoes changes in substitution rates per site overtime. Many studies have shown that in fact this may well be routinely the case (Galtier, 2001; Huelsenbeck, 2002; Wang et al., 2007). Therefore ignoring this fact in phylogenetic studies, especially when the phylogenetic signal is weak, could lead to arbitrary phylogeny reconstructions (Chapter 2 and Lockhart et al., 1998; Kolaczkowski and Thornton, 2004; Wang et al., 2008).

In this study, we reassessed the phylogeny of the SecA gene in light of new protein data and taking into account possible artifacts as a result of heterogeneous composition and heterotachous evolution.
3.2 Material and Methods

3.2.1 Sequences and alignment

As we focused on determining whether plastid SecA is monophyletic we collected homologs from plastids, from the putatively closest prokaryote group, Cyanobacteria, and an also from Firmicutes, a close non-photosyntetic prokaryote outgroup. We searched SecA sequences using NCBI protein-protein BLAST service (Madden et al., 1996) on relevant taxonomy groups: Viridiplantae for green plastids (11 full-sequence non-redundant hits), Rhodophyta for primary red endosymbions (6), Haptophytes, Cryptophytes and Heterokots for secondary red endosymbions (8), Cyanobacteria (38) and Firmicutes (100+). Unfortunately, there was no significant hit among Glaucophytes. This group allegedly diverged from other plastids right after the single primary endosymbiotic event. We reduced the Cyanobacteria and Firmicutes subsets to 12 and 10 sequences with emphasis in phylogenetic diversity without including relatively fast evolving sequences as to limit possible long-branch attraction (LBA) artifacts.

We aligned amino acid sequences using Muscle v3.2 (Edgar, 2004). Then, we removed dubious alignment sites using Gblocks v0.91 program (Castresana, 2000) allowing for gaps. We set up the minimum number of sequences for flanking regions to the minimum required (half the number of taxa + 1) since using the default option produced an alignment too short to obtain reliable bootstrapped LogDet corrected distances (see below). The resulting alignment included a total of 47 sequences and 779 amino acid sites.

3.2.2 Compositional bias analysis

To determine the effect of differences in residue preference across taxon groups we first analysed amino acid composition using principal component analysis as implemented in R statistical package (http://www.r-project.org). In this scenario residue frequencies are the variables whereas each sequence’s specific frequencies are the observations. We expected that taxa with more similar composition appear grouped in bilateral plots of major components that explain most of the variance observed.

Additionally, we used LogDet corrected distances (Lake, 1994; Lockhart et al., 1994) and recoding (Hrdy et al., 2004) to reconstruct phylogenies taking into account heterogeneous composition across lineages. In theory, clusters of taxon groups attracted due to convergent composition should have less support using these two methods.

3.2.3 Phylogeny reconstruction based on residue substitutions

We performed maximum-likelihood (ML) reconstruction using the PhyML program v2.4.4 (Guindon and Gascuel, 2003) based on the WAG rate-matrix (Whelan and Goldman, 2001). Model-
generator v0.82 (Keane et al., 2006) indicated that this is the best general purpose rate-matrix for the whole alignment. We modelled rate-across-site (RAS) heterogeneity using a discrete $\Gamma$ distribution (8 categories). In all ML phylogenetic analyses we run 100 non-parametric bootstrap repetitions. We used Consense, part of the Phylip package v3.6 (Felsenstein, 1989), to summarise bootstrap replicates into a consensus tree reconstructed by the majority rule method. For distance-based analyses we calculated marginal pairwise distances assuming the same model WAG+$\Gamma$ and ML estimator for the shape parameter $\alpha$. We used the Neighbour-joining (NJ) algorithm as implemented in the PAL library for Java (Drummond and Strimmer, 2001).

For analyses based on LogDet corrected distances, we used the LDDist Perl package v1.3.2 (Thollesson, 2004). LDDist generates a script file that we later fed to the Paup program v4.0b10 (Swofford, 2002) to perform the actual tree reconstruction using least-squares (LS). In this case we also generated a single best tree estimate with 100 bootstrap replicates summarised into a consensus tree. We allow LDDist to estimate the actual number of invariant sites using Steel's capture re-capture method (Steel et al., 2000) in order to account for RAS heterogeneity.

We used PhyloBayes v2.1c (Lartillot and Philippe, 2004) to perform Bayesian analyses including reconstruction using Dayhoff recoding. In the resulting recoded alignments, the 20 amino acids are reduced to a smaller cohesive set of 6 states grouping biochemically similar amino acids \{[MVIL], [STPAG], [HRK], [NDEQ], [FYW] and [C]\}.

### 3.2.4 Heterotachous models of protein evolution

#### 3.2.4.1 Time-homogeneous covarion models

Mainstream evolutionary models do not account for heterotachous evolution. To the best of our knowledge, only two programs currently implement covarion-like models for protein sequence analysis. MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003) carries out a full Bayesian tree topology search based on the covarion model presented by Huelsenbeck (Huelsenbeck, 2002) with the possibility of using an arbitrary rate-matrix. Procov (Wang et al., 2007) performs a ML search of optimal model parameters for three covarion models: Galtier (2001)(herein named COV$_G$), Huelsenbeck (2002) (COV$_H$) and a contributed generalisation of the former two (COV$_{GH}$). Unfortunately this program currently supports only the JTT rate-matrix (Jones et al., 1992) and an elementary tree search. Therefore the topology should be known in advance. In this manuscript we call these: *time-homogeneous* covarion models as changes in substitution rates across sites and lineages occur at an uniform steady pace. That is, these models do not account for sudden changes in evolutionary constraints, covarion shift "hot spots" or subsequent heterogeneous covarion rates across lineages or sites.
3.2.4.2 Time-heterogeneous models of functional divergence

Gu proposed a model of functional divergence that allowed for sudden changes in substitution rates reputedly due to changes in functional constraints (Gu, 1999, 2001). In order to quantify covariation of rates across major groups of taxa in the dataset we extended the probabilistic model described by Gu (2001). Under that model a percentage of sites evolve at the same relative rate across different taxon groups, whereas other sites evolve under different relative rates across groups. These substitution rate shifts are interpreted as changes in evolutionary constraints and, arguably, also changes in functional constraints. This kind of functional divergence is commonly named type-I divergence.

Here we follow the notation introduced in section 2.2.1. For each predefined taxon group \(i\) we denote the percentage of sites with the same substitution rate as at the immediate ancestral node as \(\theta_i\). The complementary percentage of sites that have undergone substitution rate shifts on the branch leading to that group right after the divergence event is \(\bar{\theta}_i = 1 - \theta_i\). Similarly, we also define \(\theta\) values at deeper nodes of the tree allowing for substitution shifts at beginning of branches leading to common ancestors of multiple taxon groups (Fig. 3.1). Consequently, the pairwise \(\theta_{ij}\) between two groups is equivalent to the product of \(\theta\) values along the path connecting both clusters. For example, in Figure 3.1 \(\theta_{XZ} = \theta_X \theta_B \theta_Z\).

![Figure 3.1: Example of type-I divergence model four taxon groups \(\{W, X, Y, Z\}\). Substitution rate shifts may occur at beginning of branches right after divergence events (internal nodes \(A\) and \(B\)). The percentage of sites that do not change rate is denoted with the corresponding conservation coefficient \(\theta\). Due to the lack of an additional outgroup beyond \(W\), \(\theta_W\) and \(\theta_A\) are in practice mutually unidentifiable. Therefore we may just work with \(\theta_W = \theta_A \theta_W\).](image)

Gu’s model assumes that the RAS distribution, \(\Gamma(\alpha)\), is common across all groups. Nevertheless, it is not surprising to find datasets where RAS distributions are different between lineages.
Stiller et al. (2003); Inagaki et al. (2004). Accordingly, we added the possibility that relative site substitution rates are drawn from different RAS distributions across groups. We can use the likelihood-ratio test (LRT) to double-check that additional group-specific shape parameters result in a significant improvement of the model-data fit. We count an additional degree of freedom (df) for each new independent shape parameter introduced. Thus in this model, $\theta_{XY}$ is not the percentage of sites with identical relative rates between groups $X$ and $Y$ (probably none if $\alpha_X \neq \alpha_Y$) but the percentage of sites with identical relative rate category or rank within the corresponding rate distribution, $\Gamma(\alpha_X)$ and $\Gamma(\alpha_Y)$ respectively. From this point we will call this model DIV.

We can compute a ML estimator for the pairwise rate conservation coefficient between two groups, $\theta_{XY}$, knowing their rooted subtree topologies and branch lengths, $T_X$ and $T_Y$, and residue patterns, $P_X$ and $P_Y$. The likelihood formula for a single site is then:

$$L_{DIV}(\{P_X, P_Y\}|\{T_X, T_Y\}, \theta_{XY}, \alpha_X, \alpha_Y) = \theta_{XY}|C|^{-1} \sum_c \{L(P_X|T_X, \alpha_X, c) L(P_Y|T_Y, \alpha_Y, c) \}$$

$$+ \theta_{XY}|C|^{-2} \sum_{c,d} \{L(P_X|T_X, \alpha_X, c) L(P_Y|T_Y, \alpha_Y, d) \} \quad \text{ (3.1)}$$

$$L(P|T, \alpha, c) \equiv \text{ standard likelihood formula for a pattern } P$$

under a topology and branch lengths $T$ scaled using the relative rate for category $c$ in $\Gamma(\alpha)$.

$C \equiv \text{ set of substitution rate categories.}$

Notice that if we are just interested in estimating $\theta_{XY}$ we can do so only knowing the unrooted topology and branch lengths for groups $X$ and $Y$. The location of the subtree roots and length of the branch connecting both groups are not relevant. This would be mathematically equivalent to fix the length of that internal branch, wherever placed, to infinity ($\infty$).

Additionally, it is worthwhile to consider the possibility that covariotide evolution does not affect all positions equally. Intuitively, alterations of functional constraints should involve some sites more often than others. Here we added an extra parameter, $\phi$, that indicates the percentage of sites susceptible to covarion shifts. We will call this extended model DIV$\phi$. In this case the likelihood calculates as:

$$L_{DIV}\phi(P|T, \theta_{XY}, \phi, \alpha_X, \alpha_Y) = (1 - \phi)|C|^{-1} \sum_c \{L(P_X|T_X, \alpha_X, c) L(P_Y|T_Y, \alpha_Y, c) \}$$

$$+ \phi L_{DIV}(P_X, P_Y|T_X, T_Y, \theta_{XY}, \alpha_X, \alpha_Y) \quad \text{ (3.2)}$$

Likelihood equations for three groups would consist of $2^3$ terms representing the possible combinations of substitution rate shift and conservation across all three clusters. For four or
more groups, in addition to a consistently increased number of terms, we need to know deep phylogenetic relationships between clusters. In this case, apparently convergent site substitution rate changes may have occurred at common ancestral nodes.

3.2.5 Phylogeny reconstruction based on site substitution rate divergence

Most methods use statistical signal concealed in patterns of residue change across taxa to infer a phylogeny. Nevertheless, shifts in relative substitution rates per site across lineages can analogically carry phylogenetic information too. At the very least, we can use these differences to quantify functional divergence between different groups using a tree like representation. More concretely, we can use pairwise $\theta$ values to reconstruct a functional tree using distance methods (Functional Distance Analysis in Wang and Gu, 2001). In order to do so, we need to transform these into tree additive distances: correct them for multiple substitution rate changes. As it is done with residue differences, we can use the Poisson correction:

$$u_{XY} = -\ln \theta_{XY}$$  \hspace{1cm} (3.3)

If we consider that only a proportion of sites $\phi$ may undergo covarian shifts we have to take this into account:

$$u_{XY} = -\phi \ln \left(1 - \phi^{-1}\theta_{XY}\right)$$  \hspace{1cm} (3.4)

In addition, we can compare inter-group topologies using the divergence model described above and the ML criterion based on the multiple-group versions of Equations (3.1) and (3.2). Subsequently, we can transform the resulting ML estimators of $\theta$ values at internal nodes into distances using Equation (3.3) or (3.4) respectively.

In this study, for pairwise distance based reconstructions we used the Fitch program (also part of the Phylip package) to calculate the topology using the weighted least-squares method (WLS) where weights are set to the inverse of the squared distance, $w_{XY} = u_{XY}^2$. In the ML estimation of $\theta$ values we optimised branch lengths and RAS distribution shape parameters using the Weka framework (http://weka.sourceforge.org) for the Java programming language. In order to speed up calculations we fixed each group subtree topology to the best ML tree (PhyML) for that group.
3.2.6 Divergent evolution bias analysis

3.2.6.1 Substitution rate shift bias analysis

We divided the dataset in cohesive sequence groups based on major taxonomy and or monophyly as shown in Results. Then, we used a graphical test described in Chapter 2 to detect the presence of unbalanced covariotide structure that may affect the resolution of deep branches of the tree between these groups. We use the average point position to determine the significance of the deviation observed.

In order to quantify covarion shifts between different groups we fitted the data to the DIV model described above. First we took only two groups at a time to obtain pairwise $\theta$ values. We considered results assuming two different RAS distributions (one per group) if that improved the model fit significantly (verified using LRT). Then we reconstructed a functional divergence tree using the resulting pairwise substitution rate divergence distance values among taxon groups (Eq. 3.3 and 3.4) and the LS algorithm. The choice between $v$ and $u$ depended on whether DIV or $\text{DIV}_i$ showed a better model-data fit.

Finally, we used parametric bootstrapping to confirm that the covariotide shift observed can indeed explain the bias towards the topology yield using methods and models that do not account for covarion shifts. We computed the simulation parameter values from the SecA dataset by fitting the divergence DIV model above based on the ML topology obtained using PhyML. We removed phylogenetic signal due to the residue synapomorphies by compressing the relevant inter-group branches to length zero.

3.2.6.2 Compositional shift bias analysis

A site has undergone type-II divergence or compositional shift when it has accumulated functionally important residue changes (Gu, 1999). Intuitively, functional divergence type-II should induce attraction in convergent groups (due to both homoplasies and conserved synapomorphies) and repulsion between divergent groups (due to an increasing number of different residues that reflect change in function).

In practice, type-II divergence is mostly amenable to detection only when it takes place in conserved regions: sites that are constant or very slow evolving but with chemically different residues across groups. In contrast, in moderate to fast evolving regions is not easy to tell between compositional changes due to divergence type-II or regular evolutionary drift.

Accordingly, we only considered strong changes at sites between groups that are conserved. We determine whether a site is slow evolving in a group by the percentage of identity, $I$, among its sequences. As in larger taxon groups it is increasingly difficult to find constant sites, to overcome this we relax the identity condition by recoding the alignment using Dayhoff 6 chemical groups. Then we compress the alignment to a consensus sequence per group where non-
3.3 Results

3.3.1 Phylogenetic reconstruction using traditional models

We first built the bootstrap consensus ML tree using the homostachous RAS model WAG+Γ. The resulting tree topology consists of 5 distinct monophyletic groups: Firmicutes (F), Cyanobacteria, Green plants α, Red plastids, and Green plants β. As shown on the corresponding branch, maximum likelihood gives a 65% bootstrap support for the monophyly of green plants α and Cyanobacteria whereas distance method gives it a 78% support and Bayesian analysis a 100% probability. Using ML and Bayesian methods *Gloeobacter* and *Synechococcus* placement basal in Cyanobacteria is not highly supported.

Conserved characters are considered unknown as to mark them as non-informative. Then, we used maximum-parsimony (MP) to reconstruct a phylogenetic tree with a minimum number of steps, so that composition shift convergent taxon groups cluster in the output topology. We performed this analysis with $I$ set to 100%, ≥75% and ≥50%.
teria (C), red and brown algae plastids (primary and secondary endosymbions) (R) and two paralogous groups in green plants: Go and Gβ1. The location of green β at the root of the tree next to Firmicutes homologs indicates this second copy in green plants could have resulted from a lateral gene transfer event (LGT) or an ancient paralogy of the SecA gene at the root of Cyanobacteria and later convergent gene losses of β homologs at the root of the red-brown and extant Cyanobacteria groups (Fig. 3.3). We never recovered a red-brown and green β cluster discarding the possibility of a reciprocal gene loss of SecA homologs across both types of chloroplast. Inclusion of a broader selection of bacterial taxa incorporating α-proteobacteria (putative origin of mitochondria) only separated green β group from other Cyanobacterial and plastid SecA homologs when we used methods prone to long-branch attraction artifacts (NJ and Poisson-corrected pairwise distances). Therefore there is no clear evidence supporting the LGT alternative. This is further discussed in Appendix D.

Regarding the relationship between Cyanobacteria, red and green α, ML reconstruction yields weak support (bootstrap support is less than 70%) for the polyphyly of plastids where green α and Cyanobacteria cluster excluding red plastid SecA homologs. In contrast, NJ reconstruction based on WAG+Γ pairwise distances yields a significantly high support (78%) for the same grouping as illustrated in Figure 3.2. Bayesian reconstruction also yields a significantly high support for this group (100%).

This is in agreement with conclusions in Valentin (1997) and contrary to the well-established theory of a single primary endosymbiosis event, as long as we assume that Cyanobacteria, red plastid and green α are indeed orthologs.

### 3.3.2 Compositional bias analysis

Barbrook and colleagues (Barbrook et al., 1998) suggested that Valentin’s high support for a polyphyletic plastid SecA was probably due to amino acid compositional bias induced by different nucleotide frequencies in the coding DNA sequences. Accordingly, we double checked whether composition is responsible for the apparent plastids paraphyly. Principal component analysis revealed that indeed composition is heterogeneous across groups as reported by the bilateral plot based on the two major components (Fig. 3.4). The first component clearly segregates red plastid homologs from the rest. This group has a higher preference for amino acid coded by AT-rich codons as previously reported by Barbrook and colleagues (Barbrook et al., 1998). This is in agreement with the fact that red and brown algae homologs are the only ones encoded in plastid genomes. Other groups are non-clearly distinguishable using the first component but

---

1 In this text we use Greek letters α and β to refer to both varieties of SecA in green plants. The α variety was the first to be annotated and therefore is considered as the SecA locus in green plants in the literature. The β variety was annotated later and is described only as a SecA related gene. Results of this study indicate that this secondary status is quite questionable.
the second as Firmicutes and green plants homologs (especially $\beta$) seat at both extremes of the Y-axis. Bilateral plots using other components do not reveal any other clear pattern.

Despite the compositional heterogeneity present in the dataset, there is no obvious convergent trend between any two groups. Even though one group is compositionally different, red plastids, there is no sign of potential attraction by Firmicutes or green $\beta$ outgroups and consequent clustering of Cyanobacteria and green $\alpha$ homologs.

In order to attest that, we used two phylogeny reconstruction methods supposed to correct for non-stationary composition in the dataset: LogDet corrected distances and recoding. As shown in Figure 3.5, both methods reconstructed well supported a Cyanobacteria and green $\alpha$ lineage. Therefore, composition does not seem to be the main cause for the incongruence between SecA and the species phylogeny.

We investigated also the possible effect of compositional shifts (divergence type-II) using recoding and parsimony as explained in Methods. Results pinpoint the existence of important compositional changes between groups in the dataset. However, there is no indication that these may induce a strong phylogenetic bias as the best three tree topologies show a surprisingly matched number of parsimony steps (Table 3.1 and Figure 3.6).
Figure 3.4: Bilateral plot of two main components (cumulative variance explained of 58%) that determines the amino acid composition amongst SecA homologs. Ellipses cluster sequences based on taxonomic group. Instances that are found away from their corresponding group are indicated with arrows. Grey letters indicate the projection of each amino acid frequency (we do not draw the corresponding vectors). Additional bilateral plots with other components do not add any relevant information.

3.3.3 Covarion shift bias analysis

3.3.3.1 Fitting time-homogeneous covarion models to the data

We used the Procov program that implements a range of covarion models based on the JTT rate-matrix. JTT+Γ results in a worse model-data fit when compared to the WAG+Γ model.
3.3 Results

(a) Dayhoff recoding

(b) LogDet

73
3 Reassessing the evolutionary story of the SecA gene

<table>
<thead>
<tr>
<th>Topology</th>
<th>100% conserved</th>
<th>≥75% conserved</th>
<th>&gt;50% conserved</th>
</tr>
</thead>
<tbody>
<tr>
<td>(F,Gβ,((Ga,R),C))</td>
<td>96</td>
<td>176</td>
<td>294</td>
</tr>
<tr>
<td>(F,Gβ,((Ga,C),R))</td>
<td>96</td>
<td>176</td>
<td>296</td>
</tr>
<tr>
<td>(F,Gβ,((C,R),Ga))</td>
<td>96</td>
<td>176</td>
<td>293</td>
</tr>
</tbody>
</table>

Table 3.1: Divergence type-II analysis. We compared the number of synapomorphies/homoplasies (figures above) between the three most parsimonious topologies. We only take into account drastic residue changes in conserved sites. We show results considering sites between groups where all, 3/4 and 1/2 of sequences present chemically similar residues within each group. Bold numbers indicate the most parsimonious topologies unless there is a triple tie.

Figure 3.6: Parsimony tree of compositional shifts between major taxon groups. We built this tree using recoded consensus sequences with identity 100% within groups. The most parsimonious tree (less number of compositional shifts) is a trichotomy.

(ΔlnLk>100, see Table 3.2) but it does not seem to show qualitative differences concerning the resolution of the deep branches of the SecA tree. We evaluated the model-data fit increase for the three covarion models implemented in Procov on the three topologies resulting from rearranging the red, green α and Cyanobacteria groups in the ML tree in Figure 4.1. LRT comparisons indicate that the best model is the most general COV_qh. The difference between the most likely green α and Cyanobacteria cluster, and the putatively correct one joining red plastid and green α does not decrease after adding the possibility of time-homogeneous covarions.

Additionally, we used the tree searching based on the Huelsenbeck covarion model COV_H implemented in MrBayes hoping that this could recover the reputedly correct green α and red plastid monophyly. Based on Procov’s output the general COV_{GH} model is bearly significantly better than COV_H (ΔlnLk = 3.08, Δdf = 2, p-value=0.046). If the dataset lacks genuine phylogenetic signal for this group, the expectation is that, using the covarion model, the reconstructed phylogeny would remove at least part of the confidence given to the Cyanobacteria and green α cluster. Nevertheless it also yielded a significantly high posterior probability for the Cyanobacteria and green α cluster (BP=1.0).
Table 3.2: Changes in likelihood of the SecA phylogeny when we allow for covariation like evolution. Bold numbers highlight the likelihood of the best hypotheses when compared using LRT. Numbers within parentheses indicate likelihood difference when compared with the best topology under the same model. In all cases, the maximum-likelihood tree is the one that clusters green α and Cyanobacteria. Nevertheless likelihood differences are rather small between topologies reflecting the weak support for this cluster obtained using ML (Fig. 4.1).

3.3.3.2 Fitting time-heterogeneous functional divergence models to the data

Regarding sudden substitution rate shifts, we first used a graphical test to detect covariotide structure that may well induce bias towards clustering certain groups. Since this algorithm takes as input four groups at a time we performed the analysis five times leaving out one group at each go. The test only gives positive results in quartets where Cyanobacteria and green α groups are present. As featured in Figure 3.7, the favored topology in all three cases is the one that joins these two groups. This indicates that covariotide bias may induce attraction of these two groups exclusively as removing one of them seems to render the test not significant. In other words, convergent or divergent substitution rate patterns in these two groups may be responsible for the artificial clustering whereas the presence or absence of other groups must not have major repercussions in the topology. Unfortunately, we cannot determine whether plastid SecA is monophyletic or polyphyletic without any of these two groups; we could only dispense with an outgroup (either green α or Firmicutes homologs).

We fitted the site-specific substitution rate divergence models described in Methods to the data in order to quantify covariotide shifts among groups. First we determined whether RAS Γ shape parameter values are significantly different across taxon groups using the DIV model. We obtained the best fit when Cyanobacteria and green α evolve under a convergent RAS model \( \Gamma(\alpha_{CGA} = 0.355) \), but divergent from the other three groups (Table 3.3). These would also converge to a single common RAS process \( \Gamma(\alpha_{RFGS} = 0.778) \). Results in previous studies suggest that in such scenario the expectance is that groups with convergent processes attract

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>Topology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(F,Gβ₁(R,(C,Gα)))</td>
</tr>
<tr>
<td>JTT+Γ</td>
<td>92</td>
<td>-33759.67 (+ 0.00)</td>
</tr>
<tr>
<td>COV₉</td>
<td>94</td>
<td>-33645.39 (+ 0.00)</td>
</tr>
<tr>
<td>COV₈</td>
<td>94</td>
<td>-33601.12 (+ 0.00)</td>
</tr>
<tr>
<td>COV₉₈</td>
<td>96</td>
<td>-33598.04 (+ 0.00)</td>
</tr>
<tr>
<td>WAG+Γ</td>
<td>92</td>
<td>-33618.17 (+ 0.00)</td>
</tr>
<tr>
<td>DIV+Γ₂</td>
<td>98</td>
<td>-33499.52 (+ 0.00)</td>
</tr>
<tr>
<td>DIVφ+Γ₂</td>
<td>99</td>
<td>-33496.63 (+ 0.00)</td>
</tr>
</tbody>
</table>
Figure 3.7: Triangles representing site plots for the 5 possible quartets out of the 5 taxon groups. Triangles are divided in 3 areas that are closest to a vertex and therefore to the corresponding unrooted topology (depicted adjacently). Inscribed numbers indicate the percentage increase or decrease in density in each area (percentage of plotted site points enclosed) as compared to simulated RAS compliant datasets. Bold letters highlight significant p-values and the hypothesis quartets that are hypothetically favoured by the covariotide structure present. Only the 3 quartets including simultaneously Cyanobacteria and green α subsets show significant evidence of an unbalanced distribution of site rate change across groups.

each other when the model assumed is stationary (Inagaki et al., 2004) nevertheless, in some cases, this may induce weak repulsion between convergent groups as shown in Chapter 2.

Quantification of pairwise divergence type-I between all groups assuming the DIV+Γ2 model reveals that Cyanobacteria and green α are also the most divergent (Table 3.4 and Figure 3.8a). Simulations showed that when substitution rate divergent groups are also the slowest evolving ones, as is the case in this dataset, there is a tendency to cluster them using probabilistic models that do not take into account substitution rate divergence (Figure 2.6 in Chapter 2).

Then we calculated the maximum-likelihood of the three best topologies using the DIV and DIVφ models with two RAS distributions Γ2(αCGα, αRFGβ). Under these two models also the best tree is the one that clusters Cyanobacteria and green α groups (Table 3.2). Nevertheless its likelihood is not significantly better (using pairwise RELL tests).

Both, time-homogeneous and time-heterogeneous models improve significantly the model-data fit as compared to ordinary RAS counterparts. Time-homogeneous COVGH produced the best improvement (ΔlnLk = 161.63). Nevertheless the time-heterogeneous DIVφ model is the one that reduced most the likelihood difference between the best and putatively correct topologies (ΔlnLk = 0.64).

A tree like representation of type-I divergence using least-squares and pairwise u distances
3.3 Results

<table>
<thead>
<tr>
<th>Model</th>
<th>lnLk</th>
<th>df</th>
<th>C</th>
<th>Gα</th>
<th>R</th>
<th>F</th>
<th>Gβ</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIV+Γ₁</td>
<td>-38782.92</td>
<td>88</td>
<td></td>
<td>0.629</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIV+Γ₂</td>
<td>-38733.33</td>
<td>89</td>
<td></td>
<td>0.355</td>
<td>0.778</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIV+Γ₅</td>
<td>-38733.02</td>
<td>92</td>
<td>0.352</td>
<td>0.360</td>
<td>0.786</td>
<td>0.746</td>
<td>0.820</td>
</tr>
</tbody>
</table>

Table 3.3: Differences in Γ shape parameter α across groups. Rows represent different models allowing for a single shape parameter, DIV+Γ₁(α), two, DIV+Γ₂(αCGa, αRFGb), or as many as groups, DIV+Γ₅(αC,αGa,αR,αF,αGβ). Columns indicate the maximised likelihood, number of degrees of freedom and shape parameter ML estimator per group under each model. Bold indicates the best model using the LRT criterion. We calculated maximum likelihoods using the marginally best subtree for each group and setting inter-group branches to ∞.

indicates that substitution rate divergence in Cyanobacteria and green α taxon groups is more coincidental than expected at random (Figure 3.8a). This is also the case under the DIV model (Figure 3.8b). The θ(CGa) value at the inner-branch separating Cyanobacteria and green α from the rest is significantly different than zero under this model (ΔlnLk=3.96, Δdf=1, p-value=0.0049). If we interpret these substitution rate shifts as true synapomorphies that would support existence of a Cyanobacteria and green α groups. Nevertheless the DIVφ model’s ML estimator for θCGa is zero (Figure 3.8c) indicating that apparent convergence is due to heterogeneous substitution rate shifts across sites. The DIVφ model is significantly better than the DIV model (ΔlnLk=2.89, Δdf=1, p-value=0.0162).

<table>
<thead>
<tr>
<th>C</th>
<th>Gα</th>
<th>R</th>
<th>F</th>
<th>Gβ</th>
</tr>
</thead>
<tbody>
<tr>
<td>θGaC = 0.684</td>
<td>uCGa=0.677</td>
<td>uCR=0.482</td>
<td>uCF=0.734</td>
<td>uCGβ=0.396</td>
</tr>
<tr>
<td>θRC = 0.727</td>
<td>uGaR=0.677</td>
<td>uGaF=1.119</td>
<td>uGaGβ=0.358</td>
<td></td>
</tr>
<tr>
<td>θFC = 0.675</td>
<td>uRGF=0.176</td>
<td>uRGβ=0.0688</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gβ = 0.754</td>
<td>uGBGβ=0.768</td>
<td>uGBR=0.937</td>
<td>uGBF=0.928</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.4: Pairwise θ value matrix between different subgroups. Entries in the bottom triangular matrix indicate the corresponding pairwise θij values whereas the above the diagonal we include the corresponding tree additive uij values (eq. 3.4). θ values were estimated by maximizing the marginal likelihood only taking two taxon groups at a time. We used DIV+Γ₂ model between Cyanobacteria or green α and other taxon groups; and the DIV+Γ₁ for all other combinations according to the best model-data fit obtained in Table 3.3. We fixed the middle branch length to ∞. To calculate uij distances we fixed φ to the ML estimator obtained under the DIVφ+Γ₂ model (φ = 0.38).

We performed phylogenetic reconstruction using simulated data with the same covariotide structure found in the real dataset based on the time-heterogeneous DIV model. In this case
Figure 3.8: Tree-like representation of divergence type-I between taxon groups. (a) based on LS on tree additive distances in Table 3.4. (b) based on ML estimators of $\theta$ at branches using the DIV+$\Gamma_2$ model. (c) based on ML estimates using the DIV+$\phi$+$\Gamma_2$ model instead. Notice that trees are at different scales. Figures indicate the estimated average number of substitution rate change per site. Trees (a) and (b) reflect the existence of convergent functional divergence type-I between Cyanobacteria and green $\alpha$ homologs as compared with other groups. In contrast in tree (c) using the DIV+$\phi$ model there are no substitution rate changes on this hypothetical branch.

distance based methods (herein NJ) result in a clear high bootstrap support for the green $\alpha$ and Cyanobacteria cluster (200 replicates, 93%) larger than obtained on the real dataset. This suggests that the high support in the original analysis of Valentin (1997) could easily be the product of substitution rate shifts and divergent RAS heterogeneity.

In contrast, ML reconstruction does not produce a high parametric bootstrap for the same grouping (43%). Although greater than the expected 33% it is not significantly so (200 replicates, $\chi^2 = 0.182$, p-value = 0.67). Moreover very short or zero inner-most branch lengths indicate that ML is not sensible to the amount of covarion shifts recovered using the model described in Methods. This observation is in agreement with the weak support obtained for the green $\alpha$ and Cyanobacteria in the real dataset and the widely accepted view that ML is more robust to departures from the evolutionary model assumed. This evidence adds up to previous observations in other chloroplast gene phylogenetic analyses (Lockhart et al., 2006).

As with maximum-likelihood, Bayesian analysis of the simulated data also produced an unresolved trichotomy of green plant $\alpha$, red plastid and Cyanobacteria homologs. Surprisingly
this method reported the strongest support for the Cyanobacteria and green α grouping in the real dataset (Figure 3.2) even considering covarion evolution (COV$_{\text{H}}$). Therefore, if plastid SecA is indeed monophyletic this dataset would constitute an empirical example where ML outperforms Bayesian analysis. Nevertheless, an alternative explanation based on a LGT event between Cyanobacteria and green plants is always possible turning this statement around.

3.4 Discussion

3.4.1 Functionally divergent genes as phylogenetic markers

Barbrook and colleagues already argued against the use of SecA as supporting evidence for a polyphyletic origin of endosymbiotic plastids (Barbrook et al., 1998) on the basis of heterogeneous compositional bias using DNA data. Additionally we found that, firstly, green plants have two SecA homologs, and secondly, even organisms with a single copy, in this case Cyanobacteria, show strong evidence of non-orthologous evolutionary patterns. This is indeed an issue at least for mainstream models of evolution that assume constant evolutionary constraints.

Most popular time-homogeneous covarion models allow for steadily paced changes in evolutionary constraints overtime. Several studies have proven that these can fit the data substantially better than ordinary homotachous models with only adding a few extra parameters (Galtier, 2001; Huelsenbeck, 2002; Wang et al., 2007). Moreover, whether covarion evolution of this kind is taken into account or not may lead to different conclusions in phylogenetic studies (Shalchian-Tabrizi et al., 2006). Nevertheless, these models do not allow for punctual episodes of dramatic covarion shifts that may occur as a result of gene subfunctionalisation or changes in the environment. Based on our findings using the functional divergence models described in Methods, this is clearly the case in the dataset analysed here in at least two occasions prior to the radiation of Cyanobacteria and green plants (only affecting the α paralog): first, rates across sites are more heterogeneous in these two lineages, and second, functional divergence type-I is also stronger in these two groups. Additionally, these evolved significantly slower that the other three clusters.

Time-homogeneous covarion- and time-heterogeneous divergence models used in this work account separately for different aspects of the evolutionary process achieving significant improvement of model-data fit (Table 3.2). This indicates that the real evolutionary process probably encompasses both aspects: sudden dramatic covarion shifts combined with progressive changes in constraints or background drift.

Several studies, including this one, have shown that phylogeny reconstruction based on datasets that exhibit non-stationary evolution patterns may lead to conflicting or wrong phylogenies (Lockhart et al., 1994, 1998; Inagaki et al., 2004; Susko et al., 2004; Kolaczkowski and Thornton, 2004; Lockhart et al., 2006). Therefore, in systematic studies we suggest to avoid
using datasets that show evidence of complex heterotachous evolution. In the SecA chloroplast gene phylogeny, leaving covarion evolution aside, our results indicate that this dataset probably lacks the phylogenetic signal that is required to determine with confidence the polyphyly or monophyly of primary endosymbiosis or to have ever challenged the now well-establish single primary endosymbiosis theory.

3.4.2 Paralogy of SecA in green plants

Most prokaryotes and chloroplasts only have a single copy of the SecA gene. Nevertheless previous studies revealed the presence of two copies in pathogenic genera *Mycobacterium*, *Staphylococcus*, *Streptococcus* and *Listeria*. These findings catalog SecA and its client proteins as potential targets for drug and vaccine development programs due to the role of SecA2 in secretory pathways linked to virulence (Braunstein et al., 2003; Houben et al., 2006; Hinchey et al., 2007).

It is doubtful that the presence of two copies in the green plant nuclear genome is related to pathogenicity. Nevertheless, this paralogy may provide a valuable contrasting model of SecA subfunctionalisation for in-depth studies of their counter-parts in pathogenic strains. The presence of orthologs in higher plants and basal green algae indicates that both copies have been around for a long time and therefore that they are most probably essential genes. Apparently green β evolutionary process is more similar to the one governing the evolution of red plastid and Firmicutes homologs as compared to green α and Cyanobacteria groups. The former three groups have followed an overall identical heterogeneous process of evolution (with the same RAS distribution $\Gamma(\alpha_{RGF})$ in Table 3.3) and present a similar averaged substitution rate (tree height in Figure 3.2). Moreover a smaller percentage of sites have undergone functional divergence type-I across these groups (Fig. 3.1). Amino acid composition patterns are clearly different in red plastids (Fig. 3.4) but this seems to be due to the fact that this group is the only one encoded in a chloroplast genome. Accordingly, a chloroplast specific rate-matrix, cpREV' (Adachi et al., 2000), has a better marginal fit for this subset as compared to the more general WAG model that offers a better fit for other groups and the whole dataset.

In contrast, green α seems to have undergone more functional divergence type-I than all other groups (Fig. 3.1), including the also divergent Cyanobacteria homologs, thus indicating that it has probably sub- or neofunctionalised into a role that imposes more restrictive evolutionary constraints.

3.4.3 Phylogenetics using rates-across-sites variability

In this study, we built tree-like representations of functional divergence between taxon groups (Figure 3.8) using least-squares. These are based on substitution rate shifts across major taxon groups. As is done in phylogenetic analyses based on residue patterns, substitution rate synapo-
morphies could be used to build a phylogenetic trees. In this particular SecA dataset, green α and Cyanobacteria SecA initially presented patterns of substitution rate convergence. Nevertheless using more robust ML approach revealed that substitution rate shift heterogeneity across sites is the most likely cause. The fact that both groups are also the most functionally divergent leads us to formulate this artifact as a new form of long-branch attraction (LBA); although they are the most conserved groups in terms of residue substitutions, they are also the most divergent ones when we consider substitution rate differences.

Generally, using site relative substitution rate shifts instead of residue changes for phylogenetic reconstruction presents two clear disadvantages. Residues are typically unambiguously known at extant sequences and the set of possible states is also well-known and universal (4 for DNA, 20 for amino acid and 61+3 for codons). In contrast, variability and relative rate of evolution per site is uncertain at tree tips (a hidden state). Moreover, the nature of the RAS distribution is uncertain. In most cases, we can only hope that the selected model is reasonably accurate. The most common choices are a discrete $\Gamma$ distribution, a percentage of variant and invariant sites, denoted $+I$, or both combined, $\Gamma+I$. Nevertheless this assumption is not true in general. For instance, Mayrose and colleagues have shown that the model-data fit can be improved significantly using a mixture of several discrete $\Gamma$ distributions instead (Mayrose et al., 2005).

A priori, these uncertainties give a head-start to inference based on extant residue patterns. In spite of that, modelling pitfalls can equally hamper inference based on both types of data. For example, here we have assumed that a shift between substitution rates is equally probable between all site substitution rate categories for simplicity. Nevertheless, it seems reasonable to allow for arbitrary transition rates between categories instead. In residue-pattern based inference, this would be equivalent to comparing the simple Jukes and Cantor (1969) model with the complex General Time Reversible (GTR) model (Tavare, 1986). In this study we have corrected for a percentage of variable and invariable sites (DIV$\phi$). But perhaps a more accurate substitution rate shift process would involve a number of site categories that change rate at different relative paces such as in the discrete $\Gamma$ RAS model for residues.

Taking into account these parallelisms between residues and substitution rates, it may be worth considering reconstructing deep branches using substitution rate shifts, especially in multi-gene family genealogies when sites are saturated across groups or inter-group branches are very short.

3.5 Acknowledgments

We would like thank Dr. Jennifer Commins for proof-reading this chapter. Science Foundation Ireland and National Development Plant have supported this work.
Chapter 4

Neutral fixation of hetero-oligomerism in the archaeal CCT

This chapter is an update of a previously published manuscript: Ruano-Rubio V., M. A. Fares Testing the Neutral Fixation of Hetero-oligomerism in the Archaeal Chaperonin CCT. Mol. Biol. Evol., 2007, 24:1384–96. Appendix E contains the Supplementary Information for that publication. Both authors wrote the manuscript jointly.

Abstract. The evolutionary transition from homo-oligomerism to hetero-oligomerism in multimeric proteins and its contribution to function innovation and organism complexity remain to be investigated. Here, we undertake the challenge of contributing to this theoretical ground by investigating the hetero-oligomerism in the molecular cytosolic chaperonin containing tailless complex polypeptide 1 (CCT) from archaea. CCT is amenable to this study because, in contrast to eukaryotic CCTs where subfunctionalisation after gene duplication has been taken to completion, archaeal CCTs present no evidence for subunit functional specialisation so far. Our analyses yield additional information to previous reports on archaeal CCT paralogy by identifying new duplication events. Analyses of selective constraints show that amino acid sites from one subunit have fixed slightly deleterious mutations at inter-subunit interfaces after gene duplication. These mutations have been followed by compensatory mutations in nearby regions of the same subunit and in the interface contact regions of its paralogous subunit. The strong selective constraints in these regions after speciation support the evolutionary entrapment of CCTs as hetero-oligomers. In addition, our results unveil different evolutionary dynamics depending on the degree of CCT hetero-oligomerism. Archaeal CCT protein complexes comprising 3 distinct
classes of subunits present two evolutionary processes. First, slightly deleterious and compensatory mutations were fixed neutrally at inter-subunit regions. Second, subfunctionalisation may have occurred at substrate-binding and adenosine triphosphate-binding regions after the second gene duplication event took place. CCTs with 2 distinct types of subunits did not present evidence of subfunctionalisation. Our results provide the first \textit{in silico} evidence for the neutral fixation of hetero-oligomerism in archaeal CCTs and provide information on the evolution of hetero-oligomerism toward subfunctionalisation in archaeal CCTs.

4.1 Introduction

Protein dimerisation and oligomerisation is a universal phenomenon in organisms and proteins with different degrees of oligomerisation and is present in the most evolutionarily conserved protein complexes (Marianayagam et al., 2004). Dimerised and oligomerised proteins are present in many important pathways including functionally important proteins as previously reported (Amoutzias et al., 2004; Lockhart et al., 2006; Marianayagam et al., 2004; Rönnstrand, 2004; Simon and Goodenough, 1998). Whether the transition from homo-oligomeric to hetero-oligomeric protein complexes is responsible for the emergence of new functions and metabolic pathways remains largely unanswered. Some authors have proposed a link between structural stability and self-interaction of oligomers (Dunbar et al., 2004). Others have demonstrated an important effect of the abundance of oligomers on the protein-protein interaction networks geometry (Ispolatov et al., 2005). Examples for homo-oligomeric and hetero-oligomeric protein complexes coexist in organisms irrespective of their complexity. For example, heat shock proteins (Hsps), also called molecular chaperones and chaperonins such as the homo-oligomeric GroEL and the hetero-oligomeric cytosolic chaperonin containing tailless complex polypeptide 1 (CCT), perform equally complex functions and both interact with proteins showing different levels of structure complexity. The fact that Hsps interact with a significant number of proteins in the cell and that the numbers of known protein clients for the different Hsps are qualitatively different make them ideal candidates to test the relationship between the degree of hetero-oligomerism and organisms complexity.

Chaperonins are double back-to-back oriented ringed proteins that assist de novo protein folding in most cellular compartments (Ellis, 1997; Bukau and Horwich, 1998; Frydman, 2001; Hartl, 1996; Hartl and Hayer-Hartl, 2002). Two groups of chaperonins have been characterised, Group I with GroEL being the best-studied protein and Group II represented by CCTs. Both groups of chaperonins are known to share a common 60-kDa protein ancestor that branched at the base of the Tree of Life. Group I evolved along the bacterial lineage and eukaryotic organelles (Bukau and Horwich, 1998; Ellis and Hartl, 1999), whereas Group II evolved along the archaeal and eukaryal lineages (Gutsche et al., 1999; Leroux and Hartl, 2000). Both types of multisubunit
chaperonins are highly similar in their structural as well as domain organisation. Both complexes contain a central cavity that binds unfolded polypeptides, and each subunit comprises three structurally and functionally distinguishable domains, namely, the apical, equatorial, and intermediate domains (Ditzel et al., 1998; Gutsche et al., 2000; Llorca et al., 1998, 1999, 2000; Nitsch et al., 1998; Schoehn et al., 2000a,b).

Despite their structural similarities, important functional differences exist between both chaperonin groups. Unlike GroEL that utilises the ring-shaped co-chaperonin GroES to discharge bound proteins to the cavity of GroEL (Hartl, 1996; Buckle et al., 1997; Kad et al., 1998; Grantcharova et al., 2001), CCTs use a helical protrusion at the apical domain that emulates the function of the co-chaperone GroES (Klumpp et al., 1997; Ditzel et al., 1998; Nitsch et al., 1998; Llorca et al., 1999). CCT also functions in conjunction with the hetero-hexameric chaperone prefoldin (Geissler et al., 1998; Vainberg et al., 1998). In contrast to GroEL, CCTs do not release proteins into the cavity of the complex, but proteins remain bound to the chaperonin (Llorca et al., 2001a). Similarly to GroEL that binds a wide range of different proteins sharing a specific GroES-like motif (Stan et al., 2006), CCT also binds a wide range of distinct proteins including cytoskeletal proteins, such as actin and tubulin, and many non-cytoskeletal proteins, including luciferase (Frydman et al., 1994), G-α transducin (Farr et al., 1997), hepatitis B virus capsid protein (Lingappa et al., 1994), cyclin E (Yang et al., 2005), the Ebsterin-Barr nuclear antigen 1 viral protein (Kashuba et al., 1999), myosin (Srikakulam and Winkelmann, 1999), and the Von Hippel-Lindau tumor suppressor VHL (Feldman et al., 1999). Studies based on proteomic approaches (Gavin et al., 2002; Ho et al., 2002) have extended this list.

One of the most intriguing features of eukaryotic CCTs is their versatility in binding substrate proteins compared with GroEL (Leroux and Hartl, 2000). For example, CCTs have the ability to establish hydrophobic interactions with partially folded proteins (Ditzel et al., 1998; Klumpp et al., 1997) and yet establish specialised non-hydrophobic specific interactions with the cytoskeleton protein actin (Hynes and Willison, 2000; Llorca et al., 1999, 2000; McCormack et al., 2001). This flexibility seems to be correlated with the hetero-oligomerism and subfunctionalisation of eukaryotic CCTs subunits (Llorca et al., 2001b; Fares and Wolfe, 2003). For instance, unlike the homo-tetradecamer GroEL protein, CCTs may present up to 9 subunits per ring with different degrees of sequence divergence (Grantham et al., 2000; Gutsche et al., 1999; Liou and Willison, 1997; Llorca et al., 1999, 2000; Sigler et al., 1998; Valpuesta et al., 2002).

The repetitive lineage-specific gene duplication and conversion in archaeal CCTs (Archibald and Roger, 2002a,b), with most CCTs presenting only 2 or 3 different subunits in each ring (Archibald et al., 1999) argue against duplication as a mean for subfunctionalisation. In their insightful work, Archibald et al. (1999) proposed instead the neutral fixation of evolutionarily trapped CCT hetero-oligomers as a more plausible explanation for such a scenario. To certify the neutral fixation of hetero-oligomerism, 4 conditions have to be met under our point of view:
4Neutral fixation of hetero-oligomerism

1) accelerated fixation rates of amino acid replacements in both copies should have followed
gene duplication as a result of relaxed selective constraints in at least one of the CCT subunits.
Additionally, the average fixation rate of one subunit should have been greater than that for its
paralogous subunit. 2) These substitutions should have cumulated in inter-subunit interfaces at
amino acid sites that became constrained afterward leading to an evolutionary entrapment of
CCT as hetero-oligomers. We distinguished two types of mutated sites depending on whether
they are located in contact regions between subunits in the same ring (within-ring inter-subunit
regions) or between subunits in different rings at their equatorial domains (between-ring inter-
subunit regions). 3) Fixation of compensatory mutations at accelerated rates or by adaptive
evolution in inter-subunit interfaces should have occurred to mitigate the effects of slightly dele-
terious mutations (e.g., mutations that were fixed at high rate despite their slight negative effect
on the biological fitness of organisms) on the molecule structure and or function. 4) Compen-
satory and slightly deleterious mutations should have coevolved to maintain the structural and
functional characteristics of the ancestral homo-oligomer.

We tested all these evolutionary conditions in a dataset including sequences from the
archaean groups Crenarchaeotes and Euryarchaeotes. We investigate this question by conducting
a comprehensive in silico analysis of the selective constraints governing the evolution of the
archaean chaperonin CCT. The results obtained from this approach contribute to building the
theoretical ground for the origin and evolution of hetero-oligomerism.

4.2 Material and methods

4.2.1 Sequence alignment and phylogenetic analyses

We searched for all protein sequences Hsp60 homologs within the Archaea domain using the
NCBI BLAST service (Madden et al., 1996). We filtered out those sequences that appear
to be partial (less than 500 residues long) or redundant (very similar to other sequences in
fully annotated genomes). We also re-annotated the γ subunit sequence for Methanococcoloides
burtonii as the annotated sequence seems to be missing its amino-terminal region. The resulting
subset included 100 sequences. Then we excluded 4 eubacterial GroEL homologs probably
acquired by lateral gene transfer (LGT). Therefore, we constructed an alignment based on a
full dataset of 96 sequences. Nonetheless, phylogenetic analyses revealed a lack of consistency in
lineage resolution due to possible phylogenetic artifacts caused by long branches. Accordingly, we
excluded those groups of sequences, that we did not consider vital for the intended downstream
analyses, suspicious of presenting long-branch attraction (LBA), non-functionalisation, or other
artifacts. Consequently, the final reduced dataset included just 72 sequences. For both datasets,
we aligned protein sequences using the program Muscle v3.6 (Edgar, 2004) with the default
settings. We then aligned nucleotide sequences arranging codons following the protein sequence alignment.

Regarding phylogenetic analyses, first we used ProtTest 1.3 (Abascal et al., 2005) and ModelGenerator v0.82 (Keane et al., 2006) to determine the best candidate substitution rate matrix for maximum likelihood (ML) inference. Both programs pinpointed RtREV + G + F as first option by many criteria despite the fact that this model was specifically devised for retro-transcriptase phylogenies (Dimmic et al., 2002). The second best model based on a different rate matrix was WAG + G (Whelan and Goldman, 2001). Because both models produced similar ML topologies considering just well supported lineages, we decided to use the latter from that point on.

We ran the program Phyml v2.4.4 (Guindon and Gascuel, 2003) upon the full and the reduced datasets to obtain a single ML tree estimate and 1,000 nonparametrical bootstrapped topologies. Then, we used the program Consense from the PHYLIP v3.6 package (Felsenstein, 1989) to summarise those bootstrap replicates in a single fully resolved consensus tree using the extended majority rule method. In either dataset, the consensus tree seemed more reliable than the single ML estimate considering previous publications on the phylogeny of archaea (Brochier et al., 2005a,b). Accordingly, we used consensus topologies for all downstream analyses and figures.

Additionally, we resolved an incompatibility in the order of speciation events between subunits in Halobacterales species by subunit alignment concatenation, ML reconstruction, and bootstrapped consensus tree building.

4.2.2 Testing non-functionalisation of very fast-evolving sequences

We used the program MEGA v3.1 (Kumar et al., 2004) to calculate synonymous ($d_S$) and non-synonymous ($d_N$) change distances between very fast and moderately evolving sequences in Methanomicrobia and Halobacteria clades. We calculated the transition/transversion ratio using the estimator ($\kappa = 1.48$) returned by the program Codeml in the PAML package v3.15 (Yang, 1997) and applied the Nei and Gojobori modified method (Nei et al., 2000). If a sequence or group of sequences are pseudogenes, the ratio $d_N/d_S$ observed between them or against any other sequence outside the group must approach 1. In any case, it must be clearly greater than the ratio obtained between putative functional genes specially when measured within orthologs where there is little room for functional divergence. We also used MEGA to calculate mean residue identity between accelerated, moderately evolving archaea sequences and eukaryotic CCT.
4.2.3 Testing the constancy of amino acid substitution rates after gene duplications

To test whether changes in substitution rates occurred after the different duplication events in archaean CCTs, we used the two-cluster test implemented in the program Lintree (Takezaki et al., 1995). The two-cluster test examines the equality of substitution rates for two clusters linked by a node on the phylogenetic tree. In order to determine what cluster is accelerated with respect to the other, we needed to use an outgroup cluster for the analyses. Because of the high divergence levels between CCTs from eukaryotes and those from archaea, we performed this test dividing the archaea data into two subsets, with the first set including sequences from Crenarchaeotes and the second set including sequences from Euryarchaeotes (Fig. 4.1 shows the phylogenetic tree with the different groups of sequences). We used representative sequences from each CCT subunit of Crenarchaeotes as outgroup sequences for the two-cluster test analysis in the Euryarchaeotes dataset and vice versa.

4.2.4 Identifying lineages and protein regions under selective constraints

To detect accelerated rates of evolution simultaneously in specific lineages of the tree and regions of the sequence alignment, we used the sliding-window-based approach (Fares et al., 2002) implemented in the program SWAPSC v1.0 (Fares, 2004). Briefly, the software slides a statistically optimum window size along the sequence alignment to detect selective constraints and estimates the probability of replacements per non-synonymous sites ($d_N$) and substitutions per synonymous sites ($d_S$). The window size is optimised by means of using a number of simulated datasets. The standard way to measure the intensity of selection when analyzing DNA variability is by comparing $d_S$ with $d_N$ (Akashi, 1999; Kimura, 1977; Posada and Crandall, 1998; Sharp, 1997; Thulasiraman et al., 1999). The ratio between the two rates ($\omega = d_N/d_S$) helps to elucidate if the gene has been fixing amino acid replacements neutrally, under adaptive evolution or

Figure 4.1 (following page): Consensus phylogeny of the archaean CCT sequences dataset composed of 72 sequences. Numbers at relevant internal nodes indicate bootstrap support out of 1,000 repetitions. Different colours indicate paralogy. Notice that these are reused in different taxonomic groups. Sequences grouped within the same colour box or lineages are putative orthologs. Subunit names are arbitrary as usually used in literature and are also independently reassigned across taxonomic groups (i.e. Thermococci $\alpha$ is in no way more similar to Thermoplasmata $\alpha$ than to Thermoplasmata $\beta$). Arrowheads indicate most probable gene duplication events considering topology and pattern of paralogy down toward leaves. The most remarkable difference with respect to the full dataset (Figure E.1) is that Methanomicrobia becomes monophyletic.
4.2 Material And Methods
purifying selection as described in section 1.1.1.2. It has been shown, however, that \( \omega \) is a poor indicator of the action of adaptive evolution (Posada and Crandall, 1998; Sharp, 1997). The reason is that detection of episodic positive selection may be hampered by the strong purifying selection against the majority of amino acid mutations throughout most of the evolutionary time resulting in \( \omega \ll 1 \). Thus, \( \omega \) is in general a conservative detector of adaptive evolution. Additionally, possible selective pressure upon synonymous substitutions may also obscure the interpretation of \( \omega \) (Mayrose et al., 2007).

We used 1,000 simulated datasets in our analysis obtained using the program Evolver from the PAML package (Yang, 1997). To perform the simulations, we took as initial parameters the average \( \omega \) value, transition-to-transversion rates ratio, \( \kappa \), and codon-usage table generated under the Goldman and Yang (G&Y) model (Goldman and Yang, 1994), using the real sequence alignment as input. The program then slides the window along the real sequence alignment and estimates \( d_N \) and \( d_S \) by the method of Li (Li, 1993). Then it determines the significance of these estimates assuming a Poisson distribution of nucleotide substitutions along the alignment. Along with adaptive evolution, SWAPSC also tests for accelerated rates of amino acid substitutions without the restriction of \( \omega > 1 \) (e.g., due to the fixation of slightly deleterious mutations by genetic drift), saturations of synonymous sites, and hot spots (where both \( d_S \) and \( d_N \) are significantly high but where \( \omega < 1 \)).

To figure out whether mutations fixed at inter-subunit interfaces have selectively trapped CCTs into hetero-oligomerism, we investigated the mutational dynamic of these sites in the lineages after-duplication-before-speciation (ADBS) compared with that in the lineage after-duplication-after-speciation (ADAS). If hetero-oligomers were selectively advantageous, we would expect relaxed constraints ADBS permitting the fixation of changes that increase the affinity between distinct monomers. We should then detect strong selective constraints ADAS at these same sites due to their importance to maintain hetero-oligomers.

### 4.2.5 Identifying sites under positive Darwinian selection by ML

We examined whether sites surrounding within-ring inter-subunits or between-ring contacts were under adaptive evolution as to compensate for slightly deleterious mutations fixed at these regions. The combination of both sets of mutations, however, would have no effect on the protein's function or structure (neutral mutations). Compensatory mutations could be also under accelerated rates of evolution, and hence, they did not need to be strictly under adaptive evolution. Selective constraints were hence tested using several codon substitution models implemented in the program Codeml in the PAML package v3.15 (Yang, 1997).

As we wanted to find out whether post-duplication lineages were under specific selective constraints, we compared the fitness of the data with three evolutionary models in two steps.
First, we compared the fitness of the data with the G&Y model (Goldman and Yang, 1994) that assumes a single $\omega$ value for all lineages and sites against the occurrence of different categories of $\omega$ values per site shared across all lineages. Both models are implemented in the program Codeml from the PAML package v3.15 (Yang, 1997) as M0 and M3, respectively. Then, we compared the outcome of M3 with models where each post-duplication branch was allowed to have some sites under a particular $\omega$ value potentially different to the $\omega$ considered for the rest of the tree (the so-called branch-site model B [BSB]) (Yang and Nielsen, 2002; Zhang et al., 2005). M0 is a special case of M3 where all site categories have the same $\omega$ value. Similarly, BSB is an extension of M3 as long as just two site categories are considered (Yang et al., 2000). Therefore, both comparisons can be carried out using the likelihood ratio test.

Finally, to identify possible overestimated $\omega$ values due to saturation or to selection on synonymous sites we applied a sliding-windows analysis (Fares et al., 2002) using the program SWAPSC (Fares, 2004).

### 4.2.6 Analysis of intramolecular coevolution

To test the hypothesis of coevolution between protein regions involved in CCT within-ring intersubunits and between-ring contacts or between these regions and nearby amino acid sites, we used the nonparametric method based on the mutual information criterion (MIC) developed by Korber et al. (Korber et al., 1993) (hereon called MICK) as well as a parametric method coevolution analysis using protein sequences (CAPS) developed by Fares and Travers (2006).

The MI is represented by the entropies that involve the joint probability distribution of states (nucleotide or amino acids) at two potentially coevolving positions in the multiple sequence alignment (Eq. 1.25). The MI-generated values range between 0, indicating independent evolution, and a positive value whose magnitude depends on the amount of covariation. In contrast, CAPS compares the correlated variance of the evolutionary rates at two sites corrected by the time since the divergence of the two sequences they belong to. This method compares the transition probability scores between two sequences at two particular sites, using the blocks substitution matrix (Henikoff and Henikoff, 1992). Variable positions included in the alignment for both types of analyses were those that are parsimony informative (i.e., they contain at least two types of amino acids and at least two of them occur with a minimum frequency of 2). We assessed the significance of the MI values and the CAPS correlation values by randomisation of pairs of sites in the alignment, calculation of their MI or CAPS correlation values, and comparison of the real values with the distribution of one million randomly sampled values. To correct for multiple nonindependent tests, we implemented the step-down permutation procedure in both methods and corrected the probabilities accordingly (Westfall and Young, 1993). MICK is implemented in the program PIMIC (available from the corresponding author on request), and
CAPS is implemented in the program CAPS (Fares and McNally, 2006).

Coevolution analyses were performed in each dataset containing the groups of paralogs highlighted in the phylogenetic tree of Figure 4.1. Each sub-alignment contained at least 9 sequences to ensure a minimum acceptable sensitivity for the methods used (Fares and Travers, 2006). When no more than 8 sequences was available, we added outgroup sequences to perform the analysis. To link specific coevolutionary events (pairs of coevolving amino acid sites) to one or the other paralogous group we followed the procedure described in Fares and Travers (2006). Briefly we first run the coevolution analysis for the entire alignment. Then we repeat the analysis removing one paralogous group at a time. We considered as paralog-specific coevolving pairs those that were no longer detected after removing the corresponding paralogous group.

4.3 Results

4.3.1 Reinvestigating paralogy in the phylogeny of archaeal CCTs

Prior to the analysis of selective constraints in the archaeal CCTs, we reinvestigated the phylogenetic distribution of gene duplications and compared our study with that by Archibald et al. (2001). After applying some quality filters, we assembled a dataset including 96 genuine archaeal CCTs distributed unevenly among species, ranging from a single copy to up to 5 genes in *Methanosarcina acetivorans* that also count with a GroEL homolog. When all sequences were considered, the tree resolution was poor at deep branches of the Euryarchaeotes group probably caused by artifacts due to fast-evolving and poorly sampled clades indicating the possible presence of LBA effects (Appendix E.1). Gene gain and loss seem to be a recurrent phenomenon in the phylogeny of archaea, as pointed out previously (Archibald et al., 2001). We also noticed the occurrence of within-archaea LGT as the most plausible explanation of the deep paralogy of both gene copies in *Methanosphaera stadtmanae* species.

There are 4 sequences remarkably accelerated: 2 gene copies in *M. acetivorans* species and another pair of orthologs in close related species *Natronomonas pharaonis* and *Haloarcula marismortui*. There is no evidence of non-functionalisation in both cases as synonymous versus non-synonymous substitution ratios within accelerated pairs are much lower than 1 ($d_N/d_S \approx 0.1$ and 0.5 for Methanomicrobia and Halobacteriales, respectively). These values are also in the range obtained within related paralogous groups that are putatively functional. Sequence identity indicates that these four sequences are more divergent from the remaining "moderately" evolving archaeal sequences than are eukarotic homologs (26% identity within the domain archaea compared with 35% across domains). Additionally, these sequences belong to species with maximum number of CCT gene copies (4 or 5). These facts indicate that these genes may well constitute separate full-fledged chaperoning systems.
The final dataset comprised 72 sequences that included those genes that belong to well-supported lineages including more than 2 taxa and that allow for a tree resolution consistent with recently published phylogenies for archaeal domain speciation (Brochier et al., 2005a,b). Nonetheless, the support for the deep branches of the tree remained weak (Fig. 4.1). The final tree included a total of 22 sequences for Crenarchaeotes group and comprised species with 2 and 3 subunits. Species with 3 subunits belong to the genera *Sulfolobus*, and previous reports pinpoint the evolution of CCT protein complexes into 9-member rings in this group (Archibald et al., 1999). The paraphyletic nature of the α subunit of Crenarchaeotes attests to the occurrence of a second duplication event in the gene coding for subunit α, resulting in α' and γ (Fig. 4.1).

The Euryarchaeotes group included 50 sequences and comprised two distinct cases of single duplication events (2 subunits) in Thermoplasma and Thermococci clades and two groups of species with 3 subunits, Methanomicrobia and Halobacteriales clades (Fig. 4.1).

### 4.3.2 Gene duplication followed by accelerated rates of evolution

One of the first questions we aimed at answering was whether gene duplication was followed by accelerated fixation rates of amino acid replacements indicating changes in selective constraints. Application of the two-cluster test using the F-corrected amino acid distances between sequences supports that gene duplication was almost always followed by accelerated rates of evolution (Table 4.1). The only exception was the case of the Thermoplasma lineage where, despite the greater fixation rate of mutations in the β subunit compared with the α subunit, the difference was not significant. Our results, however, indicate that one of the subunit clusters has undergone greater acceleration of amino acid replacement rates than the other and therefore duplication altered selective constraints.

### 4.3.3 Accelerated evolution of inter-subunit interfaces after gene duplication

To test the hypothesis of the origin of hetero-oligomerism as a result of accelerated rates of evolution in inter-subunit contacts after gene duplication, we applied the program SWAPSC v1.0 to detect selective constraints. The advantage of using this program over others is that, together with the analysis of $d_N$ and $d_S$, it allows testing saturation or selection of synonymous sites and thus removing them from subsequent analyses. Our first approach was to analyse species presenting CCTs formed by 2 divergent subunits (namely, α and β; note that the nomenclature of the subunit is totally arbitrary). In order to distinguish species-specific selective constraints from those constraints imposed after gene duplication, we focused on duplicates present in more than one species (ancestral paralogy). Selective constraints analyses in CCTs present two different evolutionary scenarios in species comprising 2 distinct types of subunits compared with those
4 NeUTRAL fixation of hetero-oligomerism

<table>
<thead>
<tr>
<th>Clade</th>
<th>Subunits</th>
<th>δ±SE</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euryarchaeotes</td>
<td>Methanomicrobia</td>
<td>(1,2) &lt; 3</td>
<td>0.174 ± 0.030</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 &lt; 2</td>
<td>0.041 ± 0.017</td>
</tr>
<tr>
<td></td>
<td><strong>Halobacteriales</strong></td>
<td>α &lt; β</td>
<td>0.108 ± 0.028</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(α,β) &lt; γ</td>
<td>0.157 ± 0.019</td>
</tr>
<tr>
<td><strong>Thermococci</strong></td>
<td>α &gt; β</td>
<td>0.032 ± 0.010</td>
<td>3.039**</td>
</tr>
<tr>
<td></td>
<td>(β,β') &lt; (α,α',γ)</td>
<td>0.105 ± 0.027</td>
<td>3.834**</td>
</tr>
<tr>
<td></td>
<td>(α,α') &lt; γ</td>
<td>0.217 ± 0.026</td>
<td>8.422**</td>
</tr>
</tbody>
</table>

Table 4.1: Detection of accelerated rates of amino acid substitution using the two-cluster test method. The first two columns indicate the clades and subunit tested. Grouped subunits are listed within brackets. Less or greater inequalities indicate what subunit or group has evolved faster. Subunit names are arbitrary as usually used in literature. δ is the absolute difference between the mean branch lengths of the compared subunits. Z designates the normal distribution value for δ. Significant values at error type I set to 5% and 1% are indicated with (*) and (**) respectively. Here we show only significant asymmetry cases.

with 3 different types of subunits (Fig. 4.1).

Analysis of the duplication events that took place on the ancestor of the Thermoplasmata clade after gene duplication uncovers events of accelerated rates of evolution in the branch leading to subunits α and β at sites directly involved in inter-subunits contacts (Table 4.2; taking as reference the sequence of Thermoplasma acidophilum, accession number NP_394733, for which the three-dimensional structure is available). Sites having undergone accelerated rates of evolution not directly involved in inter-subunit interfaces were located significantly close (less than 8Å distance) to within-ring inter-subunits contact regions that present evidence of accelerated rates of evolution (Table 4.2 and green space-fill structured sites in ). In Figure 4.2a, we rotated the three-dimensional structure to make possible the visualisation of sites under selective constraints. We also verified whether CCTs became evolutionarily trapped in hetero-oligimerism by testing whether these regions were accelerated in branches ADAS as well. All the within-ring inter-subunits contact regions presented significantly greater ω values in the branches ADBS compared with those values in the branches ADAS. In fact, SWAPSC detected these regions as significantly accelerated ADBS and under strong purifying selection ADAS. Differences thus suggest that sites that fixed amino acid replacements immediately after duplication became highly constrained after speciation. Some of the accelerated regions were detected in substrate-binding sites (Fig. 4.2a). Nevertheless these regions are also directly involved or are close to inter-subunit contacts (Ditzel et al., 1998) and they are surrounded by compensatory mutations.

In the Thermococci clade, subunit α was much more constrained than subunit β. The latter
Table 4.2: Amino acid sites detected as having undergone accelerated fixation rates of evolution in the archaeal chaperonin CCT. Amino acids’ one-letter code are followed by a number indicating the position of sites taking the amino acid sequence of *Thermoplasma acidophilum* as reference (accession number NP_394733). This table includes only subunits with significant results within each group of the tree of Figure 4.1. Here we show only sites showing accelerated rates of evolution detected by SWAPSC in regions neighbouring or located in within ring inter-subunit interfaces (black letters), between rings inter-subunit interfaces (black underlined letters) and substrate-binding sites (grey boxed letters). There is no evidence of acceleration for any other functional category. Sites simultaneously involved in substrate-binding and inter-subunit contact are here annotated as substrate-binding sites. Classification of the sites in each one of the three functional groups is based on previous reports (Ditzel et al., 1998).

<table>
<thead>
<tr>
<th>Clade</th>
<th>Subunit</th>
<th>Amino acid site under accelerated evolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermoplasmata</td>
<td>α</td>
<td>I208, **E270, D297, M298, S505, T507, I511, M512</td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>S89, F195, A399, G403, A404</td>
</tr>
<tr>
<td>Thermococci</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>T142, K163, S164, D170, G403</td>
</tr>
<tr>
<td>Methanomicrobia</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>S141, K143, K203, T333</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>E124, M128, A133, K134, V201, I217, **M298, Q300, I334, S339, F360, V361, A450</td>
</tr>
<tr>
<td></td>
<td>(1,2)</td>
<td><strong>V278, D279, S283</strong></td>
</tr>
<tr>
<td>Halobacteriales</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(β,γ)</td>
<td><strong>K24, S123, A133, V167, K202, T333</strong></td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>I503, M512</td>
</tr>
<tr>
<td></td>
<td>γ</td>
<td>P43, R44, S130, V136, S141, I208, E336, I337, <strong>K24</strong>, E25, <strong>R275-M277</strong></td>
</tr>
<tr>
<td>Crenarchaeotes</td>
<td>(α,α′,γ)</td>
<td>I511, M512</td>
</tr>
<tr>
<td></td>
<td>γ</td>
<td>L54, G55, <strong>E270, A299, Q300, P446, R447</strong></td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>E132, T142, A146, N272, M277, E377, T378**</td>
</tr>
</tbody>
</table>

Presented several regions affected by accelerated rates of fixation of amino acid replacements in the lineage ADBS (Table 4.2). These sites had significantly lower $\omega$ values in the branch ADAS (data not shown). Evolutionarily accelerated sites not directly involved in within-ring inter-subunits contacts were neighbouring regions that are involved in inter-subunit contacts (Fig. 4.2b).

Unlike clades of species having 2 different CCT subunits, groups of species presenting 3 different CCT subunits in Euryarchaeotes and Crenarchaeotes showed a more complex evolutionary dynamic. In these groups, regions affected by accelerated rates of fixation in branches ADBS...
4 Neutral fixation of hetero-oligomerism

included sites involved in within-ring inter-subunit and between-ring interfaces and sites responsible for substrate binding (Table 4.2 and Figures 4.2c-e). We also detected accelerated regions close to sites involved in substrate binding, within-ring and between-ring contacts (Fig. 4.2c and e). Other accelerated regions always surrounded most of the accelerated sites in substrate-binding regions. However, some other regions involved in substrate binding also showed adaptive Darwinian selection and were not compensated by neighbouring accelerated regions (see next section). The results on CCT presenting 3 classes of subunits suggest that subfunctionalisation may have occurred in these cases in sites involved in substrate binding.

4.3.4 Episodic Darwinian selection of compensatory mutations

The next step to detect compensatory mutations was to identify mutations fixed by adaptive evolution at regions close to accelerated amino acid sites. We calculated the likelihood of the data under G&Y, M3, and the BSB codon models (Appendix E.2). In all the comparisons, the M3 improved significantly the log-likelihood value with respect to G&Y. This indicates that CCT subunits have fixed changes under heterogeneous evolutionary constraints across codons in general.

To test for the presence of heterogeneous constraints and adaptive evolution in ADBS lineages, we compared the BSB model for each of the lineages with the M3 model. In all cases, the BSB significantly improved the log-likelihood value of the M3 model even after correcting by Bonferroni (Appendix E.2). Positive selection was detected in almost all lineages examined especially in sites neighbouring within-ring inter-subunits contacts (between 4 and 8Å distance) that underwent accelerated fixation rates of amino acid substitutions (Fig. 4.2a-e and Table 4.3). Subunit α from the Thermoplasmata clade showed evidence for Darwinian selection in a

Figure 4.2 (following page): Three-dimensional localisation of amino acid regions having undergone accelerated fixation rates of amino acid substitutions and positive Darwinian selection. We use the structure of Thermoplasma acidophilum (PDB file: 1A6D) that illustrates two adjacent subunits. Because of the homogeneous distribution of selective constraints in CCTs with 3 different types of subunits, only examples for sites under constraints in two (α and β) or combination of two subunits (e.g., the cluster α-β) were plotted in the structure. For example, where 3 subunits were present in a given organism, we only plotted the results of coevolution corresponding to two clusters. In the structure we emphasise important regions only. Regions under constraints are shown as solid sphere structures for Thermoplasmata (a), Thermococci (b), Methanomicrobia (c), Halobacteriales (d), and Crenarchaeotes (e). Regions under accelerated rates of evolution involved in within-ring inter-subunit interfaces, between-ring interfaces, substrate binding, and proximal amino acids to any of these regions are labeled in yellow, pink, blue, and green, respectively. Regions under positive Darwinian selection for the same functional categories are labeled in brown, red, black, and green.
4.3 Results

Accelerated sites
- Intra-ring interactions
- Inter-ring interactions
- Non-functional
- Substrate binding

Sites under positive selection
- Intra-ring interactions
- Inter-ring interactions
- Non-functional
- Substrate binding
single amino acid site located in the substrate-binding region, although this single site was close to a site under adaptive evolution that is located outside the substrate-binding region, indicating compensatory effects. Apart from this example, adaptive evolution in CCTs comprising 2 distinct classes of subunits was always associated to sites neighbouring inter-subunit contact regions (Fig. 4.2a and b). In contrast, adaptive evolution was massively detected in substrate-binding sites in CCTs showing 3 different types of subunits in addition to inter-subunit regions (Fig. 4.2c-e and Table 4.3).

4.3.5 Amino acids coevolution supports compensation of neutrally fixed slightly deleterious mutations

The last condition a mutation should meet as to be considered a compensatory mutation is to present evidence of coevolution with slightly deleterious mutations. This analysis is particularly complex because of the many different intermingled coevolutionary effects, including functional, structural, and interaction coevolution. Furthermore, in many cases sites do not meet the statistical criteria as to be considered in the analysis of coevolution. For example, many of the sites accelerated or under adaptive evolution were discarded because they were not parsimony informative. In spite of these complications, our analysis shows clear intra-subunit coevolution in all CCT clusters examined in Euryarchaeotes and Crenarchaeotes. We also found that most of the coevolution has happened among amino acid sites three-dimensionally proximal. Figure 4.3 presents examples of these compensatory mutations. In this figure, we rotated three-dimensional structures conveniently as to show clearly sites under coevolution. Coevolution occurred among sites belonging to substrate-binding domains, inter-subunit contacts, and neighbour sites in physical contact (less than 8Å distant). The degree of compensation (marked by combinations of at least one site fixing slightly deleterious mutations and a number of coevolving sites under positive selection) however, differs between site categories and clades. In addition to the compensatory relationship between mutations, we also detected coevolution among sites belonging to inter-subunit contacts in all the groups examined (Fig. 4.3). Species with 3 classes of subunit

Table 4.3 (following page): Amino acid sites under adaptive evolution detected using SWAPSC and or the branch-site model implemented in PAML. Only subunits presenting evidence of adaptive evolution at particular amino acid sites are shown. Here we show only sites that are located in important functional regions. Sites located within or neighbouring inter-subunit, inter-ring communication regions, substrate-binding sites and ATP-binding site are in black, black underlined, grey box and grey box underlined, respectively. Sites simultaneously involved in substrate-binding and inter-subunit contact are here annotated as substrate-binding sites. Classification of the sites in each one of the three functional groups is based on previous reports (Ditzel et al., 1998).
<table>
<thead>
<tr>
<th>Model</th>
<th>Subunit</th>
<th>lnLk</th>
<th>Amino acid sites under adaptive evolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>G&amp;Y</td>
<td>-</td>
<td>-82846.864</td>
<td>-</td>
</tr>
<tr>
<td>M3</td>
<td>-</td>
<td>-80906.659</td>
<td>-</td>
</tr>
<tr>
<td><strong>BSB (foreground branches):</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thermoplasmata</td>
<td>α</td>
<td>-80895.128</td>
<td>A207, E271</td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>-80884.337</td>
<td>F90, K134, Y177, L400, G499</td>
</tr>
<tr>
<td>Thermococci</td>
<td>α</td>
<td>-80882.896</td>
<td>V167</td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>-80879.584</td>
<td>V167, A168, K171, A404</td>
</tr>
<tr>
<td>Methanomicobia</td>
<td>(1,2)</td>
<td>-80899.282</td>
<td>S283</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-80889.646</td>
<td>S18, E384</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-80894.809</td>
<td>I137, D138, I144, G145, K202, L240-A242, H301, R275, E276</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-80888.479</td>
<td>R127, D241, N272-R275, Q292, S304, A306, D335, E444, I445</td>
</tr>
<tr>
<td>Methanomicobia</td>
<td>(1,2)</td>
<td>-80899.282</td>
<td>S283</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>-80889.646</td>
<td>S18, K134, S159, V167, K203</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-80894.809</td>
<td>I137, D138, I144, G145, K202, L240-A242, H301, R275, E276</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-80888.479</td>
<td>R127, D241, N272-R275, Q292, S304, A306, D335, E444, I445</td>
</tr>
<tr>
<td>Halobacteriales</td>
<td>α</td>
<td>-80859.259</td>
<td>S18, K78, K134, S159, V167, K203</td>
</tr>
<tr>
<td>Crenarchaeotes</td>
<td>(α,α',γ)</td>
<td>-80892.373</td>
<td>E271, A397, R497, A502</td>
</tr>
<tr>
<td></td>
<td>γ</td>
<td>-80888.914</td>
<td>P226, Q300, R312, R313, I445, A448</td>
</tr>
<tr>
<td></td>
<td>(β,β')</td>
<td>-80893.860</td>
<td>G145, A146, Y191, V201, D209, E271, N272, E379, H380, I390, Q425, V488, Q501</td>
</tr>
</tbody>
</table>
also present some examples of coevolution in substrate-binding regions (Fig. 4.2c-e). However there are also instances of substrate-binding sites under positive selection that neither present evidence of coevolution nor are proximal to sites under acceleration indicating possible subfunctionalisation (e.g. single substrate-binding site under positive selection in Fig. 4.2e subunit α-α'-γ).

Analysis of compensatory mutations in Crenarchaeotes presented very similar results to those of the CCT with 3 different types of subunits of Halobacteriales species group. Crenarchaeotes CCT subunits have undergone fixation of coevolving amino acid replacements at regions neighbouring within-ring inter-subunits contacts (Fig. 4.3e). It is worth noticing that accelerated sites were often detected to coevolve with positively selected sites (Fig. 4.3). When the restriction of parsimony was removed, most of the sites detected in previous analyses as accelerated and under adaptive evolution at inter-subunit contacts were reported as coevolving. In summary, coevolution has been occurring mainly between accelerated sites and sites under adaptive evolution supporting the compensatory role hypothesis for positively selected amino acid substitutions in CCTs presenting 2 distinct classes of subunits. In contrast, CCTs with 3 distinct classes of subunits present cases of non-compensatory Darwinian selection in substrate-binding regions.

4.4 Discussion

In this study, we propose a role for gene duplication and selection in protein functional innovation. Our results studying the hetero-oligomeric molecular chaperonin CCT in archaea yield interesting information about the evolution of protein complexity in terms of hetero-oligomerism versus homo-oligomerism. Initially, there is no structural reason as to why hetero-oligomerism should be more advantageous than homo-oligomerism unless this hetero-oligomerism is linked...
Sites under positive selection
- Intra-ring interactions
- Inter-ring interactions
- Non-functional
- Substrate binding

Coevolving sites
- Intra-ring interactions
- Inter-ring interactions
- Non-functional
- Substrate binding
neutral fixation of hetero-oligomerism

to some functional role. Example of such role is the previously shown correspondence between hetero-oligomerism and subfunctionalisation in the eukaryotic CCT (Fares and Wolfe, 2003). This is strongly supported by the fact that all 8 gene copies have been kept because the rapid successive gene duplications of CCTs took place in the most recent ancestor of all eukaryotes. As ours and others' results indicate, archaean CCT show a strikingly contrasting evolutionary scenario compared with eukaryotic CCTs: recurrent paralogy followed by gene non-functionalisation and disintegration has been a recurrent phenomenon (Archibald and Roger, 2002a,b). These observations combined with evidence for gene conversion in archaean CCTs yield information supporting the role of neutral evolution in the emergence of protein complexity.

Previous work presented the coevolved interdependence between subunits as the most appealing possibility to explain the patterns of recurrent paralogy in archaean CCTs (Archibald et al., 1999). These authors proposed that ancestral archaean CCTs were homo-oligomers but that gene duplication was followed at some evolutionary stage by the neutral fixation of slightly deleterious mutations in one of the CCT subunits. These mutations may have been responsible for a decrease in the fitness (stability) of homo-oligomers therefore favoring fixation of hetero-oligomers. Fixation of mutations in the other paralog may have increased the stability of the hetero-oligomer outcompeting homo-oligomeric CCTs. As a result, archaean CCTs became evolutionarily trapped as hetero-oligomers. A likely evolutionary explanation to such an entrapment is that the first set of mutations fixed in the one paralog was slightly deleterious or nearly neutral, whereas the second set of mutations in the other paralog was fixed by adaptive evolution conditional to the fixation of slightly deleterious mutations (apparently non-neutral mutations). Both sets of mutations had no selective value on the function of the protein when combined and were hence neutrally fixed.

Several lines of evidence in our study support the neutral fixation of archaean CCT hetero-oligomerism and argue against the role of selection through subfunctionalisation or neofunctionalisation. First, we found that most of the mutations fixed in archaean CCTs with 2 different classes of subunits are located within-ring inter-subunit or between-ring interfaces. Second, these regions show a significantly accelerated fixation rate of mutations. Third, accelerated fixation rates of mutations and mutations fixed by adaptive evolution have also affected peptide regions in the structure neighbouring those accelerated interfaces. Finally, these two sets of mutations present evidence of coevolution, indicating a compensatory relationship between them.

Interestingly, we found contrasting evolutionary scenarios when comparing archaean CCTs with 2 different classes of subunits to archaean CCTs showing 3 different classes of subunits. Several mutations in genes coding for CCTs with 3 distinct subunits have affected substrate-binding regions and ATP-binding sites. Some of these mutations were fixed by adaptive evolution in one of the subunits and present no evidence of secondary compensatory changes supporting their possible involvement in subfunctionalisation. These differences between both groups of
CCTs provide evidence in support of a more complex hypothesis to explain the evolution of hetero-oligomerism. A plausible evolutionary explanation for these conclusions is that once hetero-oligomerism became neutrally fixed, distinct CCT subunits started accumulating specializing mutations leading to subfunctionalisation. Our data suggest that subfunctionalisation may depend on the number of distinct subunits available in the complex. In this context, eukaryotic CCTs constitute an excellent example of such evolutionary process taken to completion. On the other hand, dispensability of several subunits in the hyperthermophilic archaeon suggests functional redundancy of some of the subunits and argues against subfunctionalisation by formation of CCTs comprising 3 distinct subunits with a fixed geometry (Kagawa et al., 2003). Analysis of the dispensability of the different CCT genes (cct1, cct2, and cct3) in *Haloferax volcanii* demonstrates dispensability of at least 2 out of the 3 genes (Kapatai et al., 2006). These authors also show that the CCT protein complex in this archaeon is a double ring with 8-fold symmetry but that the rings are mixed complexes of different subunits (Kapatai et al., 2006). Although, this study does not support a fixed geometry for a CCT complex formed from 3 distinct subunits, the different use of combinations of 2 subunits suggests that at least one of the subunits may have undergone subfunctionalisation after gene duplication, something also supported by the fact that CCT3 (called CCT γ subunit in this work) cannot support growth on its own (Kapatai et al., 2006). The variability in the dispensability of the different CCT subunits in different organisms and our limited knowledge about the flexibility of the CCT complex geometry preclude stating definitive conclusions and demonstrate that there is still much to be learnt about CCT complexes' structural and evolutionary properties in different archaeal lineages.

Coevolution between CCTs and client proteins in eukaryotic cells may have fuelled a subfunctionalisation process in eukaryotic CCTs compared with archaeal CCTs. For instance, there are several examples of coevolution between eukaryotic CCTs and hetero-oligomeric proteins, also generated at the origin of the eukaryotic cell, such as actin and tubulin (Horwich and Willison, 1993; Llorca et al., 1999, 2000). Tubulin binds to 5 CCT subunits in two different arrangements, utilizing hence the 8 CCT subunits (Llorca et al., 2000). Llorca et al. (Llorca et al., 1999, 2000) proposed subfunctionalisation leading to protein-binding specialisation in CCT to explain the binding and folding model with actin and tubulin. Greater number of distinct subunits would permit greater possible arrangements in which protein clients could bind CCTs and greater protein client's versatility. This is only possible through the coevolution between these protein clients and CCTs. Then, it follows that CCTs hetero-oligomerism and protein client's versatility are two tightly linked phenomena. In conclusion, hetero-oligomerism may well be related to the gain of cell protein complexity. Establishing a ground theory for the origin and evolution of hetero-oligomerism may unearth breakthrough information on the origin and evolutionary factors responsible for the emergence of cell complexity in eukaryotes.
4.5 Acknowledgments

This work has been supported by Science Foundation Ireland. We would also like to thank the editor and the reviewers of the related publication for their helpful suggestions.
Chapter 5

Detecting concerted evolution in yeast after genome duplication

The materials in this chapter and Appendix F have recently been submitted for review to a peer-reviewed journal.

Abstract Non-independent sequence evolution has important repercussions on the evolutionary analyses within multi-gene family members. Often, such analyses rely on the simplistic, yet mathematically convenient, assumption of independent evolution of paralogs. Also concerted evolution is an important aspect of functional evolution of paralogous genes. Here we assess the occurrence of concerted evolution in yeast and other related species that diverged relatively soon after a whole genome duplication (WGD) event that took place 100Myr ago. We find that concerted evolution correlates with differences in inter-species patterns of gene loss, gene function and the species under analysis. We apply detection methods previously used on yeast together with a new distance based method we have developed. We also make important remarks on probable pitfalls when applying statistical tests to detect concerted evolution and propose procedures to avoid false discovery.

5.1 Introduction

Homologous sequences are the product of several alternative rounds of speciation (orthologies) and intra-specific duplication events (paralogies). Typically, in most evolutionary analyses, investigators assume that each paralogous sequence follows a totally independent evolutionary process after duplication. Although this supposition is convenient for data modelling, it may still lead to reliable conclusions as long as there is a reasonable level of independent evolution between homologs.
This is in fact almost certain in most strict orthologies (single gene family homologies) except when inter-specific genetic exchange (lateral gene transfer, recombination, and so forth) or convergent changes in functional and evolutionary constraints are common. In contrast, diverse molecular selective processes, such as concerted evolution (introduced in section 1.2.3.4), may ligate the evolution of intra-specific paralogs.

Concerted evolution results from intra-genomic recombination mechanisms, such as gene conversion and unequal crossing-over (Zimmer et al., 1980), and population scale genetic processes or molecular drive (Dover et al., 1982a,b). Nevertheless, concerted evolution patterns may also arise from selective pressures. For example, “fail-safe” redundancy or increased gene dosage may be preserved by purifying selection or convergent coevolution between homologs. Consequently, these processes would result in functional conservation and more similar intra-genomic sequences. Only in cases of recent or ongoing concerted evolution we can differentiate between these two possible causes using synonymous distances. Typically purifying selection should have a stronger influence on non-synonymous substitutions as compared to synonymous substitutions whereas molecular mechanisms such as gene conversion should homogenise every homologous nucleotide site. Nevertheless, selection may also influence strongly the fixation of synonymous substitutions complicating the distinction between these two causes (Mayrose et al., 2007).

Additionally, other genome specific parameters that shape the evolution of groups of genes may contribute to sequence homogenisation. Compositional biases, such as variable GC content or codon preference, reduce the effective number of possible states or residues at each site and therefore the number of pairwise differences or other naive similarity measures (Pesole et al., 1995). Although these parameters are quite relevant in genome evolution studies (Akashi, 1994; Bentley and Parkhill, 2004), they may be regarded as neutral in the context of concerted evolution. That is, similarities between inter-specific paralogs explained only by compositional bias alone are not in fact extraordinary. Thus, by themselves, they are not a definitive indicator of coevolution of homologous sequences. Accordingly, throughout this chapter we consider the detection of genuine concerted evolution patterns only caused by active molecular drive and or selective pressures for functional conservation.

A countless number of studies have proposed many different statistical methods to determine the presence of concerted evolution (by looking for evidence of gene conversion, recombination or sequence convergence in general). Drouin and colleagues (Drouin et al., 1999) proposed a full gene conversion analysis strategy including an excellent summary of methods that were available up to that date. Nevertheless other methods have been proposed since then (Wiuf, 2000; Archibald and Roger, 2002a; Gloor et al., 2005; Ezawa et al., 2006; Kedzierska and Husmeier, 2006; Sugino and Innan, 2006).

In this chapter we investigate the occurrence of concerted evolution in a particularly interesting collection of paralogous gene pairs resulting from a whole genome duplication (WGD).
This event took place in the yeast lineage 100Myr ago (Wolfe and Shields, 1997; Kellis et al., 2004). We considered this dataset for two reasons: first, we have a special interest in it as an excellent “substrate” for divergent paralogous evolution research and second there is a vast amount of accurate information available on WGD paralogs making analyses easier and results more conclusive. We focused in the distribution of concerted evolution cases among different patterns of gene loss across species and functional categories.

5.2 Material and Methods

5.2.1 Sequences and alignments

We retrieved protein and nucleotide sequences from relevant sequencing projects or databases: Saccharomyces Genome Database (SGD) (Cherry et al., 1998) for Saccharomyces cerevisiae, Saccharomyces bayanus, Saccharomyces castellii (we used protein sequences as re-annotated in the Yeast Gene Order Browser mentioned below) and Saccharomyces kluyveri, Kellis and colleagues’ supplementary material for Kluyveromyces waltii (Kellis et al., 2004), Centre de Bioinformatique de Bourdeaux (CBI, http://cbi.labri.fr) for Candida glabrata and Kluyveromyces lactis and the NCBI RefSeq database (Maglott et al., 2000) for Candida albicans and Debaryomyces Hansenii.

We aligned protein sequences using Muscle v3.52 (Edgar, 2004). We subsequently designed coding nucleotide sequences as indicated by the corresponding protein alignment. To ensure alignment quality we removed badly aligned areas and gaps using the program Gblocks v0.91b (Castresana, 2000). We ran Gblocks on nucleotide alignments with default parameters and setting the data type to codons. We translated the resulting nucleotide alignments back into amino acids for further analyses based on protein sequences. We also removed partial sequences that would suppose a substantial loss of information due to the complete gap deletion policy.

5.2.1.1 WGD duplicates dataset

Byrne and Wolfe developed a web based Yeast Gene Order Browser (YGOB) that dynamically reconstructs genomic homologous regions across closely related species of yeast (Byrne and Wolfe, 2005). As we carried out this study, the public version of the browser had information on 7 species: 4 pre-WGD (A. gossypii, K. lactis, K. waltii and S. kluyveri) and 3 post-WGD (S. cerevisiae, S. castellii and C. glabrata) that diverged before and after the WGD event, respectively. The latest version of the browser includes a new post-WGD species, Kluyveromyces polysporus. Nevertheless, we decided to exclude this species from our analysis as the resolution of paralogies and orthodoxies remains relatively sketchy as compared to the resolution between other post-WGD species (cf. Scannell et al., 2007).

In YGOB terminology a pillar is the set of homologous sequences encoded by the ancestral
locus and resulting duplicated loci. Absences (gene losses) are also annotated as such. They
classified those pillars with at least one surviving duplicate per each post-WGD species based
on differential gene loss patterns: which species have lost a duplicate and whether losses are
orthologs or paralogs across species. Here we only considered those categories with at least
one surviving pair of paralogs, as just then a concerted evolution detection test is applicable.
We re-classified them in four categories (Figure 5.1): those with two copies present in all three
post-WGD species, class C0.0 (250 cases in total), with one single loss in all post-WGD species,
class C1.0 (149), and a single extant paralogous pair where lost homologs were orthologs, class
C2.0 (152), or paralogs, class C1.1 (36)\(^1\).

In those analyses that required \textit{C. albicans} or \textit{D. hansenii} homologs, we searched them using
hidden Markov model (HMM) as implemented in the package HAMMER v2.3.2 (Eddy, 1998).
We built HMMs using full protein alignments (before filtering codon positions using Gblocks)
of post- and pre-WGD species. Finally, we selected the best significant hit (E-value < \(10^{-5}\)) in
\textit{C. albicans} and \textit{D. hansenii} if available.

5.2.1.2 Control dataset: inter-specific single gene families

Apart from the WGD duplicates, we also built single gene families alignments including homologs
species. We did so with two objectives in mind: first, to determine the species phylogeny
of post-WGD species, and second, to build a control dataset to check on the performance of
concerted evolution detection methods when the null hypothesis is true. This way we could
determine whether these are too liberal or conservative on real sequence alignments. Unexpected
statistical structures present in real data can challenge the apparent accuracy obtained using
simulated or permuted datasets.

We used the standalone version of NCBI BLAST engine v2.2.2 (Madden et al., 1996). We
excluded those gene homologs with non-self intra-specific BLAST hits with an E-value ≤ 0.05.
Then we built maximal orthologous sets from different genomes so that each member of the
group can find all other members. This upper-bound E-value may seem too high as we were
carrying out multiple tests. Nevertheless, we performed further analyses just on groups of 6 or
more sequences ensuring a reasonable credibility level of the resulting homologous sequence set.
This selection process generated a total of 413 sequence alignments lacking concerted evolution;
at least as long as we ignore the possibility of inter-specific recombination events.

\(^{1}\)Byrne and Wolfe originally named these classes C0, C1, C2(D-F) and C2(A-C) respectively (Byrne and
Wolfe, 2005).
5.2 Material and Methods

Figure 5.1: Extended Venn diagram that indicates the number of pillars in each category, species and combinations of both. Each circle represents a post-WGD species; names are located adjacent. Intersections represent pillar subsets with common extant paralogous pairs across species. When there is only one extant paralogous pair in a single species (and two gene losses in the other two species) we classify pillars in two subgroups within the corresponding non-intersecting area: reciprocal gene losses (class C1.1) and orthologous gene losses (class C2.0). Figures in brackets indicate the total number of pillars that count with all post-WGD paralogous pairs after leaving out short sequences. Top-right corner boxes inscribe the total number of pillars per each category. Arrows represent possible transitions between categories based on single gene loss events. A single loss in the common ancestor of \textit{S. cerevisiae} and \textit{C. glabrata} is enough to pass directly from C0.0 to C2.0.

5.2.1.3 Intra-specific multi gene families pairwise distances

To determine whether there are different genome-wide levels of conservation of nearly identical gene duplicates across species, we grouped genes into families considering non-necessarily mutual intra-specific BLAST hits (E-value \(\leq 10^{-20}\)). We merged together in families those genes that have some common hit in an inclusive fashion (first, second hand hit and so forth). Then we calculated all possible pairwise distances within each gene family. Big families produce a quadratically larger number of pairwise distances \(O(n^2)\) where \(n\) is the family size. Thus, in order to produce an unbiased pairwise distance distribution plot we used weighted bootstraps (10000 repeats each graph). The probability of drawing a distance from a particular gene family is proportional to the total number of genes that belong to that family.

We classified pairwise distances depending on whether the paralogs existed before, or were the result of the WGD event. In order to do that we used synteny information, taking \textit{K. waltii} as scaffold species as posted in the YGOB (downloadable synteny score file).
5 Detecting concerted evolution in yeast

5.2.2 Concerted evolution and gene conversion tests

Several studies have previously looked for and reported concerted evolution in yeast. On one hand, some addressed concerted evolution as a digression or did not perform a corroborative statistical test (Kellis et al., 2004; Pyne et al., 2005). On the other hand, some carried out a formal test to identify and characterise concerted evolution cases. For example, Drouin (2002) used Sawyer’s approach (Sawyer, 1989) based on distribution of pairwise changes between paralogous pairs to determine the occurrence and length of gene conversion regions within multi-gene families in yeast. Gao and Innan (2004) used nucleotide synonymous distances, $d_S$, and tree building to unveil as to what extent concerted evolution could artifactually increase gene duplication rate estimates as reported by molecular clock based methods. Lin and colleagues proposed their own distance-based approach to check on the correspondence between gene conversion and codon usage bias in species *Saccharomyces sensus stricto* (Lin et al., 2006). Sugino and Innan used maximum-likelihood (ML) and parsimony to measure the duration of concerted evolution in yeast WGD duplicates (Sugino and Innan, 2005, 2006).

In this study we utilised consensus tree drawing, heterogeneous distribution of changes and our own distance based method to detect the presence of on-going or past concerted evolution in post-WGD species. We also used an ML approximation to Sugino and Innan approach to further compare concerted evolution in different pillar classes.

5.2.2.1 Heterogeneous distribution of changes

Sawyer’s test of gene conversion roughly consists of looking for regions of homology between two sequences of a size greater than expected by chance taking into account just polymorphic sites (those that vary in the containing homologous sequence alignment) and the overall similarity between both sequences. Sawyer initially proposed this method in his original manuscript (Sawyer, 1989) although there is a more recent version implemented in the software program GENECONV (http://www.math.wustl.edu/~sawyer/geneconv/).

Several people have applied this method to a variety of taxonomic groups. However not all of those studies took into account potential artifacts caused by non-random patterns present in the data that are not gene conversion by-products. For instance, the possible impact of mutational cold and hot spots, especially with the lack of outgroup homologous sequences, was briefly mentioned by the author himself in the original manuscript (Sawyer, 1989) and later by Semple and Wolfe (1999). Accordingly, arbitrarily large segments of homology may be found within contiguous regions with a low mutational rate. In that case, these may well be reported as putatively convergent.

Drouin, in his work on yeast protein coding sequences, considered all nucleotide positions rather than just silent sites (Drouin, 2002). Since different amino acid regions are subject of
alternative functional and evolutionary constraints, non-silent substitutions are probably fixed at different rates. Drouin argues (personal communication) that the risk of obtaining false positives, probably resulting in significant short fragments, is outweighed by the gain in power to detect genuine gene conversion based on a greater amount of data.

In this study, in order to assess the percentage of potential false positives using Sawyer's approach, we looked for concerted evolution on strict single gene families assembled as described in Methods. Gene conversion or any other form of concerted evolution must not be present in these datasets due to inter-species barriers. We assessed the performance of GENECONV v1.81 using amino acid alignments (-SeqType=PROTEIN option), nucleotide sequences (-SeqType=NUCLEOTIDES option) and codon sequences (-SeqType=SILENT option). We allowed for a maximum of two mismatches per fragment (-gscale=2 option). Then we proceeded to apply this method on the WGD paralogous dataset using set-ups that do not produce a significantly high number of false positives in the control dataset.

5.2.2.2 Distance based methods

Lin and colleagues used their own distance based method to determine the presence of putative concerted evolution between WGD paralogs of close related yeast sensus stricto species: \textit{S. cerevisiae}, \textit{S. bayanus}, \textit{S. mikatae} and \textit{S. paradoxous} (Lin et al., 2006). To do so, they looked for regions in the alignment where the pairwise distance between paralogs (number of nucleotide differences) is significantly smaller than the average distance within orthologous groups. Despite its apparent simplicity, this approach must have performed quite well on their data as the time-lapse between gene duplications and the earliest speciation event (giving raise to the \textit{S. bayanus} lineage) is considerably large (\approx 80Mya). Under those conditions, even relatively fast and saturation prone silent nucleotide sites can be very informative. Nevertheless, by the same token application on earlier divergent species as is the case with \textit{S. cerevisiae}, \textit{C. glabrata} and \textit{S. castellii} may not perform as well as expected because duplication and speciation are more contemporaneous. Moreover, possible relaxation of the molecular clock could make paralogs significantly less divergent than orthologs without the intervention of concerted evolution.

We have independently developed a distance-based method that although taken from a different angle, turned out to "generalise" Lin's approach without the molecular clock-like restriction. In contrast to Lin's method, here we do not test whether the pairwise distance between paralogs is smaller than the average between orthologs but rather whether it is as great as expected considering all other pairwise distances. In this study, we did not consider scanning particular regions of the alignment for concerted evolution as Lin and Sawyers methods actually do. Nonetheless it may be easily extended to incorporate a sliding-window or some other region scanning mechanism.
Let us consider the unrooted quartet of sequences \((a, b, c, d)\) as depicted in Figure 5.2. The four-point condition (Buneman, 1974) states that the sum of actual pairwise distances of both pairs of sister taxa in the true topology is the smallest of the three possible unrooted topologies:

\[
(D_{ab} + D_{cd}) \leq (D_{ac} + D_{bd}) = (D_{ad} + D_{bc})
\]

where \(D_{xy}\) is the pairwise distance between sequences \(x\) and \(y\). Consequently, the inner-most branch’s length, \(l_{5}\), is easily calculable as:

\[
(D_{ac} + D_{bd}) - (D_{ab} + D_{cd}) = (D_{ad} + D_{bc}) - (D_{ab} + D_{cd}) = 2l_{5}
\]

Assuming that concerted evolution can only occur between sequences \(a\) and \(c\), then the only pairwise distance that may be affected is \(D_{ac}\) with an observed value \(D'_{ac}\) systematically smaller depending on the intensity and duration of the sequence homogenisation. If the difference between the divergence expected under independent evolution and the observed distance is high enough, \(E[D_{ac} - D'_{ac}] > 2l_{5}\), at least the distance based phylogenetic reconstruction method based on the four-point condition (Eq. 5.1) would report inconsistent trees where taxa \(a\) and \(c\) cluster together.

The second part of the inequality (5.1) reveals an alternative estimation for \(D_{ac}\) that does not depend on \(D'_{ac}\):

\[
\tilde{D}_{ac} = D'_{ad} + D'_{bc} - D'_{bd}
\]

From this point on we will refer to this alternative estimator, \(D'_{ac}\), as the expected or concerted-free distance between paralogs. Then in order to find out whether paralogs have evolved in concert we will test whether observed- and expected distances are significantly different. We generalised this method in order to handle phylogenies with more than 4 taxa and several concerted evolving candidate pairs simultaneously as further developed in Appendix F.1.

![Figure 5.2: Unrooted quartet \(((a, b, c, d)\). Asymptotically, every pairwise distance \(D_{xy}\) must correspond to the sum of branch length, \(l_i\), along the path between taxa \(x\) and \(y\). E.g. \(D_{ac} \rightarrow l_1 + l_5 + l_3\) (dashed line).](image-url)
In the control dataset based on single gene families is not clear what sequence pair to use as candidate for concerted evolution as there is no actual paralogy. For each alignment, we selected a post-WGD species sequence (\textit{S. cerevisiae}, \textit{C. glabrata} or \textit{S. castellii} in this order) and a pre-WGD sequence from \textit{S. kluyveri} or \textit{K. waltii} (closest pre-WGD relatives); actual WGD paralogies must be younger. As regarding the yeast WGD real duplicates dataset, we took every pair of surviving paralogs as amenable to concerted evolution. We also added pre-WGD species homologs to increase accuracy as typically more data decreases estimator variance. In fact, this last step is required if both post-WGD gene losses are orthologous (class C2.0) as this is the only way to build quartets where paralogs are situated at alternative sides of the inner-most branch as depicted in Figure 5.2. We will discuss alternative ways to calculate the significance of the test within the discussion of results on the control dataset below in Results.

We analysed amino acid and nucleotide data separately. In analyses based on nucleotide sequences, we initially filtered out sites prone to introduce compositional bias artifacts. Typically researchers do indiscriminately remove all third codon positions as they amount for most of the synonymous mutations present in the dataset. First and second codon position are normally left intact although they may also contain synonymous differences. Here we took a more selective approach. We considered a nucleotide site in the analyses if: all sequences have different amino acid residues at the containing codon, or sets of sequences that encode the same residue have only two members, or, if there is more than two members, they all have the same nucleotide at that position. This way we eliminated only sites that contain a mixture of identity and synonymous mutations between sequences coding for the same amino acid. Nevertheless, we also analysed full nucleotide alignments in order to demonstrate the artifacts introduced when we use all the positions of the alignment.

We developed the software required for this test using the Java programming language borrowing functionality implemented in the PAL library v1.5 (Drummond and Strimmer, 2001). We also developed some gluing Perl and XSLT scripts using Bioperl v1.5 (Stajich et al., 2002) and Saxon v8.8J package (http://saxon.sourceforge.net) respectively.

### 5.2.2.3 Phylogenetic tree drawing

Another simple and very popular method to detect concerted evolution is phylogenetic tree reconstruction. Paralogs evolving in concert should appear topologically closer than expected. As a matter of fact we can claim as evidence of concerted evolution only a strongly supported monophyly of both paralogs; lack of phylogenetic signal can explain an unresolved topology.

Therefore this approach guarantees consistent results only if the concert is strong enough and affects the whole sequence alignment. Weak but sequence wide concerted evolution may show up just as a barely noticeable tendency to join both paralogs. Strong but partial sequence
concerted evolution may have more or less influence depending on how many sites are affected and how fast these evolve.

Clade support is typically quantified using bootstrap replicates. However tempting is to interpret bootstrap support as p-values or the probability that the clade exists, previous studies revealed them as poor estimators for such figures (Holmes, 2005). In the high support range, they are typically an underestimation of the actual clade probability. Many take 70% as significant support for a node roughly equivalent to a type I error of 0.05 (Hillis and Bull, 1993). Here we used this simplistic rejection criterion although some more elaborated and mathematically sound but less popular methods are also available (Felsenstein, 2004; Holmes, 2005).

In this study we used weighted least-squares (WLS) distance based inference alongside of maximum-likelihood (ML). For WLS we used the implementation in FITCH program, part of the PHYLIP v3.6 package (Felsenstein, 1989). For ML trees reconstruction we used the fast heuristic search implemented in PhyML v2.4.4 (Guindon and Gascuel, 2003). In both cases we produced 100 bootstrapped repeats that we summarised into consensus topologies using the majority-rule extended method implemented in CONSENSE program, also part of the PHYLIP package.

5.2.2.4 Heterogeneous duration of concerted evolution

Sugino and Innan investigated the duration of concerted evolution in yeast WGD duplicates (Sugino and Innan, 2005). They assumed that paralogs have evolved in concert from duplication up to an arbitrary point in time after which they have become too divergent as to homogenise.

The lack of concerted evolution is then a special case where the duration was in effect zero (paralogs diverged right after duplication). They concluded that concerted evolution was quite common in yeast WGD paralogs as the data fitted better a model where each paralogous pair has undergone concerted evolution for an arbitrary amount of time exponentially distributed. They also claimed that the WGD event took place right after or during the speciation that gave raise to the K. waltii lineage. Nevertheless, there is contrasting phylogenetic and genomic evidence that gives support to the occurrence of several rounds of speciation between these two events (Wong and Wolfe, 2006). Consequently, we took the later statement with caution.

Following the scheme depicted in Figure 5.3 and assuming that there was no significant change in evolutionary rates between WGD paralogs and pre-WGD homologs (molecular-clock), we can quantify an upper-bound value, $\tau$, to the proportion of time expended in concert after the WGD event:

$$\tau = 1 - \frac{b}{c} \quad (5.4)$$

where $b$ is proportional to the apparent age of the paralogy, $b = b_1 = b_2$, and $c$ the age of the K.
5.2 Material and Methods

*waltii* speciation event. In a later work, Sugino and Innan investigated the correlation between the duration of concerted evolution and gene expression levels (Sugino and Innan, 2006). That time they approximated $\tau$ per each WGD yeast paralog using parsimony distances.

Here we did so per each WGD paralogous pair and species in order to compare distributions of this statistic across pillar classes. We used ML branch length estimators instead of parsimony distances. We tested the molecular clock assumption using the CODEML program, included in the PAML v3.15 package (Yang 1997), and the likelihood-ratio test (LRT) (Hoel, 1962). We built quartets composed of each paralogous pair, a close pre-WGD species homolog, *K. waltii* or *S. kluyveri*, and a rooting outgroup from *D. hansenii* or *C. albicans* in this order.

Unfortunately a significant number of genes rejected the molecular clock hypothesis ($\approx$65% overall). We believe that this high percentage of rejection in comparison to Sugino and Innan data is due to the fact that they only considered very conserved regions ($\geq$90% identity in first and second codon positions only) as to ensure alignment quality. In the absence of a molecular clock, we proceeded instead calculating $\tau'$ as:

$$
\tau' = \frac{a}{a + b'}, \quad b' = \frac{(b_1 + b_2)}{2}
$$

(5.5)

where $a$ is proportional to the apparent amount of time between speciation and paralogy as illustrated in Figure 5.3. Again, we used branch lengths as estimated by CODEML.

![Figure 5.3: Rooted ultrametric quartet tree assumed to calculate the duration of concerted evolution. It consists of two WGD extant paralogs in a single post-WGD species and two outgroup homologs, *K. waltii* (*S. kluyveri* in its absence) and *D. hansenii* (*C. albicans* in its absence). If the WGD event took place right after *K. waltii* speciation (arrow), the duration of concerted evolution of paralogous copies is proportional to the length of the middle branch, a. In the absence of concerted evolution, tip branch lengths $b_1$, $b_2$ and $c$ are approximately equal.](image)
5.2.3 Molecular evolution models

In this study, whenever we needed to specify a probabilistic molecular evolution model, we used WAG (Whelan and Goldman, 2001) and HKY (Hasegawa et al., 1985) instantaneous change rate matrices for amino acid and nucleotide sequences respectively. We modelled heterogeneous rates across sites using a discrete $\Gamma(\alpha)$ distribution (16 categories) and a percentage of invariants sites, $I$, for both types of data. We fixed nucleotide base frequencies as observed in the input alignment (empirical frequencies). In pairwise distance based analyses, we used PhyML to optimise site substitution rate shape parameters, $\alpha$ and $I$, branch lengths and the transition/transversion ratio, $\kappa$, (nucleotides only) based on the actual topology as resolved by the syntenic information available in the YGOB. We used the resulting model parameters to estimate all pairwise distances using again the ML criterion. We did similarly for concerted evolution duration analyses but using the CODEML program instead. In this case, however, we did not use invariants since this program does not support them.

5.2.4 Post-WGD species phylogeny

We concatenated single gene family amino acid alignments from the control dataset and removed dubious regions using Gblocks but this time allowing for gaps (gene absences in some species). We optimised topology, branch lengths and model parameters simultaneously using PhyML. Just for this tree reconstruction we used a 4 category discrete $\Gamma(\alpha)$ RAS distribution in order to speed up calculations. The resulting topology (Fig. 5.4) is consistent with recent reports using sequence-only evidence (Fitzpatrick et al., 2006; Jeffroy et al., 2006). Nevertheless, Scannell and colleagues noticed a greater number of common orthologous losses shared between \textit{C. glabrata} and \textit{S. cerevisiae} than between \textit{S. cerevisiae} and \textit{S. castellii} (Scannell et al., 2006). They carried out further analyses that revealed a number of genomic rearrangements that give further support to an alternative topology where \textit{S. castellii} is outgroup to a \textit{C. glabrata} and \textit{S. cerevisiae} lineage. Based on those evidences, they concluded that probably the topology recovered using sequence information only is result of systematic bias. In this study we grouped \textit{C. glabrata} and \textit{S. cerevisiae} in all our posterior analyses.

5.2.5 Functional analysis

5.2.5.1 Functional enrichment of ontology terms in concerted evolution paralogous pairs

We determined which functional groups of genes most frequently evolve in concert using gene ontology (GO) term annotations. GO terms represent biological concepts including biological processes, molecular functions and cellular components. In order to identify ontologies that
are over- or under-represented among paralogous pairs that evolved in concert, we used the FatiGO+ tool, part of the BABELOMICS suite (Al-Shahrour et al., 2006). As there is not specific functional annotation for \( C. \) glabrata or \( S. \) castellii genes, we used their \( S. \) cerevisiae homolog's annotations as a proxy. To reduce the number of misleading false positives we only considered terms that were found significant after correcting for multiple tests using the false discovery rate (FDR) method implemented in FatiGO+.

Not surprisingly, paralogous genes share many GO annotations. Therefore handling them as independent entities may yield artifactual over- or under-representations; shared GO terms are selected twice for each paralogous pair in concert. As a workaround, we replaced each paralogous gene pair with a chimeric supergene annotated once with all GO terms result of the union of the annotations of the original genes. Due to this peculiar requirement we needed to use FatiGO+'s "your annotation" facility. We downloaded yeast genes annotations from SGD and the ontology term hierarchical information from The Gene Ontology Consortium (http://www.geneontology.org) as they were posted on 20th April 2007.

Unfortunately FDR is a conservative multiple test correction approach due to positive and negative correlations between ontologies. In Appendix F.4 we include the most significant over- or under-representations regardless of whether they are still significant after applying FDR.

5.2.5.2 Yeast genes viability data

In order to investigate whether concerted evolving genes are dispensable, we examined the viability information as posted in March 2007 in the Comprehensive Yeast Genome Database (CYGD) Gülüdener et al. (2005) by the Munich Information Center for Protein Sequences (MIPS).
This database collects results from low- and high-throughput single gene knockout experiments in yeast. They classify each gene depending on the resulting mutant phenotype. Based on their annotation, yeast genes are 15.4% (lethal/essential), 73.7% (viable/dispensable), 2.6% (contradicting results/ambiguous) and 8.3% (not tested/unknown) genome-wide.

5.2.6 Dealing with compositional bias

Mainstream probabilistic models of molecular sequence evolution assume that residue composition is stationary and common to all lineages. Consequently, if it changes across lineages, sequences with convergent composition may become apparently closer related (Mooers and Holmes, 2000). This results in shorter pairwise distances and false clustering in phylogeny reconstructions. For this reason we decided to rely only on results obtained based on amino acid or filtered nucleotide sequences (without mixture of synonymous and non-synonymous changes). This data, although not completely bias free, must be less susceptible to compositional bias artifacts.

Nevertheless, we also wanted to investigate the effect of base and codon usage bias among WGD paralogs using all nucleotide positions. This is of a great methodological interest as several concerted evolution detection methods and studies are based on full DNA sequences or synonymous nucleotide sites only datasets (e.g. codouble method (Balding et al., 1992)). To this end, we utilised paralinear distances and synonymous codon resampling as explained below. We also calculated codon usage indices for paralogous pairs in order to further demonstrate the relation between codon usage bias and false concerted evolution detection.

5.2.6.1 LogDet correction

Paralinear or LogDet distances (Lake, 1994; Lockhart et al., 1994; Steel, 1994) are a simple but effective approach to account for heterogeneous composition between sequences, that is, to produce consistent tree additive distances even when residue composition is not stationary (Tamura and Kumar, 2002; Thollesson, 2004). We calculated LogDet corrected distances using the formulation in Appendix B.4.

Unfortunately, LogDet is in general inconsistent under heterogeneous RAS models. This can be corrected to some extent performing a “constant sites removal” or CSR-LogDet (Penny et al., 1999; Waddell, 1995). Accordingly we removed a percentage of constant sites prior to calculating LogDet distances. We estimated the actual number of invariant sites fixing the shape parameter $\alpha = \infty$ (homogeneous change rate across variable sites) and the tree topology to the actual one in PhyML.
5.2 MATERIAL AND METHODS

5.2.6.2 Synonymous codon resampling procedure

Even in the absence of nucleotide compositional bias, preferential use of a particular set of codons in different species can make intra-specific paralogs apparently more similar as compared to phylogenetically closer inter-specific homologous sequences. LogDet does not offer a doable option for handling heterogeneous codon composition as the amount of data needed to properly estimate all the entries in the 61x61 matrices required is huge. Alternative codon based evolutionary models may offer a mathematically sound solution to this problem. Nevertheless, here we preferred to apply a more simple approach based on shuffling synonymous codons as explained below.

Let us consider a particular paralogous pair that shows no evidence of concerted evolution when using amino acid sequence or filtered nucleotide sequence data. However, the same gene pair is detected as to have undergone concerted evolution if we use all nucleotide positions. Then, these extra data must contain a great deal of the actual concert signal. Consequently, the expectation is that as soon as such additional signal is somehow removed, the test must result negative again.

In order to remove the statistical signal at newly added synonymous nucleotide positions, we resampled codons based on each sequence's empirical codon frequencies: for each one of the twenty amino acids we calculated codon frequencies and used these to generate a chimeric sequence that translates into the same protein product. The resulting alignment is identical to the original at the amino acid level, yet synonymous nucleotide sites are virtually saturated. Nevertheless, in those sequences where codon usage bias is strong, resampling would have a limited effect; the same codons are picked up again just by compositional convergence. Moreover, if two sequences have a similar and strong codon usage preference, the pairwise silent site distance, $d_S$, would not change significantly after resampling.

Consequently, a substantial proportion of additional positive cases of concerted evolution (not detected before using all nucleotide sites) that do not vanish after performing a synonymous codon resampling, would indicate potential false concerted evolution identification due to codon usage bias artifacts. In other words, codon composition would be as good explanation as concerted evolution to the patterns observed. Thus, the later cannot be confirmed neither rejected. In order to determine whether this ratio is significantly high we compared it to the percentage of positive cases using filtered nucleotide sequences in all genes as a generous upper limit. Using the expected 5% limit of false positives (type I error) as the null rate may actually be too optimistic as filtered nucleotide sites or amino acid may well have already contributed a significant part of the concert signal that would persist after resampling.
5.2.6.3 Codon usage bias and gene expression data

There are several metrics to quantify codon usage bias. The most popular, at least in the concerted evolution and yeast literature herein cited, is the codon adaptation index (CAI) (Sharp and Li, 1987). The CAI value of a gene indicates how similar its codon composition is compared to the most dominant trend in the genome, typically dictated by highly expressed genes. High CAI values (closer to 1) indicate stronger codon usage bias and higher levels of expression. Low CAI (closer to 0) indicates just the opposite.

We used the CodonW program v1.4.4 (http://codonw.sourceforge.net) to measure codon usage bias in the post-WGD dataset. First, we calculated the optimal (translation efficient) codon tables for *S. castellii* and *C. glabrata*. In order to verify these tables, we checked that the top 5% genes, that were used to build each table, have high constitutive expression levels. We used yeast homolog's levels as published by Holstege and colleagues (Holstege et al., 1998) as a proxy to the actual expression levels in *S. castellii* and *C. glabrata*. Regarding *S. cerevisiae*, we used the built-in table provided in CodonW. Finally we run CodonW a second time to calculate CAI values for each post-WGD dataset sequence using the corresponding table.

5.3 Results and Discussion

5.3.1 Control dataset

Concerted evolution is rather improbable between orthologous genes from different species. However, concert like patterns could still emerge due to occasional convergence of evolutionary constraints. In order to minimise such cases we used only yeast single gene families as a real-data control dataset. The expectation is that, the number of “paralogous” pairs of sequences that reject the null hypothesis is close to the targeted type I error.

5.3.1.1 A method based on the distribution of changes

Sawyer's method, as implemented in GENECONV, detects the occurrence of gene conversion with pairs of sequences in a multiple sequence alignment. It generates pairwise p-values (one for each pair analysed) together with a single corrected p-value for the whole alignment (global p-value) as to account for multiple tests: all possible pairwise comparisons within the alignment. We first checked on the performance of this approach just taking into consideration silent nucleotide polymorphisms. We set up the design type I error to 5%. Global p-values do not yield a significantly different percentage of false positives as indicated in Table 5.1. Nevertheless, pairwise detections do not reach the 5% mark. Therefore this configuration is rather conservative with an actual type I error of 3.2% for this control dataset. As shown below, posterior
application of this set-up on the WGD paralogous dataset does also confirm a lack of power as anticipated by Drouin and colleagues (Drouin et al., 1999).

In contrast, this method in both pair- and alignment-wise flavors seems to become too liberal if we use either amino acids or all nucleotide polymorphisms with false discovery rates of up to 22% (Table 5.1). Heterogeneous fixation of non-synonymous changes due to evolutionary constraints is the most plausible alternative explanation to the conversion like patterns observed. In contrast, silent substitutions should be more uniformly distributed, as they are not subject to selection on the performance of an unchanged protein product.

In order to reduce the percentage of false positives, we investigated whether there is a positive correlation between genetic distance and the false discovery rate. If false positives are just an issue for highly divergent sequence pairs, we could still have safely discarded cases below a conservative identity threshold. Unfortunately, this does not seem to be the case as the small difference between pairwise genetic distances of false positive and true negative cases indicates that false discovery is in fact more likely for shorter distances (not shown). Additionally, we managed to bring down the number of false positives changing some GENECONV options or the format of the input data (e.g. decreasing the number of mismatches allowed or reducing alignments into triplets and consequently also the number of polymorphic sites). Nevertheless, this sort of modifications are supposed to increase the type II error (thus reduce the power of the test) never decrease the type I error as it seems to be the case here. Therefore, we refrained from performing any tuning of this sort in order to correct for the elevated number of false positives. We applied all nucleotide polymorphism on the yeast WGD dataset but rejecting the null hypothesis only in cases with pairwise p-value ≤ 0.02 as in the control dataset this threshold matched the targeted 5%.

### 5.3.1.2 Distance based method

We first checked on the null performance of this method where we determined the significance of the test, directly comparing the observed pairwise distance, $D'_{xy}$, and the alternative concert-free estimator $\hat{D}_{xy}$. In order to do this we calculated a distribution based on one of these two estimators, the reference distance, and then checked whether the other does or does not fall into the appropriate one-tail rejection zone ($D'_{xy}$ is significantly lower than $\hat{D}_{xy}$ or vice versa). When we used the expected distance $\hat{D}_{xy}$ as the reference, the distribution was approximated using parametric bootstrapping. In contrast, when we used the observed pairwise distance $D'_{xy}$ as the reference we built the distribution using non-parametric bootstrapping. We call the resulting significances $p_{value_{DP}}$ and $p_{value_{Dnp}}$ respectively.
<table>
<thead>
<tr>
<th>Data</th>
<th>total cases</th>
<th>Protein positives</th>
<th>Protein percent.</th>
<th>Protein p-value</th>
<th>Nucleotide positives</th>
<th>Nucleotide percent.</th>
<th>Nucleotide p-value</th>
<th>Silent (Nucl.) positives</th>
<th>Silent (Nucl.) percent.</th>
<th>Silent (Nucl.) p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alignments</td>
<td>441</td>
<td>73</td>
<td>16.6%</td>
<td>1 ***</td>
<td>99</td>
<td>22.4%</td>
<td>1 ***</td>
<td>20</td>
<td>4.5%</td>
<td>0.3786146</td>
</tr>
<tr>
<td>Pair of species</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. gossypii vs C. glabrata</td>
<td>300</td>
<td>14</td>
<td>4.7%</td>
<td>0.463</td>
<td>28</td>
<td>9.3%</td>
<td>0.999 **</td>
<td>15</td>
<td>5%</td>
<td>0.432</td>
</tr>
<tr>
<td>A. gossypii vs K. lactis</td>
<td>334</td>
<td>13</td>
<td>3.9%</td>
<td>0.215</td>
<td>30</td>
<td>9%</td>
<td>0.999 **</td>
<td>11</td>
<td>3.1%</td>
<td>0.09</td>
</tr>
<tr>
<td>S. castellii vs S. kluyveri</td>
<td>215</td>
<td>13</td>
<td>6%</td>
<td>0.809</td>
<td>18</td>
<td>8.4%</td>
<td>0.988 *</td>
<td>7</td>
<td>3.3%</td>
<td>0.153</td>
</tr>
<tr>
<td>S. cerevisiae vs S. kluyveri</td>
<td>254</td>
<td>27</td>
<td>10.6%</td>
<td>1 ***</td>
<td>30</td>
<td>11.8%</td>
<td>1 ***</td>
<td>8</td>
<td>3.1%</td>
<td>0.108</td>
</tr>
<tr>
<td>Any two species</td>
<td>7913</td>
<td>618</td>
<td>7.8%</td>
<td>1 ***</td>
<td>769</td>
<td>9.7%</td>
<td>1 ***</td>
<td>258</td>
<td>3.3%</td>
<td>2.66 · 10^-14 ***</td>
</tr>
</tbody>
</table>

Table 5.1: Performance summary of Sawyer’s approach applied to the empirical control dataset. Columns indicates the total number of tests and, for each data-type, false positives counts (absolute and percentage) and its significance (p-value) where we expect approximately 5% of false positive by chance. We calculated p-values as the two-tail cumulative probability assuming the p-value is drawn from a binomial where the rate is set to 0.05. First data row includes totals based on (alignment-wise) global corrected p-values. Additionally, there is a few example rows for species pairwise comparisons (out of 28). The last row is the total/average based on all species pairwise comparisons (shown and not shown).
5.3 RESULTS AND DISCUSSION

\begin{align*}
p_{\text{value}}_{Dp} & = \Pr \left( [\hat{D}_{xy} + \varepsilon_p] \leq D'_{xy} \right) \quad (5.6) \\
p_{\text{value}}_{Dnp} & = \Pr \left( \bar{D}_{xy} \leq [D'_{xy} + \varepsilon_{np}] \right) \quad (5.7) \\
\varepsilon_p & \equiv \text{Parametric bootstrap sampling error} \\
\varepsilon_{np} & \equiv \text{Non-parametric bootstrap sampling error}
\end{align*}

All bootstrappings consisted of 10000 replicates. With both protein and nucleotide distances, an empirical cumulative distribution plot of the \( p_{\text{value}}_{Dp} \) indicated that this test is actually very conservative as illustrated in Figure 5.5 (only nucleotide results are shown). For nucleotides, an intended type I error of 1% and 5% actually rejected the null hypothesis only in 0.24% and 1.24% of cases respectively. Using amino acids does not improve significantly with 0.47% and 1.78% instead. This is in fact due to a dependency between \( D'_{xy} \) and \( D_{xy} \) as both are based on the same data. Using non-parametric bootstrap and taking \( D'_{xy} \) as the reference distance showed the same tendency.

In order to achieve the designed type I error, we needed to quantify such dependency on a case-by-case basis. To this end, we used the signed difference of both distance estimators, \( I'_{xy} = D'_{xy} - \bar{D}_{xy} \). Then we tested whether this difference is significantly smaller than the expected by chance. Again, we tried two alternative approaches to determine the significance of the test based on parametric bootstrapping (\( p_{\text{value}}_{Ip} \)) and non-parametric bootstrapping (\( p_{\text{value}}_{Inp} \)). This time however, we needed to perform simulations in order to obtain an estimate of the mean difference, \( \bar{I}_{xy} \), if the null hypothesis is true. This mean is not necessarily zero:

\begin{align*}
p_{\text{value}}_{Ip} & = \Pr \left( [\bar{I}_{xy} + \delta_p] \leq I'_{xy} \right) \quad (5.8) \\
p_{\text{value}}_{Inp} & = \Pr \left( \bar{I}_{xy} \leq [I'_{xy} + \delta_{np}] \right) \quad (5.9) \\
\bar{I}_{xy} & \equiv \text{Mean difference obtained by parametric simulation} \\
\delta_p & \equiv \text{Parametric bootstrap sampling error} \\
\delta_{np} & \equiv \text{Non-parametric bootstrap sampling error}
\end{align*}

For all parametric simulations we assumed branch lengths as estimated using WLS discarding pairwise distances amenable to concerted evolution as explained in Appendix F.1. In all cases we generated bootstrapped samples composed of 10000 repeats. We only applied this approach with nucleotides as amino acid pairwise distance estimation using ML was computationally too demanding for analyses involving so many alignments. Unfortunately, empirical cumulative distribution plots for, \( p_{\text{value}}_{Ip} \) and \( p_{\text{value}}_{Inp} \) indicated that a test based on either alternative would be too liberal when applied to real data (Fig. 5.5). For example, intended type I errors 1% and 5% produced a total of 3.15% and 9.69% respectively using parametric bootstrapping.
A logical explanation for this phenomenon is that simulated data presents lower variance than does real data.

Interestingly, these results parallel the optimistic tendency of Sawyer's approach using all nucleotide polymorphisms. Accordingly, for the WGD paralogous dataset, we applied the parametric bootstrapping but considering the critical p-value obtained from the empirical control dataset for a designed type I error 5% (pvalue_{crit} < 0.02). We also used the more conservative alternative pvalue_{DP}.

![Cumulative p-value distribution](image)

Figure 5.5: Empirical cumulative p-value distribution based on filtered nucleotide alignments and the four alternatives purposed to calculate the distance test significance (Eqs. 5.6 to 5.9). The diagonal grey line highlights the ideal uniform cumulative distribution.

### 5.3.2 Yeast WGD duplicates dataset

#### 5.3.2.1 Performance of different detection methods

Search of convergent regions based on the distribution of changes seems to be less powerful than other approaches used in this study. Moreover, silent polymorphism evidence did not report a significant number of concerted evolution cases (Figure F.2 and Table 5.6 in Appendices). This is in fact not surprising if we take into account that this set-up caused a conservative result to be produced from the control dataset above. Nevertheless, this method already reported a significant number of joint concerted evolution across all post-WGD species among pillars belonging to category C0.0.

In contrast, using all nucleotide positions reported a significant percentage of positives, detecting a similar percentage of cases in all combinations of species and pillar categories (around
5.3 Results and Discussion

10%). Additionally, we also observed a significant number of dubious positive cases of inter-specific gene conversion (7.3% overall with a p-value $\ll 0.001$). Unless specified otherwise, GENECONV looks for conversion between all homologous pairs in the alignment. Therefore, part of the apparent power gain is in fact due to an increase in type I error.

Tree reconstruction based on ML and WLS indicated the existence of concerted evolution in both *Saccharomyces* species but not in *C. glabrata* except for category C1.1 using ML trees. Category C0.0 marginal (per species) positive cases do not exceed the number expected if the null hypothesis were true. Nevertheless, the number of joint cases is significantly high. Results indicate that the ML tree based test is slightly more powerful than the WLS tree based test. This may be partly due to the fact that we allowed ML searches to optimise model parameters alongside tree topology and branch lengths, whereas WLS input distances were based on sub-optimal model parameter values that maximised the likelihood of the actual topology (based on syntenic information).

Regarding distance comparison based methods, application of $p_{value}^{DP}$ and $p_{value}^{DP}$ showed evidence of concerted evolution in all three species, although with a clearly lower incidence among *C. glabrata* paralogs. In contrast to tree reconstruction, category C0.0 showed a significant number of cases of concerted evolution for each species. According to results in the control dataset, $p_{value}^{DP}$ is conservative when compared with $p_{value}^{DP}$. Numerically, $p_{value}^{DP}$ seems to be roughly as powerful as ML tree reconstruction, based on the total number of hits. On the other hand, $p_{value}^{DP}$ reports approximately 15% more cases than the other two. Ontology enrichment analysis of pillars detected as evolving in concert validates results based on tree reconstruction and distance approaches as discussed below.

Although our results suggest that the distance approach is more powerful than tree reconstruction, this depends heavily on critical values chosen to determine the significance of the test. In the distance method based on parametric estimator difference bootstrapping ($p_{value}^{DP}$) we used the value obtained from an empirical control dataset. Regarding tree bootstrapping, the critical support index must depend on the tree shape (topology and branch lengths) and the model. Parametric bootstrapping could indicate a specific critical value for each pillar phylogeny. However, simulated data normally behaves better than does real data (less noisy). Consequently this may well yield low critical bootstrap values (a liberal test).

5.3.2.2 Concerted evolution and compositional bias

Codon usage bias, gene conversion and intra-specific functional conservation can explain paralogous sequence convergence. Lin and colleagues already noticed that it is difficult to tell whether concerted evolution patterns are a product of compositional bias or genuine gene conversion (Lin et al., 2006). Compositional heterogeneity, if not taken into account, can produce
Figure 5.6: Summary of numerical results of concerted evolution detection methods applied on nucleotide alignments. For each post-WGD species and category we show the percentage of positive cases and its significance as detected using several methods. Each row represents a species. The last row is the average of results in all species. Contribution for each species is weighted based on the total number of cases in each category. There is one graph column for each method. For tree building and distance methods we show here results based on filtered nucleotide alignments as described in Methods. Each graph is divided in four bars, one for each pillar category as indicated by labels at the bottom. The bar height (width) represents the percentage of positives whereas its colour tells us the significance level of that percentage. This figure computes as the cumulative probability of obtaining less positives cases assuming a binomial with rate set to 5% (type I error). Dashed lines mark the weighted average percentage across pillar categories for each species and detection method. Numerical values are available in Table F.2.

underestimated pairwise distances between sequences with more similar base or codon composition. Accordingly, tree reconstruction based on mainstream evolutionary models also would yield topologies where taxa are clustered due to composition and not true phylogenetic signal. However, in spite of the risk of obtaining too many false positives, it may seem worthwhile to
use all nucleotide positions as some cases of concerted evolution may be only detectable if we take into account silent changes (e.g. gene conversion events in very conserved regions).

Tentatively, we applied distance and ML tree reconstruction methods to the WGD paralogs dataset using all nucleotides sites (without filtering for mixtures of synonymous and non-synonymous substitutions). Both methods reported an increased number of positive detections, especially using distances on *S. castellii* paralogs where this figure was nearly doubled as shown in Table 5.2.

Although this may be regarded as resulting from a power increase (reduction of type II error as there is more data available) we double checked whether it could be also explained by compositional bias. Firstly, we used CSR-LogDet corrected nucleotide distances (Eq. B.9) in order to account for overall differences in GC content between homologous sequences. Qualitative results did not change much hence indicating that the nucleotide composition in newly added sites (mostly third codon positions) does not explain the trend observed. Secondly, we repeated the analysis after resampling synonymous codons per each sequence using sequence specific empirical frequencies as described in Methods. A considerable number of additional positive cases (detected only when we used all nucleotide positions) were still reported as putative concerted evolution. Consequently, we could not tell whether it was due to a genuine test power increase or simply to convergent codon preferences. Moreover, CAI indices are significantly higher in additional positive cases retained after codon resampling as compared to all other genes in the dataset (Table 5.2).

Looking at results using different methods, ML tree reconstruction seems to be more robust against false positives caused by compositional bias. This is especially obvious in the *S. castellii* subset where the number of hits using distances is nearly doubled. This observation is in agreement with the popular view that ML tree reconstruction is robust to deviations from the model assumed. However, at least in this dataset, using all nucleotide positions with either method is a very likely source of false positives. Moreover, the resampling procedure herein described cannot filter out possible cases where codon usage bias is heterogeneous along the alignment (Akashi, 1994; Iida and Akashi, 2000) and therefore the actual number of false positives introduced may be underestimated. If this is the case, concerted evolution detection based on heterogeneous distribution of changes, e.g. using GENECONV, may also report an increase number of false positives.

Additionally, translation optimisation is known to influence amino acid composition (Akashi, 1994; Akashi and Gojobori, 2002). For example, highly expressed genes show a tendency towards using codons more efficiently translated due to an heterogeneous tRNA pool. This may bias results based on amino acid or filtered nucleotide sites. Accordingly, we tried to apply CSR-LogDet to protein alignments. Unfortunately, this correction requires long sequences in order to calculate accurately the 20x20 entry matrices required to compute Equation B.9. Most
alignments in the control and WGD paralogous dataset do not seem to provide enough data, specially using constant site removal and performing bootstrapping (as indicated by copious infinite pairwise distances). Here, we did not follow up on this issue any further although it may constitute an interesting target for future work.

5.3.2.3 Concerted evolution and patterns of gene loss

From this point on, unless otherwise stated, we draw conclusions from empirical results obtained using our distance method based on parametric bootstrapping of estimator differences ($p_{value, p}$). The extended Venn diagram in Figure 5.7 summarises concerted evolution detected at different gene categories and post-WGD species analysed. Concerted evolution is present in all three species although is less prevalent in *C. glabrata* (10.7% compared to an averaged 16.6% in *Saccharomyces* species).

Despite all pillar categories showing significant evidence of concerted evolution in at least one species, there are still clear quantitative differences between categories. Concert is more frequent in C1.0 and C1.1 (23.7% and 33.3% of cases respectively). Moreover, loci in these classes seem to have expended more time in concert (Figure 5.8). More concretely, the mixed distribution of the $\tau$ statistic in *Saccharomyces* species and categories C1.0 or C1.1 is significantly shifted to higher values as compared to categories C0.0 and C2.0 (Table 5.3). In contrast, *C. glabrata* paralogs do no exhibit significant differences across categories.

Category C1.1 includes exclusively cases of reciprocal gene loss. Scannell and colleagues concluded that reciprocal gene loss may have had major implications in the reproductive isolation and speciation of post-WGD species (Scannell et al., 2006). They also stated that reciprocal gene loss must indicate symmetrical functional redundancy, as either copy is equally dispensable. Likewise, concerted evolution is also believed to be responsible for species differentiation (Dover, 1982; Elder and Turner, 1995). Additionally, it is only logic to think that concerted evolution is also a good indicator for functional redundancy. Moreover, reciprocal gene loss and concerted evolution are complementary markers: the latter is not applicable if we lack surviving paralogous pairs due to exhaustive gene loss (category C3 and C4 in Byrne and Wolfe, 2005) whereas the former is not applicable to pillars with no more than a single gene loss (categories C0.0 and C1.0 in this text).

Then it follows that positive evidence of concerted evolution in category C2.0 can be explained also by functional redundancy. Late orthologous gene losses, in most part, reflect gene subfunctionalisation and convergent loss of one copy across species: it is more probable to lose the paralogous gene that is less essential. However, since functional redundancy may randomly result in two orthologous gene losses (50% chances), the number of “under-cover” symmetric redundancy cases in C2.0 must roughly match the total number of observed cases in C1.1 dataset.
<table>
<thead>
<tr>
<th>Species</th>
<th>Filtered nucleotides</th>
<th>All nucleotides</th>
<th>All nucleotides resampled</th>
<th>CAI values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>total lost gain revert retain p-value</td>
<td>retain mean median</td>
<td>rest of genes mean median p-value</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Distance based method (p-value&lt;sub&gt;dp&lt;/sub&gt;)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>54</td>
<td>76   8 30 21 9</td>
<td>0.0327 *</td>
<td>0.324 0.19 0.238 0.152 0.0273 *</td>
</tr>
<tr>
<td>S. castellii</td>
<td>57</td>
<td>140  4 87 51 36</td>
<td>5.92 · 10^(-7)***</td>
<td>0.417 0.368 0.410 0.34 0.0132 *</td>
</tr>
<tr>
<td>C. glabrata</td>
<td>25</td>
<td>43   5 23 15 8</td>
<td>7.6 · 10^(-4)***</td>
<td>0.421 0.26 0.304 0.236 0.0512</td>
</tr>
<tr>
<td><strong>Distance based method (p-value&lt;sub&gt;fp&lt;/sub&gt;)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>67</td>
<td>75   17 25 19 6</td>
<td>0.114</td>
<td>0.303 0.19 0.239 0.152 0.0857</td>
</tr>
<tr>
<td>S. castellii</td>
<td>82</td>
<td>160  11 79 46 33</td>
<td>1.23 · 10^(-4)***</td>
<td>0.425 0.387 0.41 0.341 0.00708 **</td>
</tr>
<tr>
<td>C. glabrata</td>
<td>36</td>
<td>47   10 21 10 11</td>
<td>3.68 · 10^(-7)***</td>
<td>0.449 0.391 0.302 0.235 5.31 · 10^(-5) ***</td>
</tr>
<tr>
<td><strong>ML tree reconstruction method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>51</td>
<td>52   7 9 7 2 2</td>
<td>0.14</td>
<td>0.256 0.25 0.24 0.152 0.0589</td>
</tr>
<tr>
<td>S. castellii</td>
<td>52</td>
<td>83   7 38 18 20</td>
<td>1.67 · 10^(-10)***</td>
<td>0.447 0.4 0.409 0.341 0.00211 **</td>
</tr>
<tr>
<td>C. glabrata</td>
<td>19</td>
<td>28   4 13 9 4 4</td>
<td>0.0024 **</td>
<td>0.554 0.6435 0.303 0.235 0.00242 **</td>
</tr>
</tbody>
</table>

Table 5.2: Summary of the effect of codon usage bias in the number of positive detections of concertated evolution for all three post-WGD species using distance and ML tree reconstruction. Numerical columns from left to right indicate: the number of positive cases using filtered nucleotide alignments, number of positives using all positions, number concerted evolution cases that are no longer detected after using all positions (lost), number of new positive cases (gained), number of gains that are no longer detected after resampling synonymous codons (reverted), number of gains that still are positive after resampling (retained) and the significance of the ratio between retained and reverted cases, mean and median of the CAI values for retained paralogs, mean and median of the CAI values for the rest of WGD duplicate genes, and the significance of the difference between these two distributions. The p-values for the number of retained genes is the result of a Fisher exact test between the retained property and detection using the resampled alignment. The CAI distribution p-value results from a Mann-Whitney median test between both distributions.
Figure 5.7: Following the layout in Figure 5.1, this diagram summarises the occurrence of concerted evolution in WGD paralogs. Figures are based on applying the distance base method on filtered nucleotide alignments. It includes the percentage of positives. In each Venn diagram sub-area, boxed percentages indicate the percentage of joined positives between all species involved. Adjacent non-boxed percentages indicate the marginal percentage of positive cases for each species. Different font style tell us whether the number of positives is significant, that is, higher than expected under the null hypothesis (lack of concerted evolution). We calculated the significance as the probability of observing a greater number of positives using a binomial distribution and setting the rate to the expected type I error, 5%. Italic, underlined, bold and bold underlined percentages indicate whether the number of positives is non-significant, significant at 5%, significant at 1% or significant at 0.1% respectively. Grey filled areas highlight the presence of significantly high figures. Species marginal percentages are located under the species name, whereas marginal values for each category are included in the corresponding top-right box.

(Scannell et al., 2006). Accordingly, after subtracting the number of positive cases in C1.1 for each species in C2.0 we do not obtain a significant high number of concerted evolution positive cases amongst the remaining (4.3%, p-value = 0.705).

Most positive concerted evolution cases in category C0.0 are common to more than one post-WGD species with nearly a third being shared by all three. Results indicate that in this category either paralogous genes are sub- or neofunctionalised and both have been maintained (they carry out important functions), or copies are at least partially redundant and have been retained due to pressure to produce “more of the same product” (Sugino and Innan, 2006).
5.3 Results and Discussion

Figure 5.8: Box-and-whisker diagram that features concert duration distributions based on a modified Sugino and Innan statistic, τ (Eqs. 4 and 5). Each vertical line separates data from different pillar categories. There is one box per species and category. Labels S.cer, C.gla and S.cas stand for *S. cerevisiae*, *C. glabrata* and *S. castellii* respectively.

<table>
<thead>
<tr>
<th></th>
<th>C0.0</th>
<th>C1.0</th>
<th>C1.1</th>
<th>C2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1.0</td>
<td>0.0435 *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1.1</td>
<td>0.00478 **</td>
<td>0.0874</td>
<td></td>
<td>0.0202 *</td>
</tr>
<tr>
<td>C2.0</td>
<td>0.12</td>
<td>0.00727 **</td>
<td>0.00178 **</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3: Comparison of τ distributions across gene loss patterns. The figures shown are p-values resulting of two-tail Mann-Whitney test between gene classes combining τ values of *S. cerevisiae* and *S. castellii* in order to improve the power of the test. Results over the diagonal are based only in pairs for which the molecular clock hypothesis could not be rejected. Results under the diagonal use all pairs available and τ’statistic instead.

5.3.2.4 Functional analysis of paralogs evolving in concert

We looked for gene ontology terms over- or under-represented among paralogs that have evolved in concert considering the overall incidence of concerted evolution in post-WGD species (Table 5.4). In this case the null hypothesis states that concerted evolution is a normally occurring process and that it does not depend on the functional role of the gene pair. Our results rejected this hypothesis.

The most robust trend is the over-representation of genes involved in ribosome function and structure. Previous work has already highlighted this fact (Kellis et al., 2004; Pyne et al., 2005; Sugino and Innan, 2006). Additionally, there are a few other terms also significantly over- or
under-represented after correcting for multiple tests. For example, at first glance, genes involved in telomere structure and length maintenance seemed also prone to have evolved in concert, at least within the *S. cerevisiae* proteome. Nevertheless, most of these genes are also annotated with ribosomal and translation activity related terms. This may be the result of indirect genetic interactions rather than direct involvement in telomere biology (Edmonds et al., 2004). Also the under-representation of nuclear genes (and therefore membrane-bound organelles genes) is not surprising since most ribosomal proteins fulfill their role outside the nucleus. Over-represented cytosolic large and small ribosomal subunit entries in Table 5.4 attest to this fact. Cell cycle controlling processes are also under-represented. In this case, the percentage of positive cases is neither significantly lower nor higher than the rate expected in the absence of concerted evolution (5%) and therefore we conclude that concerted evolution does not seem to shape the evolution of WGD paralogous pairs involved in cell cycle control.

Interestingly, glucosidase activity and chitin- and beta-glucan-containing cell wall terms are distinctly overrepresented in *C. glabrata*. Most genes annotated with either of these two terms are involved in processes that control biosynthesis and maintenance of cell and spore walls. Therefore they have an important role in morphology, cell growth and mating processes in fungi (Adams, 2004; Klis et al., 2006).

This is indeed quite remarkable considering that *C. glabrata* is the post-WGD species with the greatest number of WGD paralogous losses together with the lowest incidence of concerted evolution based on other results. *C. glabrata* shows signs of concerted evolution in all four paralogous pairs annotated as showing glucosidase activity and half of the surviving paralogous pairs annotated as expressed in chitin- and beta-glucan cell wall component. Both *Saccharomyces* species also exhibit a high percentage of concerted evolution in these groups of genes yet not to the same extent as *C. glabrata*. Several gene knockout experiments conducted on yeast homologs provide experimental evidence of functional overlapping, collaboration and specialisation in different life cycle stages within some of these paralogous pairs (Larriba et al., 1995; Cappellaro et al., 1998; Mrsa and Tanner, 1999).

As regarding the functional dispensability of paralogs that have evolved in concert, we intuitively expected that if one copy is pseudogenised (ceases to work or to be expressed), the other copy must be able to take over most of the lost paralog’s functions. The lethal phenotype is already under-represented amongst WGD duplicated genes with only 5.9% non-viable mutants in contrast with the genome-wide 15.4%. This is in fact something expected as single gene families, obviously excluded from the WGD paralogous subset, should be in average more indispensable than multi-gene family homologs. In total, there is just a single gene putatively evolving under concerted evolution that has been found to be essential: the GTP binding protein GSP1. However, in spite of the reduced percentage of lethal losses (<1%), this is not statistically significant due to the already low overall lethal rate among WGD paralogs. (Fisher one-tail p-value=0.102).
<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Concert</th>
<th>Species</th>
<th>Inc./Dec.</th>
<th>S. cerevisiae</th>
<th>S. castellii</th>
<th>C. glabrata</th>
<th>Any</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0003735</td>
<td>structural constituent of ribosome</td>
<td>+</td>
<td></td>
<td></td>
<td>3.34·10⁻¹⁴***</td>
<td>1.96·10⁻¹²***</td>
<td>1</td>
<td>5.28·10⁻²³***</td>
</tr>
<tr>
<td>GO:0015926</td>
<td>glucosidase activity</td>
<td>+</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>0.0308*</td>
<td>0.904</td>
</tr>
</tbody>
</table>

### Molecular Functions

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Concert</th>
<th>Species</th>
<th>Inc./Dec.</th>
<th>S. cerevisiae</th>
<th>S. castellii</th>
<th>C. glabrata</th>
<th>Any</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006412</td>
<td>translation</td>
<td>+</td>
<td></td>
<td></td>
<td>2.89·10⁻¹⁶***</td>
<td>2.92·10⁻¹²***</td>
<td>0.778</td>
<td>3.01·10⁻²²***</td>
</tr>
</tbody>
</table>

### Biological Process

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Concert</th>
<th>Species</th>
<th>Inc./Dec.</th>
<th>S. cerevisiae</th>
<th>S. castellii</th>
<th>C. glabrata</th>
<th>Any</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0022402</td>
<td>cell cycle process</td>
<td>−</td>
<td></td>
<td></td>
<td>1</td>
<td>0.962</td>
<td>1</td>
<td>0.0111*</td>
</tr>
<tr>
<td>GO:0000723</td>
<td>telomere maintenance</td>
<td>+</td>
<td></td>
<td></td>
<td>0.0321*</td>
<td>1</td>
<td>1</td>
<td>0.153</td>
</tr>
<tr>
<td>GO:0042274</td>
<td>ribosomal small subunit biogenesis and assembly</td>
<td>+</td>
<td></td>
<td></td>
<td>0.0438*</td>
<td>1</td>
<td>1</td>
<td>0.0871</td>
</tr>
<tr>
<td>GO:0042257</td>
<td>ribosomal subunit assembly</td>
<td>+</td>
<td></td>
<td></td>
<td>0.0438*</td>
<td>1</td>
<td>1</td>
<td>0.23</td>
</tr>
</tbody>
</table>

### Cellular Component

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Concert</th>
<th>Species</th>
<th>Inc./Dec.</th>
<th>S. cerevisiae</th>
<th>S. castellii</th>
<th>C. glabrata</th>
<th>Any</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0005842</td>
<td>cytosolic large ribosomal subunit (sensu Eukaryota)</td>
<td>+</td>
<td></td>
<td></td>
<td>9.09·10⁻⁹***</td>
<td>2.28·10⁻⁷***</td>
<td>1</td>
<td>6.87·10⁻¹³***</td>
</tr>
<tr>
<td>GO:0005843</td>
<td>cytosolic small ribosomal subunit (sensu Eukaryota)</td>
<td>+</td>
<td></td>
<td></td>
<td>3.42·10⁻⁶***</td>
<td>2.43·10⁻⁵***</td>
<td>0.19</td>
<td>1.67·10⁻⁹***</td>
</tr>
<tr>
<td>GO:0009277</td>
<td>chitin- and beta-glucan-containing cell wall</td>
<td>+</td>
<td></td>
<td></td>
<td>1</td>
<td>0.0597**</td>
<td>0.324</td>
<td></td>
</tr>
<tr>
<td>GO:0043231</td>
<td>intracellular membrane-bound organelle</td>
<td>−</td>
<td></td>
<td></td>
<td>0.0323</td>
<td>5.22·10⁻⁴***</td>
<td>1</td>
<td>6.9·10⁻⁵***</td>
</tr>
<tr>
<td>GO:0005634</td>
<td>nucleus</td>
<td>−</td>
<td></td>
<td></td>
<td>0.226</td>
<td>0.0699</td>
<td>1</td>
<td>7.74·10⁻³***</td>
</tr>
</tbody>
</table>

Table 5.4: List of GO terms that are significantly over- or under-represented (+ or - in first data column) amongst paralogs that have evolved in concert. P-values display on the table are false discovery rate (FDR) adjusted as reported by Fatigo+. For each individual post-WGD species we consider positive concert cases as reported by the distance method on nucleotides (p-value, < 0.02). The last column is based on positive hits in any post-WGD species after correcting for multiple tests by Bonferroni. Grey background indicates that the presence of that term can be explained by other enclosed ontologies and therefore is in fact redundant. Terms that show bias but that have significant ontological successors or parts at every species column were already excluded for the same reason.
Although genes under concerted evolution may seem to be dispensable in the laboratory, deletions may still reduce the fitness of the individual under certain environmental conditions. Only this way “redundant” copies would be preserved overtime due to purifying selection. Consequently, changes in selective pressures must be the most likely cause of gene loss in WGD paralogs (e.g. triggered by changes in the environment and genetic drift).

5.3.2.5 Severe loss of post-WGD gene redundancy in C. glabrata

*C. glabrata* does present a distinguishable lower number of extant WGD paralogous pairs and positive cases of concerted evolution (Figure 5.7) compared to other post-WGD species. Furthermore, there is no significant difference between τ statistic distributions in all four categories in *C. glabrata* despite some clear on-going full sequence homogenisation (paralogous pairs with τ = 1) in category C0.0 (Figure 5.8). Only in this class, *C. glabrata* has an equivalent or slightly greater proportion of concerted evolution positive cases compared to the other two post-WGD species analysed. In category C1.0, most single losses involve *C. glabrata* gene copies (88%). Within this subset there are 16.8% cases of convergent concert in *Saccharomyces* species (over 50% of the maximum possible taking into account the marginal number of positives per each species). Therefore a *C. glabrata* single gene loss seems to be also a good indicator of functional redundancy together with reciprocal gene loss.

A plot of intra-specific paralogous pairwise distance distribution based on whole proteomes shows qualitative differences between post-WGD *Saccharomyces* species, *C. glabrata* and other pre-WGD species such as *K. waltii* (Figure 5.9). All species present a characteristic bimodal distribution with a variable number of zero or quasi-zero pairwise distances (less than 0.05 changes per site). Both *Saccharomyces* species are the ones that exhibit a greater percentage of short paralogous distances. Moreover, WGD paralogies are the major contributor in *S. castellii* and nearly as important as post-WGD paralogies in *S. cerevisiae* if we consider relative densities rather than frequencies (not shown). This fact indicates that the WGD event, despite being 100Myr old, is a major contribution of genetic redundancy in these species. In contrast, *C. glabrata* does show the smallest percentage of zero or quasi-zero pairwise distances amongst all species analysed. Redundancy buffered in WGD paralogies has been lost in *C. glabrata*. This species shows a profile more similar to other pre-WGD species instead.

In view of these results we can conclude that *C. glabrata* has undergone a post-speciation gene redundancy reversing process and stabilizing evolutionary dynamics have maintained redundancy in *S. cerevisiae* and *S. castellii*. This contrasting feature between *C. glabrata* and the other two post-WGD species studied, parallels phenotypic peculiarities (e.g. pathogenicity) and the molecular phylogenetic exclusion of *C. glabrata* from the yeast lineage.

Already Dujon and collaborators noticed an increment of the gene loss rate and smaller
fraction of (quasi-) identical paralogs in *C. glabrata* as compared with other Hemiascomycetous yeast species (Dujon et al., 2004). Moreover, in later work, Fisher and colleagues found that *C. glabrata*, *C. albicans* and *D. hansenii* show evidence of accelerated genome evolution (genomic instability and multiple rearrangements) (Fischer et al., 2006). The first two are the major causal agent of candidiasis in humans whereas the later has occasionally been related to disease. As pathogens, fixation of genetic changes may be boosted by reduced population sizes and consequent genetic drift. Nonetheless, the distribution of pairwise distances in *C. albicans* does not match the genetic economy present in *C. glabrata*. Although *D. hansenii* shows a slightly more similar profile, this feature cannot clearly be linked with pathogenicity.

Figure 5.9: Amino acidic pairwise distance profiles between paralogs in some yeast species based on 10000 bootstrap replicates (see Methods for details). Distances are corrected for multiple changes (Eq. B.3). Each bar includes a range of 0.05 changes per site wide. Except for *C. albicans* and *D. hansenii*, different grays indicate the contribution of different types of paralogous pairs: WGD paralogs (product of the whole genome duplication), pre-WGD paralogs (between loci existent before the WGD event) and other paralogies (involving any locus that is not product the WGD event neither existed previously). The dashed line separates pre- and post-WGD species. Arrowheads indicate two clear modes at each frequency plot.
5 Detecting concerted evolution in yeast

5.4 Conclusions

If the possibility that paralogous genes have evolved in concert is not taken into account, certain types of analyses may yield spurious results. For instance, attempts to date duplication events would identify paralogies as more recent than they actually are (Gao and Innan, 2004). Additionally, concerted evolution is an important process that drives the evolution and functional differentiation or redundancy in some multi-gene families (Elder and Turner, 1995). Therefore it is important to use statistical tests in order to detect convergent evolution. In this study, we have proposed a new distance-based method to detect concerted evolution. We used this method, together with other approaches previously used in multi-loci analyses in yeast to find patterns of concerted evolution amongst WGD paralogies.

Our findings underline the importance to carry out control analyses to assess the performance of detection methods when the null hypothesis is true. Concert-free empirical datasets are very useful to this end as unknown statistical structures present in real data could hamper methods based on simple model assumptions. Unfortunately, simulated or randomised data is usually too “well-behaved”. In this study, we used single gene family orthologs as the control dataset since convergent evolution should be rare within these homologies. To the best of our knowledge there are not many precedents of empirical type I error assessment in the literature. One example is Posada’s work that evaluated the sensitivity and specificity of recombination methods using empirical data (Posada, 2002). Our results using the single-gene family control dataset indicate that noise in real data can result in significant deviations from the intended type I error. Therefore re-calibration of statistical tests may well be needed in order to avoid false positives.

Actual intra-specific paralogous sequences may well evolve under divergent intra-specific parameters that could result in artifactual convergence signal. Unfortunately, a control dataset composed of inter-specific homologs will never capture this complexity. Here we proposed and used a simple method based on synonymous codon resampling to detect false positives due to codon preferences. We found that if we take all nucleotide position into account divergent codon frequencies cause an increased rate of false positives, especially among S. castellii paralogs. Thus, a possible gain in test power using more data is paralleled by systematic errors. Therefore, we recommend caution when interpreting results obtained using methods that strongly rely on synonymous polymorphism.

Results from the yeast WGD paralogous dataset indicate that distance- and tree reconstruction based methods are in general more powerful than searching for heterogeneous distribution of substitutions. The first two approaches take advantage of phylogenetic information available based on synteny information (Byrne and Wolfe, 2005). This allows them to detect patterns of concerted evolution that involve changes homogeneously distributed along the sequence. These
5.5 Acknowledgements

We would like to thank members of the Molecular Evolution and Bioinformatics Laboratory in the Smurfit Institute of Genetics led by Prof. Kenneth H. Wolfe for providing us with early access to results in their work on *Kluyveromyces polysporus* and specially Dr. Kevin Byrne for showing us how to make the most of the YGOB. Finally, we also want to express our gratitude to Dr. Guy Drouin for clarifications about his work in gene conversion in yeast and Dr. Jose Castresana for additional indications as to how to use Gblocks. This work has been supported by Science Foundation Ireland, grant number (04/BR/B0441).
Chapter 6

General discussion

Preface. In Chapters 2 to 5 above, I have described some forms of complex non-stationary molecular evolution and contributed methods for detecting and exploding their occurrence to reach biologically relevant conclusions. In this final chapter, I summarise these conclusions and achievements, and connect this work with prospective lines of further research.

As highlighted in Chapter 1, molecular evolution presents non-stationary peculiarities that allow us to link homologous sequence data based observations to higher biological aspects such as genome-scaled evolution, gene function and biological systems dynamics. This complexity permits us to reach meaningful conclusions from homologous sequence analyses. In my view, this constitutes the good edge of the molecular primary sequence evolution complexity “sword” used by molecular evolutionist. Nevertheless this complexity also presents a bad edge as departures from assumptions may lead to spurious results specially in phylogeny dependent studies as non-stationary and episodic evolution may well result in erroneous tree reconstructions.

In Chapter 2, I dealt with phylogenetic artifacts caused by heterotachous evolutionary patterns. I showed that changes in substitution rates across groups of taxa not taken into account by the phylogenetic reconstruction method can result in spurious outcome phylogeny. These results add up to other studies that describe homologous effects in other alternative heterotachous scenarios (Chang, 1996; Inagaki et al., 2004; Kolaczkowski and Thornton, 2004; Philippe et al., 2005). The novelty of this approach is that I used a more realistic model of gene functional divergence. This type of heterotachy may result from (a) neo- or subfunctionalisation of paralogous genes after gene duplication or (b) simply sudden changes in evolutionary constraints due to other changes in the molecular and functional context of that gene (Gu, 1999, 2006; Siltberg and Liberles, 2002).

Consistent resolution of multi-gene family phylogenies is key to phylogeny-dependent downstream analyses such as: evolutionary constraint analysis that pinpoint important functional
sites under adaptive evolution (e.g. Fares and Wolfe, 2003), sorting in time the occurrence of
gene or genome duplication and speciation events (Fares et al., 2006; Langkjaer et al., 2003), or
studying gene gain and loss dynamics (Gao and Innan, 2004; Cotton and Page, 2006).

Findings confirmed the existence of phylogenetic artifacts when genuine phylogenetic signal
is weak due to a brief time elapse between successive duplications (short internal branch length)
or to later fast evolution of lineages. I also found that the presence of bias depends on multiple
non-independent factors such as tree shape, height, percentage of functionally diverging sites,
and the intensity of rate heterogeneity. Moreover, other major sources of bias such as convergent
and concerted evolution (addressed in Chapter 5) or compositional bias aggravates this situation
making difficult to establish the specific cause of bias as is shown in the SecA gene case (Chapter
3).

Several studies have already proposed statistical tests to detect covariotide structure in the
data. On the one hand covarion-like and standard probabilistic models fit to the data can be
compared using the likelihood-ratio test (LRT) and other model selection criteria such as Akaike
and Bayesian information criterion (AIC and BIC respectively) (Galtier, 2001; Huelsenbeck,
2002; Wang et al., 2007). On the other hand some have developed ad-hoc approaches such as the
parametric test proposed by Ané and colleagues (Ané et al., 2005) and an earlier non-parametric
approach described by Lockhart and coworkers (Lockhart et al., 1998). Nevertheless, none of
these methods attempts to determine whether such covarion structure can result in biased phy-
logeny reconstructions. Some have addressed this problem using diverse strategies. For example,
Huelsenbeck (1997) proposed a method to determine whether MP reconstruction would fall into
a LBA trap (Felsenstein, 1978) using simulations. Nevertheless his method assumes that a tree
reconstructed using maximum-likelihood is correct. This assumption is not true in general as
shown in this work (Chapters 2, 3 and 5) and elsewhere. In a series of articles, Lyons-Weiler
and coworkers proposed a phylogeny independent all-encompassing method, relative apparent
synapomorphy analysis (RASA), that would be able to determine what lineages and sites are re-
sponsible for LBA among other useful observation for phylogenetic analyses (Lyons-Weiler et al.,
1996; Lyons-Weiler and Hoelzer, 1997; Lyons-Weiler and Milinkovitch, 1997; Lyons-Weiler et al.,
1998). Nevertheless, several groups showed later compelling evidence against these claims (Far-
riss, 2002; Faivovich, 2002; Simmons et al., 2002). The presence of compositional heterogeneity
can be detected using an standard \( \chi^2 \) test. However, since homologous sequences are not inde-
pendent observations, the phylogeny must be known.

In Chapter 2, I designed a test to detect possible phylogenetic bias in quartet of well-resolved
subtrees. It is based on the triangle geometry where each vertex represents one of the three pos-
sible unrooted topologies that could link those four groups. Quartet analysis is quite common in
phylogenetic theoretical studies as is the smallest dataset that presents unrooted phylogenetic
uncertainty (Felsenstein, 2004). The resulting test was evaluated satisfactorily against biased
and non-biased simulated datasets. Later I applied this method to real problematic data: the SecA gene in plastids. Formerly, this phylogeny constituted evidence against the now consolidated single primary endosymbiosis theory of plastid acquisition by photosynthetic eukaryotes (Valentin, 1997). Later Barbrook and colleagues strongly suggested that the apparent polyphyly of plastid SecA was result of compositional bias (Barbrook et al., 1998). Nevertheless, the reassessment of this phylogeny shows that the actual cause is covariotide structure and heterogeneous RAS distributions across major taxon groups probably due to functional divergence. Moreover, apparently SecA is not even a single gene family in green plants, a group that has two paralogous copies of this gene.

These results complement other works that show the susceptibility of single or few-genes studies to this type of artifacts (Lockhart et al., 1998; Inagaki et al., 2004; Lockhart et al., 2006), thus advocating for the use of more robust multi-gene phylogenomic studies towards resolving the Tree of Life (e.g. Fitzpatrick et al., 2006; Rodríguez-Ezpeleta et al., 2007). Nevertheless it would be interesting to test whether certain total evidence techniques are susceptible to covariotide bias as well. More concretely, standard multi-gene alignment concatenation based analyses assume a single set of branch lengths common across genes (linked model). In fact many studies have already shown that this is not true in general (Pupko et al., 2002; Douzery and Huchon, 2004; Takishita et al., 2005; Simpson et al., 2006). Thus, under this approach, well-supported taxon groups may result from correlated lineage lengths and not true phylogenetic signal (Chapter 2). Although intuitively this would be the case more often between truly close related lineages (resulting in a "good" bias), most researchers may well prefer a honest "do not know" or "not that sure" answer. Moreover, this could give linked models on concatenated alignments an unfair advantage upon other supermatrix and supertree approaches in terms of how much data is required to obtain a reliable (apparently high supported) phylogeny.

Correction of phylogenetic artifacts such as the ones caused by covariotide or compositional bias consists in closing the gap between the data used in the analysis and the optimal criteria recruited for the tree reconstruction. Some have proposed selective criteria to choose data that a priori should be less problematic. For example, Brinkmann, Rodríguez-Ezpeleta and colleagues argued that elimination of fast evolving sites and lineages alleviates LBA effects (Brinkmann et al., 2005; Rodríguez-Ezpeleta et al., 2007). Additionally, Fares and colleagues noticed that phylogenetic bias in yeast whole genome duplication (WGD) paralogs results is more acute in asymmetrically evolving duplicates (Fares et al., 2006). These phylogenies were used earlier (Langkjæer et al., 2003) to challenge the now well-established WGD theory (Wolfe and Shields, 1997). The test based on quartets described in Chapter 2 and functional divergence model fit used in Chapter 3 can be incorporated to this set of tools. Although I have not chased this any further, one way to move forward is the definition of phylogeny-independent criteria to sort sites based on their signal to noise ratio (e.g. Pisani, 2004). We could employ computational
optimisation techniques such as genetic algorithms to perform weighted bootstrapping if the
signal to noise ratio depends on the sample as a whole rather than individual sites. Alternatively,
using more realistic models of evolution may also reduce the bias (Pupko et al., 2002; Shalchian-
Tabrizi et al., 2006; Lartillot et al., 2007). Along these lines, I think that developing methods
that account for an arbitrary number of rate shifts and heterogeneous covarion structure across
the tree would be a step forward.

Despite the handicap that non-stationarity imposes on the evolutionary analysis of primary
sequence data, this peculiarity conceals important aspects of the evolution of biological
molecules. For example, the CCT complex is a particular example of adaptation across life
domains. CCT are essential heat-shock molecules with an important role in the correct fold-
ing of nascent protein chains into their functional three-dimensional conformation. Eukaryotic
CCT is known to fold molecular elements of the cytoskeleton in a very specialised way through
hydrophilic specific interactions (Llorca et al., 2001a; McCormack et al., 2001). The correlation
between number of CCT genes and cell architecture complexity across life domain leads to the
hypothesis that CCT hetero-oligomerism was a decisive factor in the emergence of complex and
diverse cytosolic structures in eukaryotic cells (Table 6.1).

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Archaeobacteria</th>
<th>Eukaryota</th>
</tr>
</thead>
<tbody>
<tr>
<td>strict homo-oligomerism</td>
<td>variable hetero-oligomerism</td>
<td>strict hetero-oligomerism</td>
</tr>
<tr>
<td>(1 gene)</td>
<td>(1-5 genes)</td>
<td>(8-9 genes)</td>
</tr>
</tbody>
</table>

Archaeal CCT homologs show an intermediate stage between homo- and hetero-oligomerism
with recurrent independent gains of additional CCT genes in diverse lineages. Archibald and
colleagues noticed this peculiarity and put forward a functionally neutral model of evolution
where hetero-oligomers are retained due to compensatory coevolution between subunits without
functional advantage gain (Archibald et al., 1999). In Chapter 4, I put this hypothesis to test
using in silico analyses based in four conditions to be met in order to corroborate Archibald’s
functionally neutral theory of CCT hetero-oligomerism evolution: (1) the presence of acceler-
ated and possibly asymmetric evolution of inter-specific paralogs after duplication, (2) that these
changes concentrate in inter-subunit interface or proximal regions, (3) the presence of compen-
satory mutations (acceleration or positive selection) in proximal regions and (4) evidence that
compensated and compensating sites coevolve.

Results indicate that lineages with only 2 genes encoding different CCT subunits show evo-
olutionary traits compatible with the neutral theory. However, in lineages with 3 subunits some
changes in substrate-binding sites have been fixed under adaptive or accelerated evolution and
do not seem to compensate for other changes in inter-subunit interfaces. This observation and
the prevalence of functional divergence in eukaryotic CCT (Fares and Wolfe, 2003) indicates that functionally neutral fixation of hetero-oligomerism is more unlikely with an increasing number of subunit types. Intuitively, adding one more diverging subunit requires coincidental compensatory changes in all other paralogs, thus the probability of successfully fix these changes with no functional gain diminishes with more subunits types present. Further analogous studies on other hetero-oligomeric complexes would reveal whether this trend is just a peculiar aspect of CCT evolution in Archeabacteria or can be generalised.

Other analyses based on non-stationary models can help in characterising the evolutionary fate of gene duplicates even in the absence of three-dimensional structure. **Saccharomyces** species present an interesting case as an ancestral WGD event resulted in over 450 extant concurrent paralogies in this model species (Wolfe and Shields, 1997). Thus, it is not surprising that this dataset has been focus of extensive research because (a) it presents a great opportunity to analyse multiple gene duplications like a system rather than independent events and (b) we can remove the timing factor from the equation since we are certain that all duplications occurred simultaneously.

Nevertheless, convergent evolution is an important obstacle for this kind of studies as most mathematical tools available assume that paralogous genes evolve independently after duplication. After evaluating the applicability of some existing methods to detect concerted evolution, I decided to develop a new method based on a simple distance-based approach and synteny information collected in Yeast Gene Order Browser (YGOB) (Byrne and Wolfe, 2005). I highlighted the importance of controlling for a possible elevated number of false positives due to non-specific statistical signal present in real datasets. I found that we can effectively use strict single gene families as a negative control dataset for concerted evolution. Simulated or randomly permuted data is far too well behaved and may lead to a liberal test design. Nevertheless, this technique cannot correct for genome specific evolutionary parameters that may result in sequence convergence patterns such as compositional bias due to codon usage preference. I showed that in general it is a good practice to exclude nucleotide positions that present mixtures of synonymous and non-synonymous substitutions.

Finally, I used this new method on the WGD paralogs dataset to determine to what degree convergence has shaped the evolutionary history of these paralogs. I found that the degree of concert depends on several factors: the species under analysis, gene redundancy (differential gene loss patterns across species), and gene function as reported through ontology terms annotations. I discovered a new functional category that presents a strikingly high occurrence of concerted evolution on **Candida glabrata**; this species is the one with a lower level of genetic redundancy. These genes are linked to cell wall morphology and secretion mechanism and may have an important role in the pathogenicity of this species. However, fully exploring this hypothesis requires further in-depth analyses including experimental corroboration.
Once we can pinpoint the presence of confounding inter-dependent evolution of gene duplicates we can progress to analyse their sequences for more clues that allow us to determine their evolutionary fate. Although it has not been approached in this work, future lines of research include exploring ways to extract more information out of primary sequence data. For example, so far in literature asymmetric sequence evolution has been quantified using only whole sequence substitution-rates metrics such as the average number nucleotide or amino acid substitutions or selective pressures (measured using the $\omega$ ratio) (Conant and Lewis, 2001; Fares et al., 2006; Byrne and Wolfe, 2007). Nevertheless changes in background substitution- and recombinatorial rates can also result in asymmetric evolution not bound to functional divergence. Other functional divergence models proposed in other studies (e.g. Gu, 1999; Siltberg and Liberles, 2002) and the ones described in Chapters 2 and 3 may offer an alternative to assess evolutionary asymmetry where the statistic of interest is not the length of key tree branches but the proportion and functional importance of sites that diverged after duplication. As far as I am aware of, this kind of models have been applied to a single or few genes datasets only, never to a sizable dataset.

Recently, Tirosh and coworkers have question the efficacy of analysis based only on genome sequence in the study of functional divergence of paralogies (Tirosh and Barkai, 2007; Tirosh et al., 2007). They rightfully argued that differential gene expression patterns constitute a strong determinant of functional divergence. This is in agreement with the DDC subfunctionalisation model (described in section 1.1.1.4). They also propose a novel method to compare co-expression across species which they applied to the WGD yeast paralogs using pathogenic *Candida albicans* as outgroup (Tirosh and Barkai, 2007). The lack of substantial expression data for a closer pre-WGD yeast relative species presented a limitation to their analysis. Nevertheless that study highlights the importance of taking into account several heterogeneous sources of information to fully determine the nature of functional and molecular diverging evolution of paralogies. These data types include gene coding- and promoter regions sequence, expression profiles, interaction networks and three-dimensional structure. In turn, the integrative and joint analysis of heterogeneous biological data sources seems to be the best way forward to produce more realistic models of molecular evolution. However the ultimate aim is not only the full characterisation of the evolution single homologies, but also the evolution of systems as a unit.
Appendix A

Genetic code

A.1 The twenty amino acids

Protein main chains is constituted of a succession of twenty possible amino acids characterised by a variable side chain or residue. Each amino acid has an associated single letter and three letter code to make notation more compact specially for multi sequence alignments.

<table>
<thead>
<tr>
<th>full name</th>
<th>single letter</th>
<th>three letter</th>
<th>full name</th>
<th>single letter</th>
<th>three letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>methionine</td>
<td>M</td>
<td>Met</td>
<td>asparagine</td>
<td>N</td>
<td>Asn</td>
</tr>
<tr>
<td>valine</td>
<td>V</td>
<td>Val</td>
<td>aspartic acid</td>
<td>D</td>
<td>Asp</td>
</tr>
<tr>
<td>isoleucine</td>
<td>I</td>
<td>Ile</td>
<td>glutamic acid</td>
<td>E</td>
<td>Glu</td>
</tr>
<tr>
<td>leucine</td>
<td>L</td>
<td>Leu</td>
<td>glutamine</td>
<td>Q</td>
<td>Gln</td>
</tr>
<tr>
<td>serine</td>
<td>S</td>
<td>Ser</td>
<td>histidine</td>
<td>H</td>
<td>His</td>
</tr>
<tr>
<td>threonine</td>
<td>T</td>
<td>Thr</td>
<td>arginine</td>
<td>R</td>
<td>Arg</td>
</tr>
<tr>
<td>proline</td>
<td>P</td>
<td>Pro</td>
<td>lysine</td>
<td>K</td>
<td>Lys</td>
</tr>
<tr>
<td>alanine</td>
<td>A</td>
<td>Ala</td>
<td>phenylalanine</td>
<td>F</td>
<td>Phe</td>
</tr>
<tr>
<td>glycine</td>
<td>G</td>
<td>Gly</td>
<td>tyrosine</td>
<td>Y</td>
<td>Tyr</td>
</tr>
<tr>
<td>cystine</td>
<td>C</td>
<td>Cys</td>
<td>tryptophan</td>
<td>W</td>
<td>Trp</td>
</tr>
</tbody>
</table>

Table A.1: Amino acid single- and three letter codes. Amino acid are grouped as used for the Dayhoff 6 groups recoding. These classes are based in Dayhoff log-odds matrix (Dayhoff et al., 1978). Dayhoff 6 groups are \{AGPST, C, FWY, HRK, MILV and NDEQ\} whereas Dayhoff 4 groups, as defined by Rodriguez-Ezepeleta and colleagues (2007), as combining aromatic (FWY) with hydrophobic (MILV) residues and discarding "rare" cystines (C) as unknown characters.
A.2 Standard Genetic Code

The standard code is the one used by most eukaryota, bacteria, archaea and chloroplast. Translation typically starts at codon ATG. Other codons can also initiate translation (e.g. TTG or CTG) although this is quite rare in eukaryota. In any case, start codons are invariably translated into a methionine (Met) regardless of their corresponding amino acid on the code table.

<table>
<thead>
<tr>
<th>Codon</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT</td>
<td>Phe</td>
</tr>
<tr>
<td>TTC</td>
<td>Phe</td>
</tr>
<tr>
<td>TTA</td>
<td>Leu</td>
</tr>
<tr>
<td>TTG</td>
<td>Leu</td>
</tr>
<tr>
<td>CTT</td>
<td>Leu</td>
</tr>
<tr>
<td>CTC</td>
<td>Leu</td>
</tr>
<tr>
<td>CTA</td>
<td>Leu</td>
</tr>
<tr>
<td>CTG</td>
<td>Leu</td>
</tr>
<tr>
<td>ATT</td>
<td>Ile</td>
</tr>
<tr>
<td>ATC</td>
<td>Ile</td>
</tr>
<tr>
<td>ATA</td>
<td>Ile</td>
</tr>
<tr>
<td>ATG</td>
<td>Met</td>
</tr>
<tr>
<td>GTT</td>
<td>Val</td>
</tr>
<tr>
<td>GTC</td>
<td>Val</td>
</tr>
<tr>
<td>GTA</td>
<td>Val</td>
</tr>
<tr>
<td>GTG</td>
<td>Val</td>
</tr>
</tbody>
</table>

Table A.2: Standard genetic code. Different column, row or line within a cell represent changes in second, first and third codon position respectively. "Ter" denotes non-sense terminal codons.

As described in section 1.1.1.2, most nucleotide changes at at first and second codon positions result in an amino acid replacements. Exceptions include non-sense substitution TAA→TGA, leucine synonymous substitutions: TAA→CTA and TTG→CTG; and arginine synonymous substitutions: CGA↔AGA and CGG↔AGG.
Appendix B

Evolutionary models and distances formulae

B.1 DNA evolution models

The following table is a compilation of Q-matrices for most common evolutionary stationary models used for DNA phylogenetic analyses. Columns represent the original base whereas rows represent the resulting base. Both dimensions are indexed with A, G, C and T in this order.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(-\kappa\pi_G) , \pi_C , \pi_T)</td>
<td>(-\kappa\pi_G) , \pi_C , \pi_T)</td>
<td>(-\pi_G) , \pi_C , \pi_T)</td>
</tr>
<tr>
<td>(\kappa\pi_A) , \pi_G , \pi_T)</td>
<td>(\kappa\pi_A) , \pi_C , \pi_T)</td>
<td>(\pi_A) , \pi_C , \pi_T)</td>
</tr>
<tr>
<td>(\pi_A) , \pi_G , \kappa\pi_T)</td>
<td>(\pi_A) , \pi_G , \lambda\pi_T)</td>
<td>(\pi_A) , \pi_C , \pi_T)</td>
</tr>
</tbody>
</table>

Table B.1: Popular DNA models Q matrices. Columns represent the original base whereas rows represent the resulting base. Both dimensions are indexed with A, G, C and T in this order. The stationary frequency for base \(x\) is denoted with \(\pi_x\). Other Greek letters represent other model specific parameters. Diagonal elements omitted (\(-\)) are equal to the negative sum of other elements in the row.
B.2 Codon-based models

In the case of codon based models the corresponding Q matrix has 61x61 entries. Most of this are in fact zero as typically only transition between codons that involve a single nucletide change are considered. To easy notation, models are presented using a "cases" construct:

\[
Q_{ij} = \begin{cases} 
0, & \text{both codons differ at two or more positions} \\
\pi_j, & \text{synonymous transversion} \\
\kappa \pi_j, & \text{synonymous transition} \\
\omega \pi_j, & \text{nonsynonymous transversion} \\
\omega \kappa \pi_j, & \text{nonsynonymous transcription}
\end{cases}
\]  

(B.1)

The model above is the one published by Nielsen and Yang (1998) that simultaneously takes into account selective presure ($\omega$), an arbitrary transition/trasversion ratio ($\kappa$) and different codon frequencies ($\pi_i$).

B.3 Model-based pairwise distances measures

The most simple distance mecasure is the percentage of differences between two sequences also named $p$-distance:

\[
p = \frac{\text{number of differences}}{\text{number of sites}}
\]  

(B.2)

If we assume that all sites evolve under the same substitution rate, multiple substitution can be corrected using the Poisson distance:

\[
d_{\text{Poisson}} = -C \ln(1 - C^{-1} p)
\]  

(B.3)

\[
C = \frac{(1 - |A|) / |A|}{\text{number of possible states}}
\]

(B.4)

However, when substitution rates are heterogeneous across sites as drawn from a continous $\Gamma$ distribution with tied shape and scale parameter $\alpha$ and $\beta = \alpha^{-1}$ the following correction is needed instead:

\[
d_r = -C \alpha \left[ 1 - (1 - C^{-1} p)^{-1/\alpha} \right]
\]  

(B.5)
So far we have assumed that all substitutions are equiprobable (JC model). Nevertheless is well known that this is not true in general in biological data sequences. For example the pairwise distance based in the K2P model when rates are homogeneous across sites can be calculated as:

\[
\begin{align*}
d_{K2P} &= -0.25 \cdot \ln \left[ \left(1 - 2P - Q\right)^2 \left(1 - 2Q\right) \right] \quad \text{(B.6)} \\
P &= \text{observed transitions/number of sites} \\
Q &= \text{observed transversions/number of sites}
\end{align*}
\]

The ration transition/transversion \( \kappa \) is estimated as:

\[
\kappa = 2 \cdot \frac{\ln \left(1 - 2P - Q\right)}{\ln \left(1 - 2Q\right)} - 1 \quad \text{(B.7)}
\]

The corresponding continuous \( \Gamma \) correction is:

\[
\begin{align*}
d_{K2P+\Gamma} &= -0.5 \cdot \left[ 1.5 - \left(1 - 2P - Q\right)^{-1/\alpha} + 0.5 \cdot (1 - 2Q)^{-1/\alpha} \right] \\
\kappa_{\Gamma} &= 2 \cdot \frac{1 - \left(1 - 2P - Q\right)^{-1/\alpha} + 0.5 \cdot (1 - 2Q)^{-1/\alpha}}{1 - (1 - 2Q)^{-1/\alpha}}
\end{align*}
\]

These equations become more complex with other models. These can be found elsewhere (e.g. Felsenstein, 2004).

### B.4 LogDet correction

The following logDet formula consists in the original logDet correction proposed by Steel but taking into account possible compositional heterogeneity across taxa and avoiding zero determinants using residue frequencies, \( \pi_{\text{trace}} \), due to poor sampling in small datasets:
\[ \hat{D}_{xy} = \frac{1}{N-1} \left( \ln \det (J_{xy}) - 0.5 \ln \det (F_x F_y) \right) \]  

\( N \equiv \) number of character in the alphabet  

(e.g. 4 for DNA or 20 for protein)  

\[
J_{xy}(i,j) = \begin{cases} 
\pi_{\text{trace}}, & i = j \wedge (\pi_{xi} = 0 \lor \pi_{yi} = 0) \\
\pi_{xij}, & \text{otherwise}
\end{cases}
\]

\[
F_x(i,j) = \begin{cases} 
\max(\pi_{\text{trace}}, \pi_{xi}), & i = j \\
0, & i \neq j
\end{cases}
\]

\( \pi_{xi} \equiv \) normalised frequency of character \( i \) in sequence \( x \)

\( \pi_{xij} \equiv \) joint frequency of character \( i \) and \( j \) in sequences \( x \) and \( y \)

\( \pi_{i} = (\pi_{xi} + \pi_{yi})/2 \), averaged frequency of character \( i \)

\( \pi_{\text{trace}} \equiv \) normalised minimal frequency to avoid zero determinants
Appendix C

Covarion bias test equation demonstration

Under an RAS model of evolution and considering subtree with similar tree shapes, the bias plot proposed here must look the same (present the same density of points) from any hypothesis (tree topology) perspective. Let us define the pair \((X_H, Y_H)\) as the mean point location of the plot when we set the origin of coordinates on the triangle centre and when the y-axis is set to pass through the hypothesis \(H\) vertex based on the layout depicted in Figure 2.3. Then based on the symmetry assumption, variance should be equal for all three coordinate systems:

\[
\begin{align*}
VAR[X_H] &= VAR[X_{H_1}] = VAR[X_{H_2}] \\
VAR[Y_H] &= VAR[Y_{H_1}] = VAR[Y_{H_2}]
\end{align*}
\] (C.1)

It is easy to transform a point \((X,Y)\) based on a coordinate system to another that is a simple rotation of the former with angle \(\lambda\):

\[
\begin{pmatrix}
X' \\
Y'
\end{pmatrix} =
\begin{pmatrix}
X \cos \lambda + Y \sin \lambda \\
-X \sin \lambda + Y \cos \lambda
\end{pmatrix}
\] (C.2)

Knowing that:

\[
VAR[aX + bY] = a^2VAR[X] + b^2VAR[Y] + 2abCOV[X,Y]
\] (C.3)

and using the transformation Equation (C.2), the variance for hypotheses \(H_0\) or \(H_2\) (\(\lambda = 2\pi/3\) and \(-2\pi/3\) respectively) can be expressed as a function of the variance for hypothesis \(H_1\) as follows:
where:

\[ \text{VAR}[X_h] = V(\lambda, X_{H_1}, Y_{H_1}) + C(\lambda, X_{H_1}, Y_{H_1}) \]
\[ \text{VAR}[Y_h] = V(\lambda, X_{H_1}, Y_{H_1}) - C(\lambda, X_{H_1}, Y_{H_1}) \]  
(C.4)

\[ \text{VAR}[X_h] = V(-\lambda, X_{H_1}, Y_{H_1}) + C(-\lambda, X_{H_1}, Y_{H_1}) \]
\[ \text{VAR}[Y_h] = V(-\lambda, X_{H_1}, Y_{H_1}) - C(-\lambda, X_{H_1}, Y_{H_1}) \]

\[ V(\lambda, X, Y) = \cos^2 \lambda V[\text{VAR}[X]] + \sin^2 \lambda V[\text{VAR}[Y]] \]  
(C.5)

\[ C(\lambda, X, Y) = 2 \cos \lambda \sin \lambda \text{COV}[X, Y] \]

As \( \sin \lambda = -\sin \lambda \) and \( \cos \lambda = \cos -\lambda \), it follows that:

\[ V(\lambda, X, Y) = V(-\lambda, X, Y) \]  
(C.6)

\[ C(\lambda, X, Y) = -C(-\lambda, X, Y) \]

From Equations (C.1), (C.4) and (C.6) we can deduce that the covariance term, \( C \), must be null for all three hypotheses:

\[ \text{VAR}[X_{H_0}] = \text{VAR}[X_{H_2}] \]
\[ V(\lambda, X_{H_1}, Y_{H_1}) + C(\lambda, X_{H_1}, Y_{H_1}) = V(-\lambda, X_{H_1}, Y_{H_1}) + C(-\lambda, X_{H_1}, Y_{H_1}) \]  
(C.7)

\[ V(\lambda, X_{H_1}, Y_{H_1}) + C(\lambda, X_{H_1}, Y_{H_1}) = V(\lambda, X_{H_1}, Y_{H_1}) - C(\lambda, X_{H_1}, Y_{H_1}) \]
\[ 2C(\lambda, X_{H_1}, Y_{H_1}) = 0 \]

This last equality is satisfied if and only if \( \lambda \) is multiple of \( \pi/2 \) (not the case) or when the covariance is zero. Taking this into account we can easily prove that the variance of \( Y_{H_1} \) must be the same as that of \( X_{H_1} \) to satisfy the symmetry condition:

\[ \text{VAR}[X_{H_1}] = \cos^2 \lambda \text{VAR}[X_{H_1}] + \sin^2 \lambda \text{VAR}[Y_{H_1}] + 0 \]
\[ 1 = \cos^2 \lambda \frac{\text{VAR}[X_{H_1}]}{\text{VAR}[X_{H_1}]} + \sin^2 \lambda \frac{\text{VAR}[Y_{H_1}]}{\text{VAR}[X_{H_1}]} \]  
(C.8)

This equation only holds if:

\[ \frac{\text{VAR}[X_{H_1}]}{\text{VAR}[Y_{H_1}]} = 1, \quad \text{VAR}[X_{H_1}] = \text{VAR}[Y_{H_1}] \]  
(C.9)

because \( \cos^2 \lambda + \sin \lambda = 1 \).
Here we consider the fact that for any angle $\lambda$ both coordinates have equal variances and their covariances are equal to zero. Assuming that the site position obtained constitutes an independent sample, the central limit theorem (CLT) ensures that the distribution of the means approaches a Gaussian and, because it has the same variance from every angle $x$-axis or $y$-axis, it must approximate a circular 2-D bell centred on the triangle barycentre where the standard error is $\sigma_x/\sqrt{N}$, $N$ being the sample size. Nonetheless, because site point coordinates are calculated based on parameter estimators obtained using the same site sample, it is not certain that the independence condition for CLT is actually met with a limited amount of data. Consequently, we need to confirm any conclusion below through simulations.

We can determine the formula for the cumulate probability of the distance to the centre taking as starting point the bivariate normal distribution or the chi-square. Here we develop the former. As the covariation is null, the mean is zero and the deviation is the same from every axis:

$$N(X, Y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \quad (C.10)$$

For further simplification, we integrate the probability volume along the axis $y = 0$, within certain distance $x$ using the circle perimeter equation as follows:

$$\Pr(D \leq d) = \int_0^d N(x, 0) L(x) dx$$

$$N(x, 0) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

$$L(x) = 2\pi x, \quad \equiv \text{perimeter}$$

$$\Pr(D \leq d) = \int_0^d \frac{1}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx$$

$$= \frac{1}{\sigma^2} \int_0^{d^2} \sqrt{\frac{\sigma^2}{2\pi}} \exp\left(-\frac{z^2}{2\sigma^2}\right) \frac{dz}{\sqrt{2\pi}}$$

$$= \frac{1}{2\sigma^2} \left[ -\frac{\sigma^2}{2\sigma^2} \exp\left(-\frac{z^2}{2\sigma^2}\right) \right]_0^{d^2}$$

$$= \frac{1}{2\sigma^2} \exp\left(-\frac{d^2}{2\sigma^2}\right)$$

We confirmed these results performing simulations and testing the goodness-of-fit of the p-value and mean point distributions using Kolmogorov-Smirnov test against a uniform and bivariate normal distribution, respectively. The simulations show that this constitutes a good approximation even with asymmetric data sets (different subtree heights and number of taxa).
up to 10,000 sites in length. Here the distance must be taken from the mean point centre estimated through parametric bootstrapping. However, using general samples, mean or median comparison tests like Wilcoxon test are more recommendable in these cases.
Appendix D

Investigating the origin of the green plastid SecA $\beta$ homolog

To determine whether green plastid SecA $\beta$ homologs resulted from a lateral gene transfer event (LGT) or an ancestral paralogy (Fig. 3.3) we incorporated SecA homologs from taxon groups that are supposed to be close relatives of Cyanobacteria (Gupta and Griffiths, 2002) to the phylogeny in Figure 3.2. Additionally we included $\alpha$-proteobacteria sequences to test the possibility that SecA $\beta$ has a mitochondrial origin. The resulting phylogeny (Fig. D.1) always presents a cluster including all plastid and Cyanobacteria homologs therefore not showing clear evidence supporting the LGT hypothesis.

Nevertheless it is always possible either that the potential gene transfer donor species was a close Cyanobacteria relative or that $\alpha$ and $\beta$ homologs do not follow the species phylogeny due to ancestral LGT events involving bacterial groups not present in this phylogeny. Unfortunately an exhaustive phylogenetic reconstruction of the full bacterial SecA homologous set using robust reconstruction methods is computationally prohibitive (over 800 sequences). Therefore we first reconstructed the full bacterial phylogeny using a simple and quick method (neighbour-joining using Poisson-corrected pairwise distances). We found that $\beta$ no longer constituted a cluster with $\alpha$ homologs (data not shown). Nevertheless when we used a robust method (maximum-likelihood based on WAG+$\Gamma$ model) on a subset of the data (chloroplasts, Cyanobacteria, Firmicutes, and close relatives of green $\beta$ in the full bacterial phylogeny) the Cyanobacteria and plastid monophyly appeared again. This pinpoints the presence of phylogenetic artifacts using simple reconstruction methods. In conclusion, there is no irrefutable evidence of a LGT event as the origin of SecA $\beta$ in green plastids. Therefore we cannot discard the ancestral paralogy hypothesis.
Figure D.1: Extended SecA phylogeny incorporating some bacterial taxon groups closely related to Cyanobacteria. We reconstructed this tree with maximum-likelihood and the WAG+Γ model. Numbers at relevant internal nodes indicate the non-parametric bootstrap support from a total of 100 repeats.
Figure E.1 represents the phylogeny of 96 non-partial and non-redundant CCT homologs found in archaeal genomes. Numbers at relevant internal nodes indicate bootstrap support out of 1000 repetitions. Different colours indicate paralogy. Notice that these are reused in different taxonomic groups. Sequences grouped within the same colour box or lineages are putative orthologs. Arrowheads indicate most probable gene duplication events considering topology and pattern of paralogy down towards leaves. Question marks indicate those possible duplication events most questionable considering lack of phylogenetic support. Dashed lines highlight apparent deep paralogies (possibly caused by lateral gene transfer, early gene duplication before taxonomic group divergence or simply phylogenetic reconstruction artifacts).
E Hetero-oligomerism in Archaeal CCT: Supplementary material

Figure E.1: Full archaeal CCT phylogeny.
E.2 Fit comparison for diverse codon-based models

We compared different codon-based models of DNA sequence evolution using the LRT criterion. The site-heterogeneous and time-homogeneous model M3 was significantly better than the stationary Goldman and Yang (1994) model (G&Y) as shown in Table E.1. In turn, branch-site models (BSB) at each branch leading to a major CCT subunit cluster also had a better model-data fit.

<table>
<thead>
<tr>
<th>Models</th>
<th>2ΔlnLk</th>
<th>Δdf</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>G&amp;Y vs M3</td>
<td>3880.41</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>BSB (Thermoplasmata α) vs M3</td>
<td>23.06</td>
<td>2</td>
<td>9.83 · 10⁻⁶</td>
</tr>
<tr>
<td>BSB (Thermoplasmata β) vs M3</td>
<td>44.64</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>BSB (Thermococci α) vs M3</td>
<td>47.53</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>BSB (Thermococci β) vs M3</td>
<td>54.15</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>BSB (Methanomicrobia 1-2) vs M3</td>
<td>14.76</td>
<td>2</td>
<td>6.23 · 10⁻⁴</td>
</tr>
<tr>
<td>BSB (Methanomicrobia 1) vs M3</td>
<td>34.03</td>
<td>2</td>
<td>4.08 · 10⁻⁸</td>
</tr>
<tr>
<td>BSB (Methanomicrobia 2) vs M3</td>
<td>23.70</td>
<td>2</td>
<td>7.14 · 10⁻⁶</td>
</tr>
<tr>
<td>BSB (Methanomicrobia 3) vs M3</td>
<td>36.36</td>
<td>2</td>
<td>1.27 · 10⁻⁸</td>
</tr>
<tr>
<td>BSB (Halobacteriales α) vs M3</td>
<td>94.80</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>BSB (Halobacteriales β-γ) vs M3</td>
<td>108.67</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>BSB (Halobacteriales β) vs M3</td>
<td>315.21</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>BSB (Halobacteriales γ) vs M3</td>
<td>253.18</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>BSB (Crenarchaeotes α-α'-γ) vs M3</td>
<td>28.57</td>
<td>2</td>
<td>6.25 · 10⁻⁷</td>
</tr>
<tr>
<td>BSB (Crenarchaeotes β-β') vs M3</td>
<td>25.59</td>
<td>2</td>
<td>2.77 · 10⁻⁶</td>
</tr>
<tr>
<td>BSB (Crenarchaeotes γ) vs M3</td>
<td>35.49</td>
<td>2</td>
<td>1.97 · 10⁻⁸</td>
</tr>
</tbody>
</table>

Table E.1: Likelihood ratio test for the comparison between codon-based models. Goldman and Yang model (G&Y) was compared to the M3 model. M3 model was also compared to branch-site model for each lineage after duplication using the likelihood ratio test. Twice the difference in the log-likelihood values for nested models (2ΔlnLk) were approximated to a χ² distribution, with the degrees of freedom being the difference in the number of parameters between the models compared.
Appendix F

Least-squares branch length estimation ignoring dubious pairwise distances

F.1 Method explained

First consider a quartet of species $V = \{a, b, c, d\}$. There is three possible full-resolved unrooted topologies $T = \{((a, b), c, d), ((a, c), b, d), ((a, d), b, c)\}$. Least-square tree search would select the topology and branch lengths that yield the minimal sum of squared differences between the observed pairwise distances and the ones induced by topology and branch lengths estimators. For instance, in the tree depicted in Figure 5.2, induced pairwise distances are solved by the following equation system:

$$
\begin{bmatrix}
D_{ab} \\
D_{ac} \\
D_{ad} \\
D_{bc} \\
D_{bd} \\
D_{cd}
\end{bmatrix} =
\begin{bmatrix}
l_1 + l_2 \\
l_1 + l_3 + l_5 \\
l_1 + l_4 + l_5 \\
l_2 + l_3 + l_5 \\
l_2 + l_4 + l_5 \\
l_3 + l_4
\end{bmatrix}
$$

(T.1)

Typically, least-squares optimisation estimates the set of branch length, $l_i$, 5 in total, as to minimise the sum of squared differences between observed distances, $\{D'_{xy} | x \neq y, yx \in V^2\}$, and their corresponding tree path induced distances, $\{D_{xy} | x \neq y, yx \in V^2\}$, is closest to zero. In weighted least squares (WLS) every difference is given a weight, $w_{xy}$, taking into account the difference between estimator variances:
F Least-squares branch length estimation ignoring dubious pairwise distances

\[ Z = \sum_{x,y} w_{xy} (D_{xy} - D'_{xy})^2 \]  

(F.2)

The number of pairwise distances equalises the number of parameters: 5 branch length distances and one topology. Therefore if the topology is known, there is a redundant pairwise distance. A closer examination of the equation system (F.1) reveals that any pairwise distance between non-sister taxa (those including the inner-most branch of the tree) can be rewritten as a combination of the other 3 remaining distances across between non-sister taxa (eq. 5.3). In contrast, pairwise distances within sister taxa pairs are not dispensable, as we need them to exactly determine the length of the inner-most branch.

In spite of this "extra" datum, if distances are consistent (observed distances converge to the real values) the least square sum would approach zero for the correct topology and branch lengths. This must hold even if we leave out a "redundant" pairwise distance. In this case, there is always an exact least-square solution \( Z = 0 \) for each topology although it may include biologically meaningless negative values.

Therefore, if we know the correct topology, we can quantify the incompatibility between a sixth distance with the other 5. For instance, if we believe that concerted evolution may be responsible for a reduced sequence divergence between paralogs, we may test whether the amount of evolution expected considering all other data is greater than the observed pairwise distance. This can be extended to handle more than one suspect pair simultaneously, as long as for each paralog there is at least one extra ortholog or more generally, an homolog that is phylogenetically closer than the other paralog. In other words, the quartet schema in Figure 5.2 can be drawn for each targeted candidate paralogous pair.

In this case, there is no simple single equation solution such as (5.3) to calculate alternative concert-free distance estimators. Instead we used weighted least-square fitting (WLS) to obtain branch length estimates for the whole tree. WLS is preferable over ordinary least-squares in order to downweigh the statistical noise introduce by highly divergent sequences. We set weights to the inverted square of the observed distance \( w_{xy} = (D'_{xy})^{-2} \). Occasional zero pairwise distances were given the same weight at the smallest non-zero pairwise distance present in the alignment in order to avoid infinite weights. Then, we used the branch length sum along the tree path between candidate pair tips as the alternative concert-free pairwise distance estimator, \( \hat{D}_{xy} \).

Regarding the empirical WGD duplicates we excluded all paralogous pairwise distances as depicted in Figure F.1. If we have just a single post-WGD homolog for a particular gene copy (so there is just a single pair of paralogs to test) we still can proceed by using at least one pre-WGD homolog as a surrogate. In practice the more taxa available (pre and post-WGD species) the better; esmaller estimator variances.
Figure F.1: Illustrative unrooted tree including post-WGD duplicates and pre-WGD homologous sequences for an imaginary C.0.0 pillar where there has been no gene loss. Pairwise distances between paralogs, dashed lines, can be left out from WLS branch length inference analysis. Other pairwise distances between post-WGD orthologs and pre-WGD homologs provide enough data to estimate the length of branches involved.

It is important to notice that, this method can only detect concerted evolution that occurred after the split of post-WGD included in the dataset. Any other previous concert is mathematically equivalent to a shorter internal branch. For instance, if we wanted to detect occurrence of concerted evolution in the ancestor of C. glabrata and S. cerevisiae, we should be using only one of these two species together with S. castellii homologs. In this study, as we used all homologs available, we may have missed some cases of concerted evolution that did not last until the split of C. glabrata and S. cerevisiae and definitely any that took place before the speciation of S. castellii.
F.2 Real control dataset results using GENECOV

<table>
<thead>
<tr>
<th>Data</th>
<th>Total cases</th>
<th>Protein positives</th>
<th>Protein percent.</th>
<th>Protein p-value</th>
<th>Nucleotide positives</th>
<th>Nucleotide percent.</th>
<th>Nucleotide p-value</th>
<th>Silent (Nucl.) positives</th>
<th>Silent (Nucl.) percent.</th>
<th>Silent (Nucl.) p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alignments</td>
<td>441</td>
<td>73</td>
<td>16.6%</td>
<td>1</td>
<td>99</td>
<td>22.4%</td>
<td>1</td>
<td>20</td>
<td>4.5%</td>
<td>0.379</td>
</tr>
<tr>
<td>Pair of species</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. gossypii vs C. glabrata</td>
<td>300</td>
<td>14</td>
<td>4.7%</td>
<td>0.463</td>
<td>28</td>
<td>9.3%</td>
<td>0.999</td>
<td>15</td>
<td>5%</td>
<td>0.432</td>
</tr>
<tr>
<td>A. gossypii vs K. lactis</td>
<td>334</td>
<td>13</td>
<td>3.9%</td>
<td>0.215</td>
<td>30</td>
<td>9%</td>
<td>0.999</td>
<td>11</td>
<td>3.1%</td>
<td>0.09</td>
</tr>
<tr>
<td>A. gossypii vs K. waltii</td>
<td>346</td>
<td>33</td>
<td>9.5%</td>
<td>1</td>
<td>26</td>
<td>7.5%</td>
<td>0.984</td>
<td>8</td>
<td>2.3%</td>
<td>0.00923</td>
</tr>
<tr>
<td>A. gossypii vs S. bayanus</td>
<td>318</td>
<td>11</td>
<td>3.5%</td>
<td>0.126</td>
<td>30</td>
<td>9.4%</td>
<td>1</td>
<td>4</td>
<td>1.6%</td>
<td>3.41 \times 10^{-4}</td>
</tr>
<tr>
<td>A. gossypii vs S. castellii</td>
<td>301</td>
<td>18</td>
<td>6%</td>
<td>0.821</td>
<td>25</td>
<td>8.3%</td>
<td>0.995</td>
<td>6</td>
<td>2%</td>
<td>0.00637</td>
</tr>
<tr>
<td>A. gossypii vs S. cerevisiae</td>
<td>348</td>
<td>14</td>
<td>4%</td>
<td>0.244</td>
<td>23</td>
<td>6.6%</td>
<td>0.928</td>
<td>10</td>
<td>2.9%</td>
<td>0.0371</td>
</tr>
<tr>
<td>A. gossypii vs S. kluyveri</td>
<td>272</td>
<td>20</td>
<td>7.4%</td>
<td>0.664</td>
<td>17</td>
<td>6.3%</td>
<td>0.861</td>
<td>8</td>
<td>2.9%</td>
<td>0.0702</td>
</tr>
<tr>
<td>C. glabrata vs K. lactis</td>
<td>271</td>
<td>18</td>
<td>6.6%</td>
<td>0.911</td>
<td>30</td>
<td>11.1%</td>
<td>1</td>
<td>7</td>
<td>2.6%</td>
<td>0.0369</td>
</tr>
<tr>
<td>C. glabrata vs K. waltii</td>
<td>288</td>
<td>23</td>
<td>8%</td>
<td>0.983</td>
<td>23</td>
<td>8%</td>
<td>0.99</td>
<td>12</td>
<td>4.2%</td>
<td>0.316</td>
</tr>
<tr>
<td>C. glabrata vs S. bayanus</td>
<td>258</td>
<td>28</td>
<td>11%</td>
<td>1</td>
<td>37</td>
<td>14.3%</td>
<td>1</td>
<td>11</td>
<td>4.3%</td>
<td>0.358</td>
</tr>
<tr>
<td>C. glabrata vs S. castellii</td>
<td>247</td>
<td>24</td>
<td>9.7%</td>
<td>0.999</td>
<td>28</td>
<td>11.3%</td>
<td>1</td>
<td>11</td>
<td>4.5%</td>
<td>0.418</td>
</tr>
<tr>
<td>C. glabrata vs S. cerevisiae</td>
<td>283</td>
<td>31</td>
<td>11%</td>
<td>1</td>
<td>34</td>
<td>12%</td>
<td>1</td>
<td>6</td>
<td>2.1%</td>
<td>0.0114</td>
</tr>
<tr>
<td>C. glabrata vs S. kluyveri</td>
<td>218</td>
<td>16</td>
<td>7.3%</td>
<td>0.952</td>
<td>28</td>
<td>12.9%</td>
<td>1</td>
<td>6</td>
<td>2.8%</td>
<td>0.0775</td>
</tr>
<tr>
<td>K. lactis vs K. waltii</td>
<td>311</td>
<td>24</td>
<td>7.7%</td>
<td>0.986</td>
<td>31</td>
<td>10%</td>
<td>1</td>
<td>12</td>
<td>3.9%</td>
<td>0.218</td>
</tr>
<tr>
<td>K. lactis vs S. bayanus</td>
<td>288</td>
<td>22</td>
<td>7.6%</td>
<td>0.981</td>
<td>31</td>
<td>12.8%</td>
<td>1</td>
<td>10</td>
<td>3.5%</td>
<td>0.144</td>
</tr>
<tr>
<td>K. lactis vs S. castellii</td>
<td>274</td>
<td>23</td>
<td>8.4%</td>
<td>0.994</td>
<td>20</td>
<td>7.3%</td>
<td>0.964</td>
<td>9</td>
<td>3.3%</td>
<td>0.118</td>
</tr>
<tr>
<td>K. lactis vs S. cerevisiae</td>
<td>310</td>
<td>27</td>
<td>8.7%</td>
<td>0.998</td>
<td>29</td>
<td>9.4%</td>
<td>1</td>
<td>5</td>
<td>1.6%</td>
<td>0.00163</td>
</tr>
<tr>
<td>Pair of species</td>
<td>total cases</td>
<td>Protein positives</td>
<td>percent.</td>
<td>p-value</td>
<td>Nucleotide positives</td>
<td>percent.</td>
<td>p-value</td>
<td>Silent (Nucl.) positives</td>
<td>percent.</td>
<td>p-value</td>
</tr>
<tr>
<td>----------------</td>
<td>-------------</td>
<td>-------------------</td>
<td>----------</td>
<td>---------</td>
<td>----------------------</td>
<td>----------</td>
<td>---------</td>
<td>--------------------------</td>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td><em>K. lactis</em> vs <em>S. kluyveri</em></td>
<td>240</td>
<td>16</td>
<td>6.7%</td>
<td>0.904</td>
<td>27</td>
<td>11.3%</td>
<td>1</td>
<td>12</td>
<td>5%</td>
<td>0.576</td>
</tr>
<tr>
<td><em>K. waltii</em> vs <em>S. bayanus</em></td>
<td>295</td>
<td>23</td>
<td>7.8%</td>
<td>0.986</td>
<td>31</td>
<td>10.8%</td>
<td>1</td>
<td>14</td>
<td>4.7%</td>
<td>0.489</td>
</tr>
<tr>
<td><em>K. waltii</em> vs <em>S. castellii</em></td>
<td>279</td>
<td>30</td>
<td>10.8%</td>
<td>1</td>
<td>29</td>
<td>10.4%</td>
<td>1</td>
<td>8</td>
<td>2.9%</td>
<td>0.059</td>
</tr>
<tr>
<td><em>K. waltii</em> vs <em>S. cerevisiae</em></td>
<td>326</td>
<td>19</td>
<td>5.8%</td>
<td>0.796</td>
<td>25</td>
<td>7.7%</td>
<td>0.986</td>
<td>11</td>
<td>3.4%</td>
<td>0.107</td>
</tr>
<tr>
<td><em>K. waltii</em> vs <em>S. kluyveri</em></td>
<td>256</td>
<td>28</td>
<td>10.9%</td>
<td>1</td>
<td>26</td>
<td>10.2%</td>
<td>1</td>
<td>5</td>
<td>2%</td>
<td>0.0107</td>
</tr>
<tr>
<td><em>S. bayanus</em> vs <em>S. castellii</em></td>
<td>258</td>
<td>26</td>
<td>10.1%</td>
<td>1</td>
<td>33</td>
<td>12.8%</td>
<td>1</td>
<td>15</td>
<td>5.8%</td>
<td>0.777</td>
</tr>
<tr>
<td><em>S. bayanus</em> vs <em>S. cerevisiae</em></td>
<td>305</td>
<td>25</td>
<td>8.2%</td>
<td>0.994</td>
<td>23</td>
<td>7.5%</td>
<td>0.98</td>
<td>10</td>
<td>3.3%</td>
<td>0.101</td>
</tr>
<tr>
<td><em>S. bayanus</em> vs <em>S. kluyveri</em></td>
<td>237</td>
<td>22</td>
<td>9.3%</td>
<td>0.998</td>
<td>28</td>
<td>11.8%</td>
<td>1</td>
<td>9</td>
<td>3.8%</td>
<td>0.249</td>
</tr>
<tr>
<td><em>S. castellii</em> vs <em>S. cerevisiae</em></td>
<td>281</td>
<td>30</td>
<td>10.7%</td>
<td>1</td>
<td>39</td>
<td>13.9%</td>
<td>1</td>
<td>8</td>
<td>2.8%</td>
<td>0.056</td>
</tr>
<tr>
<td><em>S. castellii</em> vs <em>S. kluyveri</em></td>
<td>215</td>
<td>13</td>
<td>6%</td>
<td>0.809</td>
<td>18</td>
<td>8.4%</td>
<td>0.988</td>
<td>7</td>
<td>3.3%</td>
<td>0.153</td>
</tr>
<tr>
<td><em>S. cerevisiae</em> vs <em>S. kluyveri</em></td>
<td>254</td>
<td>27</td>
<td>10.6%</td>
<td>1</td>
<td>30</td>
<td>11.8%</td>
<td>1</td>
<td>8</td>
<td>3.1%</td>
<td>0.108</td>
</tr>
</tbody>
</table>

Table F.1: Summary of the performance of Sawyer’s approach on the control dataset. Columns include the total number of tests and for each data-type false positives counts (absolute and percentage) and significance of the difference with the expected (5%). The latter datum is calculated as the two-tail cumulative probability based on a binomial distribution where the rate is set to 0.05. First data row summarises results based on global corrected p-values. Additionally, there is one row for each possible pair of species, indicating marginal pairwise results. The last row is the total based on all pairwise comparisons. These figures argue that using protein or all nucleotide polymorphisms is rather a liberal choice whereas only silent polymorphism is conservative for pairwise comparisons.
### F.3 Numeric results among yeast WGD paralogs

<table>
<thead>
<tr>
<th>Species</th>
<th>C0.0 pos.</th>
<th>C0.0 p-value</th>
<th>C1.0 pos.</th>
<th>C1.0 p-value</th>
<th>C2.0 pos.</th>
<th>C2.0 p-value</th>
<th>C1.1 pos.</th>
<th>C1.1 p-value</th>
<th>Total pos.</th>
<th>Total p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em></td>
<td>5.6%</td>
<td>0.266</td>
<td>6.2%</td>
<td>0.204</td>
<td>0%</td>
<td>0.46</td>
<td>5.3%</td>
<td>0.296</td>
<td>5.6%</td>
<td>0.244</td>
</tr>
<tr>
<td><em>S. castellii</em></td>
<td>4.5%</td>
<td>0.581</td>
<td>7%</td>
<td>0.118</td>
<td>6.7%</td>
<td>0.171</td>
<td>3.5%</td>
<td>0.62</td>
<td>5%</td>
<td>0.449</td>
</tr>
<tr>
<td><em>C. glabrata</em></td>
<td>4.4%</td>
<td>0.598</td>
<td>3.9%</td>
<td>0.473</td>
<td>0%</td>
<td>0.37</td>
<td>6.9%</td>
<td>0.175</td>
<td>4.4%</td>
<td>0.628</td>
</tr>
<tr>
<td>Averaged</td>
<td>4.8%</td>
<td>0.5</td>
<td>6.1%</td>
<td>0.158</td>
<td>2.8%</td>
<td>0.543</td>
<td>4.6%</td>
<td>0.492</td>
<td>5%</td>
<td>0.438</td>
</tr>
<tr>
<td>All species</td>
<td>0.4%</td>
<td>3.6 $\cdot 10^{-4}$</td>
<td>0.9%</td>
<td>0.418</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td></td>
</tr>
</tbody>
</table>

**GENECONV - silent polymorphism only (type I error <5%, critical p-value= 0.05)**

<table>
<thead>
<tr>
<th>Species</th>
<th>C0.0 pos.</th>
<th>C0.0 p-value</th>
<th>C1.0 pos.</th>
<th>C1.0 p-value</th>
<th>C2.0 pos.</th>
<th>C2.0 p-value</th>
<th>C1.1 pos.</th>
<th>C1.1 p-value</th>
<th>Total pos.</th>
<th>Total p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em></td>
<td>9.2%</td>
<td>0.00177</td>
<td>11.5%</td>
<td>8.14 $\cdot 10^{-4}$</td>
<td>16.7%</td>
<td>0.0196</td>
<td>10.5%</td>
<td>0.0397</td>
<td>10.3%</td>
<td>3.08 $\cdot 10^{-6}$</td>
</tr>
<tr>
<td><em>S. castellii</em></td>
<td>10.5%</td>
<td>1.32 $\cdot 10^{-4}$</td>
<td>10.5%</td>
<td>0.00468</td>
<td>20%</td>
<td>0.00547</td>
<td>4.7%</td>
<td>0.421</td>
<td>9.8%</td>
<td>8.98 $\cdot 10^{-6}$</td>
</tr>
<tr>
<td><em>C. glabrata</em></td>
<td>9.2%</td>
<td>0.00186</td>
<td>3.9%</td>
<td>0.473</td>
<td>0%</td>
<td>0.37</td>
<td>10.3%</td>
<td>0.0548</td>
<td>8.26%</td>
<td>0.00382</td>
</tr>
<tr>
<td>Averaged</td>
<td>9.7%</td>
<td>6.14 $\cdot 10^{-8}$</td>
<td>9.8%</td>
<td>2.11 $\cdot 10^{-4}$</td>
<td>13.9%</td>
<td>0.00834</td>
<td>7.2%</td>
<td>0.0798</td>
<td>9.5%</td>
<td>2.48 $\cdot 10^{-11}$</td>
</tr>
<tr>
<td>All species</td>
<td>0.8%</td>
<td>0.00142</td>
<td>1.4%</td>
<td>0.160</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td></td>
</tr>
</tbody>
</table>

**GENECONV - all nucleotide polymorphism (type I error ≈ 5%, critical p-value = 0.02)**

<table>
<thead>
<tr>
<th>Species</th>
<th>C0.0 pos.</th>
<th>C0.0 p-value</th>
<th>C1.0 pos.</th>
<th>C1.0 p-value</th>
<th>C2.0 pos.</th>
<th>C2.0 p-value</th>
<th>C1.1 pos.</th>
<th>C1.1 p-value</th>
<th>Total pos.</th>
<th>Total p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em></td>
<td>5.2%</td>
<td>0.365</td>
<td>20.8%</td>
<td>5.38 $\cdot 10^{-11}$</td>
<td>25%</td>
<td>0.00224</td>
<td>10.5%</td>
<td>0.0397</td>
<td>11%</td>
<td>2.32 $\cdot 10^{-7}$</td>
</tr>
<tr>
<td><em>S. castellii</em></td>
<td>4.5%</td>
<td>0.582</td>
<td>21.9%</td>
<td>6.69 $\cdot 10^{-11}$</td>
<td>40%</td>
<td>3.52 $\cdot 10^{-6}$</td>
<td>11%</td>
<td>0.0107</td>
<td>11%</td>
<td>6.24 $\cdot 10^{-8}$</td>
</tr>
<tr>
<td><em>C. glabrata</em></td>
<td>6%</td>
<td>0.188</td>
<td>3.9%</td>
<td>0.473</td>
<td>11%</td>
<td>0.0712</td>
<td>6.9%</td>
<td>0.163</td>
<td>5.9%</td>
<td>0.182</td>
</tr>
<tr>
<td>Averaged</td>
<td>5.2%</td>
<td>0.348</td>
<td>18.3%</td>
<td>0</td>
<td>27.8%</td>
<td>9.12 $\cdot 10^{-7}$</td>
<td>9.9%</td>
<td>0.0041</td>
<td>9.6%</td>
<td>1.21 $\cdot 10^{-11}$</td>
</tr>
<tr>
<td>All species</td>
<td>2%</td>
<td>9.99 $\cdot 10^{-15}$</td>
<td>11%</td>
<td>6.24 $\cdot 10^{-6}$</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td></td>
</tr>
<tr>
<td>Species</td>
<td>C0.0</td>
<td></td>
<td>C1.0</td>
<td></td>
<td>C2.0</td>
<td></td>
<td>C1.1</td>
<td></td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
<td>---</td>
<td>------</td>
<td>---</td>
<td>------</td>
<td>---</td>
<td>------</td>
<td>---</td>
<td>-------</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
</tr>
<tr>
<td>ML nucleotide trees (Type I error = 5%, critical bootstrap = 70%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>4.8%</td>
<td>0.477</td>
<td>24.6%</td>
<td>7.33 · 10(^{-15})</td>
<td>25%</td>
<td>0.00224</td>
<td>10.5%</td>
<td>0.0397</td>
<td>11.9%</td>
<td>5.38 · 10(^{-9})</td>
</tr>
<tr>
<td>S. castellii</td>
<td>4.5%</td>
<td>0.582</td>
<td>25.4%</td>
<td>4.22 · 10(^{-14})</td>
<td>20%</td>
<td>0.00547</td>
<td>10.5%</td>
<td>0.0107</td>
<td>11.3%</td>
<td>2.51 · 10(^{-8})</td>
</tr>
<tr>
<td>C. glabrata</td>
<td>5.6%</td>
<td>0.271</td>
<td>3.9%</td>
<td>0.473</td>
<td>22.2%</td>
<td>0.00836</td>
<td>3.6%</td>
<td>0.412</td>
<td>5.6%</td>
<td>0.252</td>
</tr>
<tr>
<td>Averaged</td>
<td>5%</td>
<td>0.477</td>
<td>21.4%</td>
<td>0</td>
<td>22.2%</td>
<td>5.35 · 10(^{-5})</td>
<td>9.2%</td>
<td>0.00952</td>
<td>9.9%</td>
<td>6.23 · 10(^{-13})</td>
</tr>
<tr>
<td>All species</td>
<td>2.8%</td>
<td>0</td>
<td>17.1%</td>
<td>1.35 · 10(^{-9})</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
</tbody>
</table>

Distance method based on estimator comparison (Type I error < 5%, critical p-value\(_{DP} = 0.05\))

<table>
<thead>
<tr>
<th>Species</th>
<th>C0.0</th>
<th></th>
<th>C1.0</th>
<th></th>
<th>C2.0</th>
<th></th>
<th>C1.1</th>
<th></th>
<th>Total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>7.2%</td>
<td>0.0458</td>
<td>22.3%</td>
<td>1.73 · 10(^{-12})</td>
<td>33.3%</td>
<td>1.84 · 10(^{-4})</td>
<td>7.9%</td>
<td>0.12</td>
<td>12.6%</td>
<td>2.56 · 10(^{-10})</td>
</tr>
<tr>
<td>S. castellii</td>
<td>6.5%</td>
<td>0.115</td>
<td>23.7%</td>
<td>1.84 · 10(^{-12})</td>
<td>33.3%</td>
<td>5.28 · 10(^{-5})</td>
<td>10.5%</td>
<td>0.0107</td>
<td>12.3%</td>
<td>1.94 · 10(^{-10})</td>
</tr>
<tr>
<td>C. glabrata</td>
<td>8%</td>
<td>0.0149</td>
<td>3.9%</td>
<td>0.473</td>
<td>11%</td>
<td>0.0712</td>
<td>7.1%</td>
<td>0.163</td>
<td>7.4%</td>
<td>0.0208</td>
</tr>
<tr>
<td>Averaged</td>
<td>7.2%</td>
<td>0.00778</td>
<td>19.6%</td>
<td>&quot;</td>
<td>27.8%</td>
<td>9.12 · 10(^{-7})</td>
<td>9.2%</td>
<td>0.00952</td>
<td>11.1%</td>
<td>0</td>
</tr>
<tr>
<td>All species</td>
<td>3.6%</td>
<td>1.11 · 10(^{-15})</td>
<td>11.6%</td>
<td>6.0 · 10(^{-6})</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
</tbody>
</table>

Distance method based on estimator difference (Type I error ≈ 5%, critical p-value\(_{DP} = 0.02\))

<table>
<thead>
<tr>
<th>Species</th>
<th>C0.0</th>
<th></th>
<th>C1.0</th>
<th></th>
<th>C2.0</th>
<th></th>
<th>C1.1</th>
<th></th>
<th>Total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
<td>pos.</td>
<td>p-value</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>10.8%</td>
<td>6.15 · 10(^{-5})</td>
<td>25.4%</td>
<td>1.11 · 10(^{-15})</td>
<td>33.3%</td>
<td>1.84 · 10(^{-4})</td>
<td>7.9%</td>
<td>0.12</td>
<td>15.6%</td>
<td>8.47 · 10(^{-6})</td>
</tr>
<tr>
<td>S. castellii</td>
<td>11.3%</td>
<td>2.07 · 10(^{-5})</td>
<td>29.8%</td>
<td>0</td>
<td>46.7%</td>
<td>1.83 · 10(^{-7})</td>
<td>15.1%</td>
<td>9.15 · 10(^{-5})</td>
<td>17.7%</td>
<td>1.49 · 10(^{-6})</td>
</tr>
<tr>
<td>C. glabrata</td>
<td>12%</td>
<td>3.61 · 10(^{-6})</td>
<td>5.9%</td>
<td>0.253</td>
<td>11.1%</td>
<td>0.0712</td>
<td>7.1%</td>
<td>0.163</td>
<td>10.7%</td>
<td>3.26 · 10(^{-4})</td>
</tr>
<tr>
<td>Averaged</td>
<td>11.4%</td>
<td>1.22 · 10(^{-12})</td>
<td>23.7%</td>
<td>0</td>
<td>33.3%</td>
<td>9.48 · 10(^{-9})</td>
<td>11.8%</td>
<td>2.29 · 10(^{-4})</td>
<td>15.1%</td>
<td>0</td>
</tr>
<tr>
<td>All species</td>
<td>3.3%</td>
<td>1.89 · 10(^{-10})</td>
<td>14.8%</td>
<td>7.62 · 10(^{-8})</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
<td>&quot;</td>
</tr>
</tbody>
</table>

Table F.2: Summary of numerical results of concerted evolution detection methods applied on nucleotide alignments
The following tables gather FatiGO+ individual statistical test results for enrichment of GO terms for each yeast species studied and all simultaneously as indicated in captions. We have conveniently abbreviated ontology descriptions. Full descriptions can be found in public GO web resources. We only show terms with p-values equal or lower than 0.01.

### Table F.3: Full interspecies GO term enrichment analysis

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Convergent % (num.)</th>
<th>Non-Conv. % (num.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Molecular Function</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0003735</td>
<td>structural constituent of ribosome</td>
<td>37.04 (40)</td>
<td>1.46 (7)</td>
<td>$9.5 \cdot 10^{-26}$</td>
</tr>
<tr>
<td>GO:0005198</td>
<td>structural molecule activity</td>
<td>37.96 (41)</td>
<td>3.97 (19)</td>
<td>$8.1 \cdot 10^{-20}$</td>
</tr>
<tr>
<td>GO:0008135</td>
<td>transl. factor act., nucl. acid bind.</td>
<td>4.63 (5)</td>
<td>0.21 (1)</td>
<td>0.001</td>
</tr>
<tr>
<td>GO:0045182</td>
<td>translation regulator activity</td>
<td>4.63 (5)</td>
<td>0.21 (1)</td>
<td>0.001</td>
</tr>
<tr>
<td>GO:0004672</td>
<td>protein kinase activity</td>
<td>0 (0)</td>
<td>5.85 (28)</td>
<td>0.0048</td>
</tr>
<tr>
<td>GO:0016772</td>
<td>transf. on phosphorus-cont. grps.</td>
<td>2.78 (3)</td>
<td>10.44 (50)</td>
<td>0.0089</td>
</tr>
<tr>
<td></td>
<td>Biological Process</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0006412</td>
<td>translation</td>
<td>43.52 (47)</td>
<td>3.55 (17)</td>
<td>$2.4 \cdot 10^{-25}$</td>
</tr>
<tr>
<td>GO:0009059</td>
<td>macromolecule biosynthetic proc.</td>
<td>50.93 (55)</td>
<td>7.72 (37)</td>
<td>$3.1 \cdot 10^{-23}$</td>
</tr>
<tr>
<td>GO:0009058</td>
<td>biosynthetic process</td>
<td>60.19 (65)</td>
<td>15.03 (72)</td>
<td>$1.3 \cdot 10^{-20}$</td>
</tr>
<tr>
<td>GO:0044260</td>
<td>cellular macromolecule meta. proc.</td>
<td>54.63 (59)</td>
<td>21.50 (103)</td>
<td>$3.4 \cdot 10^{-11}$</td>
</tr>
<tr>
<td>GO:0044267</td>
<td>cellular protein metabolic proc.</td>
<td>50.93 (55)</td>
<td>19.62 (94)</td>
<td>$2.3 \cdot 10^{-10}$</td>
</tr>
<tr>
<td>GO:0019538</td>
<td>protein metabolic process</td>
<td>50.93 (55)</td>
<td>20.25 (97)</td>
<td>$4.3 \cdot 10^{-10}$</td>
</tr>
<tr>
<td>GO:0044238</td>
<td>primary metabolic process</td>
<td>77.78 (84)</td>
<td>55.11 (264)</td>
<td>$1.2 \cdot 10^{-05}$</td>
</tr>
<tr>
<td>GO:0043170</td>
<td>macromolecule metabolic proc.</td>
<td>67.59 (73)</td>
<td>44.47 (213)</td>
<td>$1.7 \cdot 10^{-05}$</td>
</tr>
<tr>
<td>GO:0008152</td>
<td>metabolic process</td>
<td>79.63 (86)</td>
<td>58.25 (279)</td>
<td>$2.6 \cdot 10^{-05}$</td>
</tr>
<tr>
<td>GO:0044237</td>
<td>cellular metabolic process</td>
<td>77.78 (84)</td>
<td>56.58 (271)</td>
<td>$3.3 \cdot 10^{-05}$</td>
</tr>
<tr>
<td>GO:0022402</td>
<td>cell cycle process</td>
<td>1.85 (2)</td>
<td>13.99 (67)</td>
<td>$9.6 \cdot 10^{-05}$</td>
</tr>
<tr>
<td>GO:0007049</td>
<td>cell cycle</td>
<td>2.78 (3)</td>
<td>14.82 (71)</td>
<td>0.00019</td>
</tr>
<tr>
<td>GO:0065007</td>
<td>biological regulation</td>
<td>15.74 (17)</td>
<td>31.52 (151)</td>
<td>0.0009</td>
</tr>
<tr>
<td>GO:0022403</td>
<td>cell cycle phase</td>
<td>1.85 (2)</td>
<td>11.48 (55)</td>
<td>0.00097</td>
</tr>
<tr>
<td>GO:0022618</td>
<td>protein-RNA complex assembly</td>
<td>6.48 (7)</td>
<td>0.84 (4)</td>
<td>0.001</td>
</tr>
<tr>
<td>GO:0042274</td>
<td>ribo. small sub. biogene. &amp; assem.</td>
<td>3.70 (4)</td>
<td>0 (0)</td>
<td>0.0011</td>
</tr>
<tr>
<td>GO:0006259</td>
<td>DNA metabolic process</td>
<td>1.85 (2)</td>
<td>11.66 (53)</td>
<td>0.0015</td>
</tr>
<tr>
<td>GO:0000723</td>
<td>telomere maintenance</td>
<td>16.67 (18)</td>
<td>6.89 (33)</td>
<td>0.0023</td>
</tr>
<tr>
<td>GO:0032200</td>
<td>telomere organization &amp; biogene.</td>
<td>16.67 (18)</td>
<td>6.89 (33)</td>
<td>0.0023</td>
</tr>
</tbody>
</table>
Table F.3: Full interspecies GO term enrichment analysis

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Convergent % (num.)</th>
<th>Non-Conv. % (num.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0016568</td>
<td>chromatin modification</td>
<td>0 (0)</td>
<td>6.26 (30)</td>
<td>0.0029</td>
</tr>
<tr>
<td>GO:0007165</td>
<td>signal transduction</td>
<td>0.93 (1)</td>
<td>8.35 (40)</td>
<td>0.003</td>
</tr>
<tr>
<td>GO:0000278</td>
<td>mitotic cell cycle</td>
<td>1.85 (2)</td>
<td>10.02 (48)</td>
<td>0.0036</td>
</tr>
<tr>
<td>GO:0050794</td>
<td>regulation of cellular process</td>
<td>12.96 (14)</td>
<td>25.89 (124)</td>
<td>0.0037</td>
</tr>
<tr>
<td>GO:0050789</td>
<td>regulation of biological process</td>
<td>13.89 (15)</td>
<td>26.93 (129)</td>
<td>0.0042</td>
</tr>
<tr>
<td>GO:0042257</td>
<td>ribosomal subunit assembly</td>
<td>3.70 (4)</td>
<td>0.21 (1)</td>
<td>0.0047</td>
</tr>
<tr>
<td>GO:0042255</td>
<td>ribosome assembly</td>
<td>3.70 (4)</td>
<td>0.21 (1)</td>
<td>0.0047</td>
</tr>
<tr>
<td>GO:0006950</td>
<td>response to stress</td>
<td>4.63 (5)</td>
<td>14.20 (68)</td>
<td>0.0055</td>
</tr>
<tr>
<td>GO:0007154</td>
<td>cell communication</td>
<td>1.85 (2)</td>
<td>9.60 (46)</td>
<td>0.0056</td>
</tr>
<tr>
<td>GO:0022613</td>
<td>ribonuc. prot. complx. biog. &amp; asm.</td>
<td>11.11 (12)</td>
<td>3.76 (18)</td>
<td>0.0057</td>
</tr>
<tr>
<td>GO:0006414</td>
<td>translational elongation</td>
<td>2.78 (3)</td>
<td>0 (0)</td>
<td>0.0061</td>
</tr>
<tr>
<td>GO:0000028</td>
<td>ribo. small sub. asm. and mainten.</td>
<td>2.78 (3)</td>
<td>0 (0)</td>
<td>0.0061</td>
</tr>
<tr>
<td>GO:0046165</td>
<td>alcohol biosynthetic process</td>
<td>4.63 (5)</td>
<td>0.63 (3)</td>
<td>0.0068</td>
</tr>
<tr>
<td>GO:0019219</td>
<td>regulation of nucl. metabolic proc.</td>
<td>5.56 (6)</td>
<td>14.82 (71)</td>
<td>0.0074</td>
</tr>
<tr>
<td>GO:0016310</td>
<td>phosphorylation</td>
<td>0 (0)</td>
<td>5.43 (26)</td>
<td>0.0078</td>
</tr>
<tr>
<td>GO:0009719</td>
<td>response to endogenous stimulus</td>
<td>0 (0)</td>
<td>5.64 (27)</td>
<td>0.0083</td>
</tr>
<tr>
<td>GO:0043687</td>
<td>post-transl. protein modification</td>
<td>1.85 (2)</td>
<td>8.98 (43)</td>
<td>0.0085</td>
</tr>
<tr>
<td>GO:0016051</td>
<td>carbohydrate biosynthetic proc.</td>
<td>7.41 (8)</td>
<td>2.09 (10)</td>
<td>0.0088</td>
</tr>
<tr>
<td>GO:0006464</td>
<td>protein modification process</td>
<td>2.78 (3)</td>
<td>10.65 (51)</td>
<td>0.0089</td>
</tr>
<tr>
<td>GO:0044249</td>
<td>cellular biosynthetic process</td>
<td>16.67 (18)</td>
<td>7.72 (37)</td>
<td>0.0093</td>
</tr>
</tbody>
</table>

Cellular Component

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>% (num.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0044445</td>
<td>cytosolic part</td>
<td>38.89 (42)</td>
<td>2.5 \cdot 10^{-25}</td>
</tr>
<tr>
<td>GO:0005830</td>
<td>cytosolic ribo. (sensu Eukaryota)</td>
<td>37.96 (41)</td>
<td>3.7 \cdot 10^{-25}</td>
</tr>
<tr>
<td>GO:0005840</td>
<td>ribosome</td>
<td>44.44 (48)</td>
<td>2.6 \cdot 10^{-23}</td>
</tr>
<tr>
<td>GO:0030529</td>
<td>ribonucprot. cmplx.</td>
<td>45.37 (49)</td>
<td>4.9 \cdot 10^{-22}</td>
</tr>
<tr>
<td>GO:0005829</td>
<td>cytosol</td>
<td>42.59 (46)</td>
<td>1.6 \cdot 10^{-17}</td>
</tr>
<tr>
<td>GO:0015934</td>
<td>large ribosomal subunit</td>
<td>21.30 (23)</td>
<td>1.4 \cdot 10^{-14}</td>
</tr>
<tr>
<td>GO:0005842</td>
<td>cytosolic large ribo. sub.</td>
<td>21.30 (23)</td>
<td>1.4 \cdot 10^{-14}</td>
</tr>
<tr>
<td>GO:0016282</td>
<td>eukaryotic 43S preinit. cmplx.</td>
<td>17.59 (19)</td>
<td>1.2 \cdot 10^{-11}</td>
</tr>
<tr>
<td>GO:0043228</td>
<td>non-membrane-bound organelle</td>
<td>48.15 (52)</td>
<td>1.4 \cdot 10^{-11}</td>
</tr>
<tr>
<td>GO:0043232</td>
<td>intracell. non-membr-bound. organ.</td>
<td>48.15 (52)</td>
<td>1.4 \cdot 10^{-11}</td>
</tr>
<tr>
<td>GO:0015935</td>
<td>small ribosomal subunit</td>
<td>16.67 (18)</td>
<td>6.5 \cdot 10^{-11}</td>
</tr>
<tr>
<td>GO:0050843</td>
<td>cytosolic small ribo. sub.</td>
<td>16.67 (18)</td>
<td>6.5 \cdot 10^{-11}</td>
</tr>
</tbody>
</table>
Table F.3: Full interspecies GO term enrichment analysis

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Convergent % (num.)</th>
<th>Non-Conv. % (num.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0016283</td>
<td>eukaryotic 48S initiation complex</td>
<td>16.67 (18)</td>
<td>0.84 (4)</td>
<td>6.5 \cdot 10^{-11}</td>
</tr>
<tr>
<td>GO:0032991</td>
<td>macromolecular complex</td>
<td>52.78 (57)</td>
<td>22.76 (109)</td>
<td>2.7 \cdot 10^{-09}</td>
</tr>
<tr>
<td>GO:0043234</td>
<td>protein complex</td>
<td>48.15 (52)</td>
<td>19.62 (94)</td>
<td>5.3 \cdot 10^{-09}</td>
</tr>
<tr>
<td>GO:0044444</td>
<td>cytoplasmal part</td>
<td>76.85 (83)</td>
<td>49.27 (236)</td>
<td>1.2 \cdot 10^{-07}</td>
</tr>
<tr>
<td>GO:0043227</td>
<td>membrane-bound organelle</td>
<td>45.37 (49)</td>
<td>69.52 (333)</td>
<td>3.7 \cdot 10^{-06}</td>
</tr>
<tr>
<td>GO:0043231</td>
<td>intracell. membr-bound. organ.</td>
<td>45.37 (49)</td>
<td>69.52 (333)</td>
<td>3.7 \cdot 10^{-06}</td>
</tr>
<tr>
<td>GO:0044446</td>
<td>intracellular organelle part</td>
<td>54.63 (59)</td>
<td>36.12 (173)</td>
<td>0.00047</td>
</tr>
<tr>
<td>GO:0044422</td>
<td>organelle part</td>
<td>54.63 (59)</td>
<td>36.12 (173)</td>
<td>0.00047</td>
</tr>
<tr>
<td>GO:0005634</td>
<td>nucleus</td>
<td>20.37 (22)</td>
<td>38 (182)</td>
<td>0.00048</td>
</tr>
<tr>
<td>GO:0005737</td>
<td>cytoplasm</td>
<td>87.04 (94)</td>
<td>72.44 (347)</td>
<td>0.0013</td>
</tr>
<tr>
<td>GO:0044452</td>
<td>nucleolar part</td>
<td>3.70 (4)</td>
<td>0.21 (1)</td>
<td>0.0047</td>
</tr>
<tr>
<td>GO:0031967</td>
<td>organelle envelope</td>
<td>1.85 (2)</td>
<td>9.60 (46)</td>
<td>0.0056</td>
</tr>
<tr>
<td>GO:0031975</td>
<td>envelope</td>
<td>1.85 (2)</td>
<td>9.60 (46)</td>
<td>0.0056</td>
</tr>
<tr>
<td>GO:0005732</td>
<td>small nucleolar. ribonuclprot. cmplx.</td>
<td>2.78 (3)</td>
<td>0 (0)</td>
<td>0.0061</td>
</tr>
</tbody>
</table>

Table F.4: *Saccharomyces cerevisiae* GO term enrichment analysis

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Convergent % (num.)</th>
<th>Non-Conv. % (num.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular Function</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0003735</td>
<td>structural constituent of ribosome</td>
<td>43.28 (29)</td>
<td>3.59 (13)</td>
<td>6 \cdot 10^{-17}</td>
</tr>
<tr>
<td>GO:0005198</td>
<td>structural molecule activity</td>
<td>43.28 (29)</td>
<td>6.08 (22)</td>
<td>1.5 \cdot 10^{-13}</td>
</tr>
<tr>
<td>GO:0008135</td>
<td>transl. factor act., nucleic acid bind.</td>
<td>5.97 (4)</td>
<td>0.28 (1)</td>
<td>0.0024</td>
</tr>
<tr>
<td>GO:0045182</td>
<td>translation regulator activity</td>
<td>5.97 (4)</td>
<td>0.28 (1)</td>
<td>0.0024</td>
</tr>
<tr>
<td>Biological Process</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009059</td>
<td>macromolecule biosynthetic proc.</td>
<td>61.19 (41)</td>
<td>8.56 (31)</td>
<td>1.9 \cdot 10^{-20}</td>
</tr>
<tr>
<td>GO:0006412</td>
<td>translation</td>
<td>50.75 (34)</td>
<td>4.97 (18)</td>
<td>4.5 \cdot 10^{-19}</td>
</tr>
<tr>
<td>GO:0009058</td>
<td>biosynthetic process</td>
<td>68.66 (46)</td>
<td>16.30 (59)</td>
<td>2.9 \cdot 10^{-17}</td>
</tr>
<tr>
<td>GO:004260</td>
<td>cell. macromol. metabolic proc.</td>
<td>64.18 (43)</td>
<td>23.48 (85)</td>
<td>3.3 \cdot 10^{-10}</td>
</tr>
<tr>
<td>GO:004267</td>
<td>cellular protein metabolic proc.</td>
<td>56.72 (38)</td>
<td>21.82 (79)</td>
<td>2.6 \cdot 10^{-08}</td>
</tr>
<tr>
<td>GO:0019538</td>
<td>protein metabolic process</td>
<td>56.72 (38)</td>
<td>22.38 (81)</td>
<td>6.3 \cdot 10^{-08}</td>
</tr>
<tr>
<td>GO:0008152</td>
<td>metabolic process</td>
<td>86.57 (58)</td>
<td>59.67 (216)</td>
<td>1.3 \cdot 10^{-05}</td>
</tr>
<tr>
<td>GO:0044238</td>
<td>primary metabolic process</td>
<td>83.58 (56)</td>
<td>56.08 (203)</td>
<td>1.8 \cdot 10^{-05}</td>
</tr>
</tbody>
</table>

170
### Table F.4: *Saccharomyces cerevisiae* GO term enrichment analysis

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Convergent % (num.)</th>
<th>Non-Conv. % (num.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0043170</td>
<td>macromolecule metabolic proc.</td>
<td>73.13 (49)</td>
<td>45.30 (164)</td>
<td>3 \cdot 10^{-05}</td>
</tr>
<tr>
<td>GO:0044237</td>
<td>cellular metabolic process</td>
<td>83.58 (56)</td>
<td>57.73 (209)</td>
<td>3.5 \cdot 10^{-05}</td>
</tr>
<tr>
<td>GO:0022618</td>
<td>protein-RNA complex assembly</td>
<td>8.96 (6)</td>
<td>0.55 (2)</td>
<td>0.00026</td>
</tr>
<tr>
<td>GO:0007223</td>
<td>telomere maintenance</td>
<td>20.90 (14)</td>
<td>6.08 (22)</td>
<td>0.00033</td>
</tr>
<tr>
<td>GO:0032200</td>
<td>telomere organiz. and biogene.</td>
<td>20.90 (14)</td>
<td>6.08 (22)</td>
<td>0.00033</td>
</tr>
<tr>
<td>GO:0042274</td>
<td>ribosomal subunit assembly</td>
<td>5.97 (4)</td>
<td>0 (0)</td>
<td>0.00055</td>
</tr>
<tr>
<td>GO:0042255</td>
<td>ribosome assembly</td>
<td>5.97 (4)</td>
<td>0 (0)</td>
<td>0.00055</td>
</tr>
<tr>
<td>GO:0022613</td>
<td>ribonucplt. cmlx. biog. &amp; assm.</td>
<td>14.93 (10)</td>
<td>3.59 (13)</td>
<td>0.00096</td>
</tr>
<tr>
<td>GO:0006073</td>
<td>glucan metabolic process</td>
<td>8.96 (6)</td>
<td>1.10 (4)</td>
<td>0.0015</td>
</tr>
<tr>
<td>GO:006259</td>
<td>DNA metabolic process</td>
<td>0 (0)</td>
<td>10.77 (39)</td>
<td>0.0018</td>
</tr>
<tr>
<td>GO:0046165</td>
<td>alcohol biosynthetic process</td>
<td>5.97 (4)</td>
<td>0.28 (1)</td>
<td>0.0024</td>
</tr>
<tr>
<td>GO:006364</td>
<td>rRNA processing</td>
<td>5.97 (4)</td>
<td>0.28 (1)</td>
<td>0.0024</td>
</tr>
<tr>
<td>GO:0000028</td>
<td>ribo. small sub. assmb. &amp; maint.</td>
<td>4.48 (3)</td>
<td>0 (0)</td>
<td>0.0037</td>
</tr>
<tr>
<td>GO:0006059</td>
<td>gluconeogenesis</td>
<td>4.48 (3)</td>
<td>0 (0)</td>
<td>0.0037</td>
</tr>
<tr>
<td>GO:0046364</td>
<td>monosaccharide biosyt. proc.</td>
<td>4.48 (3)</td>
<td>0 (0)</td>
<td>0.0037</td>
</tr>
<tr>
<td>GO:0019319</td>
<td>hexose biosynthetic process</td>
<td>4.48 (3)</td>
<td>0 (0)</td>
<td>0.0037</td>
</tr>
<tr>
<td>GO:006355</td>
<td>regul. of transc., DNA-depend.</td>
<td>1.49 (1)</td>
<td>12.43 (45)</td>
<td>0.0045</td>
</tr>
<tr>
<td>GO:006357</td>
<td>regul. transc. RNA pmerase II</td>
<td>0 (0)</td>
<td>8.84 (32)</td>
<td>0.0049</td>
</tr>
<tr>
<td>GO:0042644</td>
<td>cellul. polysacch. metabolic proc.</td>
<td>8.96 (6)</td>
<td>1.66 (6)</td>
<td>0.005</td>
</tr>
<tr>
<td>GO:005976</td>
<td>polysacch. metabolic proc.</td>
<td>8.96 (6)</td>
<td>1.66 (6)</td>
<td>0.005</td>
</tr>
<tr>
<td>GO:0019219</td>
<td>regul. of nucls. metabolic proc.</td>
<td>2.99 (2)</td>
<td>14.92 (54)</td>
<td>0.0051</td>
</tr>
<tr>
<td>GO:0042254</td>
<td>ribosome biogene. and assmbl.</td>
<td>11.94 (8)</td>
<td>3.31 (12)</td>
<td>0.0062</td>
</tr>
<tr>
<td>GO:0045823</td>
<td>negative regul. of cellular proc.</td>
<td>0 (0)</td>
<td>8.29 (30)</td>
<td>0.0079</td>
</tr>
<tr>
<td>GO:0045819</td>
<td>negative regul. of biologic. proc.</td>
<td>0 (0)</td>
<td>8.29 (30)</td>
<td>0.0079</td>
</tr>
<tr>
<td>GO:0032774</td>
<td>RNA biosynthetic process</td>
<td>2.99 (2)</td>
<td>14.09 (51)</td>
<td>0.008</td>
</tr>
<tr>
<td>GO:006351</td>
<td>transcription, DNA-dependent</td>
<td>2.99 (2)</td>
<td>14.09 (51)</td>
<td>0.008</td>
</tr>
<tr>
<td>GO:0016501</td>
<td>carbohydrate biosynthetic proc.</td>
<td>8.96 (6)</td>
<td>1.93 (7)</td>
<td>0.0082</td>
</tr>
<tr>
<td>GO:0019580</td>
<td>ener. deriv. by oxid. of org. cmpds.</td>
<td>17.91 (12)</td>
<td>7.18 (26)</td>
<td>0.0087</td>
</tr>
</tbody>
</table>

#### Cellular Component

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Convergent % (num.)</th>
<th>Non-Conv. % (num.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0044445</td>
<td>cytosolic part</td>
<td>44.78 (30)</td>
<td>3.87 (14)</td>
<td>2.4 \cdot 10^{-17}</td>
</tr>
<tr>
<td>GO:0005830</td>
<td>cytosolic ribo. (sensu Eukaryota)</td>
<td>43.28 (29)</td>
<td>3.87 (14)</td>
<td>1.7 \cdot 10^{-16}</td>
</tr>
<tr>
<td>GO:0005840</td>
<td>ribosome</td>
<td>52.24 (35)</td>
<td>7.73 (28)</td>
<td>2.4 \cdot 10^{-16}</td>
</tr>
</tbody>
</table>
### Table F.4: *Saccharomyces cerevisiae* GO term enrichment analysis

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Convergent % (num.)</th>
<th>Non-Conv. % (num.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0030529</td>
<td>ribonucleoprotein complex</td>
<td>53.73 (36)</td>
<td>8.84 (32)</td>
<td>$6.1 \cdot 10^{-16}$</td>
</tr>
<tr>
<td>GO:0005829</td>
<td>cytosol</td>
<td>49.25 (33)</td>
<td>10.50 (38)</td>
<td>$3.4 \cdot 10^{-12}$</td>
</tr>
<tr>
<td>GO:0043228</td>
<td>non-membrane-bound organelle</td>
<td>56.72 (38)</td>
<td>17.68 (64)</td>
<td>$1.6 \cdot 10^{-10}$</td>
</tr>
<tr>
<td>GO:0043232</td>
<td>intracell. non-membr-bound. organ.</td>
<td>56.72 (38)</td>
<td>17.68 (64)</td>
<td>$1.6 \cdot 10^{-10}$</td>
</tr>
<tr>
<td>GO:0015934</td>
<td>large ribosomal subunit</td>
<td>23.88 (16)</td>
<td>1.93 (7)</td>
<td>$2.4 \cdot 10^{-09}$</td>
</tr>
<tr>
<td>GO:0005842</td>
<td>cytosolic large ribo. sub.</td>
<td>23.88 (16)</td>
<td>1.93 (7)</td>
<td>$2.4 \cdot 10^{-09}$</td>
</tr>
<tr>
<td>GO:0032991</td>
<td>macromolecular complex</td>
<td>58.21 (39)</td>
<td>23.76 (86)</td>
<td>$5.7 \cdot 10^{-08}$</td>
</tr>
<tr>
<td>GO:0016282</td>
<td>eukaryotic 43S preinit. complex</td>
<td>19.40 (13)</td>
<td>1.66 (6)</td>
<td>$1.4 \cdot 10^{-07}$</td>
</tr>
<tr>
<td>GO:0015935</td>
<td>small ribosomal subunit</td>
<td>19.40 (13)</td>
<td>1.66 (6)</td>
<td>$1.4 \cdot 10^{-07}$</td>
</tr>
<tr>
<td>GO:0005843</td>
<td>cyto. small ribo. sub.</td>
<td>19.40 (13)</td>
<td>1.66 (6)</td>
<td>$1.4 \cdot 10^{-07}$</td>
</tr>
<tr>
<td>GO:0016283</td>
<td>eukaryotic 48S init. complex</td>
<td>19.40 (13)</td>
<td>1.66 (6)</td>
<td>$1.4 \cdot 10^{-07}$</td>
</tr>
<tr>
<td>GO:0043234</td>
<td>protein complex</td>
<td>50.75 (34)</td>
<td>19.61 (71)</td>
<td>$3.3 \cdot 10^{-07}$</td>
</tr>
<tr>
<td>GO:0044444</td>
<td>cytoplasmic part</td>
<td>79.10 (53)</td>
<td>55.80 (202)</td>
<td>0.00037</td>
</tr>
<tr>
<td>GO:0044446</td>
<td>intracellular organelle part</td>
<td>56.72 (38)</td>
<td>34.53 (125)</td>
<td>0.00091</td>
</tr>
<tr>
<td>GO:0044422</td>
<td>organelle part</td>
<td>56.72 (38)</td>
<td>34.53 (125)</td>
<td>0.00091</td>
</tr>
<tr>
<td>GO:0043227</td>
<td>membr-bound organelle</td>
<td>47.76 (32)</td>
<td>68.51 (248)</td>
<td>0.0019</td>
</tr>
<tr>
<td>GO:0043231</td>
<td>intracell. membr-bound. organ.</td>
<td>47.76 (32)</td>
<td>68.51 (248)</td>
<td>0.0019</td>
</tr>
</tbody>
</table>

### Table F.5: *Saccharomyces castelli* GO term enrichment analysis

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Convergent % (num.)</th>
<th>Non-Conv. % (num.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0003735</td>
<td>structural constituent of ribosome</td>
<td>36.59 (30)</td>
<td>3.68 (14)</td>
<td>$3.5 \cdot 10^{-15}$</td>
</tr>
<tr>
<td>GO:0005198</td>
<td>structural molecule activity</td>
<td>36.59 (30)</td>
<td>5.79 (22)</td>
<td>$2.9 \cdot 10^{-12}$</td>
</tr>
<tr>
<td>GO:0005515</td>
<td>protein binding</td>
<td>2.44 (2)</td>
<td>11.84 (45)</td>
<td>0.008</td>
</tr>
</tbody>
</table>

#### Molecular Function

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Convergent % (num.)</th>
<th>Non-Conv. % (num.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006412</td>
<td>translation</td>
<td>42.68 (35)</td>
<td>6.05 (23)</td>
<td>$2.3 \cdot 10^{-15}$</td>
</tr>
<tr>
<td>GO:0009059</td>
<td>macromolecule biosynthetic proc.</td>
<td>48.78 (40)</td>
<td>10.53 (40)</td>
<td>7.4 $\cdot 10^{-14}$</td>
</tr>
<tr>
<td>GO:0009058</td>
<td>biosynthetic process</td>
<td>57.32 (47)</td>
<td>17.37 (66)</td>
<td>9.8 $\cdot 10^{-13}$</td>
</tr>
<tr>
<td>GO:0044267</td>
<td>cellular protein metabolic process</td>
<td>52.44 (43)</td>
<td>21.05 (80)</td>
<td>3.2 $\cdot 10^{-08}$</td>
</tr>
<tr>
<td>GO:0019538</td>
<td>protein metabolic process</td>
<td>52.44 (43)</td>
<td>21.84 (83)</td>
<td>9. $\cdot 10^{-08}$</td>
</tr>
<tr>
<td>GO:0044260</td>
<td>cellul. macromol. metabolic proc.</td>
<td>53.66 (44)</td>
<td>23.68 (90)</td>
<td>$2 \cdot 10^{-07}$</td>
</tr>
</tbody>
</table>
Table F.5: *Saccharomyces castellii* GO term enrichment analysis

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Convergent</th>
<th>Non-Conv.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0044238</td>
<td>primary metabolic process</td>
<td>79.27 (65)</td>
<td>56.32 (214)</td>
<td>9.8 · 10^{-05}</td>
</tr>
<tr>
<td>GO:0043170</td>
<td>macromolecule metabolic process</td>
<td>69.51 (57)</td>
<td>46.05 (175)</td>
<td>0.00014</td>
</tr>
<tr>
<td>GO:0008152</td>
<td>metabolic process</td>
<td>79.27 (65)</td>
<td>59.47 (226)</td>
<td>0.00063</td>
</tr>
<tr>
<td>GO:0044237</td>
<td>cellular metabolic process</td>
<td>78.05 (64)</td>
<td>57.89 (220)</td>
<td>0.00067</td>
</tr>
<tr>
<td>GO:0022402</td>
<td>cell cycle process</td>
<td>2.44 (2)</td>
<td>15 (57)</td>
<td>0.00083</td>
</tr>
<tr>
<td>GO:0007049</td>
<td>cell cycle</td>
<td>3.66 (3)</td>
<td>16.05 (61)</td>
<td>0.0023</td>
</tr>
<tr>
<td>GO:0006414</td>
<td>translational elongation</td>
<td>3.66 (3)</td>
<td>0 (0)</td>
<td>0.0054</td>
</tr>
<tr>
<td>GO:0006013</td>
<td>mannose metabolic process</td>
<td>3.66 (3)</td>
<td>0 (0)</td>
<td>0.0054</td>
</tr>
<tr>
<td>GO:0022403</td>
<td>cell cycle phase</td>
<td>2.44 (2)</td>
<td>12.11 (46)</td>
<td>0.0082</td>
</tr>
</tbody>
</table>

Molecular Component

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>% (num.)</th>
<th>% (num.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:005830</td>
<td>cytosolic ribo. (sensu Eukaryota)</td>
<td>37.80 (31)</td>
<td>4.21 (16)</td>
<td>4.4 · 10^{-15}</td>
</tr>
<tr>
<td>GO:0044445</td>
<td>cytosolic part</td>
<td>37.80 (31)</td>
<td>4.47 (17)</td>
<td>1.1 · 10^{-14}</td>
</tr>
<tr>
<td>GO:0005829</td>
<td>cytosol</td>
<td>45.12 (37)</td>
<td>8.95 (34)</td>
<td>2 · 10^{-13}</td>
</tr>
<tr>
<td>GO:0005840</td>
<td>ribosome</td>
<td>41.46 (34)</td>
<td>7.37 (28)</td>
<td>4.2 · 10^{-13}</td>
</tr>
<tr>
<td>GO:0030529</td>
<td>ribonucleoprotein complex</td>
<td>42.68 (35)</td>
<td>8.42 (32)</td>
<td>1.2 · 10^{-12}</td>
</tr>
<tr>
<td>GO:0015934</td>
<td>large ribosomal subunit</td>
<td>20.73 (17)</td>
<td>1.84 (7)</td>
<td>4.8 · 10^{-09}</td>
</tr>
<tr>
<td>GO:0005842</td>
<td>cytosolic large ribo. sub.</td>
<td>20.73 (17)</td>
<td>1.84 (7)</td>
<td>4.8 · 10^{-09}</td>
</tr>
<tr>
<td>GO:0043234</td>
<td>protein complex</td>
<td>51.22 (42)</td>
<td>19.74 (75)</td>
<td>1.9 · 10^{-08}</td>
</tr>
<tr>
<td>GO:0032991</td>
<td>macromolecular complex</td>
<td>53.66 (44)</td>
<td>23.16 (88)</td>
<td>1.4 · 10^{-07}</td>
</tr>
<tr>
<td>GO:0043228</td>
<td>non-membrane-bound organelle</td>
<td>46.34 (38)</td>
<td>18.16 (69)</td>
<td>2.3 · 10^{-07}</td>
</tr>
<tr>
<td>GO:0043232</td>
<td>intracell. non-membr-bound. organ.</td>
<td>46.34 (38)</td>
<td>18.16 (69)</td>
<td>2.3 · 10^{-07}</td>
</tr>
<tr>
<td>GO:0016282</td>
<td>eukaryotic 43S preinit. cmplx.</td>
<td>18.29 (15)</td>
<td>2.11 (8)</td>
<td>2.6 · 10^{-07}</td>
</tr>
<tr>
<td>GO:0015935</td>
<td>small ribosomal subunit</td>
<td>17.07 (14)</td>
<td>2.11 (8)</td>
<td>1.1 · 10^{-06}</td>
</tr>
<tr>
<td>GO:0005843</td>
<td>cytosolic small ribo. sub.</td>
<td>17.07 (14)</td>
<td>2.11 (8)</td>
<td>1.1 · 10^{-06}</td>
</tr>
<tr>
<td>GO:0016283</td>
<td>eukaryotic 48S initiation complex</td>
<td>17.07 (14)</td>
<td>2.11 (8)</td>
<td>1.1 · 10^{-06}</td>
</tr>
<tr>
<td>GO:0043227</td>
<td>membrane-bound organelle</td>
<td>42.68 (35)</td>
<td>67.89 (258)</td>
<td>2.6 · 10^{-05}</td>
</tr>
<tr>
<td>GO:0043231</td>
<td>intracell. membrane-bound organ.</td>
<td>42.68 (35)</td>
<td>67.89 (258)</td>
<td>2.6 · 10^{-05}</td>
</tr>
<tr>
<td>GO:0044444</td>
<td>cytoplasmic part</td>
<td>74.39 (61)</td>
<td>50.26 (191)</td>
<td>7.7 · 10^{-05}</td>
</tr>
<tr>
<td>GO:0044446</td>
<td>intracellular organelle part</td>
<td>54.88 (45)</td>
<td>34.21 (130)</td>
<td>0.00065</td>
</tr>
<tr>
<td>GO:0044422</td>
<td>organelle part</td>
<td>54.88 (45)</td>
<td>34.21 (130)</td>
<td>0.00065</td>
</tr>
<tr>
<td>GO:0005737</td>
<td>cytoplasm</td>
<td>89.02 (73)</td>
<td>71.84 (273)</td>
<td>0.00071</td>
</tr>
<tr>
<td>GO:0005634</td>
<td>nucleus</td>
<td>20.73 (17)</td>
<td>37.37 (142)</td>
<td>0.0046</td>
</tr>
</tbody>
</table>
**Table F.6: Candida glabrata GO term enrichment analysis**

<table>
<thead>
<tr>
<th>GO id</th>
<th>Description</th>
<th>Convergent</th>
<th>Non-Conv.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>% (num.)</td>
<td>% (num.)</td>
<td></td>
</tr>
<tr>
<td><strong>Molecular Function</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0015926</td>
<td>glucosidase activity</td>
<td>11.11 (4)</td>
<td>0 (0)</td>
<td>0.00011</td>
</tr>
<tr>
<td>GO:0004553</td>
<td>hydrolase on O-glycosyl compds.</td>
<td>11.11 (4)</td>
<td>0 (0)</td>
<td>0.00011</td>
</tr>
<tr>
<td>GO:0016798</td>
<td>hydrolase on glycosyl bonds</td>
<td>11.11 (4)</td>
<td>0.33 (1)</td>
<td>0.00051</td>
</tr>
<tr>
<td>GO:0003735</td>
<td>structural constituent of ribo.</td>
<td>8.33 (3)</td>
<td>0.66 (2)</td>
<td>0.0096</td>
</tr>
<tr>
<td><strong>Biological Process</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0009059</td>
<td>macromolecule biosynt. proc.</td>
<td>27.78 (10)</td>
<td>7.62 (23)</td>
<td>0.00086</td>
</tr>
<tr>
<td>GO:0006412</td>
<td>translation</td>
<td>16.67 (6)</td>
<td>2.98 (9)</td>
<td>0.0024</td>
</tr>
<tr>
<td>GO:0043284</td>
<td>biopolymer biosynthetic proc.</td>
<td>11.11 (4)</td>
<td>0.99 (3)</td>
<td>0.003</td>
</tr>
<tr>
<td>GO:0007047</td>
<td>cell wall organiz. and biogene.</td>
<td>25 (9)</td>
<td>7.62 (23)</td>
<td>0.0031</td>
</tr>
<tr>
<td>GO:0045229</td>
<td>ext. encap. struct. organ. &amp; biog.</td>
<td>25 (9)</td>
<td>7.62 (23)</td>
<td>0.0031</td>
</tr>
<tr>
<td>GO:0009058</td>
<td>biosynthetic process</td>
<td>33.33 (12)</td>
<td>13.25 (40)</td>
<td>0.0052</td>
</tr>
<tr>
<td>GO:0044264</td>
<td>cellular polysac. metab. proc.</td>
<td>11.11 (4)</td>
<td>1.66 (5)</td>
<td>0.0093</td>
</tr>
<tr>
<td>GO:0005976</td>
<td>polysaccharide metabolic proc.</td>
<td>11.11 (4)</td>
<td>1.66 (5)</td>
<td>0.0093</td>
</tr>
<tr>
<td>GO:0006073</td>
<td>glucan metabolic process</td>
<td>11.11 (4)</td>
<td>1.66 (5)</td>
<td>0.0093</td>
</tr>
<tr>
<td><strong>Cellular Component</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005618</td>
<td>cell wall</td>
<td>22.22 (8)</td>
<td>2.65 (8)</td>
<td>5.3 (10^{-05})</td>
</tr>
<tr>
<td>GO:0009277</td>
<td>chitin- &amp; beta-glue-cont. cell wall</td>
<td>22.22 (8)</td>
<td>2.65 (8)</td>
<td>5.3 (10^{-05})</td>
</tr>
<tr>
<td>GO:0030312</td>
<td>external encapsulating structure</td>
<td>22.22 (8)</td>
<td>2.65 (8)</td>
<td>5.3 (10^{-05})</td>
</tr>
<tr>
<td>GO:0016282</td>
<td>eukaryotic 43S preinit. complex</td>
<td>8.33 (3)</td>
<td>0.33 (1)</td>
<td>0.0041</td>
</tr>
<tr>
<td>GO:0015935</td>
<td>small ribosomal subunit</td>
<td>8.33 (3)</td>
<td>0.33 (1)</td>
<td>0.0041</td>
</tr>
<tr>
<td>GO:0005843</td>
<td>cytosolic small ribo. sub.</td>
<td>8.33 (3)</td>
<td>0.33 (1)</td>
<td>0.0041</td>
</tr>
<tr>
<td>GO:0016283</td>
<td>eukaryotic 48S initiation complex</td>
<td>8.33 (3)</td>
<td>0.33 (1)</td>
<td>0.0041</td>
</tr>
<tr>
<td>GO:0030529</td>
<td>ribonucleoprotein complex</td>
<td>19.44 (7)</td>
<td>4.97 (15)</td>
<td>0.0045</td>
</tr>
<tr>
<td>GO:0005840</td>
<td>ribosome</td>
<td>16.67 (6)</td>
<td>3.97 (12)</td>
<td>0.0069</td>
</tr>
<tr>
<td>GO:0005759</td>
<td>mitochondrial matrix</td>
<td>8.33 (3)</td>
<td>0.66 (2)</td>
<td>0.0096</td>
</tr>
<tr>
<td>GO:0031980</td>
<td>mitochondrial lumen</td>
<td>8.33 (3)</td>
<td>0.66 (2)</td>
<td>0.0096</td>
</tr>
<tr>
<td>GO:0005830</td>
<td>cytosolic ribosome</td>
<td>8.33 (3)</td>
<td>0.66 (2)</td>
<td>0.0096</td>
</tr>
</tbody>
</table>
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


188


BIBLIOGRAPHY


BIBLIOGRAPHY


