Estimating and modelling rates of evolution with applications to phylogenetics and codon selection

Rachel Bronwen Bevan

Doctor of Philosophy

Department of Computer Science

McGill University

McGill Centre for Bioinformatics

Montreal, Québec

October 12, 2006

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Your file Votre référence
ISBN: 978-0-494-32149-2
Our file Notre référence
ISBN: 978-0-494-32149-2

DEDICATION

This thesis is dedicated to my family. To Trevor, Kirk, Frances, Mom and Dad. If only I were eloquent enough to properly express my gratitude. But I am not, and so you have simply my thanks.

'There is a time in every man's education
when he arrives at the conviction that envy is ignorance;
that imitation is suicide;
that he must take himself for better, for worse,
as his portion;
that though the wide universe is full of good,
no kernel of nourishing corn can come to him
but through his toil
bestowed on that plot of ground
which is given to him to till.
The power that resides in him is new in Nature,
and none but he knows what that is which he can do,
nor does he know until he has tried.'

Ralph Waldo Emerson

Acknowledgements

There are many people whom I would like to thank for the help and support they have provided over the years. Many thanks to both of my supervisors, Dr. David Bryant and Dr. B. Franz Lang for the invaluable guidance you have given me throughout this PhD. I would also like to thank the members of my advisory committee, Russ Steele, Mathieu Blanchette and the chair of the committee Mike Hallett, for useful suggestions and guidance.

Tina, Michelle, Becky, Scott and Markus, thanks for the encouragement and support. Thanks to the group at MCB for providing a relaxing and fun work environment. Thanks also for the sushi and poker nights. Robin and Audrey, having family in Montreal has made all the difference. Thanks for translating the abstract. Finally, Trevor, thank you for always listening.

Preface

This thesis was written according the McGill University requirements as found on the Faculty of Graduate and Post–doctoral Studies website (www.mcgill.ca/gps). I have chosen to write a manuscript based thesis. According to McGill University guidelines, 'as an alternative to the traditional thesis format, the thesis can consist of a collection of papers of which the student is an author or co-author. These papers must have a cohesive, unitary character making them a report of a single program of research'. Furthermore, these papers can consist of 'the text of one or more papers submitted, or to be submitted, for publication, or the clearly-duplicated text (not the reprints) of one or more published papers.'

Contribution of Authors

Chapter 2: **Bevan, R.B.**, Lang, B.F., and Bryant D. (2005) Calculating the Evolutionary Rates of Different Genes: A Fast, Accurate Estimator with Applications to Maximum Likelihood Phylogenetic Analysis. *Systematic Biology.* 54(6): 900–915.

The idea for this article was suggested by Dr. David Bryant, and an initial solution proposed. The candidate solved the problem, implemented all code, designed all analyses and wrote the manuscript. Dr. B Franz Lang provided the empirical data for analysis. Both Dr. David Bryant and Dr. B. Franz Lang provided helpful comments on the manuscript.

Chapter 3: **Bevan, R.B.**, Bryant, D. and Lang, B.F. (2006) Accounting for gene rate heterogeneity in phylogenetic inference. *Accepted, Systematic Biology.*

The idea for one of the methods analyzed in this article was initially proposed by Dr. Joseph Felsenstein. The candidate implemented all code, designed all experiments and analyses for the article, and wrote the manuscript. Dr. David Bryant and Dr. Franz Lang provided helpful comments on the manuscript.

Chapter 4: **Bevan, R.B.**, Lang, B.F. (2006) Using evolutionary models to detect selection. *To be submitted, Molecular Biology and Evolution.*

The idea to detect synonymous codon selection was proposed by Dr. B. Franz Lang. The candidate designed the approach to detecting synonymous selection and implemented all code and analyses for the paper. The candidate wrote the manuscript with detailed comments from Dr. B. Franz Lang.

Bevan, R.B. and Lang, B.F. (2004) Mitochondrial genome evolution: the origin of mitochondria and of eukaryotes. *Topics in Current Genetics*, 8:1-35.

The idea for the manuscript was proposed by Dr. B. Franz Lang. The candidate wrote the majority of the manuscript, with detailed help from Dr. B. Franz Lang. This manuscript was written in the course of the PhD and is included in Appendix 2 with kind permission of Springer Science and Business Media.

Abstract

This thesis addresses two problems that have applications in evolution and phylogenetics: (i) estimating and accounting for evolutionary rate heterogeneity in a phylogenetic context (Chapters 2 and 3); (ii) detecting synonymous selection upon a set of codons (Chapter 4).

Chapter 2 presents a fast algorithm (DistR) to estimate gene/protein evolutionary rates based on pairwise distances between pairs of taxa derived from gene/protein sequence data. Simulation studies indicate that this algorithm accurately estimates rates and is robust to missing data. Moreover, by including evolutionary rates estimated by the DistR algorithm as additional parameters into a phylogenetic model, a significantly improved fit over the concatenated model is obtained as measured by the Akaike Information Criterion (AIC).

However, allowing every gene/protein to have its own evolutionary rate – termed the n–parameter approach – is only one method of accounting for gene rate heterogeneity in phylogenetic inference. Under the α–parameter approach, a Γ distribution is fit to the gene rates in order to account for rate heterogeneity, a method that is much slower than the n–parameter approach. Comparison of the n–parameter to the α–parameter approaches (Chapter 3) indicates that the n–parameter method provides a better fit over the concatenated model than the α–parameter approach. Interestingly, improved model fit over the concatenated model is highly correlated with the presence of a gene that has a slow relative rate.

Chapter 4 addresses the question of detecting synonymous selection on sets of codons using parametric codon models. Parametric codon models are used to simulate data under the null hypothesis that there is no synonymous selection on a particular codon; codons that have unexpected synonymous usage in empirical data, compared to the null distribution, are classified as Highly Selected Codons (HSCs). Two different data sets are analyzed to identify HSCs: nuclear genes of various *Saccharomyces* species that are well–known to undergo translational selection; mitochondrial genes of several *Reclinomonas* species that are highly A+T biased. Eleven *Saccharomyces* codons are determined to be under synonymous selection (HSCs). Nine of these codons were previously identified as undergoing translational selection. Similarly, 10 *Reclinomonas* codons are identified as undergoing synonymous selection. Comparison to traditional non–parametric approaches shows that these methods do not identify any *Reclinomonas* codons as under synonymous selection due to the high A+T bias of the genes.

Abrégé

Cette thèse pose deux problems qui ont des applications liées à l'évolution et la phylogénétique : (i) l'estimation et la prise en compte de l'hétérogénéité du taux d'évolution dans un contexe phylogénétique (Chapitres 2 et 3) ; (ii) la détection de la sélection synonyme parmi un ensemble de codons (Chapitre 4).

Le Chapitre 2 présente un algorithme rapide (DistR) pour estimer le taux d'évolution d'un gène ou d'une protéine basé sur les distances entre les paires de taxa dérivé de données de séquence gène ou protéine. Des études de simulations indiquent que cet algorithme estime correctement les taux et qu'il est robuste vis-à-vis des données manquantes. De plus, en incluant les taux d'évolution estimés par l'algorithme DistR comme des paramètres additionnels dans un modèle phylogénétique, une amélioration significative de l'ajustement par rapport au modèle concaténé est obtenu selon le Critère de l'Information d'Akaike (Akaike Information Criterion (AIC)).

Une méthode permettant de prendre en compte l'hétérogénéité du taux d'évolution des gènes lors de l'inférence phylogénétique consiste á permettre à chaque gène d'avoir son propre taux d'évolution. Cette méthode est appelée approache á n–paramètres. Une autre méthode qui est plus lente que l'approche á n–paramètres est l'approche á α–paramètres. Avec cette dernière approche, une distribution–Γ est ajustée aux taux d'évolution génétique pour prendre en compte l'hétérogénéité de ces taux.

La comparaison entre ces deux approches (Chapitre 3) indique que la méthode á n–paramètres donne un meilleur ajustement que le modèle concatené. Il est intéressant de noter qu'un meilleur ajustement de ces modèles comparé au modèle concatené est très corrélé á la présence d'un gène au taux relatif lent.

Le Chapitre 4 aborde la question de la détection de sélection synonyme dans des groupes de codons utilisant les modèles paramétriques du codon. Les modèles paramétriques du codon sont utilisés pour simuler les informations selon l'hypothèse nulle stipulant qu'il n'y a pas de sélection synonyme sur un codon en particulier. Les codons qui ont une utilisation synonyme extreme selon des données empiriques sont classifiés parmi les codons hautement sélectionnés (HSCs).

Deux groupes de données sont analysés pour identifier les codons HSCs : les gènes nucléaires de diverses espèces de *Saccharomyces* pour la sélection de la traduction et les gènes des mitochondries de diverses espèces de *Reclinomonas*, qui sont hautement biaisés A+T. Onze codons de *Saccharomyces* ont été identifies comme étant suject á la sélection synonyme (HSCs). Parmi ces codons, neuf ont été identifis comme 'etant suject á la sélection de la traduction.

De la même façon, dix codons de *Reclinomonas* subissent la sélection synonyme. La comparaison de l'approche traditionnelle non–paramétrique montre que ces méthodes n'identifient pas de codon sous la sélection synonyme due à l'important biais A+T des gènes.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

1.1 Charles Darwin on Natural Selection

'Let it borne in mind how infinitely complex and close–fitting are the mutual relations of all organic beings to each other and to their physical conditions of life. Can it, then, be thought improbable, seeing that variations useful to man have undoubtedly occurred, that other variations useful in some way to each being in the great and complex battle of life, should sometimes occur in the course of thousands of generations?'

Charles Darwin, The Origin of Species, 1859

1.2 Outline

Evolution is a fundamental biological theory which posits that all species are related through common descent (Darwin, 1859); molecular evolution assumes that these relationships can be elucidated through analysis of data at the molecular level. Such relationships are represented by a *phylogeny* (or tree) which can be inferred (among other techniques) using maximum likelihood *phylogenetic* models of evolution. This thesis investigates these models in two contexts. The first

context addresses improving current models to account for fundamental biological processes. The second context addresses using such models to determine if certain biological processes violate assumptions of neutral evolution.

This chapter provides a detailed introduction to two key areas that are necessary to understand this thesis: (i) a biological background including assumptions behind molecular evolution; (ii) approaches to phylogenetic inference including parsimony, neighbour joining, maximum likelihood and Bayesian statistics. Finally a short introduction of each future chapter is presented.

1.3 Genetics and Evolution

1.3.1 DNA, Amino Acids and the Genetic Code

DNA (*deoxyribonucleic acid*) stores all genetic information. There are four chemical building blocks in DNA: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Cytosine and Thymine are *pyrimidines* (denoted Y); Guanine and Adenine are *purines* (denoted R). Each of these chemical structures bonds to a *sugar* (ribose) and a *phosphate* to form a *nucleotide. Single–stranded DNA* is formed by covalent bonding of the sugar in one nucleotide, to the phosphate in another, to form a *sugar phosphate backbone. Double–stranded DNA* forms through hydrogen bonding of two strands of single–stranded DNA. Hydrogen bonds are more likely to form, and are stronger between two sets of nucleotides: A–T and C–G (Voet et al., 1999).

Nucleotide triplets (or *codons*) code for amino acids forming the standard genetic code (Table 1–2) (Voet et al., 1999). *Amino acids* are the building blocks of proteins which perform important structural/functional roles in the cell. There are twenty commonly used amino acids in most genetic systems. Both the three and one letter abbreviations are given in Table 1–2. Because there are 64 possible codons, there is redundancy built into the genetic code. For example, the amino acid valine (V or Val, Table 1–2) has four codons GT(Y|R). Here Y|R indicates that either a purine or pyrimidine is allowed in the third codon position and the codon will still code for valine. Because most amino acids (except methionine (M) and tryptophan (W)) have more than one codon, the genetic code is *degenerate.* In general, the third position in the codon is either *two–fold* or *four–fold* degenerate. Thus, either two or four different nucleotides are allowed respectively, in the third codon position, without affecting the amino acid. Because amino acids are coded for by three nucleotides, there are three possible *reading frames* for any genetic sequence. A reading frame that begins with a start codon, and stops with an end codon is an *open reading frame.* An open reading frame might code for an RNA or protein (Voet et al., 1999).

1.3.2 Mutation

There are four types of mutation that affect genes: substitution, insertion, deletion and inversion (Nei and Kumar, 2000). Substitution is the replacement of one nucleotide with another. There are two types of substitution, transitions (purine $\rightarrow$ purine and pyrimidine $\rightarrow$ pyrimidine) and transversions (purine $\rightarrow$

	T			C			A			G			
T	TTT	Phe	[F]	TCT	Ser	[S]	TAT	Tyr	[Y]	TGT	Cys	[C]	**T**
	TTC	Phe	[F]	TCC	Ser	[S]	TAC	Tyr	[Y]	TGC	Cys	[C]	**C**
	TTA	Leu	[L]	TCA	Ser	[S]	TAA	Ter	[end]	TGA	Ter	[end]	**A**
	TTG	Leu	[L]	TCG	Ser	[S]	TAG	Ter	[end]	TGG	Trp	[W]	**G**
C	CTT	Leu	[L]	CCT	Pro	[P]	CAT	His	[H]	CGT	Arg	[R]	**T**
	CTC	Leu	[L]	CCC	Pro	[P]	CAC	His	[H]	CGC	Arg	[R]	**C**
	CTA	Leu	[L]	CCA	Pro	[P]	CAA	Gln	[Q]	CGA	Arg	[R]	**A**
	CTG	Leu	[L]	CCG	Pro	[P]	CAG	Gln	[Q]	CGG	Arg	[R]	**G**
A	ATT	Ile	[I]	ACT	Thr	[T]	AAT	Asn	[N]	AGT	Ser	[S]	**T**
	ATC	Ile	[I]	ACC	Thr	[T]	AAC	Asn	[N]	AGC	Ser	[S]	**C**
	ATA	Ile	[I]	ACA	Thr	[T]	AAA	Lys	[K]	AGA	Arg	[R]	**A**
	ATG	Met	[M]	ACG	Thr	[T]	AAG	Lys	[K]	AGG	Arg	[R]	**G**
G	GTT	Val	[V]	GCT	Ala	[A]	GAT	Asp	[D]	GGT	Gly	[G]	**T**
	GTC	Val	[V]	GCC	Ala	[A]	GAC	Asp	[D]	GGC	Gly	[G]	**C**
	GTA	Val	[V]	GCA	Ala	[A]	GAA	Glu	[E]	GGA	Gly	[G]	**A**
	GTG	Val	[V]	GCG	Ala	[A]	GAG	Glu	[E]	GGG	Gly	[G]	**G**

Table 1–2: The standard genetic code.

pyrimidine and *vice versa*). The former type of substitution is more common (Voet et al., 1999). Substitution mutations affect a single codon. A change in the codon that leads to a change at the amino acid level is a *non–synonymous* mutation. One that does not lead to a change at the amino acid level is a *synonymous* mutation (Voet et al., 1999).

Insertion and deletion consist respectively of the addition or removal of (a) nucleotide(s) from the sequence. These mutations can lead to a change in the reading frame of the gene if the insertion/deletion is not a multiple of three nucleotides. In this case the codons *downstream* (after) the insertion/deletion are all shifted potentially leading to a protein that is non–functional. Indeed, insertions and deletions are more rarely fixed in the population (than substitutions) over time because they often affect more than one codon. Inversion usually affects the entire gene. When genes are inverted the double–stranded DNA is cut on both sides of the gene (or genes). The gene is then reinserted into the DNA so that it is reads in the opposite direction. Inversions are also much rarer than substitutions (Voet et al., 1999; Nei and Kumar, 2000).

1.3.3 Selection

Selection at the molecular level is defined in terms of the the rates of synonymous and non–synonymous substitution at a particular *site* (a codon that is common in many species). If the rate of non–synonymous substitution is significantly greater than the rate of synonymous substitution, then the site is under *positive selection* (Nei and Kumar, 2000). Thus, changes away from the current

amino acid are selected for in order to improve the *fitness* (function/structure) of the protein (and hence the organism). If the rate of synonymous substitution is significantly greater than the rate of non–synonymous substitution, then the site is under *purifying selection.* This means that the current amino acid is strongly selected for in order to retain the current fitness (i.e. structure/function) of the protein. Hence there is selection against mutation away from the current amino acid. If the non–synonymous and synonymous rates are equal then there is no apparent selection acting on the site.

1.3.4 The theory of neutral evolution

The theory of *neutral evolution* posits that the majority of changes within a genome are selectively neutral, in that they do not affect the fitness of the organism. The essential basis of the theory is that most mutations are *selectively neutral* and accumulate randomly over time, regardless of the fact that they present no improved fitness to the organism (Kimura, 1968; King and Jukes, 1969; Kimura, 1983). Because these selectively neutral mutations are fixed in the population, it is possible to distinguish between different species based on these neutral mutations. Species which are more closely related will have fewer mutations than those that are distantly related. These similarities and differences at a molecular level allow for the inference of evolutionary relationships between species (Nei and Kumar, 2000; Felsenstein, 2004a).

1.4 Phylogenetic Inference

The goal of phylogenetic inference is to determine a tree that best explains the evolutionary relationships between species. Such relationships can be conveniently represented by a tree or *phylogeny*, with internal nodes denoting hypothetical ancestors of current species.

Phylogenetic analysis begins with an *alignment* of sequence data from various species of interest. The data consists of *orthologous* gene or protein sequences. These are derived from the same ancestral gene/protein and have the same function in each species. It is important to avoid the use of *paralogs* and *xenologs*. The use of either types of sequence can lead to error in the phylogenetic inference (Felsenstein, 2004a). Paralogs are gene sequences that result from a duplication event rather than a speciation event. The use of a paralog can occur as follows: assume there are two paralogous genes in all species, gene A and gene B. The goal is to analyze gene A in all species: in one species gene A is selected for analysis, in another species gene B is selected for analysis (because it is incorrectly inferred to be orthologous to gene A). Thus, the evolutionary history of gene A is not properly represented. Xenologs are gene sequences that are acquired by horizontal gene transfer (the transfer of a gene from one species to another), and thus do not represent the evolutionary history of the species under analysis (Voet et al., 1999).

Orthologous sequences are aligned using standard dynamic programming techniques (Needleman and Wunsch, 1970; Smith and Waterman, 1981; Thompson et al., 1994). The result is a set of *sites* which are assumed to share a common evolutionary history (Figure 1–1). The sites in Figure 1–1 have no gap states.

Figure 1–1: Example of a DNA alignment. The alignment is made of many sites, one of which is depicted by an arrow. The site is made up of many states, one from each species. There are four states (known as nucleotides) in DNA: adenine, A; cytosine, C; guanine, G; thymine, T.

Gaps are introduced to account for possible insertion/deletion mutations. However, they often represent uncertainty in the alignment, or areas where the alignment is bad. Thus, it is common to curate the alignment by removing (eliminating) sites with many gap states before phylogenetic analysis. This can be done by hand or using a program such as Gblocks (Castresana, 2000).

1.4.1 Parsimony

The goal of the *parsimony* approach to phylogenetic inference is to obtain a tree that minimizes the number of substitutions over all sites over all possible groupings of species (Fitch, 1977; Felsenstein, 2004a). This concept of *maximum parsimony* was first introduced by Fitch (Fitch, 1971, 1977). For example, for the high–lighted site in Figure 1–1, a most parsimonious tree can be seen in Figure 1–2a. When species S1 and species S2 are grouped into a *clade* there is no change

of state necessary between these two species and the common ancestor A1. The same is true for species S3 and species S4. The only state change necessary is between the two common ancestors A1 and A2. Conversely, Figure 1–2b shows a less parsimonious grouping of the species. When species S1 and S3 are grouped together at least one mutation is necessary between the common ancestor A1 and species S1 or S3. The same is true for species S2 and S4 and common ancestor A2. Here the most parsimonious explanation of this grouping of the species is that the two ancestors A1 and A2 both had the same state derived from the ancestor at the root of the tree.

Maximum parsimony has the advantage that it is relatively fast to compute, especially when compared to model–based approaches to phylogenetic inference. However, the problem of finding the maximum parsimony tree is NP–complete (Foulds and Graham, 1982). Furthermore, parsimony does not converge to the correct tree in all cases (i.e. it is inconsistent) (Felsenstein, 1978, 2004a). *Long branch attraction (LBA)* occurs when two long branches of a tree, which have the same state by chance or *convergent* mutation, rather than common descent, incorrectly resolve into a phylogenetic grouping (or *clade*, Figure 1–3). If LBA occurs between two or more species (or clades) for a number sites, the resultant most parsimonious tree topology will be incorrect (Felsenstein, 1978).

Figure 1–3 demonstrates how LBA can occur under parsimony. The true relationships between the species are given in Figure 1–3a, and the inferred most parsimonious relationships in Figure 1–3b. In Figure 1–3a both ancestral nodes have the nucleotide A. Due to the long length of time during which change can

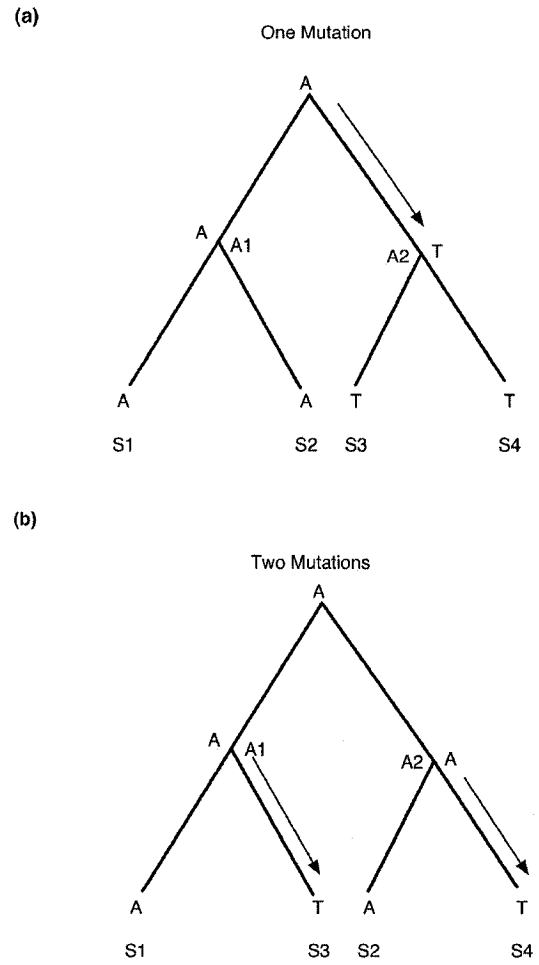


Figure 1–2: An example of (a) a most parsimonious and (b) less parsimonious grouping of species S1–S4 based upon the high–lighted site in Figure 1–1.

occur, both species S1 and S2 have nucleotide C. When parsimony is applied, species S1 and S2 are grouped incorrectly into a single clade (Figure 1–3b). This problem of inferring the incorrect tree topology can occur even with large amounts of sequence data, which is why parsimony is inconsistent (Felsenstein, 1978, 2004a).

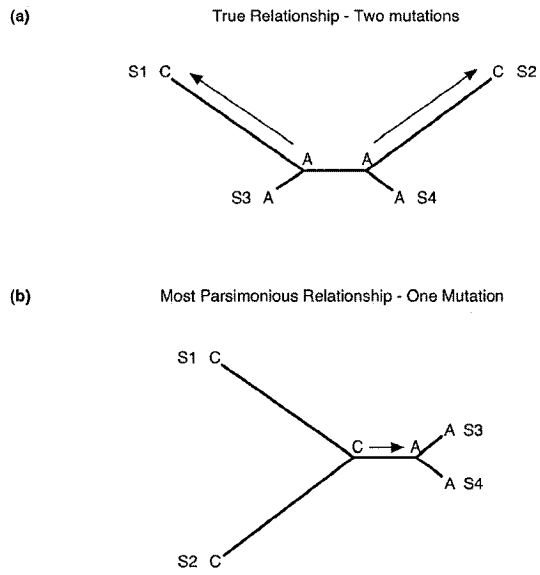


Figure 1–3: An example of how long branch attraction occurs between two species. (a) The species true relationships where species S1 and S2 with nucleotides C do not form a clade. (b) When parsimony is used to infer the relationships between species the minimum number of mutations is one, where species S1 and S2 form a clade.

1.4.2 Distance based methods – NJ and BIONJ

Neighbour–joining (NJ) is a well–known approach to inferring phylogenetic trees. It is based upon inferring pairwise distances between species and using these distances to build a tree topology (Saitou and Nei, 1987). The algorithm proceeds by finding the minimum of a criterion C between all pairs of n taxa. Let $d_{x,y}$ represent the distance between species x and y in distance matrix D. The value of the criterion $C_{x,y}$ is given as:

$$C_{x,y} = (r-2)d_{x,y} - S_x - S_y$$

for $S_k = \sum_{l=1}^{r} d_{k,l}$ and r the current dimension of the distance matrix D. At the initial step of the algorithm $r = n$. Let the two taxa that minimize the value of $C_{x,y}$ be i and j. These taxa are selected and joined together in the tree as 'neighbours' by an ancestral node u (Figure 1–4). The taxa i and j are removed from the distance matrix D and the distances from u to all other taxa m in D are calculated as

$$d_{u,m} = \frac{1}{2}(d_{i,m} + d_{j,m} - d_{i,u} - d_{j,u})$$

where $d_{i,u} = \frac{1}{2}\left(d_{i,j} + \frac{S_i - S_j}{(r-2)}\right)$ and $d_{j,u}$ is calculated symmetrically (Saitou and Nei, 1987; Gascuel, 1997). Thus the number of pairwise distances in the matrix is reduced by one ($r = r - 1$). This procedure is repeated until all species are joined into a tree. In theory, the two species which minimize the distance criterion will be most closely related to each other. However, this is only true in practice if there are no parallel or backward substitutions (mutations that occur but are unobserved) (Saitou and Nei, 1987).

BIONJ is an extension of the NJ algorithm that attempts to minimize the sampling variance of the distance matrix D (Gascuel, 1997). This is achieved by calculating the new distances $d_{u,m}$ as (Gascuel, 1997):

$$d_{u,m} = \lambda d_{i,m} + (1 - \lambda) d_{j,m} - \lambda d_{i,u} - (1 - \lambda) d_{j,u}$$

The goal of the BIONJ algorithm is to adjust the λ parameter in order to reduce the sampling variance (Gascuel, 1997). The BIONJ algorithm finds an

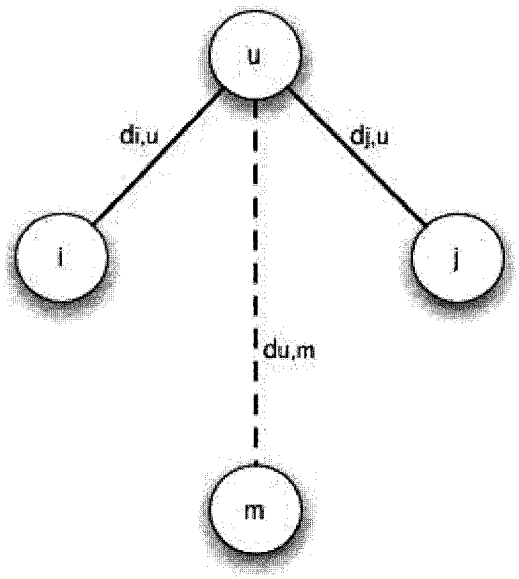


Figure 1–4: Example of how taxa are joined into clades using the NJ algorithm. Here taxa i and j minimize the criterion C. They are joined by an ancestral node u (solid lines). New distances from u to all other nodes m are calculated (dashed line) and all distances involving taxa i and j are removed from the distance matrix D.

estimate of λ that minimizes

$$\sum_{k=1,k\neq j,i}^{r} v_{u,k} = \sum_{k=1,k\neq i,j}^{r} Var[\lambda d_{i,k} + (1-\lambda)d_{j,k} - \lambda d_{i,u} - (1-\lambda)d_{j,u}]$$
$$= \sum_{k=1,k\neq i,j}^{r} Var[\lambda d_{i,k} + (1-\lambda)d_{j,k}]$$

since the distances $d_{i,u}$ and $d_{j,u}$ are considered to be constant (Gascuel, 1997). The estimate of λ which minimizes this variance can be solved for analytically. Furthermore, the variance and covariance of the distance estimates can be calculated easily for sequence data (e.g. Bulmer, 1991; Gascuel, 1997).

1.4.3 Maximum likelihood inference

One of the major techniques employed to infer phylogenetic relationships between species is that of Maximum Likelihood (ML). In ML relationships between species are inferred based on sequence data by assuming an underlying model of substitution between *states* (e.g. nucleotides) in a site. The probability of a substitution occurring in an infinitesimal amount of time is computed based on the underlying frequencies of states in the alignment or database of alignments. These probabilities are up and down-weighted as appropriate, depending on different biological processes (Felsenstein, 1981b,a).

For example, in DNA the probability of a change C↔ T or A↔ G is greater than (C|T) ↔ (A|G). This is because C and T are both pyrimidines, whereas A and T are both purines. Changing states within a chemical group (called a *transition*) is favoured biologically over changes between chemical groups (called a *transversion*). This is accounted for by including a transition:transversion ratio ($\frac{Tn}{Tv}$) in the DNA model. This is an unknown parameter that is estimated from the data.

The goal of ML phylogenetic inference is to estimate the branch lengths, tree topology and parameters which maximize the likelihood of the data. The probability of a site is calculated under a given set of parameter values of the model. In mostly widely used models each site is assumed to evolve independently, thus the probability of the data given the model is simply the product of the site–wise probabilities. The set of parameter values that maximize the likelihood of

the data are assumed to best model the history of the biological data (Felsenstein, 1981b,a).

1.4.3.1 Probability of change along a branch

The process of change from one state to another along a branch is described by a discrete Markov chain. Here the states correspond to nucleotides, amino acids or codons, depending upon the data of interest. For each state of the chain there is a probability of change to another state, and a probability of remaining in the current state. Under a discrete Markov chain the probability of being in state i at time t depends solely upon the previous state the chain was in (Ross, 2003). Thus,

$$P_{i,j} = P(S_t = i|S_{t-1} = j) = P(S_t = i|S_{t-1} = j, S_{t-2} = k, ..., S_0 = l)$$

Let the transition matrix that describes these probabilities be $\mathbf{X}$. Therefore, for an m–state Markov chain with states labelled from 1, ..., m:

$$\mathbf{X} = \begin{pmatrix} P_{1,1} & \cdots & P_{1,m} \\ \vdots & \ddots & \vdots \\ P_{m,1} & \cdots & P_{m,m} \end{pmatrix}$$

From $\mathbf{X}$ we can obtain an n–step transition matrix as: $\mathbf{X}^n = \mathbf{X}\mathbf{X}\cdots\mathbf{X}$.

The Markov chains used have two important properties: they are ergodic – if the Markov chain is run for an infinite length of time every state i in the chain is visited with non–zero probability π_i and the chain is aperiodic; they are irreducible – it is possible to reach each state from every other state in the chain

(not necessarily in one change). Furthermore, if the condition of detailed balance holds – $\pi_j X_{j,i} = \pi_i X_{i,j}$ – then the Markov chain is *time reversible* (Ross, 2003).

The ergodic property is particularly important because it means that there is a *stationary* probability of being in state i denoted π_i. Given these stationary probabilities it is possible to calculate the probability of starting in a particular state i and ending in state j after k time points (which correspond to k state transitions) as:

$$\begin{aligned} P(S_k = j, k \text{ time points}, S_0 = i) &= P(S_k = j, k \text{ time points } | S_0 = i) P(S_0 = i) \\ &= X_{i,j}^k \pi_i \\ &= R_{i,j}^k \end{aligned}$$

This discrete time Markov chain describes a process in which the time points are constant and known. However, in evolution, the changes from one state to another occur at different (non–constant) time intervals. To compute the probability of k time points under such a paradigm a Poisson distribution is used (Felsenstein, 1981b,a). Let K be the random variable describing the number of time points that have passed. Thus:

$$K \sim Po(\mu t)$$

$$\text{and } P(K = k|\mu, t) = e^{-\mu t} \frac{(\mu t)^k}{k!}$$

where μt is the expected length of time between state changes. To calculate the probability of changing from state i at time 0 to state j at time t it is necessary to sum over the possible number of k state transitions (time points), multiplied by

the probability of having that number of transitions between state i and state j. Define $P(t)$ as the probability of change from one state to the next in infinitesimal time interval t. Thus,

$$P(t) = \sum_{k=0}^{\infty} \mathbf{R}^k e^{-\mu t} \frac{(\mu t)^k}{k!} \tag{1.1}$$

$$= e^{\mu t} \sum_{k=0}^{\infty} \frac{(\mathbf{R}\mu t)^k}{k!} \tag{1.2}$$

$$= e^{\mu t} e^{\mathbf{R}\mu t} \tag{1.3}$$

$$= e^{(\mathbf{R}-\mathbf{I})\mu t} \tag{1.4}$$

where I is the identity matrix. Setting $\mathbf{Q} = (\mathbf{R} - \mathbf{I})\mu$ gives $P(t) = e^{Qt}$, where $\mathbf{Q}$ is the instantaneous rate matrix of change (Felsenstein, 1981b,a). The $\mathbf{Q}$ matrix describes the probability of change from one state to another in an infinitesimal amount of time (Felsenstein, 1981b,a).

1.4.3.2 DNA models

The general time reversible (GTR) Q matrix for DNA sequences is given as:

$$Q = \begin{pmatrix} * & a\mu\pi_C & b\mu\pi_G & c\mu\pi_T \\ a\mu\pi_A & * & d\mu\pi_G & e\mu\pi_T \\ b\mu\pi_A & d\mu\pi_C & * & f\mu\pi_T \\ c\mu\pi_A & e\mu\pi_C & f\mu\pi_G & * \end{pmatrix}$$

where π_A, π_C, π_G and π_T are the equilibrium frequencies of the nucleotides, which are either set to $\frac{1}{4}$ or to the frequencies of the data under analysis depending upon

the model used. The diagonal elements ($*$) are set so that all the rows sum to 0 (i.e. $Q_{i,i} = -\sum_{i \neq j} Q_{i,j}$). μ is a rate parameter that gives the expected number of changes per time unit t. In most cases $\mu = 1$ (Felsenstein, 1981b,a, 2004a). The values a, ..., f are parameters estimated from the data that weight the probability of change from one nucleotide to another (away from the stationary frequencies) depending upon the nucleotides under analysis. They are generally called *rate* parameters because they modify the instantaneous rate of change from one nucleotide to another (Felsenstein, 2004a). In the most complex model, the GTR model, there are 6 rate parameters $a - f$ so that the process is still time reversible (i.e. $\pi_i Q_{i,j} = \pi_j Q_{j,i}$) (Tavaré, 1986). Also, it is required that $\Sigma \pi_i Q_{i,i} = -1$ reducing the number of rate parameters to five.

Other DNA models are more restricted. For instance in the HKY model (Hasegawa et al., 1985)

$$Q = \begin{pmatrix} * & \mu\pi_C & \kappa\mu\pi_G & \mu\pi_T \\ \mu\pi_A & * & \mu\pi_G & \kappa\mu\pi_T \\ \kappa\mu\pi_A & \mu\pi_C & * & \mu\pi_T \\ \mu\pi_A & \kappa\mu\pi_C & \mu\pi_G & * \end{pmatrix}$$

Here κ is a rate parameter used to describe the difference in rates between transitions and transversions:

$$\begin{aligned} \frac{Tn}{Tv} &= \frac{\kappa\mu(\pi_G + \pi_T + \pi_A + \pi_C)}{2\mu(\pi_A + \pi_C + \pi_G + \pi_T)} \\ &= \frac{\kappa}{2} \end{aligned}$$

1.4.3.3 Protein models of evolution

Protein substitution models have 20 different states (one for each amino acid) and the substitution rate multipliers $s_{i,j}$ are estimated empirically. Thus

$$Q = \begin{pmatrix} * & s_{1,2} & \cdots & s_{1,20} \\ \vdots & \ddots & & \vdots \\ s_{20,1} & \cdots & s_{20,19} & * \end{pmatrix} diag\,(\pi_1, \cdots, \pi_{20})$$

for amino acids 1, ..., 20 with equilibrium frequencies π_1, ..., π_{20}. Here $s_{l,k}$ is set equal to $s_{k,l}$ to make the Markov chain reversible (i.e. $\pi_i Q_{i,j} = \pi_j Q_{j,i}$). The diagonal elements of Q are fixed so that the row values sum to 0. The substitution rate parameters $s_{k,l}$ are computed from a large number of substitutions in many proteins.

Two popular amino acid substitution matrices include JTT and WAG (Jones et al., 1992; Whelan and Goldman, 2001). The substitution rates for the JTT model were calculated using parsimony based on the observed substitutions in a large globular protein database (Jones et al., 1992). However, this might underestimate the number of substitutions because parsimony assumes only one substitution along a particular branch (Jones et al., 1992; Whelan and Goldman, 2001). Furthermore, the relative ratios of substitution as inferred by maximum parsimony are biased. The Q matrix may be biased as a result.

The substitution rates for the WAG model were calculated using a maximum likelihood approach (Whelan and Goldman, 2001). The best fit of the model M (which here includes the WAG matrix) was obtained based on assuming that

the neighbour–joining tree with branch lengths estimated using the JTT+F model, will give a close to optimal solution for the tree. Thus, if the set of trees T for all protein families D are assumed to be known it is possible to calculate $L(M|T,D) = \prod_{\text{protein families } i} L(M|T_i, D_i)$. Here D_i is the data in protein family i and T_i is the tree for protein family i. Thus, the maximum likelihood estimate of M is obtained over many protein families (Whelan and Goldman, 2001).

1.4.3.4 Codon models

Codon models of evolution are primarily used to estimate the non-synonymous to synonymous rate ratio ($\frac{d_N}{d_S}$) of sites within a gene. One of the reasons they are not used to infer phylogenies is the large computational power needed to solve the standard equation $P = e^{\mathbf{Q}t}$, where $\mathbf{Q}$ is a 61x61 instantaneous rate matrix for the 61 coding codons. In these models and their variants, the relative instantaneous substitution rate from codon i to codon j ($i \neq j$) is given by:

$$q_{ij} = \begin{cases} 0 & \text{if i and j differ at two or three nucleotide positions} \\ \pi_j & \text{if i and j differ by one synonymous transversion} \\ \kappa\pi_j & \text{if i and j differ by one synonymous transition} \\ \omega\pi_j & \text{if i and j differ by one non–synonymous transversion} \\ \omega\kappa\pi_j & \text{if i and j differ by one non–synonymous transition} \end{cases}$$

where ω is the rate of non–synonymous to synonymous substitution, κ is the rate of transitions to transversions and π_j is the equilibrium frequency of codon j (Yang and Nielsen, 2000).

Yang *et al.* proposed incorporation of site specific values of ω into this matrix (Yang et al., 2000), allowing for the non–synonymous rate of substitution to differ across sites. Under the extension, the codon matrix is defined for a particular site h as $q_{ij}^{(h)}$ with $\omega = \omega^{(h)}$ (Yang et al., 2000). If $\omega^{(h)} > 1$ at site h then the site is undergoing positive selection. This means that for a particular site h, the rate of non–synonymous change is significantly greater than the rate of synonymous change according to the expectation under a neutral model of evolution. Neutral and purifying selection at site h are determined by $\omega^{(h)} = 1$ and $\omega^{(h)} < 1$ respectively. The $\omega^{(h)}$ can be calculated based upon a distribution, (e.g Beta), thus reducing the number of parameters to be estimated (Yang and Nielsen, 2000).

Only allowing non–synonymous rate change across sites makes the implicit assumption that any synonymous changes are selectively neutral. This assumption is removed by allowing for synonymous rates to vary across sites (Pond and Muse, 2005). A distribution can be fit to synonymous rate across sites in the exact manner as for the distribution for non–synonymous rates across sites. However, the mean rate of synonymous substitution must be 1.0 in order to avoid identifiability problems (Pond and Muse, 2005) (where there is more than one set of parameter values that give the same probability for a particular model).

Setting $\omega^{(h)} = \beta^{(h)}$ below leads to the following instantaneous substitution matrix for site h:

$$q_{ij}^{(h)} = \begin{cases} 0 & \text{if i and j differ at two or three nucleotide positions} \\ \alpha^{(h)}\pi_j & \text{if i and j differ by one synonymous transversion} \\ \alpha^{(h)}\kappa\pi_j & \text{if i and j differ by one synonymous transition} \\ \beta^{(h)}\pi_j & \text{if i and j differ by one non-synonymous transversion} \\ \beta^{(h)}\kappa\pi_j & \text{if i and j differ by one non-synonymous transition} \end{cases}$$

where $(\alpha^{(h)}, \beta^{(h)})$ is a random vector calculated based upon a discretized bivariate distribution h parameterized by ν. It is also possible to let $\alpha^{(h)}$ and $\beta^{(h)}$ to vary independently (Pond and Muse, 2005). This model is similar to a model that allows for unique non–synonymous and synonymous rates at each site (Massingham and Goldman, 2005), but does not over–parameterize the data.

1.4.3.5 Obtaining maximum likelihood estimates of parameters

The goal of maximum likelihood inference is to obtain estimates of the model parameters that maximize the probability of the data. For phylogenetic maximum likelihood inference the likelihood function is defined as:

$$\begin{aligned} L(\theta, \lambda, T|Data) &= P(Data|\theta, \lambda, T) \\ &= \prod_{i=1}^{n} P(D_i|\theta, \lambda, T) \end{aligned}$$

for tree T, branch lengths λ, and other model parameters θ. $P(D_i|\theta, \lambda, T)$ is calculated by summing over all possible assignments of states to internal nodes of the tree T, since the topology is assumed to be known. This summation can be calculated in the time it takes to do a post–order traversal of a tree, or $\mathcal{O}(n)$. Using the notation of Felsenstein (Felsenstein, 2004a), denote $L_k^{(i)}(s)$ as the likelihood of the data at site i given that node k has state s. This is calculated as:

$$L_k^{(i)}(s) = \left(\sum_{\text{states } x} P(x|s, t_a) L_a^{(i)}(x) \right) \left(\sum_{\text{states } y} P(y|s, t_b) L_b^{(i)}(y) \right)$$

where t_a is the length of the branch connecting node k with descendent node a and t_b is the length of the branch connecting node k with descendent node b. $P(x|s,t) = P(t)_{s,x}$ is the probability that state s changed to state x in the first descendant lineage in time t. The likelihood at node k is the product of the likelihood at subtrees a and b because each branch on the tree is assumed to be independent.

Note that for species m at external node e with state z it is necessary to define

$$L_e^{(i)}(s) = \begin{cases} 1 & \text{if state } s \text{ at node } e \text{ equals } z \\ 0 & \text{otherwise} \end{cases}$$

Thus the likelihood of a site can be computed recursively, resulting in $L_r^{(i)}(s)$, the likelihood at the root r of the tree for state s. However, since the state

at the root is unknown, the overall likelihood of the tree for site i is given as $L^{(i)} = \sum_{\text{states } x} \pi_x L_r^{(i)}(x)$. Note that this can be calculated for unrooted trees by assigning the root to one of the internal nodes of the tree.

1.4.3.6 Accounting for site rate heterogeneity in DNA and protein models

Under DNA and protein evolutionary models it is necessary to account for the fact that different sites undergo different rates of evolution in order to prevent model violation problems (Yang, 1996; Waddell and Steel, 1997; Pupko et al., 2002b). Under a model of evolution that does not account for site rate heterogeneity, branch lengths of the tree are maximized over all sites. However, some sites will evolve slowly with little to no change over long periods of time, whereas other sites will evolve more quickly with many changes. Thus, site rate parameters that allow some sites to have a fast rate of evolution and others a slow rate of evolution should be incorporated into the model. Since the site is evolving faster/slower, all branch lengths are multiplied by the rate parameter causing the probability of change over the rate corrected branch lengths to increase or decrease depending on the actual rate of evolution of the site. Under such a model of evolution, the probability of change from state s to state x over time t at rate R is:

$$P(tR)_{s,x} = e^{QRt}$$

Gamma model for site rate heterogeneity

Calculating a rate parameter R for every site s leads to the problem of infinite parameterization; the model has more parameters than the amount of data available to infer the maximum parameter estimates, and thus will over–fit the data. One solution is to assume that the rates R are are drawn from some known distribution f, with parameters ν, and integrate out over the possible values of R. In this case, the number of parameters added to the model is equal to the dimension of ν.

Yang proposed that f be a Gamma distribution, which normally has two parameters α and β (Yang, 1993, 1994). However, due to the constraint that the average rate of substitution must be 1.0 (in order to avoid non–identifiability problems) β is constrained so that the mean of the distribution is one. This is achieved with the Gamma distribution by setting $\beta = \frac{1}{\alpha}$. Thus

$$P(Data|\theta, \lambda, T, \alpha) = \int_0^\infty P(Data|\theta, \lambda, T, R) f(R|\alpha) dR$$

However, computing the likelihood for all possible rate values drawn from a continuous distribution is too time intensive. Thus Yang proposed approximating the continuous Gamma distribution with a discrete approximation that has C equiprobable rate categories, R_1, ..., R_C based upon the Gamma distribution

(Yang, 1994). Under this approximation the likelihood is calculated as:

$$P(Data|\theta,\lambda,T,\alpha) = \int_0^\infty P(Data|\theta,\lambda,T,R)f(R|\alpha)dR$$
$$\approx \sum_{i=1}^{C} P(Data|\theta,\lambda,T,\alpha,R_i)\frac{1}{C}$$

Invariable Sites Mixture Model

Sites that are invariant (with no mutation) across a tree cause problems with phylogenetic inference by misleading genetic divergence estimates (Reeves, 1992; Churchill et al., 1992). However, some sites are invariant due to chance and other cannot change (termed 'invariable' sites). In a statistical framework that accounts for invariant sites the likelihood is calculated as:

$$P(Data|\theta,\alpha,\lambda,T,\psi) = \psi P(Data|\theta,\alpha,\lambda,T,invariant)$$
$$+ (1-\psi)P(Data|\theta,\alpha,\lambda,T,variant)$$

where ψ is the proportion of invariable sites. *Variant* and *invariant* are indicator variables as to whether or not the site has a mutation. When the site is variant $P(Data|\theta,\alpha,\lambda,invariant) = 0$. When the site is invariant $P(Data|\theta,\alpha,\lambda,T,invariant) = \pi_s$ where s is the character state of the invariant site. This is the probability that the site is invariant due to chance (Reeves, 1992; Churchill et al., 1992; Swofford et al., 1996)

$P(Data|\theta,\alpha,\lambda,T,variant) = P(Data|\theta,\alpha,\lambda,T)$ regardless of whether or not the site is variant. (Reeves, 1992; Churchill et al., 1992; Swofford et al., 1996).

This is due to the fact that even if a site is invariant, it is possible that there were substitutions are unobserved.

1.4.4 Bayesian phylogenetic inference

In Bayesian statistics the goal is to obtain a posterior distribution over the parameters of interest, rather than maximizing the probability of the data given the model (as in frequentist statistics). Within Bayesian inference the likelihood of the data is calculated, however it is weighted by the prior probability of the parameters of interest, and normalized by the probability of the data.

In Bayesian phylogenetic inference, the goal is to obtain the posterior distribution of trees. Let $\Omega = (T, \lambda, \theta)$ for tree T with branch lengths λ and model parameters θ. The posterior distribution over all parameters is calculated as (Larget and Simon, 1999; Huelsenbeck and Ronquist, 2001; Huelsenbeck et al., 2002):

$$\begin{aligned} P(\theta, \lambda, T|Data) &= \frac{P(Data|\theta, \lambda, T)P(\theta, \lambda, T)}{P(Data)} \\ &= \frac{P(Data|\theta, \lambda, T)P(\theta, \lambda, T)}{\int_{\Omega} P(Data|\theta, \lambda, T')P(\theta, \lambda, T')d\Omega} \end{aligned} \tag{1.5}$$

Furthermore, in order to obtain the posterior distribution of the tree T it is necessary to integrate over branch lengths λ and model parameters θ (Larget and

Simon, 1999; Huelsenbeck and Ronquist, 2001; Huelsenbeck et al., 2002). Thus,

$$P(T|Data) = \frac{\int_\lambda \int_\theta P(\theta, \lambda, T|Data)P(\theta, \lambda, T)d\theta d\lambda}{\int_\Omega P(Data|\theta, \lambda, T')P(\theta, \lambda, T')d\Omega}$$
$$= \frac{\int_\lambda \int_\theta P(\theta, \lambda, T|Data)P(\theta, \lambda, T)d\theta d\lambda}{\sum_{\text{trees}T'} \int_\lambda \int_\theta P(Data|\theta, \lambda, T')P(\theta, \lambda, T')d\theta d\lambda}$$

However, it is often quite difficult to calculate these integrals in closed form. One solution is to sample the posterior distribution of Ω, and sum over the sampled values of λ, θ and T'. Such sampling is achieved using Markov Chain Monte Carlo (MCMC). The Metropolis algorithm is a well known technique to sample from the posterior distribution of the parameters. The algorithm starts at initial estimates for Ω of Ω^0. At time t a candidate value of Ω^* is proposed and is accepted with probability $\min(r, 1)$ where $r = \frac{P(\Omega^*|Data)}{P(\Omega^{t-1}|Data)}$. The proposed value of Ω^* is based upon a jumping distribution $J_t(\Omega^*|\Omega^{t-1})$ which is symmetric. If a uniform random variable on [0,1] is less than r then the new state of the chain is accepted, otherwise it is rejected. Thus, if the probability of the parameters given the data is worse at Ω^* than Ω^{t-1}, the chain moves to the proposed state with probability r. Otherwise the chain moves to the new estimates with probability one. This is repeated until the sample size of Ω is sufficiently large.

The Metropolis algorithm samples from the posterior distribution of Ω. For example, as shown in (Gelman et al., 2000) consider two samples 1 and 2 of Ω. Assume that $P(\Omega_1|Data) > P(\Omega_2|Data)$. Now the probability of a transition from Ω_2 to Ω_1 is

$$P(\Omega^t = \Omega_1, \Omega^{t-1} = \Omega_2) = P(\Omega_2|Data)J_t(\Omega_1|\Omega_2)$$

which has acceptance probability of one, since $r = \frac{P(\Omega_1|Data)}{P(\Omega_2|Data)} > 1$. Thus, with probability one Ω_1 has a better probability under the posterior distribution. Furthermore, the probability of a transition from Ω_1 to Ω_2 is

$$\begin{aligned} P(\Omega^t = \Omega_2, \Omega^{t-1} = \Omega_1) &= P(\Omega_1|Data)J_t(\Omega_2|\Omega_1)r \\ &= P(\Omega_1|Data)J_t(\Omega_2|\Omega_1)\frac{P(\Omega_2|Data)}{P(\Omega_1|Data)} \\ &= J_t(\Omega_2|\Omega_1)P(\Omega_2|Data) \\ &= P(\Omega^t = \Omega_1, \Omega^{t-1} = \Omega_2) \end{aligned}$$

.

This is true because the jumping distribution J is symmetric. Thus the probability of a transition from sample 1 to sample 2 is the same as the probability of transition of sample 2 to sample 1. Because J is symmetric Ω_1, Ω_2 have the same marginal distribution of $P(\Omega|Data)$. Thus sampling according to the Metropolis algorithm samples from the posterior distribution $P(\Omega|Data)$.

The Metropolis–Hastings algorithm is a generalization of the Metropolis algorithm that allows for non–symmetric jumping distributions. Because of this the ratio r must be modified in order to retain the Markov property of the chain.

$$\begin{aligned} r &= \frac{P(\Omega^*|Data)/J_t(\Omega^*|\Omega^{t-1})}{P(\Omega^{t-1}|Data)/J_t(\Omega^{t-1}|\Omega^*)} \\ &= \frac{P(\Omega^*|Data)J_t(\Omega^{t-1}|\Omega^*)}{P(\Omega^{t-1}|Data)J_t(\Omega^*|\Omega^{t-1})} \end{aligned}$$

The probability of the next sample value of Ω^* given the data, is weighted by the probability of moving from the proposed sample back to the current sample

under the jumping distribution J. The probability of the current sample value Ω^{t-1} is weighted by the probability of moving from the current sample to the proposed sample. This weighting allows for the Markov property of the chain to be retained when sampling (i.e. time reversible, ergodic, irreducible). The ratio of $\frac{J_t(\Omega^{t-1}|\Omega^*)}{J_t(\Omega^*|\Omega^{t-1})}$ is known as the Hastings ratio, and is set to one under the Metropolis algorithm (because J is symmetric).

One of the nice properties of the Metropolis (Metropolis–Hastings) algorithm is that P(Data) does not need to be calculated. This is because $P(Data)$ cancels out when r is calculated:

$$\begin{aligned} r &= \frac{P(\Omega^*|Data)J_t(\Omega^{t-1}|\Omega^*)}{P(\Omega^{t-1}|Data)J_t(\Omega^*|\Omega^{t-1})} \\ &= \frac{\frac{P(Data|\Omega^*)P(\Omega^*)}{P(Data)}J_t(\Omega^{t-1}|\Omega^*)}{\frac{P(Data|\Omega^{t-1})P(\Omega^{t-1})}{P(Data)}J_t(\Omega^*|\Omega^{t-1})} \\ &= \frac{P(Data|\Omega^*)P(\Omega^*)J_t(\Omega^{t-1}|\Omega^*)}{P(Data|\Omega^{t-1})P(\Omega^{t-1})J_t(\Omega^*|\Omega^{t-1})} \end{aligned}$$

Thus, it is only necessary to compute the likelihood of the data given the parameter estimates ($P(Data|\Omega)$), the prior over Ω, the Hastings ratio (if J is not symmetric) and a uniform random variable on [0,1], in order to sample from Ω.

1.5 Tree Support

Within a frequentist paradigm, the goal is to find the maximum likelihood parameter estimates for the data, under the distribution presumed to describe the data f. If f is an unknown distribution, or the maximum likelihood estimator is

intractable, it is not possible to mathematically obtain a variance on the parameter estimates.

One approach to determining the uncertainty of the estimates is the bootstrap (Efron, 1979; Felsenstein, 2004a). Under the bootstrap the data are assumed to be identically and independently distributed according to the true underlying distribution of the data. Thus, when the data are sufficiently large, the empirical distribution $\hat{f}$ is an estimator of the true distribution f (Efron, 1979; Felsenstein, 2004a). Under this assumption, a new data set can be drawn (with replacement) from the empirical distribution $\hat{f}$ in order to obtain new estimates of the parameters of interest. From these estimates it is possible to obtain a variance on the parameter estimates.

Bootstrap is a commonly used method of inferring tree support. The sites in the alignment are assumed to be IID, and are sampled with replacement to obtain new data sets that are assumed to be drawn from the underlying distribution that describes sequence data. Tree inference (non–Bayesian, i.e. ML or parsimony) is applied to the resampled data sets in order to infer the variance on the tree topology supported by the data. The internal support for each branching is obtained based on the proportion of times the branching is seen across all bootstrap data sets. In a Bayesian context it is not necessary to calculate bootstrap estimates of the variance since these estimates can be obtained from the posterior MCMC sample of parameter estimates.

1.6 Tree Search Heuristics

1.6.1 NNI and SPR

Two common tree search heuristics include nearest–neighbour interchange (NNI) and subtree prune and regraft (SPR). The former approach exchanges subtrees that are ancestors of two internal nodes that are attached directly by an edge (Felsenstein, 2004a); the latter removes a subtree from an internal node and joins it to another internal node of the tree (Felsenstein, 2004a). Both PAUP and PHYLIP use NNI and SPR to search tree space; SPR is used for large–scale rearrangements when inserting taxa into the initial tree; NNI is used once all the taxa are in the tree for local rearrangements (Felsenstein, 2004a). Once an SPR or NNI move has been performed, the branch lengths and other parameters are optimized. If this leads to a better likelihood, the new tree is kept, otherwise a new branch swap is proposed based upon the old tree topology, branch lengths and parameter estimates. This search is repeated until no moves lead to a better tree topology.

1.6.2 PHYML

PHYML is a phylogenetic tree search algorithm that employs a fast tree search heuristic. For a bifurcating tree, there are three possible rearrangements of subtrees around an internal edge e for subtrees A, B, C, and D. Normal algorithms apply an NNI, then optimize the branch lengths and evaluate the likelihood

to determine if the swap leads to a tree of better likelihood. However, if the conditional likelihood for each of the subtrees rooted at A, B, C and D is stored it is possible to quickly compute the approximate likelihood (all branches are not optimized simultaneously) of the three possible arrangements around edge e. PHYML decides which swaps to make based upon a scoring system of these three likelihoods.

Let L_1, L_2 and L_3 respectively denote the likelihood of possible topologies 1, 2 and 3. Assume that L_1 is the likelihood of the original tree topology (tree topology 1). Let $S = L_1 - L_i$ for $i = 2$ and $i = 3$. These scores are calculated for all possible internal nodes and ranked from greatest to least. All possible swaps are performed according to this ranking as follows (Guindon and Gascuel, 2003):

1. A proportion λ initialized to 0.75 determines the number of swaps to be performed out of the total number of swaps;
2. starting with the swap of highest score each swap is performed sequentially;
3. once a swap is performed external branches and internal branches not involved in the swap are given branch length $l = l + \lambda(l_i - l)$ for branch length l of the current branch and branch length l_i of the edge e under tree topology 1;
4. if the swap leads to a worse likelihood decrease λ by dividing by 2, thus reducing the number of potential swaps tested.

A tree is left unmodified if $\lambda = 0$. This is because none of the swaps are selected, and none of the branch lengths modified. Conversely, if $\lambda = 1.0$, all swaps will be selected and performed, assuming of course that λ is not decreased as

the swaps are tested. The algorithm is guaranteed to converge to a tree topology because the potential number of swaps is decreased when the likelihood for a given swap is worse than the current likelihood (Guindon and Gascuel, 2003). PHYML can start with any input tree, but will build an initial BIONJ tree if none is provided (Guindon and Gascuel, 2003).

1.7 Thesis Contributions

This thesis addresses two questions that have applications in evolution and phylogenetics: (i) how best to account for gene or protein rate heterogeneity in a phylogenetic context; (ii) detecting synonymous selection on particular codon (or set of codons). To address the first question the DistR algorithm was developed. DistR quickly determines the relative evolutionary rates of a set of genes (Chapter 2). Furthermore, different methods of incorporating gene rate heterogeneity into phylogenetic models are investigated (Chapter 3). Finally, Chapter 4 focuses on detecting synonymous selection upon codons. Parametric methods are used to simulate data under the null hypothesis that codons do not undergo synonymous selection. Particular codons that have synonymous usage that violate this null distribution are identified in *Reclinomonas* and *Saccharomyces.*

1.7.1 The DistR Algorithm – estimating rate heterogeneity

Extensions to the basic maximum likelihood (ML) model have been made that account for site rate heterogeneity (Yang, 1993). However, gene rate heterogeneity

also plays an important role in evolution – some genes evolve more quickly than others. This problem becomes particularly important in the context of finding the correct model to fit the data. There is not enough information in one gene to correctly infer most phylogenies on more distantly related species. With greater amounts of sequence data available, genes are concatenated into large data sets and analyzed with the same models as those used for single genes. However this approach only accounts for site rate heterogeneity. It does does not adequately account for the gene rate heterogeneity of the evolutionary process. This may lead to incorrect phylogenetic inferences.

An ML approach to accounting for gene rate heterogeneity has been proposed for both DNA (Yang, 1996) and amino acid (Pupko et al., 2002b) data. However, neither implementation allows for tree space to be searched. Furthermore, the time to calculate ML gene rates is slow – a fast and accurate approximate method will provide a useful starting point for any detailed ML analysis. Chapter 2 of this thesis focuses on estimating the relative evolutionary rates of genes/proteins quickly, and incorporating these estimates into the ML framework for estimating phylogenies (Bevan et al., 2005).

1.7.2 Methods of accounting for gene rate heterogeneity

There are two approaches to incorporate gene rate heterogeneity into a phylogenetic model. The first, termed the n–parameter method, allows for each gene to have one rate of evolution. Thus, the probability of the data is calculated at a single ML estimate of gene rate for each gene (Yang, 1996; Pupko et al.,

2002b). The second, termed the α–parameter model, allows for a distribution of gene rates. Thus, the probability of the data is averaged over a set of gene rates defined by this distribution (Felsenstein, 2001, 2004b).

The former method is computationally faster, however it might suffer from the problem of infinite parameterization (i.e over–fit of the model to the data) when there are many genes in the data set. The later method is computationally much slower, since the probability of the data must be calculated for each gene rate defined by the distribution. However, because the gene rates are defined by a distribution, only the parameters of the distribution must be maximized over, eliminating the problem of infinite parameterization.

Chapter 3 compares the α–parameter and n–parameter methods of accounting for gene rate heterogeneity using the Akaike Information Criterion (AIC). This information criterion is widely used to correct the log–likelihood of the data by the number of parameters in the model. The AIC of two models can be compared to determine if one model has an improved fit compared to another. Additionally, analysis is performed to determine what properties of the data under analysis correlate with improved model fit of a gene rates model over the concatenated model.

1.7.3 Using phylogenetic models to detect codons under synonymous selection

Chapter 4 also focuses on rates of evolution. However, the goal is to detect synonymous selective pressures on a set of codons across a genome (or sets of

genes). Here synonymous selective pressure is selection for the use of a particular synonymous codon at a codon site. Normally, a site will have more than one synonymous codon because synonymous codons are hypothesized to be selectively neutral. Thus, it is assumed that there is no selective (evolutionary) advantage to using one synonymous codon versus another at a particular site.

Both non–parametric and parametric methods have been developed to analyze codon usage. Non–parametric methods use summary statistics of the gene data to determine a set of codons that are purported to be under synonymous selective pressure. However, they do not account for mutation, codon bias, gene rate evolution and other evolutionary pressures that will bias the results of the analyses. Parametric methods (phylogenetic codon models) do account for such pressures, however they are currently only used to detect sites that are under positive or purifying selection.

This chapter applies phylogenetic codon models to the question of determining a set of sites under synonymous selective pressure. Codon models are used to simulate data using ML parameter estimates from the genes under analysis. Properties of codon usage in synonymous and non–synonymous can then be compared in the simulated and real data. If the codon models are not capturing particular properties of the data (such as the usage of a codon in invariant sites versus the usage of a codon in variant sites), then it is possible that the site is under synonymous selective pressure.

Chapter 2
Calculating the Evolutionary Rates of Different Genes: A Fast, Accurate Estimator with Applications to Maximum Likelihood Phylogenetic Analysis

2.1 Background

Gene rate heterogeneity is an important property of evolution that should be accounted for in phylogenetic models. Not only is the rate of evolution of a gene important in a phylogenetic context, but it is important to understand fundamental biological processes since it is highly correlated with expression level (Drummond et al., 2005).

This chapter presents an algorithm to quickly infer rates of evolution of genes. Weighted least squares is used to infer gene rates from distance data between pairs of species in different genes. These rate estimates can be compared to estimates obtained in a maximum likelihood framework, with no missing data. Finally, the rates are included in phylogenetic models, in order to compare the fit of a model that accounts for gene rate heterogeneity, to the concatenated model which does not.

2.2 Abstract

In phylogenetic analyses with combined multi–gene or multi–protein data sets, accounting for differing evolutionary dynamics at different loci is essential for accurate tree prediction. Existing maximum likelihood (ML) and Bayesian approaches are computationally intensive. We present an alternative approach that is orders of magnitude faster. The method, Distance Rates (DistR), estimates rates based upon distances derived from gene/protein sequence data. Simulation studies indicate that this technique is accurate compared with other methods and robust to missing sequence data. The DistR method was applied to a fungal mitochondrial data set, and the rate estimates compared well to those obtained using existing ML and Bayesian approaches. Inclusion of the protein rates estimated from the DistR method into the ML calculation of trees as a branch length multiplier resulted in a significantly improved fit as measured by the Akaike Information Criterion (AIC). Furthermore, bootstrap support for the ML topology was significantly greater when protein rates were used, and some evident errors in the concatenated ML tree topology (i.e. without protein rates) were corrected.

2.3 Introduction

It is widely recognized that the analysis of multiple unlinked genes is superior to single gene analyses for phylogenetic reconstruction. These unlinked genes may, however, be evolving according to very different rules. Heterogeneity of the evolutionary process must be accounted for in phylogenetic analyses (Bull et al.,

1993; Huelsenbeck et al., 1996; Yang, 1996; Bapteste et al., 2002; Pupko et al., 2002b; Nylander et al., 2004). The concept of accounting for differing evolutionary pressures within phylogenetic analysis is not new (Yang, 1993). Site specific rates of evolution can be computed for amino acids (e.g. Rate4Site, Mayrose et al., 2004; Pupko et al., 2002a) and DNA (e.g. DNArates, Olsen et al., 1993) using both Bayesian and Maximum Likelihood approaches.

Site rates within a gene are likely to be more correlated than rates for sites in different genes. To account for this, it can be assumed that each gene evolves at a different average rate and that these gene rates are drawn from some common distribution (Felsenstein, 2001, 2004a; Cranston and Rannala, 2005). Both Bayesian (Huelsenbeck and Ronquist, 2001) and Maximum Likelihood (Yang, 1996; Pupko et al., 2002b) methods exist to estimate gene rates (or more generally, locus rates) but these are computationally expensive.

We present a fast, accurate method to estimate the relative evolutionary rates of genes/proteins. For example, when run on a data set with 63 proteins over 123 taxa the algorithm takes less than a second. The method can be applied to protein or nucleotide data, though here we focus on protein sequences. The basic idea is to use pairwise estimates of evolutionary divergence (distances) to deduce the relative rates of different proteins, even when the proteins are not all present in all of the taxa. Although this approach does not give the ML estimates for the rates (Yang, 1996; Pupko et al., 2002b), it does provide an excellent approximation.

After computing rates they are incorporated as extra parameters into the ML tree search, resulting in improved fit as measured by the AIC. The rates estimated

using the DistR procedure have been coded into PHYML version 2.2, available at http://atgc.lirmm.fr/phyml/ (Guindon and Gascuel, 2003). PHYML was used because incorporation of the rates was straightforward and because PHYML is an especially fast implementation of ML.

2.4 Methods

2.4.1 The DistR method

To begin with, the method will be explained through an example. Figure 2–1 represents three different protein alignments. Not all taxa are present in all three alignments. Suppose that the three proteins have rates r_1, r_2, and r_3. These rates will affect distances inferred from the alignments. Reversing the problem involves using the pairwise distances between species to estimate the different rates r_1, r_2, and r_3.

Figure 2–1 outlines two ways of obtaining distances from each protein. In the first method ML trees are constructed and the length of the path between two taxa in these trees is measured (referred to hereafter as *patristic ML distances*). In the second method distances are estimated directly from the alignments, as is customary in distance–based methods (referred to hereafter as *pairwise ML distances*). The end result from both methods is a distance matrix for each protein.

If the rate in one protein is twice the rate in a second protein, then the expected distance estimates from the first protein should be twice the expected

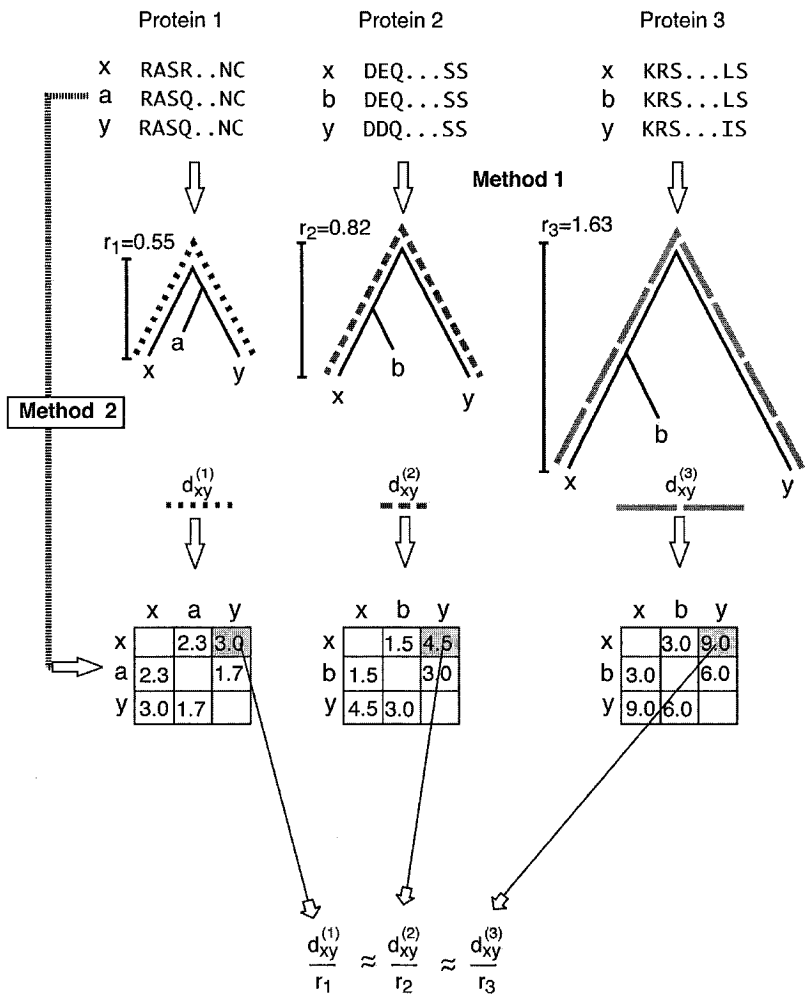


Figure 2–1: The general idea of the DistR estimation procedure. Beginning with individual protein alignments over a set of taxa (with missing data), distances between the species are estimated for each protein alignment. There are two choices of how to estimate the distances: directly from the alignment data (method 2); as the sum of the pairwise distances between taxa on a tree built from the alignment data (method 1). The result is a matrix of pairwise distances between taxa. The ratio of the pairwise distances to the rate of evolution of the protein should be approximately the same for all proteins.

distance estimates from the second protein. This should hold, approximately, for both pairwise ML distances and patristic ML distances. Equivalently, the distance

estimate from the first protein, divided by two, should be approximately the distance estimate of the second protein.

In the example (Figure 2–1), and later on, the distance between taxa x and y estimated from protein k is denoted $d_{xy}^{(k)}$, irrespective of whether it is a pairwise or patristic ML distance. Suppose that, for each k, the rate in protein k equals r_k. It follows that $\frac{d_{xy}^{(1)}}{r_1}$ will be approximately equal to $\frac{d_{xy}^{(2)}}{r_2}$ which in turn will be approximately equal to $\frac{d_{xy}^{(3)}}{r_3}$. This is denoted as

$$\frac{d_{xy}^{(1)}}{r_1} \approx \frac{d_{xy}^{(2)}}{r_2} \approx \frac{d_{xy}^{(3)}}{r_3}, \tag{2.1}$$

where '$\approx$' means 'approximately equal'. In Figure 2–1, this gives $\frac{3.0}{0.55} \approx \frac{4.5}{0.82} \approx \frac{9.0}{1.63}$.

In a sense, the distance estimates obtained from each gene are normalized so that the scale is the same. Define this normalized distance or *consensus distance* between any two taxa as p_{xy}, with the assumption that

$$p_{xy} \approx \frac{d_{xy}^{(1)}}{r_1} \approx \frac{d_{xy}^{(2)}}{r_2} \approx \frac{d_{xy}^{(3)}}{r_3}.$$

Assume that rates r_1, r_2, and r_3 in Figure 2–1 are unknown, while the distances remain known. The above approximate equality leads to

$$p_{xy} \approx \frac{3.0}{r_1} \approx \frac{4.5}{r_2} \approx \frac{9.0}{r_3}. \tag{2.2}$$

The unknowns p_{xy}, r_1, r_2, and r_3 can be solved for using a least squares approach.

The relation in equation (2.2) provides a framework to solve for the *relative rates* r_1, r_2, and r_3, given estimates for the distances $d_{xy}^{(k)}$. This is the basic idea

behind the method. The main issues are how to: (a) handle the fact that the relations are only approximate; (b) deal with missing distances; (c) compute the rate estimates quickly. These issues are addressed in the following text and in Appendix 2.

To formalize the problem, suppose that there are n proteins (or genes, etc.) over m species. The distance between species x and y derived from protein k is denoted $d_{xy}^{(k)}$. The basic assumption made is that the ratio of the estimated distance between a pair of taxa for a given protein ($d_{xy}^{(k)}$ for protein k and taxa x,y), to the rate of the protein (r_k for protein k), is approximately equal across all proteins.

The rates r_1 ,r_2, ..., r_n are unknown quantities to be estimated based upon the distance data from a given protein alignment. To do this, assume that there exists an unknown consensus distance p_{xy} such that

$$p_{xy} \approx \frac{d_{xy}^{(1)}}{r_1} \approx \frac{d_{xy}^{(2)}}{r_2} \approx \cdots \approx \frac{d_{xy}^{(n)}}{r_n},$$

where $n = 3$ for the example in Figure 2–1. All the consensus distances and rates can now be estimated using a least squares approach.

In the least squares method it is possible to incorporate measures of uncertainty about the estimated distances $d_{xy}^{(k)}$. Distance estimates with low variance should contribute more to the analysis, while distance estimates with high variance (or infinite variance in the case of missing entries) should contribute little. Let $w_{xy}^{(k)} \geq 0$ be a measure of the uncertainty in the distance estimate between taxa x and y derived from protein k. If $d_{xy}^{(k)}$ is accurate then $w_{xy}^{(k)}$ should be high. If

there is less certainty about the accuracy of $d_{xy}^{(k)}$ then $w_{xy}^{(k)}$ should be low. This is achieved using the inverse of the variance of $d_{xy}^{(k)}$, that is, $w_{xy}^{(k)} = \frac{1}{\mathrm{Var}\left(d_{xy}^{(k)}\right)}$. If protein k is not present in both x and y then $w_{xy}^{(k)} = 0$. To measure the variance of the distance estimates the approximate formula of Bulmer Bulmer (1991) is used in the implementation of DistR. Other variance estimators could also be used.

Under a weighted least squares (WLS) framework the total discrepancy between the ratios $\frac{d_{xy}^{(n)}}{r_n}$ and the consensus distances p_{xy} is measured by

$$q(\mathbf{p}, \mathbf{r}) = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} \left(p_{xy} - \frac{d_{xy}^{(k)}}{r_k} \right)^2 \tag{2.3}$$

where $\mathbf{p}$ denotes the vector $\left[p_{12}, p_{13}, ..., p_{\frac{(m-1)m}{2}}\right]^T$ and $\mathbf{r}$ denotes the vector $[r_1, ..., r_n]^T$. This is similar to the minimization function used by Lapointe and Cucumel (1997) in the average consensus method. The main difference is that they assume one rate over all proteins, whereas this method includes different rates for each protein. Note that if taxa x and y are missing from a protein k then an estimate for $d_{xy}^{(k)}$ cannot be obtained. However, this is not a problem since the weight $w_{xy}^{(k)}$ will be zero in this case.

Estimating both rates and consensus distances using $q(\mathbf{p}, \mathbf{r})$ leads to the problem of *non–identifiability*. In the absence of any error each estimated protein distance $d_{xy}^{(k)}$ is the product of the rate of the protein r_k and the consensus distance p_{xy}. Thus, a perfect fit to the equation is still achieved if all the rates are multiplied by some constant and all the consensus distances divided by the same constant. There is a problem of determining scale. Hence, equation (2.3) does not

have a well–defined minimum. To solve this problem a constraint

$$\sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} p_{xy} = \kappa \tag{2.4}$$

must be added to system, where κ is an arbitrary positive constant. The particular value of κ is irrelevant since changing κ merely causes all estimated rates to be multiplied by the same constant value. For this reason, it is possible to infer *relative rates* only. In DistR $\kappa = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} d_{xy}^{(k)}$, thus constraining the weighted estimated distances to be equal to the weighted consensus distances. This was empirically determined to minimize the variance of the DistR estimates by testing multiple constraints.

Appendix 3 describes an extremely fast algorithm for minimizing the function $q(\mathbf{p}, \mathbf{r})$ subject to the constraint in equation (2.4). The algorithm takes $O(nm^2 + n^3)$ time and $O(n^2 + m^2)$ memory. For example, when run on a data set with 63 proteins over 123 taxa, the algorithm takes less than a second. An implementation with source code is available at `http://www.mcb.mcgill.ca/~rachel`.

2.4.2 Experimental Studies

An extremely rapid method for estimating the relative rates of different genes has been proposed. The method is orders of magnitude faster than existing ML and Bayesian approaches. The most important question remaining is to what extent this increase in speed affects the accuracy of the estimates. In order to address this question, the accuracy of the new method was assessed using both simulated and empirical data.

In all the analyses PHYML (version 2.2) was used (Guindon and Gascuel, 2003) to compute ML distances and trees, with a JTT protein model, eight Gamma categories plus invariant sites and the default (BIONJ) starting tree. The Gamma shape parameter and proportion of invariant sites were estimated using default optimization routines in the program. When constructing ML trees from real data several bootstrap values were computed. As detailed below these values depend upon: whether patristic or pairwise ML distances were used in the DistR procedure; whether the rates were re-estimated for each bootstrap replicate.

For both the simulated and empirical data DistR estimates based upon patristic and ML distances were compared. This comparison was made in order to determine whether or not the additional computational effort required for estimating patristic ML distances is justified.

2.4.3 Experimental Studies—Simulated Data

The two key questions addressed through the simulation studies are:

- *Patristic versus pairwise ML distances.*— How accurate are the rate estimates using pairwise versus patristic ML distances?
- *Missing distances between taxa.*— How are DistR rate estimates affected when proteins are not present in all taxa?

To answer these questions protein alignments were simulated using Pseq–Gen (Grassly et al., 1997) with the JTT model of evolution. The initial tree and branch lengths were taken from an independent analysis of mitochondrial Atp8 proteins in 58 eukaryotes. Two types of simulations were carried out. The first, intended to

address the first question, involved construction of 20 protein trees by randomly deleting taxa from the starting tree. In total there were four protein trees with 53 taxa, four with 48 taxa, four with 43 taxa, four with 38 taxa, and four with 33 taxa. For each tree a rate was sampled from a pre–computed distribution of rates based on real data (data not shown), and protein alignments of length 100, 300, 500, and 1000 generated using Pseq–Gen (Grassly et al., 1997) (note that the average length of naturally occurring proteins is approximately 300–amino acids). The second analysis, intended to address the second question, increased the number of taxa deleted from the starting tree. In total there were seven trees with 25% of the taxa, seven with 50% of the taxa, and seven with 75% of the taxa. This resulted in twenty–one trees, seven each with 16, 30, and 44 taxa respectively. For each tree a rate was sampled from a pre–computed distribution of rates based on real data (data not shown), and protein alignments of length 1000 generated using Pseq–Gen (Grassly et al., 1997). This experiment follows a protocol proposed by (Eulenstein et al., 2004). For both experiments, and for every set of parameters, 10 replicates of the experiment were performed. See Figure 2–2 for an overview of the simulations.

Statistics measured on the simulated data, including goodness–of–fit and mean squared error, are explained in detail in Appendix 1. These statistics were used to relate the accuracy of the DistR rate estimates to the known rates at which the proteins were simulated.

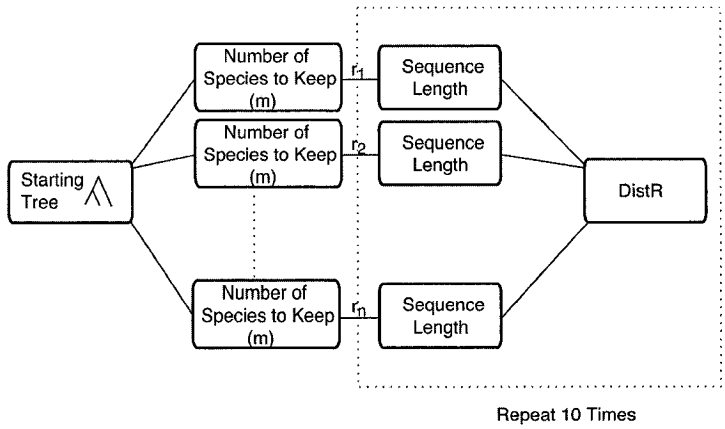


Figure 2–2: The general flow of the simulation studies. Two studies were performed, one with $n = 20$ and the other with $n = 21$ (where n is the number or proteins). The first study compared different methods of estimating distances using different alignment lengths. In the first study 20 random subtrees from an original tree of 58 species were created, four each of size $m = 33$, $m = 38$, $m = 43$, $m = 48$ and $m = 53$ (where m is the size of the taxon set for a given protein). For each tree a rate was sampled from a pre–computed distribution of rates based on real data (data not shown). Protein alignments of length 100, 300, 500, and 1000 were simulated using Pseq–Gen (Grassly et al., 1997). A second analysis compared rate estimates with increasing amounts of data. Twenty–one random subtrees from the original tree of 58 species were created, 7 each of size $m = 16$, $m = 30$ and $m = 44$ (corresponding to approximately 25%, 50%, and 75% of the species (as in Eulenstein et al., 2004)). For each tree a rate was sampled from a pre–computed distribution of rates based on real data (data not shown). Alignments of length 1000 were generated. For both studies, 10 replicates were performed for each set of parameters.

2.4.4 Experimental Studies—Empirical Data

The data analyzed in this study consist of a set of 15 aligned mitochondrial protein sequences from 29 taxa. The taxon names and accession numbers are given in Table 2–2. Protein names and alignment accession numbers appear in Table 2–5. This multi–protein data set is of moderate size, and variants thereof have been used in numerous publications (e.g. Sumida et al., 2001; Tomita et al., 2002; Lang

et al., 2002; Bullerwell et al., 2003). Furthermore, some of the species have high evolutionary rates and substitutional saturation of sites (i.e. *Smittium*), whereas others have very short branches in the resulting phylogenetic tree. Combined, these two properties can cause inaccurate grouping of the taxa due to long–branch attraction artifacts (Felsenstein, 1978).

Alignments were performed using the default settings of ClustalW (Thompson et al., 1994). Highly variable sites or those with many gaps were eliminated using Gblocks (Castresana, 2000) with the following settings: number of sequences for a flank position equal to half the number of species plus one; number of contiguous non–conserved positions equal to ten; minimum length of a block four; half the species allowed gaps. All other parameters were set to default.

The key questions addressed using real protein data are:

- *Comparison of DistR estimates to ML estimates.*— How do DistR rate estimates compare to those obtained using the ML based method COMBINE (Pupko et al., 2002b)?
- *Comparison of DistR estimates to Bayesian estimates.*— How do DistR rate estimates compare to those obtained by MrBayes (Huelsenbeck and Ronquist, 2001) under a Bayesian approach?
- *Patristic versus pairwise ML distances.*— How do rate estimates from pairwise ML distances and rate estimates from patristic ML distances compare when applied to real data?
- *Inclusion of DistR estimates into the phylogenetic tree search of PHYML.*— What is the effect of including DistR estimates in an ML tree search? Is

there a significantly improved fit? Are improved phylogenetic estimates obtained?

Species	GenBank Accession
Ascomycota	
Aspergillus nidulans	CAA33481, AAA99207, AAA31737, CAA25707, AAA31736, CAA23994, X15442, P15956, CAA23995, CAA33116, X00790, X15441, X06960, J01387, X01507
Candida albicans	AF285261
Candida glabrata	CGL511533
Hypocrea jecorina	AF447590
Penicillium marneffei	NC_005256
Pichia canadensis	NC_001762
Podospora anserina	X55026
Saccharomyces cerevisiae	AJ_011856
Schizosaccharomyces japonicus	NC_004332
Schizosaccharomyces octosporus	AF275271
Schizosaccharomyces pombe	X54421
Torrubiella confragosa	AF487277
Yarrowia lipolytica	AJ307410
Basidiomycota	
Cryptococcus neoformans	NC_004336
Schizophyllum commune	AF402141
Cantharellus cibarius [a]	
Choanoflagellida	
Monosiga brevicollis	AF538053
Chytridiomycota	
Allomyces macrogynus	U41288
Harpochytrium94	NC_004760
Harpochytrium105	NC_004623
Hyaloraphidium curvatum	AF402142
Monoblepharella	AY182007
Rhizophydium136	NC_003053
Spizellomyces punctatus	AF402142
Metazoa	
Homo sapiens	NC_001807
Metridium senile	AF000023
Zygomycota	
Smittium culisetae	AY8632133
Mortierella verticillata	AY863211
Rhizopus oryzae	AY863212

[a] Downloaded from http://megasun.bch.umontreal.ca/People/lang/FMGP/proteins.html

Table 2–2: Please see caption on following page.

Table 2–2 (caption): Names and accession numbers for protein sequences studied from Fungal species and outgroup. Fifteen proteins were down-loaded for each species (if present in the species), the names of which are in Table 2–5.

2.4.4.1 Comparison of DistR estimates to ML estimates

Note that when comparing DistR rates to those computed using COMBINE (Pupko et al., 2002b) the number of taxa and proteins had to be restricted, since COMBINE can currently only handle data sets for which all taxa are present in all proteins. Two different starting trees were included in the analysis: the ML tree from PHYML based upon the concatenated data set and the ML tree from PHYML when protein rates were incorporated. Rates were estimated under three different models: global amino acid frequencies with one Gamma distribution; local amino acid frequencies (for each protein partition) with one Gamma distribution; local amino acid frequencies with one Gamma distribution for each partition.

2.4.4.2 Comparison of DistR estimates to Bayesian estimates

Bayesian estimation of the posterior distribution of the protein rates was performed using MrBayes version 3.0 (Huelsenbeck and Ronquist, 2001). Default priors were used with the JTT model of evolution plus one Gamma distribution (8 categories), one parameter for the proportion of invariant sites, and one set of branch lengths for the entire data set. This is the same model that is used for the PHYML + protein rates analysis of the data. Two runs of four chains with 300,000 iterations were performed; the burn–in used was 30,000. A further analysis of the data was performed without protein rates (using the same model) in order to compare to the concatenated PHYML analysis. Four chains were run for 150,000 iterations, with a burn–in of 15,000. Convergence of the chains was determined empirically.

2.4.4.3 Inclusion of DistR estimates into the phylogenetic tree search of PHYML

DistR rates were incorporated into the ML framework of PHYML following the proportional approach (Yang, 1996; Pupko et al., 2002b) however optimization over the rates was not performed. ML trees over the entire data set were calculated in four different ways using this modified version of PHYML. In the first analysis, the proteins were simply concatenated (equivalent to a rate of one for each protein). In the second analysis, the estimated protein rates from the real data set (based on patristic ML distances) were used for each bootstrap replicate when computing the likelihood. In the third and fourth analyses, protein rates were estimated for each bootstrap replicate using patristic and pairwise ML distances respectively. These rates were incorporated into the likelihood computation for each bootstrap replicate. Consensus trees were computed using the CONSENSE program available in the PHYLIP package (Felsenstein, 2004d)

2.5 Results and Discussion

2.5.1 Simulated Data

2.5.1.1 Patristic versus pairwise ML distances

The first simulation study demonstrates two important results: pairwise ML distances provide equally good distance estimates as patristic ML distances to the DistR method (Figure 2–3); if the fit of the initial pairwise/patristic ML distances to the data is accurate then the DistR estimates will be accurate (Figures 2–3

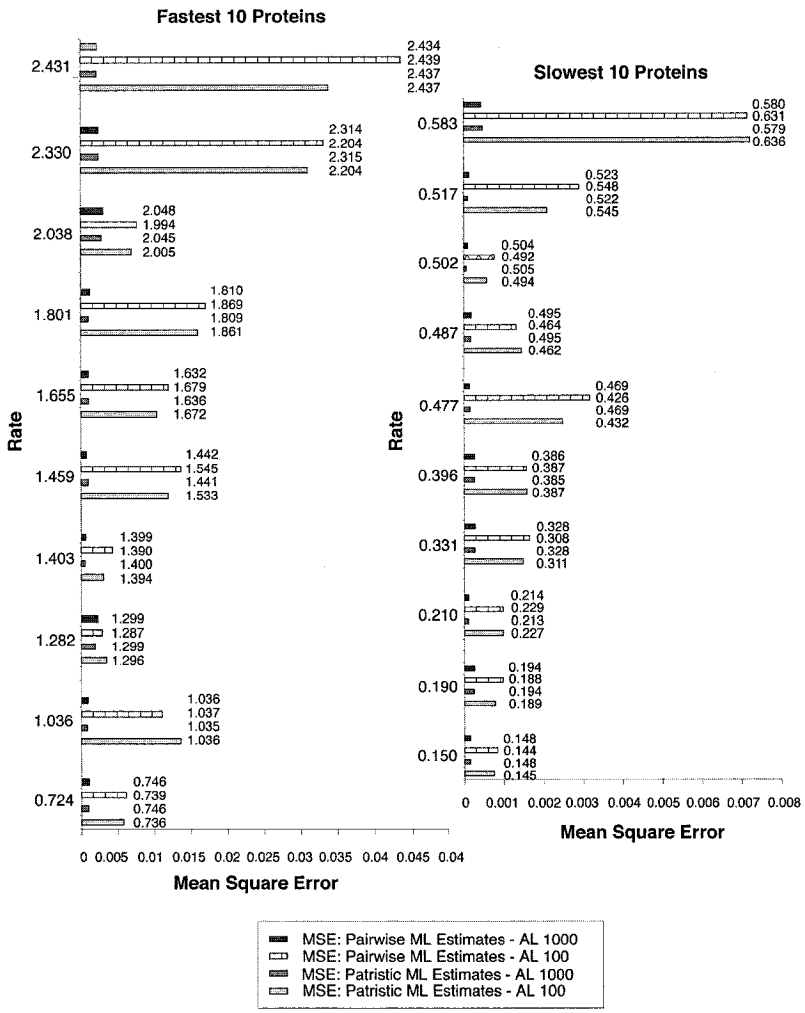


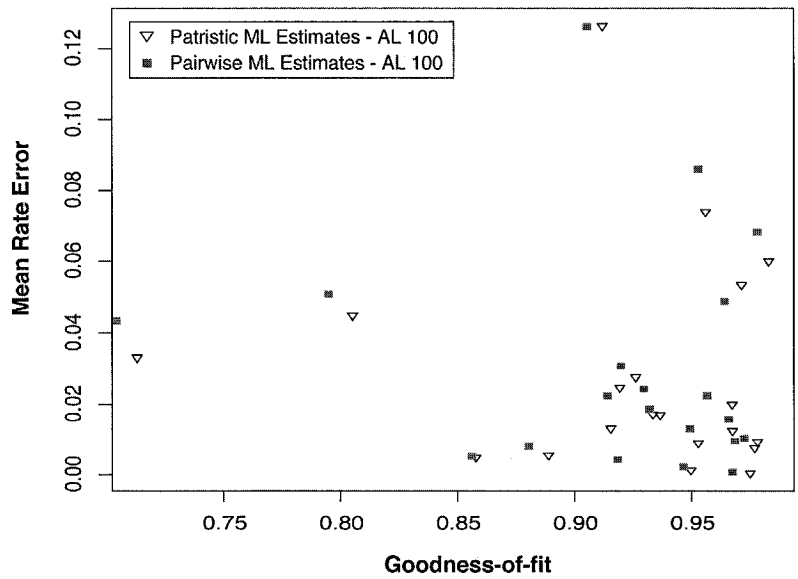
Figure 2–3: Mean squared error for different methods of distance estimation and different alignment lengths. The rates at which the data were simulated are labelled on the left hand side of the graph. The mean rate estimate for a given distance estimation method, alignment length, and rate is given on the right of the MSE bar. AL = alignment length. The ten fastest proteins are in the left–hand column. The number of species in each protein (from fastest to slowest) are: Protein 1: 53 species; Protein 2: 38 species; Protein 3: 33 species; Protein 4: 53 species; Protein 5: 38 species; Protein 6: 48 species; Protein 7: 53 species; Protein 8: 48 species; Protein 9: 43 species; Protein 10: 33 species. The ten slowest proteins are in the right–hand column. The number of species in each protein (from fastest to slowest) are: Protein 1: 33 species; Protein 2: 48 species; Protein 3: 43 species; Protein 4: 43 species; Protein 5: 48 species; Protein 6: 33 species; Protein 7: 43 species; Protein 8: 53 species; Protein 9: 38 species; Protein 10: 38 species. All rates are normalized so that the average rate is one over all 20 proteins. The total number of taxa in the data set is 58.

and 2–4). The first result is important since pairwise ML distances are very fast to compute. The second result indicates that error in the rate estimates stems principally from error in the distance estimates, rather than the DistR method itself.

The numerical results from the first experiment are summarized in Figure 2–3. The proteins are sorted in order of increasing rate, and the histogram indicates the mean squared error (MSE) over the 10 different replicates (see Appendix 1 for the exact formula used to compute MSE). Mean rate estimates are labelled to the right of each MSE bar, with the rate at which the data was simulated on the left. Results are presented only for alignments of length 100 and 1000. The results for alignments of length 300 and 500 fall in–between these two extremes. Note that the MSE increases in proportion to the rate, so results are presented on two scales.

The mean estimates for the different methods were quite close to the real rates at which the data were simulated, regardless of the alignment length, procedure used to estimate the distances, or rate at which the data was simulated (Figure 2–3). However, it is clear from the mean squared error that the DistR estimates based on shorter alignments have larger error (or greater variation), despite the fact that the mean rate estimate is often almost as accurate as that for longer alignments. Furthermore, the mean squared error tends to increase with higher rates. This is likely because the error is often in the third significant digit; for slower rates this will lead to a smaller MSE. Overall there is negligible difference between the mean and MSE statistics for a given alignment length (comparing DistR estimates based on patristic versus pairwise ML distances).

(a)



(b)

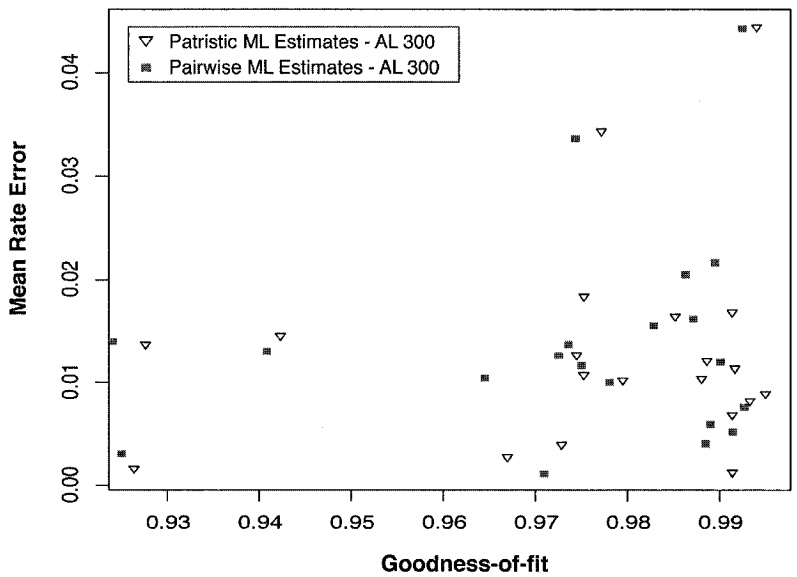


Figure 2–4: Average error of DistR rate estimates compared to goodness–of–fit of distances based upon patristic and pairwise ML distance estimates. (a) DistR rate estimates were based upon simulated proteins of length 100. (b) DistR rate estimates were based upon simulated proteins of length 300. A higher value for goodness–of–fit means that the fit of the estimated distances to the original distances is better.

Results also indicate that errors in the rate estimates are due to errors in the original distances rather than approximations introduced in the DistR method. For each protein and alignment length the error between the mean rate estimates and the real rate at which the alignments were simulated was compared to the goodness–of–fit between the estimated and true distances (Figure 2–4). This fit can be measured since the data are simulated under a known model at a particular rate. Alignments of length 100 and 300 only were examined, since the errors become negligible for longer alignments. The fit was measured using the goodness–of–fit statistic of Tanaka et al. (Tanaka and Huba, 1985), which is determined from the sum of squares error between true and estimated distances, normalized by the sum of the true distances squared. The exact formula for goodness–of–fit is presented in Appendix 1. The statistic has a maximum of one, which indicates a perfect fit.

It is expected that with longer alignments the goodness–of–fit will increase, indicating that the fit of the model to the data is better. This is clearly the case as seen when comparing goodness–of–fit for alignments of length 100 (Figure 2–4a) to that for alignments of length 300 (Figure 2–4b). The fit is further improved, and relative error reduced, with alignments of length 500 and longer (data not shown). The decrease in the goodness–of–fit (indicating a worse fit) seen with short alignment lengths indicates that the error of the method is dependent upon the error of the distance estimates and is not a property of the estimation procedure itself.

Interestingly, the error in rate estimation is in some cases less when based upon pairwise ML distances, rather than patristic ML distances. Given that the multiple sequence alignments are short (100 and 300 amino–acid residues) and include many species (at least 33 in each protein alignment), there are many trees that will fit the data equally well. Thus, there is high variation in building a ML tree to fit the original tree on which the data were simulated. Hence, estimating a ML tree with few data will likely lead to an incorrect topology. This will result in a worse fit between the original tree and the tree estimated from the alignment data. This is not true for pairwise ML distances, which do not account for topology.

2.5.1.2 Missing distances between taxa

In the previous experiment, less than half of the taxa were missing in each protein, and twenty proteins were used to estimate rates. The effects of more extreme missing taxa were also tested, where no distance estimates were present between some pairs of taxa. To achieve this, up to 75% of the taxa were removed from the starting tree. Additionally, many fewer proteins used for DistR estimation. Results indicate that the DistR method is robust to missing taxa, though having many missing taxa led to the expected increase in variance of the rate estimates.

Figure 2–5 summarizes the error in rate estimates for two simulated data sets. In the first example (Figure 2–5a) there are four protein trees, each with 16 taxa ($\approx$ 28% of the total taxon set). In the second example (Figure 2–5b) there are 8

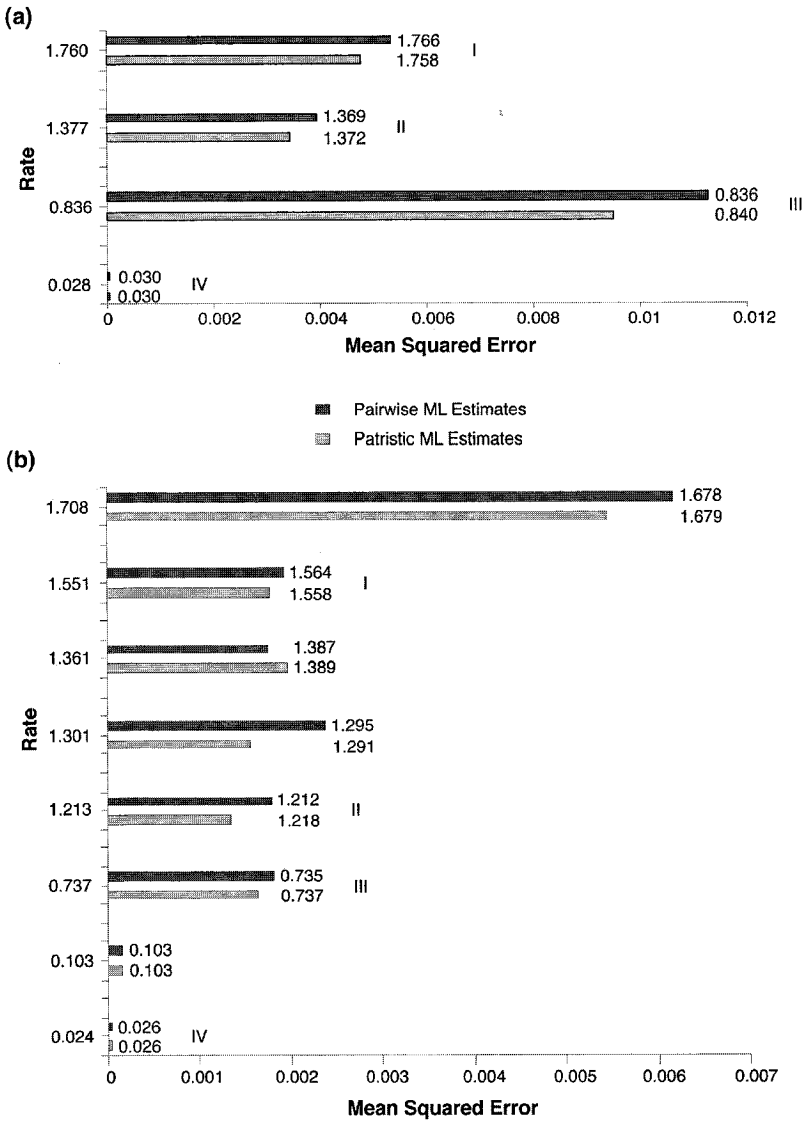


Figure 2–5: Mean squared error for different methods and different amounts of distance data. The rates at which the data were simulated are labelled on the left hand side of the graph in both (a) and (b). Mean rate estimates for both distance estimation methods are labelled on the right of the MSE bars for each protein. All rates are normalized so that the average rate is one in both (a) and (b), and are sorted from fastest to slowest. Proteins that are the same in both (a) and (b) are labelled. (a) Rate estimates based upon a data set consisting of four proteins with 16 taxa each. (b) Rate estimates based upon a data set consisting of 8 proteins; 7 with 16 taxa and one with 30 taxa.

protein trees. Seven of these have 16 taxa and the other has 30 taxa. The proteins are ordered from fastest to slowest rate in both Figure 2–5a and Figure 2–5b. Mean rate estimates are shown on the right of the MSE, and the rate at which the protein simulated (averaged to equal one) is given on the left. Simulated proteins in Figure 2–5a are labelled from I to IV. The same simulated proteins in Figure 2–5b are likewise labelled.

Once again it is evident that pairwise ML distances and patristic ML distances give almost identical average relative rate estimates (to within two or three decimal places). Furthermore, the missing data have little effect on mean rate estimates, but does have a large effect on the variance. For instance, comparing the MSE for the first protein in Figure 2–5a to that of the second protein in Figure 2–5b (it is the same simulated protein) it is clear that although the mean rate estimate is approximately as accurate with more taxa (Figure 2–5b), the MSE is clearly smaller when more distances between a pair of taxa are included in the analysis. Thus it is evident that more data in terms of pairwise distances between taxa (over multiple proteins) will reduce the error of the DistR estimate.

Calculation of the relative rates within groups of the same number of species was also performed (i.e. proteins with 16 species, proteins with 30 species, and proteins with 44 species). For each subset of proteins mean rate estimates based on pairwise ML distances were slightly worse or identical to those based on patristic ML distances (data not shown). In addition the variances were greater in general for rates estimated based on pairwise ML distances. The major difference between the three analysis was that the variance of the rate estimates was lower when more

species were included in the analysis. Furthermore the mean rate estimates were slightly more accurate for the data sets over larger taxon groups (data not shown).

Accuracy in spite of missing taxa demonstrates that the rate estimation procedure is consistent (assuming that the initial distance estimates are accurate), regardless of the number of proteins under analysis. This is because rates are not computed relative to the distance estimates of one protein. Rather, they are constrained by all the distance estimates. Thus, if one set of distance estimates is extremely biased with respect to the remainder of the distances they will not have a strong effect on the final rate estimates.

2.5.2 Empirical Data

2.5.2.1 Comparison of DistR estimates to ML estimates

Rates were calculated in a ML framework using only those proteins that are present over the entire species set (Atp6, Cob, Cox1, Cox2, and Cox3) due to a constraint of the program COMBINE (Pupko et al., 2002b). Table 2–3 shows the time for rate estimation and rate estimates based on different models under the ML framework in comparison to DistR estimates based on pairwise and patristic ML distances. Two sets of ML estimates are given for each model. The first based upon the concatenated tree, and the second on the DistR incorporated ML tree. DistR estimates are computed far more rapidly and are still accurate in comparison to ML estimates. In comparison to the 6 ML estimates the DistR rates based on patristic ML distances are slight overestimates for Cob and Cox1, and

slight underestimates for Cox2 and Cox3. The estimate for Atp6 is an average of the 6 ML estimates (Table 2–3). Notably the patristic DistR estimates for Cob and Cox1 are closest to the ML estimates based on the rate–incorporated tree using global amino acid frequencies plus the one–Gamma–distribution model. Conversely, the DistR estimates for Cox2 and Cox3 are closest to the ML estimates based on the same tree, using local amino acid frequencies and the five–Gamma–distribution model. The DistR estimates based on pairwise ML distances are quite close to those based on patristic ML distances, except for Atp6 and Cox3. Atp6 has a much higher rate—quite close to the ML estimate for the LF + 5–GAM model where the estimates were based on the rate–incorporated ML tree. However, the Cox3 estimate is quite low compared to all ML estimates; Cox3 had a higher variation in rate estimation over all proteins (Table 2–5), a case where perhaps the lack of topological information decreases the accuracy of the DistR estimate. Clearly this is not an issue for most proteins, but can be an issue for some. Overall it appears that the DistR estimates are model independent regardless of distance estimation procedure and provide excellent first approximations to the ML estimates.

2.5.2.2 Comparison of DistR estimates to Bayesian estimates

The posterior distribution of rates from MrBayes is shown in Figure 2–6. For all but three of the proteins the DistR estimates fall within the 95% posterior credible interval for the protein rate. Each of Nad6, Cox1 and Cox3 have DistR estimates that do not fall between the 95% posterior credible interval. Both Cox1

Comparison of ML rate estimates to DistR estimates

Method	Time	Atp6	Cob	Cox1	Cox2	Cox3
GF + 1–GAM	776s	1.24	0.81	0.62	0.99	1.34
		1.25	0.81	0.63	0.99	1.33
LF + 1–GAM	842s	1.35	0.80	0.61	0.94	1.31
		1.36	0.80	0.62	0.93	1.30
LF + 5–GAM	648s	1.36	0.79	0.59	0.94	1.31
		1.39	0.78	0.61	0.92	1.30
DistR Pat	0.116s	1.32	0.83	0.66	0.91	1.29
DistR Pair	0.122s	1.40	0.83	0.64	0.96	1.18

Table 2–3: Comparison of relative rate estimates and estimation time from COMBINE and DistR for five proteins (Atp6, Cob, Cox1, Cox2, and Cox3) from the fungal data set. For each model, rates based upon the maximum likelihood concatenated tree from PHYML are given on the first line, and rates based upon the maximum likelihood tree incorporating DistR rates (computed in PHYML) are given on the second. All estimates were normalized so that the average rate is one. GF = global amino acid frequencies, LF = local amino acid frequencies (calculated for each protein), 1–GAM = one Gamma distribution estimated for the entire data set, 5–GAM = one Gamma distribution for each protein, DistR Pat = DistR estimation using patristic ML distances, DistR Pair = DistR estimation using pairwise ML distances.

and Cox3 have average sequence lengths, and 29 taxa each. Nad6 is shorter at less than 100 amino acids, with only 24 species. In the case of Nad6 perhaps the short sequences length contributes to uncertainty in the DistR estimates. However it is unlikely that the Bayesian posterior distributions of the rates are accurate. This conclusion is based upon the fact that the four chains were mixing quite poorly in both runs even after 300 000 iterations (data not shown). Sampling from the posterior distribution is unlikely to be correct since the chain might be over–sampling from areas of low likelihood. Comparison of the tree of the highest likelihood from this analysis to the tree of highest likelihood based on the concatenated data indicates that MrBayes was in a suboptimal topological space

when sampling rate estimates (using the Bayesian information criterion, data not shown). Furthermore, the DistR ML tree is a significantly better fit of the model to the data based on the AIC (Felsenstein, 2004c) when compared to the likelihood of the MrBayes rate incorporated tree as computed in PHYML. Thus, although the posterior distribution of the rates appears reasonable, the chain seems to be having difficulty sampling through topology space.

Thus, it appears that the proportional model under MrBayes, when used without different parameters for each partition (as in Nylander et al., 2004), does not search tree space as well as PHYML with the rate multipliers included. Perhaps this is due to an incorrect prior on the rate parameters used. If this is the problem the DistR method can certainly be used to find a distribution of the rates of proteins, which could be used as the prior on these parameters. The discrepancy could also be due to the different search heuristics used in MrBayes. Given the computational complexity of the search, it might be difficult for the program to search for the best rate parameters while also searching for the best topology.

2.5.2.3 Patristic versus pairwise ML distances

The relative protein rates of the real data are unknown. However the variance of the rate estimates using both patristic and pairwise ML distances can be compared, a smaller estimate being preferable. Contrary to expectations, but confirming the simulation studies, rate estimates from pairwise ML distances had smaller variance than rate estimates from patristic ML distances.

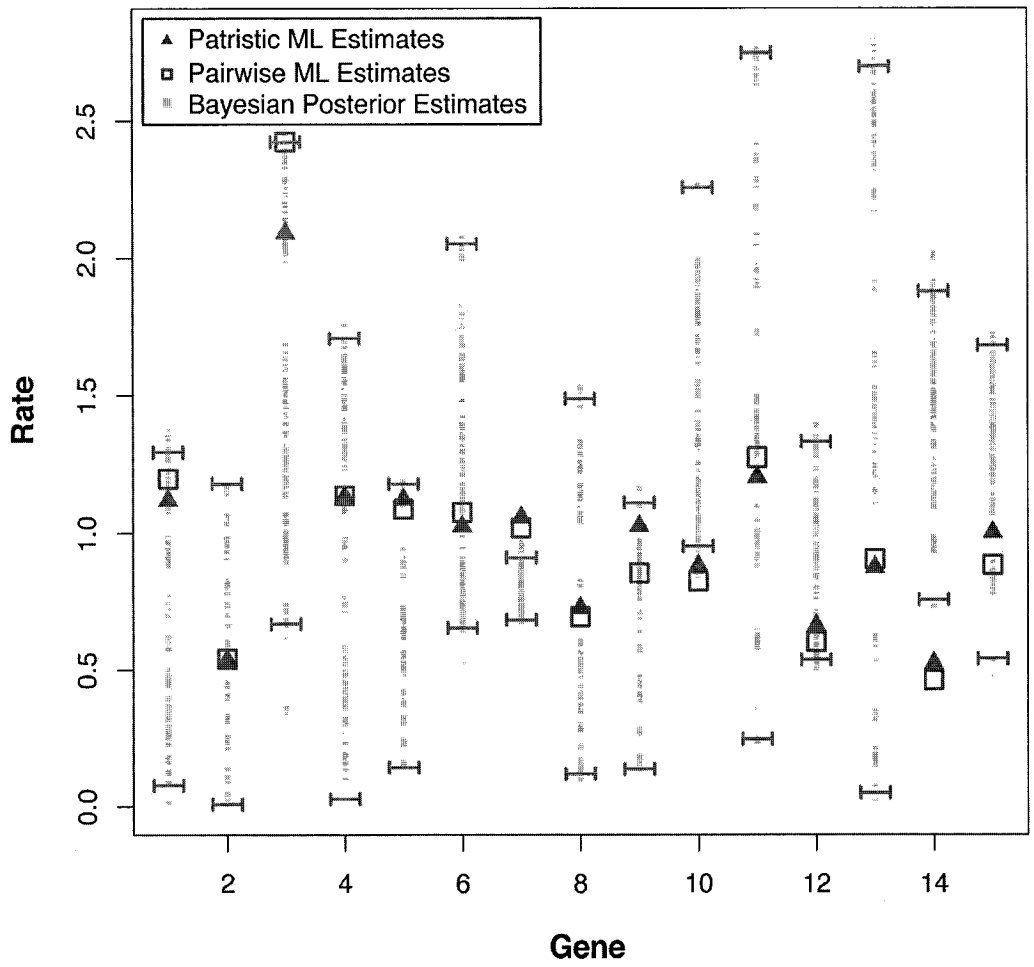


Figure 2–6: Distribution of rates from the MrBayes proportional model analysis compared to DistR estimates. Bars at either end represent the 95% credible interval. The DistR estimate based upon patristic ML distances is marked by a solid triangle. The DistR estimate based upon pairwise ML distances is marked by a square. The posterior rate estimates of MrBayes are given by a solid square. DistR estimates are normalized so that the average rate is one (as in MrBayes). Proteins are ordered from shortest to longest as follows: Atp8, Atp9, Rps3, Nad3, Nad4, Nad4L, Nad6, Atp6, Cox2, Cox3, Nad1, Nad2, Cob, Nad4, Cox1, and Nad5.

Variances of the rate values computed were estimated by non–parametric bootstrap of the protein alignments, and re–estimation of the distances and

DistR estimates for empirical data based on pairwise and patristic ML distance estimates

Protein	Acc. #	# Species	AL	Patristic Mean	Patristic Var x 10^{-3}	Pairwise Mean	Pairwise Var x 10^{-3}
Atp8	ALIGN_000885	28	32	1.08	8.68	1.15	11.8
Atp9	ALIGN_000886	26	73	0.55	5.12	0.55	4.35
Rps3	ALIGN_000900	11	77	2.02	41.1	2.33	31.5
Nad3	ALIGN_000893	24	79	1.13	8.82	1.15	10.1
Nad4	ALIGN_000894	24	424	1.14	3.52	1.10	2.76
Nad4L	ALIGN_000895	23	85	0.87	5.91	0.91	6.45
Nad6	ALIGN_000897	24	96	1.05	7.214	1.10	7.80
Atp6	ALIGN_000884	29	203	1.07	3.76	1.03	4.07
Cox2	ALIGN_000889	29	220	0.75	3.81	0.71	2.98
Cox3	ALIGN_000890	29	245	1.05	4.75	0.86	3.24
Nad1	ALIGN_000891	24	294	0.89	2.61	0.84	2.30
Nad2	ALIGN_000892	23	313	1.21	2.16	1.29	2.69
Cob	ALIGN_000887	29	375	0.67	1.17	0.61	1.04
Cox1	ALIGN_000888	29	487	0.53	1.76	0.46	.749
Nad5	ALIGN_000896	24	520	1.01	2.79	0.89	1.94

Table 2–5: Mean rate estimates and variances (Var) for rate estimates based upon bootstrap replicates over the fungal data set. Rates are normalized so that the average rate is one. Acc. # = Accession number for the alignment in EMBL. AL = alignment length. Patristic refers to rates estimated based on distances from maximum likelihood trees. Pairwise refers to rates estimated based on maximum likelihood distances.

DistR rates for each bootstrap data set. The mean and variance of the DistR estimates for pairwise and patristic ML distances show some interesting trends (Table 2–5). In general, the average rate estimates were similar, with the notable exception of Atp8, Cox3, and Rps3 (and to a lesser extent Nad2, Nad5, and Nad6). Ten of the 15 protein rates derived from patristic ML distances had greater variance than their counterparts derived from pairwise ML distances. (Table 2–5).

These results support the conclusion that introducing topology into the distance estimation procedure is not likely to lead to better distances estimates for the DistR procedure when so many taxa are involved and the alignments are short. This is a consequence of the large number of distinct trees that can fit a short alignment equally well.

2.5.2.4 Inclusion of DistR estimates into phylogenetic tree search of PHYML

The experimental results when DistR estimates are incorporated into the ML tree search demonstrate the importance of accounting for different evolutionary pressures in phylogenetic inference.

Bootstrap support values for the ML tree using concatenated data are presented in Figure 2–7a. The bootstrap support for some of the clades was quite weak. Incorporating DistR estimates based upon both patristic and pairwise ML distances into the tree search led to the same ML tree, presented in 2–7b. Overall, bootstrap support was improved in most clades when DistR estimates were incorporated into the tree search.

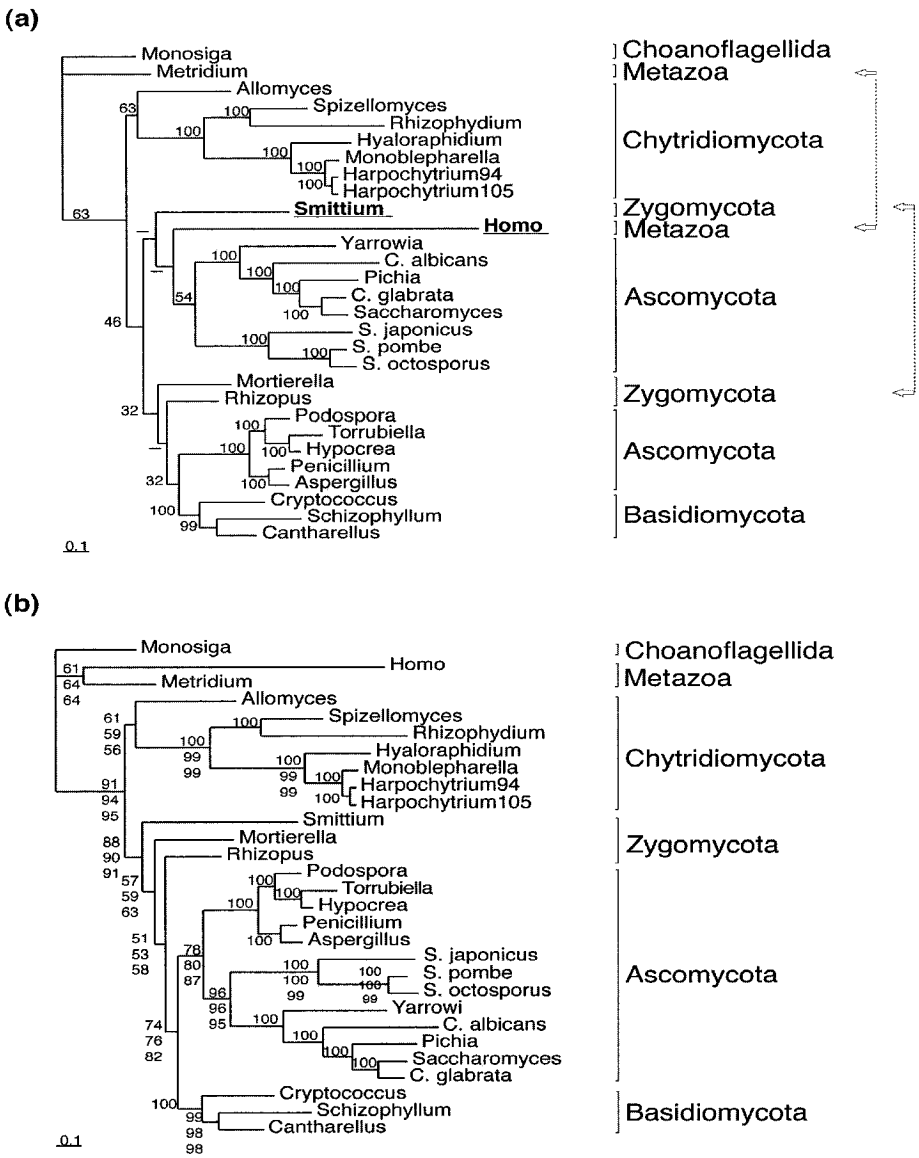


Figure 2–7: Please see caption on following page.

Figure 2–7 (caption): (a) Phylogenetic analysis based upon the mitochondrial data set. The topology shown was inferred using PHYML without DistR protein rates, using the JTT model of protein evolution, with 8 Gamma categories, and ML estimation of the alpha parameter of the Gamma distribution and the proportion of invariant sites. It was constructed using the concatenated 'unambiguously' aligned proteins. Bootstrap support for this topology was computed based upon 100 replicates. The percentage of support for each clade is given at the root of the clade. In cases where the consensus tree differed from the maximum likelihood topology a '–' is written. (b) Phylogenetic analysis based upon mitochondrial data set. The topology shown was inferred using PHYML with DistR protein rates, using the JTT model of protein evolution, with 8 Gamma categories, and ML estimation of the alpha parameter of the Gamma distribution and the proportion of invariant sites. It was constructed using the concatenated 'unambiguously' aligned proteins and protein rate estimates. The percentage of support for each clade is given. Bootstrap support for this topology was computed based upon 100 replicates, using three different methods. The top numbers give the percentage of support based upon using the patristic ML distance DistR estimates from the real data as rate values in computing the ML tree for each bootstrap replicate. The middle numbers give the percentage of support based upon re–estimating DistR estimates for each bootstrap replicate using patristic ML distances. The bottom numbers give the percentage of support based upon re–estimating DistR estimates for each bootstrap replicate using pairwise ML distances. When bootstrap support was the same for each method of incorporating rates it is given only once.

The topology of the ML concatenation–based tree does not separate Zygomycota and Ascomycota as distinct clades, which is not surprising because the Zygomycota are traditionally difficult to place. Furthermore, the out–group is incorrect since it should also contain *Homo sapiens* (which groups incorrectly with the zygomycete *Smittium* and the Ascomycota). This long–branch–attraction problem is due to the highly derived *Smittium* and *Homo* sequences. Using DistR estimates improves the bootstrap support in certain clades, and corrects the most evident topological problems, notably that Zygomycota more accurately group together (although as an unresolved paraphyletic group). Indeed, almost every branch that does not show 100% bootstrap support with the concatenated data has improved support when using protein rates. The only branching where support somewhat lessened from the concatenated to the protein–rate–based trees (and with using individual bootstrap rates) was the branching of *Allomyces* (a species that is difficult to place whatever the method or data set) with the remainder of the Chytridiomycota (Figure 2–7a and 2–7b). Bootstrap support is strongest when using protein rates based upon pairwise ML distances, where the rate estimates were re–computed for each bootstrap replicate. This is perhaps because the variation in the pairwise ML distance rate estimates was smaller than, or on the same order of magnitude as, the rate estimates based on patristic ML distances.

Both the Kishino–Hasegawa (KH) test and Akaike Information Criterion (AIC) support the ML topology with protein rates as a better fit for the model to the data than the concatenated topology. Under the Kishino–Hasegawa test (Kishino and Hasegawa, 1989; Shimodaira and Hasegawa, 2001) the concatenated

topology was significantly worse than the DistR topology ($P < 0.0001$) when the topology was computed with rate estimates calculated based on both patristic and pairwise ML distances. The AIC provides a statistical measurement of the significance of the change in log–likelihood when using two different models to fit the data. The measure compensates for the increase in the number of parameters in the rates model. When DistR estimates based on pairwise ML distances are used the AIC is 1043.65182 greater than the AIC for a single rate, concatenated analysis. When patristic ML distances are used for rate estimation the increase in AIC over the concatenated analysis is 1068.7542. Both increases in AIC are very substantial, indicating that important information in the data that is disregarded by traditional concatenated analysis, is captured by modelling protein rates.

2.6 Conclusion

A fast and accurate method to calculate the rates of partitioned data sets is presented. Although the analyses performed here are based upon protein sequence data, using nucleotide sequences should prove as effective. The error in the method is largely due to incorrect initial distance estimates for the proteins, which tend to be worse with smaller or poorly conserved sequences. Using pairwise ML distances for DistR estimation is just as accurate as using patristic ML distances. The estimates are accurate when compared to ML estimates and Bayesian posterior credible intervals for the rates. Incorporating the DistR estimates into PHYML leads to statistically better likelihood and topology.

2.7 Acknowledgements

We thank Scott Bunnell, Alain Vandal, Tal Pupko, Tim Collins, and Olivier Gascuel for helpful comments on the manuscript. Thanks to Stéphane Guindon for kindly providing the source code of PHYML v2.2 for our use. Salary and support from the Canadian Institutes of Health Research (MOP 42475; BFL), the Canadian Institute for Advanced Research (CIAR; BFL), National Science and Engineering Research Council (NSERC grant 238975–01; DB), Fonds de recherche sur la nature et les technologies (FQRNT grant 2003–NC–81840; DB), and supply of laboratory equipment and informatics infrastructure by Genome Canada is gratefully acknowledged. RBB is supported by an NSERC PGS–B scholarship.

2.8 Appendix

2.8.1 Appendix 1—Formula for mean squared error and goodness–of–fit

Mean squared error is used to describe the accuracy of rate estimates. Since only relative rates can be computed rates are normalized so that the average rate over all proteins is one. Let $\tilde{r}$ denote the true rate (that is, the rate used in simulations), and let $\hat{r}_1, ..., \hat{r}_{10}$ be the rates estimated in the 10 replicates of the experiment. The mean squared error (MSE) is defined as

$$\frac{1}{10}\sum_{i=1}^{10}(\tilde{r}-\hat{r}_i)^2.$$

Goodness–of–fit is used to measure the fit of the distance estimates to the distances in the tree used for simulation. There is a slight problem with scales since Pseq–Gen treats branch lengths as the expected number of substitutions per 100 sites while PHYML treats branch lengths as the expected number of substitutions per site. Let $\tilde{d}_{xy}^{(k)}$ be the distance between x and y in the tree used to simulate protein k, let r_k denote the rate used when simulating protein k, and let $\hat{d}_{xy}^{(k)}$ be the distance estimated by PHYML.

Given the differences in scale the goodness–of–fit measure used was

$$1.0 - \frac{\sum_{xy}\left(r_k\tilde{d}_{xy}^{(k)} - 100\hat{d}_{xy}^{(k)}\right)^2}{\sum_{xy}\left(r_k\tilde{d}_{xy}^{(k)}\right)^2}.$$

Note that the goodness–of–fit is at most one, and equals one if and only if there is a perfect fit.

2.8.2 Appendix 2—Fast algorithm for least squares estimation

This appendix shows how to quickly determine the vectors $\mathbf{p}$ and $\mathbf{r}$ that minimize the function $q(\mathbf{p}, \mathbf{r})$ in equation (2.3)

$$q(\mathbf{p}, \mathbf{r}) = \sum_{k=1}^{n}\sum_{x,y} w_{xy}^{(k)}\left(p_{xy} - \frac{d_{xy}^{(k)}}{r_k}\right)^2$$

subject to the constraint that $h(\mathbf{p}) = \kappa$, where

$$h(\mathbf{p}) = \sum_{k=1}^{n}\sum_{x,y} w_{xy}^{(k)} p_{xy}$$

and κ is an arbitrary, positive constant. In the implementation of DistR

$$\kappa = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} d_{xy}^{(k)}$$

which corresponds to the assumption that the unknown consensus distances are roughly centered on the average of the observed distances. This value can be computed in $O(nm^2)$ time for n proteins and m taxa. Any other positive constant will work, as the only effect is to change the scale of the rate estimates.

To simplify the mathematics substitute $s_k = \frac{1}{r_k}$ for each $k = 1, \ldots, n$. Let $\mathbf{s}$ denote the vector $[s_1, \ldots, s_n]^T$. Minimizing $q(\mathbf{p}, \mathbf{r})$ is then equivalent to minimizing

$$f(\mathbf{p}, \mathbf{s}) = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} \left(p_{xy} - s_k d_{xy}^{(k)}\right)^2. \tag{2.5}$$

Recall from calculus that the minimum of a one dimensional function can be found by determining where the first derivative is equal to zero. This condition extends to multi–dimensional functions with constraints. Refer to (Gill et al., 1982) for an excellent introduction to the optimization tools used here.

If $(\mathbf{p}, \mathbf{s})$ together minimize the function f, subject to the condition that $h(\mathbf{p}) = \kappa$, then there exists a real number λ such that

$$\begin{aligned} \frac{\partial f(\mathbf{p},\mathbf{s})}{\partial p_{xy}} + \lambda \frac{\partial h(\mathbf{p})}{\partial p_{xy}} &= 0 \quad \text{for all taxa } x, y \\ \frac{\partial f(\mathbf{p},\mathbf{s})}{\partial s_k} &= 0 \quad \text{for all proteins } k \\ h(\mathbf{p}) &= \kappa. \end{aligned} \tag{2.6}$$

In general, (2.6) is only a necessary condition for reaching the minimum, and not a sufficient condition. However in this case the matrix formed from the second

derivatives of $f(\mathbf{p}, \mathbf{s})$ is *positive definite*, so that the function f is convex (Gill et al., 1982). It follows that if $(\mathbf{p}, \mathbf{s})$ and λ satisfy (2.6) then $(\mathbf{p}, \mathbf{s})$ gives the global minimum.

It is possible to derive the partial derivatives of the functions f and h explicitly. To help with notation define the quantities:

$$\begin{aligned}
\alpha_k &= \textstyle\sum_{xy} 2w_{xy}^{(k)} \left(d_{xy}^{(k)}\right)^2 && \text{for all proteins } k; \\
\beta_{xy} &= 2\textstyle\sum_{k=1}^{n} w_{xy}^{(k)} && \text{for all taxa } x, y; \\
\beta_{xy,k} &= -2w_{xy}^{(k)} d_{xy}^{(k)} && \text{for all proteins } k \text{ and taxa } x, y.
\end{aligned}$$

The partial derivative of f with respect to s_k, for some protein k, is

$$\begin{aligned}
\frac{\partial f(\mathbf{p}, \mathbf{s})}{\partial s_k} &= \sum_{xy} -2w_{xy}^{(k)} \left(p_{xy} - d_{xy}^{(k)} s_k\right) d_{xy}^{(k)} \\
&= \alpha_k s_k + \sum_{xy} \beta_{xy,k} p_{xy}.
\end{aligned}$$

The partial derivatives of f and h with respect to p_{xy}, for some taxa x, y, are

$$\begin{aligned}
\frac{\partial f(\mathbf{p}, \mathbf{s})}{\partial p_{xy}} &= \sum_{k=1}^{n} 2w_{xy}^{(k)} \left(p_{xy} - d_{xy}^{(k)} s_k\right) \\
&= \sum_{k=1}^{n} \beta_{xy,k} s_k + \beta_{xy} p_{xy} \\
\frac{\partial h(\mathbf{p})}{\partial p_{xy}} &= \sum_{k=1}^{n} w_{xy}^{(k)} \\
&= \beta_{xy}/2.
\end{aligned}$$

Note from the partial derivatives that the conditions in equation (2.6) are linear equations involving the entries of $\mathbf{p}$, $\mathbf{s}$, and λ. As such, the next step is to rewrite (2.6) in terms of matrix algebra. Given that there are n proteins and m

taxa define the following: let D be the $n \times n$ matrix with $\alpha_1, \alpha_2, \ldots, \alpha_n$ down the diagonal and zeros off the diagonal; let C be the $\frac{m(m-1)}{2} \times \frac{m(m-1)}{2}$ matrix with $\beta_{12}, \beta_{13}, \ldots, \beta_{(m-1)m}$ down the diagonal and zeros off the diagonal; let B be the $\frac{m(m-1)}{2} \times n$ matrix with rows indexed by unique pairs of taxa, columns indexed by proteins, and the entry corresponding to row xy and column k equal to $\beta_{xy,k}$; let $\mathbf{v}$ be the $\frac{m(m-1)}{2}$ dimensional vector $\mathbf{v} = \frac{1}{2}[\beta_{12}, \beta_{13}, \ldots, \beta_{(m-1)m}]^T$.

The conditions in equation (2.6) can now be rewritten as

$$D\mathbf{s} + B^T\mathbf{p} = 0 \tag{2.7}$$

$$B\mathbf{s} + C\mathbf{p} + \mathbf{v}\lambda = 0 \tag{2.8}$$

$$\mathbf{v}^T\mathbf{p} = \kappa. \tag{2.9}$$

Define

$$\begin{aligned} \mathbf{u} &= B^T C^{-1}\mathbf{v} \\ \omega &= \mathbf{v}^T C^{-1}\mathbf{v}. \end{aligned}$$

Solving for $\mathbf{p}$ in (2.8) gives:

$$\mathbf{p} = C^{-1}(-B\mathbf{s} - \mathbf{v}\lambda) \tag{2.10}$$

Substituting this into (2.9) and solving for λ gives:

$$\begin{aligned} \lambda &= \frac{\kappa + \mathbf{v}^T C^{-1} B\mathbf{s}}{-\mathbf{v}^T C^{-1}\mathbf{v}}. \\ &= \frac{\kappa + \mathbf{u}^T\mathbf{s}}{-\omega}. \end{aligned}$$

Replacing λ with the above equation in (2.10) provides a solution for $\mathbf{p}$ in terms of the above defined matrices, vectors and $\mathbf{s}$ (i.e. there are no longer any unknowns except for $\mathbf{p}$ and $\mathbf{s}$):

$$\begin{aligned} \mathbf{p} &= C^{-1}\left(-B\mathbf{s} + \mathbf{v}\frac{\kappa + \mathbf{u}^T\mathbf{s}}{\omega}\right) && (2.11) \\ &= C^{-1}\left(\frac{\mathbf{v}\mathbf{u}^T}{\omega} - B\right)\mathbf{s} + \frac{\kappa}{\omega}C^{-1}\mathbf{v}. && (2.12) \end{aligned}$$

Finally, substitute (2.12) into (2.7) to get

$$\begin{aligned} 0 &= D\mathbf{s} + B^T\mathbf{p} \\ &= \left(D + \frac{\mathbf{u}\mathbf{u}^T}{\omega} - B^TC^{-1}B\right)\mathbf{s} + \frac{\kappa}{\omega}\mathbf{u}. \end{aligned}$$

Let

$$M = \left(D + \frac{\mathbf{u}\mathbf{u}^T}{\omega} - B^TC^{-1}B\right).$$

Then, $\mathbf{s}$ is found by solving the equation:

$$M\mathbf{s} = -\frac{\kappa}{\omega}\mathbf{u}. \qquad (2.13)$$

Consensus distances $\mathbf{p}$ are obtained by substituting $\mathbf{s}$ into equation (2.12).

The entire computation is summarized in Appendix 3. The running time of the algorithm is $O(nm^2+n^3)$ which is time optimal. The algorithm uses $O(n^2+m^2)$ memory in addition to the $O(nm^2)$ required to store the distance estimates $d_{xy}^{(k)}$.

There are two complications that can arise in the above calculations. Firstly, it could be the case that for a particular pair of taxa x, y, there is no single protein that contains both x and y. This means that β_{xy} is undefined, so that C is no

longer invertible. This problem is easily solved. If there is no protein with both x and y then the line in (2.6) involving the partial derivative with respect to p_{xy} is satisfied trivially. Therefore, the row and column of C, the row of B, and entry of $\mathbf{v}$ indexed by the pair x, y can be removed. The reduced problem can be solved as before, although no estimate for p_{xy} is obtained. Row removal is handled in the pseudo–code for the algorithm given in Appendix 3 by using constraints in the summations.

The second complication is that the optimization problem might have more than one solution, in which case the matrix M in (2.13) will not be invertible. This indicates that more information is required to estimate the relative rates, as would arise, for example, in a concatenation of two protein alignments over entirely different sets of taxa.

2.8.3 Appendix 3—The DistR Algorithm

Algorithm DISTR($\mathbf{d}, \mathbf{w}$)

Input: Distance estimates $d_{xy}^{(k)}$ for each pair of taxa and each protein k.
Weights $w_{xy}^{(k)}$ for each distance estimate.
Missing distances have weight zero.

Output: Rate estimates $\mathbf{r}$. Consensus distances $\mathbf{p}$.

$\kappa = \sum_{k=1}^{n} \sum_{xy} w_{xy}^{(k)} d_{xy}^{(k)}$
for k from 1 to n **do**
  $\alpha_k \leftarrow \sum_{xy} 2 w_{xy}^{(k)} (d_{xy}^{(k)})^2$
  for all taxa x, y **do**
    $\alpha_{k,xy} \leftarrow -2 w_{xy}^{(k)} d_{xy}^{(k)}$
    $\beta_{xy,k} \leftarrow -2 w_{xy}^{(k)} d_{xy}^{(k)}$
for all taxa x, y **do**
  $\beta_{xy} \leftarrow 2 \sum_{k=1}^{n} w_{xy}^{(k)}$
$\omega \leftarrow \frac{1}{4} \sum_{xy} \beta_{xy}$
for k from 1 to n **do**
  $\mathbf{u}_k \leftarrow \sum_{xy} \beta_{xy,k}$
for k from 1 to n **do**
  $\mathbf{z}_k \leftarrow -\frac{\kappa}{\omega} \mathbf{u}_k$.
  for l from 1 to n **do**
    $M_{kl} \leftarrow -\sum_{xy:\beta_{xy} \neq 0} \frac{\beta_{xy,k} \beta_{xy,l}}{\beta_{xy}} + \frac{1}{\omega} \mathbf{u}_k \mathbf{u}_l$
    if $k = l$ **then** $M_{kl} \leftarrow M_{kl} + \alpha_k$
if M is non-singular **then output** 'Insufficient data to estimate rates'
solve $M\mathbf{s} = -\frac{\kappa}{\omega} \mathbf{u}$ to obtain $\mathbf{s}$
for all taxa x, y such that $\beta_{xy} \neq 0$ **do**
  $\mathbf{p}_{xy} \leftarrow \sum_k \left(\frac{\mathbf{u}_k}{2\omega} - \frac{\beta_{xy,k}}{\beta_{xy}} \right) \mathbf{s}_k + \frac{\kappa}{2\omega}$
for k from 1 to n **do**
  $r_k \leftarrow \frac{1}{s_k}$
output $\mathbf{r}$ and $\mathbf{p}$

Chapter 3
Accounting for gene rate heterogeneity in phylogenetic inference

3.1 Background

Gene rate heterogeneity is traditionally accounted for in phylogenetic models by allowing for each gene to have a rate of evolution. However, it is possible to account for gene rate heterogeneity similarly to site rate heterogeneity. In this later approach a distribution over gene rates is assumed, which reduces the number of parameters in the model. However, this approach is computationally much slower.

This chapter compares the two methods of accounting for gene rate heterogeneity, using the Akaike Information Criterion (AIC) and Cross Validation Information Criterion (CVIC). Analysis is performed to determine: (i) which method is best according to the AIC and CVIC; (ii) the amount of data required for the two methods to converge to the same ML topology in PHYML; (iii) what properties of the data lead to an improved model fit over the concatenated model according to the AIC.

3.2 Abstract

Traditionally, phylogenetic analyses over many genes combine data into a contiguous block. Under this concatenated model all genes are assumed to evolve at the same rate. However, it is clear that genes evolve at very different rates, and that accounting for this rate heterogeneity is important if we are to accurately infer phylogenies from heterogeneous multi–gene data sets. There remain open questions regarding how best to incorporate gene rate parameters into phylogenetic models and which properties of real data correlate with improved fit over the concatenated model. In this study, two methods of accounting for gene rate heterogeneity are compared: the n–parameter method and the α–parameter methods. The former approach allows for each of the n gene partitions to have a gene rate parameter and the later fits a distribution to the gene rates. Results demonstrate that the n–parameter method is both computationally faster and in general provides a better fit over both the concatenated model than the α–parameter method. Furthermore, improved model fit over the concatenated model is highly correlated with the presence of a slow relative rate.

3.3 Introduction

The use of multi–gene data sets in phylogenetic analysis is imperative in order to resolve evolutionary relationships over large taxon sets and deep phylogenetic divergences. Multi–gene data sets have the advantage of greater resolution—with more information it is possible to find better trees (see for instance Gontcharov

et al., 2004). However the heterogeneous nature of the data does present problems. When there are many genes present in the analysis, it is necessary to account for the fact that different genes undergo different selective pressures, and that the degree of site rate heterogeneity within a gene may vary from gene to gene. The incorporation of data under different evolutionary pressures (as found in different codon positions, or different genes) should be taken into account when calculating likelihoods (Bull et al., 1993; Huelsenbeck et al., 1996; Yang, 1996; Bapteste et al., 2002; Pupko et al., 2002b; Nylander et al., 2004; Cranston and Rannala, 2005).

Determining how best to incorporate gene rates in a maximum likelihood (ML) context is a relatively unexplored area of phylogenetics research. When incorporating gene rates into maximum likelihood phylogeny estimation there are two approaches that can be taken. The first approach involves using a single rate for each gene (hereafter referred to as the n–parameter method) as initially proposed by Yang et al. for DNA sequences (Yang, 1996) and extended by Pupko et al. to protein data (Pupko et al., 2002b). This approach to accounting for gene rate heterogeneity has been shown to lead to better model fit according to the Akaike Information Criterion (AIC) (Yang, 1996; Pupko et al., 2002b; Bevan et al., 2005) and in some cases better inference of the preferred tree topology (Bevan et al., 2005). The second approach involves integrating out over all possible rates for a given gene using a discrete approximation to a continuous distribution (hereafter referred to as the α–parameter method) (Felsenstein, 2001, 2004b).

Both approaches to accounting for gene rate heterogeneity assume that a gene evolves at a particular rate of evolution. However, the n–parameter

method does not allow for any uncertainty in the rate for a particular gene, but assumes that it is valid to use ML estimates of the rate to account for gene rate heterogeneity. Conversely, the α–parameter method does account for uncertainty in the rate estimate for a gene through integration over all possible values that a rate might take. Although assuming a single rate for each gene is computationally faster than integrating, it could potentially suffer from the difficulty of 'infinite parameterization' when many genes are used in the analysis (thus over–fitting the data since there are more degrees of freedom). While the n–parameter method has been tested, and found to lead to significant improvement in maximum likelihood phylogeny estimation (Yang, 1996; Pupko et al., 2002b; Bevan et al., 2005), the α–parameter method has yet to be investigated.

In addition to determining how best to incorporate gene rate parameters, the question remains of what correlates with improved fit over the concatenated model. This paper has two goals: to determine whether the computational effort required by the α–parameter method is justified according to the Akaike Information Criterion (AIC) (and Cross Validation Information Criterion (CVIC)); to determine the properties of the data that lead to an improved fit when accounting for gene rate heterogeneity in phylogenetic models.

3.4 Materials and Methods

3.4.1 The n–parameter method

The n–parameter method is well studied and has been shown to lead to better likelihoods (Yang, 1996; Pupko et al., 2002b; Bevan et al., 2005). Consider n genes G_1, ..., G_n. Let r_1, ..., r_n denote the rates of evolution of G_1, ..., G_n, let θ denote the pair $\{T, \lambda\}$ where T is a tree topology, and λ a set of branch lengths. Also let α_s be the parameter for the distribution accounting for rates across sites. Here α_s is used instead of α in order to differentiate between rates across sites and rates across genes. The likelihood is computed as:

$$\begin{aligned} L_n(\theta, \alpha_s, r_1, ..., r_n | G_1, ..., G_n) &= P(G_1, G_2, .., G_n | \theta, \alpha_s, r_1, ..., r_n) \\ &= P(G_1|\theta, \alpha_s, r_1)P(G_2|\theta, \alpha_s, r_2) \cdots P(G_n|\theta, \alpha_s, r_n) \end{aligned} \tag{3.1}$$

The rates $r_1, ..., r_n$ have mean 1.0. The parameter θ may also include other parameters (such as the proportion of invariant sites). The formula in equation (3.1) makes the assumption that the rates of evolution of all genes are independent.

Let $G_{g,i}$ denote site i in gene g. For this site the likelihood under the n–parameter method $L_{n_{g,i}}$ is:

$$
\begin{aligned}
L_{n_{g,i}}(\theta, \alpha_s, r_g | G_{g,i}) &= P(G_{g,i}|\theta, \alpha_s, r_g) \\
&= \int_0^\infty f(R|\alpha_s) P(G_{g,i}|\theta, R, r_g) dR \\
&\approx \sum_{j=1}^{S} p(\hat{R}_j|\alpha_s) P(G_{g,i}|\theta, \hat{R}_j, r_g) \\
&= \sum_{j=1}^{S} p(\hat{R}_j|\alpha_s) P(G_{g,i}|\theta, \hat{R}_j \times r_g) \qquad (3.2)
\end{aligned}
$$

where S is the number of categories used to approximate the probability density function of the Gamma distribution ($f = \Gamma(\alpha_s, \frac{1}{\alpha_s})$) for site rate heterogeneity, $\hat{R}_j$ is the site rate for category j, and $p(\hat{R}_j)$ is the probability of this site rate category. It is possible to have one α_s over all sites in all genes. It is also possible to have one α_s for each gene (or α_{s_1}, ..., α_{s_n}). In both cases the likelihood of a site is calculated in the same way.

This model assumes that branch lengths for different genes are proportional. In effect, the branch lengths are multiplied by a value proportional to the evolutionary rate of the gene and the evolutionary rate of a site.

Since the sites are assumed to be independent, the likelihood for an entire gene $L_{n,g}$ is calculated from the product of the site likelihoods (3.2) for all sites i in gene g (i.e. sites $i \in g$) as:

$$
\begin{aligned}
L_{n_g}(\theta, \alpha_s, r_g | G_g) &= \prod_{\text{sites } i \in g} L_{n_{g,i}}(\theta, \alpha_s, r_g | G_{g,i}) \\
&= \prod_{\text{sites } i \in g} \sum_{j=1}^{S} p(\hat{R}_j|\alpha_s) P(G_{g,i}|\theta, \hat{R}_j \times r_g) \qquad (3.3)
\end{aligned}
$$

Combining (3.1) and (3.3) with the assumption of independence between genes, the log–likelihood over all sites is calculated as:

$$\begin{aligned}
\log(L_n(\theta, \alpha_s, r_1, ..., r_n | G_1, ..., G_n)) &= \log\left(\prod_{\text{genes } g} L_{n_g}(\theta, \alpha_s, r_g | G_g)\right) \\
&= \log\left(\prod_{\text{genes } g}\prod_{\text{sites } i \in g} L_{n_{g,i}}(\theta, \alpha_s, r_g | G_{g,i})\right) \\
&= \sum_{\text{genes } g}\sum_{\text{sites } i \in g} \log\left(L_{n_{g,i}}(\theta, \alpha_s, r_g | G_{g,i})\right) \\
&= \sum_{\text{genes } g}\sum_{\text{sites } i \in g} \log\left(\sum_{j=1}^{S} p(\hat{R}_j | \alpha_s) P(G_{g,i} | \theta, \hat{R}_j \times r_g)\right)
\end{aligned}$$

Thus no time or computational complexity is added when calculating the likelihood, versus computing the likelihood of the concatenated data set with no gene rates (or equivalently computing the likelihood with $r_g = 1.0$ for all genes g). The only additional computational time required is optimizing over the gene rates $r_1, ..., r_n$. However, with good starting estimates, such as those found with the DistR method (Bevan et al., 2005), this time is not too significant.

3.4.2 The α–parameter method

Define $\theta = \{\lambda, T\}$ for branch lengths λ, and a topology T, and α_s as the parameter for the rates across sites distribution. Also let ω be the parameter for

distribution h which describes rates across genes. Then the likelihood of gene g under the α–parameter method is calculated as:

$$\begin{aligned} L_{\alpha_g}(\theta, \alpha_s, \omega | G_g) &= P(G_g|\theta, \alpha_s, \omega) \\ &= \int_0^\infty h(r|\omega) P(G_g|\theta, \alpha_s, r) dr && (3.4) \\ &\approx \sum_{k=1}^{C} p(\hat{r}_k|\omega) P(G_g|\theta, \alpha_s, \hat{r}_k) && (3.5) \\ &= \sum_{k=1}^{C} p(\hat{r}_k|\omega) \prod_{\text{sites } i \in g} P(G_{g,i}|\theta, \alpha_s, \hat{r}_k) && (3.6) \\ &\approx \sum_{k=1}^{C} p(\hat{r}_k|\omega) \prod_{\text{sites } i \in g} \sum_{j=1}^{S} p(\hat{R}_j|\alpha_s) P(G_{g,i}|\theta, \hat{R}_j \times \hat{r}_k) && (3.7) \end{aligned}$$

where C is the number of categories used to approximate the probability density function h with parameter ω. Probabilities $p(\hat{r}_k)$ are used to approximate h with rates $\hat{r}_k$, where h is a density function that describes the distribution of gene rates. The best choice of distribution h will be discussed later, however the mean of the distribution must be 1.0.

In (3.4) we integrate over the parameter ω, thus computing the likelihood of the data G_g for infinitely many gene rates, weighted by the probability of the gene rate. The approximation (3.5) of the integral by a summation is made in order to reduce the number of computations involved in integrating. This involves approximating h with a discrete version of the distribution with C categories and rates $\hat{r}_1, ..., \hat{r}_C$. Without such an approximation the integration is impractical to compute. The equivalence between (3.5) and (3.6) is obtained because all sites

i in gene g are assumed to be independently evolving. The equality of (3.6) and (3.7) exists because site rate heterogeneity is accounted for as in equation (3.2). As with the n–parameter method, it is possible to have one Gamma distribution describing site rate heterogeneity, or it is possible to have n Gamma distributions, one describing the site rate heterogeneity in each gene.

Since the genes are independent the overall likelihood is

$$L_\alpha(\theta, \alpha_s, \omega | G_1, ..., G_n) = L_{\alpha_1}(\theta, \alpha_s, \omega | G_1) \cdots L_{\alpha_n}(\theta, \alpha_s, \omega | G_n)$$

Computing the log–likelihood $\log(L_\alpha(\theta|G_1, G_2, ..., G_n))$ for the α–parameter method is more complex, due to the summation over a product. See Appendix 1 for details. Under the α–parameter method it is possible to approximately compute the probability of gene g evolving at a particular rate $\hat{r}_k$, using the Gamma distribution as a prior over the possible rates as $P(\hat{r}_k|G_g, \theta, \alpha_s) \approx P(G_g|r_k, \theta, \alpha_s)P(r_k)$ (which doesn't account for the probability of the data $P(G_g)$). This can provide a sense of whether the ML rate estimate is a meaningful parameter to describe the data. Additionally, if the unnormalized probabilities are uniform accounting for rate heterogeneity for the gene of interest may not provide an improved fit over the concatenated model.

3.4.2.1 The Gamma distribution

The Gamma distribution is used to describe gene rate heterogeneity. Under the α–parameter method, reasonable starting values are chosen based upon the ML fit of the Gamma distribution to initial gene rate estimates. In phylogenetic

analyses, the expected rate over multiple genes is 1.0. Under the $\Gamma(\alpha, \beta)$ distribution this is accomplished by setting $\beta = \frac{1}{\alpha}$, since the expectation of the distribution is then $\alpha\frac{1}{\alpha}$ or 1. A log–normal distribution could also be used here (Felsenstein, 2001).

3.4.3 The DistR approach

Under both the n–parameter and α–parameter methods, using good initial estimates for the gene rates will help reduce the computation time to find maximal likelihood estimates of the gene rate parameter(s) in each method. In the case of the n–parameter method, initial estimates of the gene rates can be used directly. In the case of the α–parameter method, initial estimates of the gene rates can be used to find a maximum likelihood estimate of the α parameter of the Gamma distribution. These initial parameter estimates (either the gene rates or the initial ML estimate of α) are then further optimized to determine the maximum likelihood values.

Here, initial estimates of the gene rates $r_1, ..., r_n$ are computed beforehand using the DistR method (Bevan et al., 2005). Initial pairwise distances for the method were estimated using ML distances from PHYML (Guindon and Gascuel, 2003), with the JTT model of evolution, a proportion of invariant sites and Gamma distribution for site rate heterogeneity with 8 categories.

3.4.4 Improved fit over the concatenated model

To determine the improvement (if any) of the α–parameter and n–parameter methods over the concatenated model the Akaike Information Criterion (AIC) (Akaike, 1974) and Cross–Validation Information Criterion (CVIC) (Smyth, 2000) were used. The AIC provides a measure of the expected Kullback Leibler distance between the model of interest and the actual 'true' model. The CVIC does not rely upon data independence like the AIC (Smyth, 2000). It applies the cross–validation principle to obtain a penalized likelihood. However, it is much more computationally slow, and thus was only used on two of the smaller data sets, to validate the results.

The Likelihood Ration Test (LRT) was not used, because the concatenated model is not nested within either gene rate heterogeneity model (when gene rates are accounted for using the n–parameter or α–parameter method) when each gene has a separate Gamma distribution for site rate heterogeneity. The concatenated model is nested within the gene rate heterogeneity model (both n–parameter and α–parameter methods) with one–Gamma for site rate heterogeneity. However, the LRT does not follow a χ^2 distribution because the alternative and null models are equivalent when some parameters are fixed at the boundary of parameter space (i.e. when the value of α in the α–parameter method tends towards a large value such as 100).

3.4.4.1 Calculating the AIC and CVIC

The AIC is calculated based upon correcting the log–likelihood by some function of the number of parameters in the model of interest. Under the n–parameter method, the parameters are the gene rates for each gene, the site rate heterogeneity parameter(s), the tree topology and the proportion of invariant sites. The α–parameter method has a similar set of parameters. However, rather than one rate parameter for each gene, it has a parameter for the distribution that describes gene rate heterogeneity. The concatenated model has the same set of parameters, but no gene rate parameters and it does not allow for each gene to have separate parameters for site rate heterogeneity.

The AIC is the sum of the negative log–likelihood of the model, plus the difference in a function of the number of parameters used in each model, multiplied by two. Thus, the difference in AIC between the rates based model and concatenated model is calculated as: $\Delta AIC = 2L_r - 2L_c + 2(\Delta p)$ where L_c and L_r are the log–likelihoods of the concatenated and gene rates heterogeneity models respectively. Here Δp is the difference in a function of the number of parameters in the concatenated model and the gene rates model (e.g. either the n–parameter method or α–parameter method) and thus will be a negative number. The first order AIC does not account for sequence length and thus a second order AIC was used where the number of parameters is defined as $\frac{Kn}{(n-K-1)}$ (Burnham and Anderson, 2003) where n is the sequence length and K the number of parameters in the model of interest.

The Cross–Validation Information Criterion is useful to confirm the results of the AIC since the AIC makes the assumption of data independence. Although the concatenated model conforms to this assumption, the gene rates models do not. Under the gene rates model, each site in a gene is assumed to be under the same rate of evolution, which violates the independence assumptions of the AIC. The CVIC was designed to determine the correct number of clusters to use in a probabilistic clustering framework (i.e. components in finite mixture models) (Smyth, 2000). Thus the CVIC does not rely upon the assumption of data independence.

The CVIC for a data set is calculated by dividing the data into 2 subsets. The model of interest (concatenated or gene rates) is evaluated on one subset, obtaining ML estimates of all parameters of interest. These ML estimates are used to evaluate the likelihood of the data on the second subset, under the same model. This process is repeated b times (in this case 50), and the CVIC for model m is calculated as $CVIC_m = \frac{1}{50}\sum_{b=1}^{50} L_{2,m}$. Here $L_{2,m}$ is the likelihood of the second subset of data, evaluated under the ML estimates obtained from the first subset of data. Thus the $\Delta CVIC$ is defined as $\Delta CVIC = CVIC_r - CVIC_c$ where r and c denote the gene rates and concatenated models respectively.

If the model accounting for rate heterogeneity is preferred as a better fit to the data (versus the concatenated model), the change in AIC or CVIC between the two models (or ΔAIC, $\Delta CVIC$) will be positive, otherwise it will be negative.

3.4.5 Data Analyzed

3.4.5.1 Empirical investigation of gene rates

Under the α–parameter method it is important to choose a distribution that accurately reflects the gene rates found in empirical data. To determine if the Gamma distribution accurately reflects empirical rate estimates gene rates were calculated over a number of data sets using the DistR method (Bevan et al., 2005). The data sets used for analysis consist of: 41 data sets of size 20–40 species per gene (Harlow et al., 2004); a multi–gene data set consisting of 133 genes over 44 species (Brinkmann et al., 2005); another multi–gene data set over 37 species with 146 genes; and a 14 species data set with 106 genes (Rokas and Carroll, 2005). The first data set was prepared using automatic homology testing over 144 species, which is an extension of the analysis from Harlow et. al. (Harlow et al., 2004). The other data sets were hand curated (i.e. proteins were hand selected for analysis). In both cases initial distance estimates provided to the DistR procedure were estimated using pairwise ML distances, with eight categories for the Gamma distribution, a proportion of invariant sites, and the JTT model of evolution.

3.4.5.2 Data analyzed with n–parameter and α–parameter methods

Six protein data sets were used for analysis: a fungal mitochondrial data set with 29 species and 15 genes (Bevan et al., 2005); a eukaryotic data set with 44 species and 133 genes (Brinkmann et al., 2005); the modified Madsen alignment of placental mammals with 4 genes and 28 species (Madsen et al., 2001; Pupko

et al., 2002b) ; the modified Murphy alignment of placental mammals with 6 nuclear genes and 46 species (Murphy et al., 2001; Pupko et al., 2002b); an animal mitochondrial data set with 12 genes over 56 species (Pupko et al., 2002b); a fungal nuclear data set with 8 species and 106 genes (Rokas et al., 2003).

For each data set a modified version of PHYML was run with the default BIONJ starting tree. The JTT model of evolution was used with an estimated parameter for the proportion of invariant sites. Site rate heterogeneity was accounted for using either one Gamma distribution for all sites (hereafter denoted one–Gamma), or a separate Gamma distribution to describe site rate heterogeneity for each gene (hereafter denoted gene–Gamma). In both cases four categories were used in the discrete approximation to the distribution. Gene rate heterogeneity was accounted for using either the n–parameter or the α–parameter method as outlined above. Six equiprobable categories were used in the discrete approximation to the gene rates distribution in the α–parameter method. Gene resampling was performed on the data set over 8 fungal species and 106 genes by randomly selecting 50 gene sets of size 3, 50 gene sets of size 5, and 50 gene sets of size 10.

3.5 Results and Discussion

3.5.1 n–parameter versus α–parameter method

Five diverse data sets with differing numbers of genes and species were analyzed to determine which approach to gene rate heterogeneity results in the greatest improvement over the concatenated model based on the ΔAIC. Table 3–2

and Figure 3–2 indicate that with more genes under analysis, there is a greater average ΔAIC favouring a model that accounts for rate heterogeneity. However, based upon this data there is no clear correlation between the spread of the data (i.e. the 1st and 3rd quartiles, or α value under the α–parameter model) and improved model fit over the concatenated model.

It is evident that there is no advantage to using the α–parameter method over the n–parameter method to find a better fit to the data. According to the ΔAIC (Table 3–2), the n–parameter method has the best fit compared to the concatenated method for all data sets analyzed. This is true for both one–Gamma and gene–Gamma analyses. Thus, there is no reason to prefer the n–parameter model or α–parameter model as a better fit to the data according to the AIC.

When the CVIC was calculated on the two smallest data sets (Madsen and Murphy) the results obtained under the AIC were confirmed (Table 3–2). This provides independent corroboration that the α–parameter method does not find a better fit to the data when compared to the n–parameter method. Differences in CVIC for the gene rates model versus the concatenated model are not expected to be as large as the ΔAIC because of the way the CVIC is calculated.

This is especially interesting considering the time to find the tree under each method (Table 3–3). The n–parameter method takes longer than the concatenated model primarily due to optimization of the ML gene rates. Notably, the α–parameter method takes 2–3 times longer than the n–parameter method (Table 3–3).

Data Set	NG	NS		n–parameter		α–parameter		
			Q	one–Γ	gene–Γ	α	one–Γ	gene–Γ
Fungal	15	29	0.75, 1.07	1027.77	1152.25	6.284	893.06	1010.45
Eukaryotic	133	44	0.83, 1.14	1529.21	2474.07	8.707	1298.84	2199.74
Madsen	4	28	0.81, 1.16	154.57	427.80	4.408	149.80	423.33
$\Delta CVIC$				49.86	119.83		13.72	80.09
Madsen–PT	4	28	0.82, 1.16	163.77	436.82	4.473	152.64	426.62
$\Delta CVIC$				49.79	121.41		50.16	119.03
Madsen–nT	4	28	0.82, 1.17	149.32	422.10	4.465	140.73	414.45
Madsen–αT	4	28	0.81, 1.17	153.33	426.29	4.457	142.37	415.87
Animal	12	56	0.81, 1.21	248.87	378.14	3.587	221.21	321.0
Murphy	6	46	0.39, 1.23	188.88	293.71	1.187	186.48	281.90
$\Delta CVIC$				28.58	55.72		21.60	42.83

Table 3–2: ΔAIC values for five data sets with differing numbers of genes and species (NG = Number of Genes, NS = Number of Species). For the Madsen and Murphy data sets the $\Delta CVIC$ was calculated. It is given on the second line, after the ΔAIC values. One–Γ and gene–Γ refer to the number of Gamma distributions used to account for site rate heterogeneity: either one for the entire data set (one–Gamma), or one for each gene respectively (gene–Gamma). Q refers to the first and third quartiles, and α to the value of the α parameter for gene rate heterogeneity under the α–parameter method with one–Gamma. ΔAIC and $\Delta CVIC$ values are calculated with respect to the concatenated model. Madsen–PT refers to analysis of the Madsen data set on the 'preferred' topology. In this case all the parameters were optimized over, except for the topology which was held constant. Madsen–nT refers to analysis of the Madsen data set on the 'best' topology found under the n–parameter method with gene–Gamma (the 'best' topologies differ when searching tree space when one–Gamma versus gene-Gamma are used with the n–parameter method). Madsen–αT refers to analysis of the Madsen data set on the 'best' topology found under the α–parameter method (the topology for with one–Gamma and gene–Gamma is the same under the α–parameter method when searching topology space).

When the inferred maximum likelihood (ML) topologies of the α–parameter and n–parameter methods (with gene–Gamma) were compared, four out of the five data sets had different topologies. The eukaryotic data set did not have different topologies, however it is known to be a problematic data set in terms of long branch artifacts and heterotachy (Brinkmann et al., 2005). Thus, even when the ΔAIC indicates that there is little difference between the model fit (Table 3–2), it is possible that the α–parameter and n–parameter methods find different ML topologies (Figure 3–1).

Further investigation of the Madsen data set with gene–Gamma (Table 3–2) shows that for both methods much of the topology agrees with the topology of Murphy et al. (Murphy et al., 2001), a topology currently supported by molecular data (Figure 3–1) (Springer et al., 2004). However, the grouping within the Laurasiatheria does not correspond to the currently supported molecular hypothesis (Figure 3–1a and 3–1b) (Springer et al., 2004). The α–parameter method gives the topology for the Laurasiatheria that is closest to the Murphy topology (in terms of SPR moves), only grouping Pangolin incorrectly with Flying Fox/Round Eared Bat rather than Cat/Dog. The n–parameter method incorrectly groups Horse/Rhino and Dog/Cat into a monophyletic group, with Flying Fox/Round Eared Bat an in–group. Pangolin is also grouped incorrectly in this topology (Figure 3–1a).

Although the n–parameter method finds a slightly better fit according to the AIC, care must be taken when evaluating which method finds the best tree topology. Neither method finds the preferred Murphy topology, but this is likely

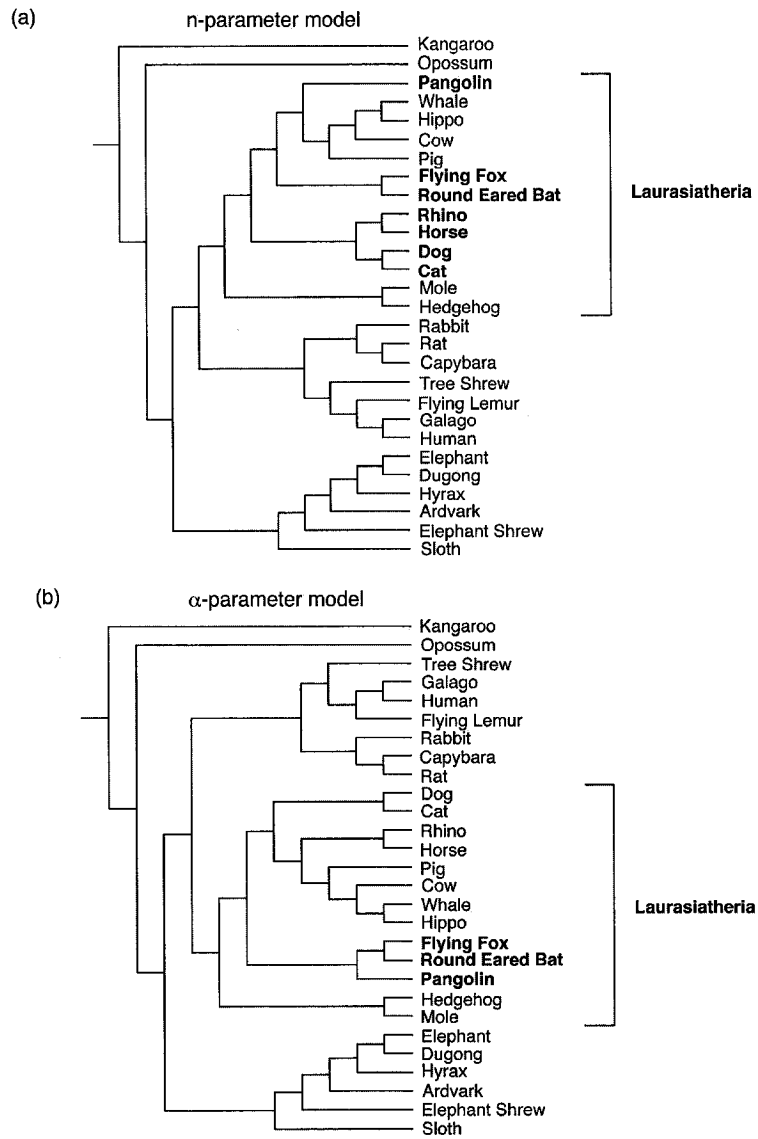


Figure 3–1: Final topologies found for the Madsen data set (Table 3–2), where branch lengths are not depicted. (a) ML topology found using the n–parameter method with gene–Gamma. (b) ML topology found using the α–parameter method. The two methods find different groupings for the Laurasiatheria species.

because only four genes were under analysis for 28 species. More data are needed to correctly resolve the phylogeny. Furthermore, when the topology found under

Data Set	n–parameter	α–parameter	concatenated
Fungal mtDNA	62m25s	220m10s	12m4s
Eukaryotic	29866m57s	6996m18s	337m0s
Madsen	23m14s	112m44s	7m32s
Animal mtDNA	120m59s	347m53s	50m0s
Murphy	33m24s	88m10s	13m41s

Table 3–3: Time for analysis using the gene rate heterogeneity and concatenated models, with the one–Gamma to account for site rate heterogeneity. For each data set, the analysis for the different models was performed on the same desktop machine. However, times across data sets are not comparable since the different data sets were all analyzed on different computers.

the α–parameter method is used to evaluate the data under the n–parameter method (and *vice versa*), according to the AIC the α–parameter method finds a better tree (Figure 3–1, Table 3–2, Madsen – αT and Madsen analyses under n-parameter method; Madsen – nT and Madsen under α–parameter method).

When the Madsen data was analyzed on the Murphy topology, optimizing for all other parameters, the α–parameter method does not find a better fit to the data than the n–parameter method (Table 3–2, Madsen-PT) according to both the ΔAIC and $\Delta CVIC$. Thus, although PHYML searches the topology space of trees differently under the α–parameter and n–parameter methods, neither method is preferred as a better fit to the data given the Murphy topology.

Figure 3–2 gives the ΔAIC values for the resampled genes data sets. These results demonstrate that: (i) both methods of accounting for gene rate heterogeneity find approximately equivalent improvement over the concatenated model; (ii) there are some data sets for which accounting for gene rate heterogeneity does not lead to an improved fit (Figure 3–2). Figure 3–2a is particularly important

because there is one data set for which accounting for gene rate heterogeneity using the α–parameter method gives a worse fit than the concatenated model ($\Delta AIC = -161.319$, one–Gamma for site rate heterogeneity), whereas the n–parameter method gives an improved fit ($\Delta AIC = 63.132$, one–Gamma for site rate heterogeneity). This is because under the α–parameter method *C. albicans* is incorrectly grouped as an in–group with *S. mikatae*, whereas under the n–parameter method the 'preferred' tree topology (Figure 3–5a) is found. This indicates that in some cases the α–parameter method has difficulty converging to the 'preferred' topology in PHYML.

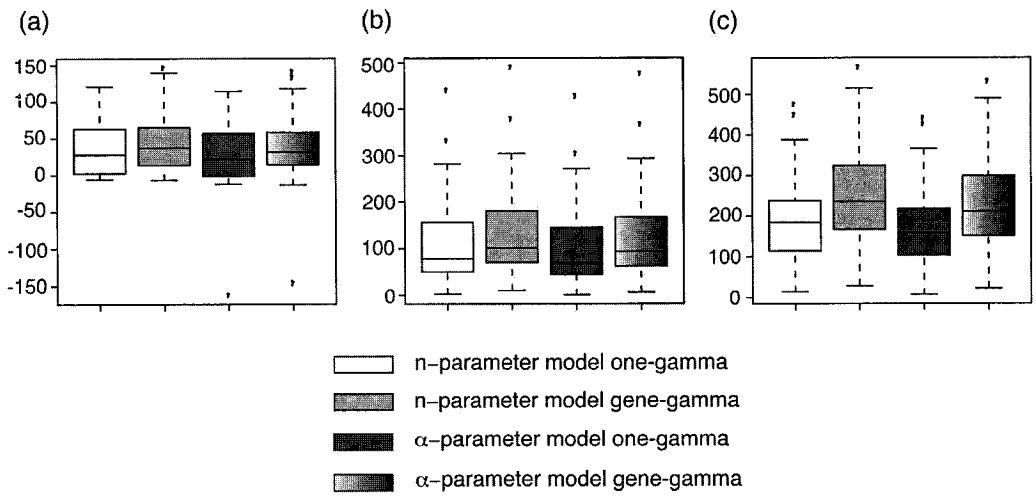


Figure 3–2: Box plots of ΔAIC values comparing the gene rates incorporated model to the concatenated model. ΔAIC values were calculated for: (a) 3 genes ; (b) 5 genes; (c) 10 genes. In each case the genes were sampled randomly from 106 genes, 50 times.

3.5.2 *Empirical rate distribution—does the Gamma distribution describe the empirical distribution of gene rates?*

The α–parameter method does not find a better fit to the data than the n–parameter method. Thus, it is important to determine if it is valid to assume that the gene rates are distributed according to the unimodal Gamma distribution. To test this assumption, the distribution of gene rates across many multi–gene data sets was determined in order to avoid the problem of sampling error. Because DistR estimates have been shown to approximate ML gene rate estimates, a large number of DistR estimates taken from multiple data sets are likely to approximate the true distribution of gene rates.

Figure 3–3a shows the distribution of all the gene rates estimated over a number of data sets. The rates were estimated using the DistR method (Bevan et al., 2005), with varied size data sets in terms of number of species, number of genes, and number of missing distances. The maximum number of missing pairwise distances was about 50%, which is fairly substantial.

The Gamma distribution provides an excellent fit to the data over many data sets (Figure 3–3a). Thus it is accurate to assume that the rate of gene evolution is distributed according to a Gamma distribution. It should be noted however, that even in the case of large data sets with many genes it is possible that the Gamma approximation will not be accurate (Figure 3–3b). In such cases it might be better to use a mixture of Gamma distributions over gene rates (as has been done for site rates in Mayrose et al., 2005). It is possible that using a better distribution will cause the α–parameter method to find a better fit to the data

than the n–parameter method. However, this option was not explored in the current analysis. It is also possible that using a better method to approximate the Gamma distribution will lead to more improvement of the fit under the α–parameter method. Thus, Laguerre integration was used to approximate the distribution. However, when this approach was used the PHYML algorithm had difficulty converging to a local optimum in some cases (data not shown). For the data sets that were successfully analyzed, this approach did not cause the α–parameter method to find a better fit to the data than the n–parameter method (data not shown).

3.5.2.1 DistR estimates versus ML estimates

DistR estimates are used as initial approximations to the gene rates in the n–parameter method, and to find an initial estimate of α in the α–parameter method. Thus, it is important to determine how accurate these initial estimates are. Figure 3–4 shows the initial DistR estimates versus the final ML estimates from the five data sets in Table 3–2, estimated using the n–parameter method with gene–Gamma. There is strong correlation between the two (Pearson's one–tailed correlation 0.904, P–value $< 2.2e^{-16}$), and the final ML parameter estimates are quite close to the starting DistR estimates. Thus, the DistR estimates provide an excellent starting point for the n–parameter method. This is especially important for large data sets in order to reduce the time spent searching rate parameter space.

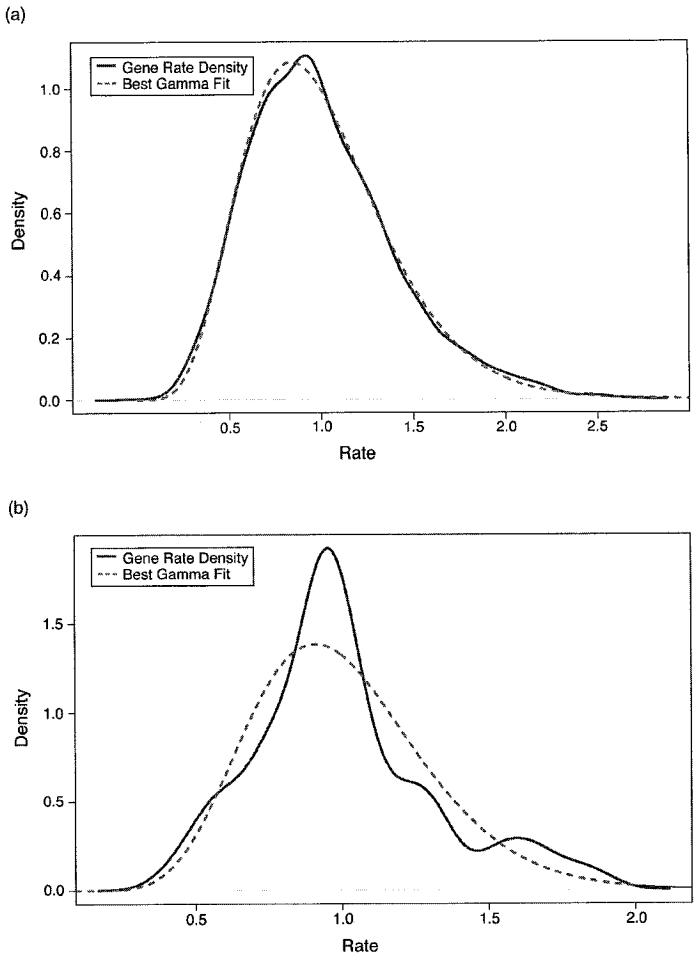


Figure 3–3: Density of estimated gene rates versus best fit of Gamma distribution. (a) Density of gene rates estimated using DistR (solid line) versus best fit of Gamma distribution (dashed line) for data described in Methods section. (b) Density of gene rates estimated using DistR (solid line) versus best fit of Gamma distribution (dashed line), for 133 genes over 44 species.

In general the initial α parameter estimates were also quite close to the final estimates under the α–parameter method with gene–Gamma. As expected, when more genes were present in the analysis the initial estimate was more accurate. For example, both the fungal and animal mtDNA data sets have more than 10

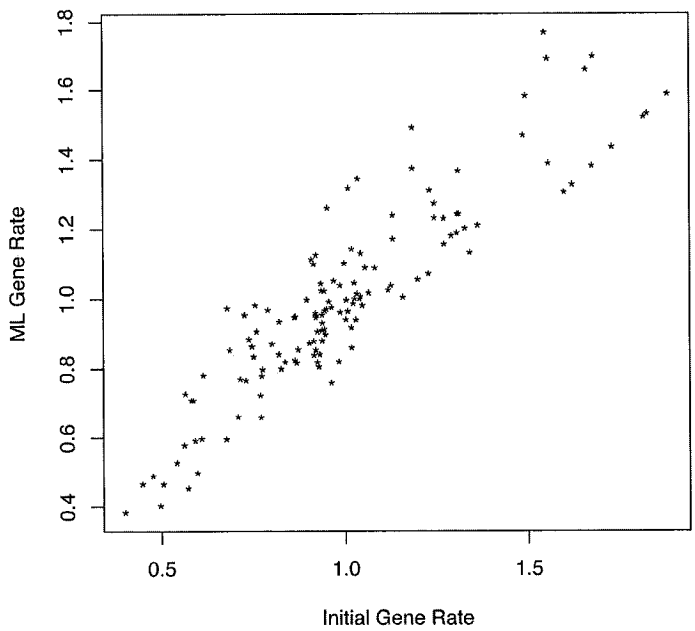


Figure 3–4: Correlation of the initial DistR gene rate estimate with the ML gene rate estimates. Maximum likelihood gene rates were estimated using the JTT model of evolution with a site rate Gamma distribution for each gene and a proportion of invariant sites. Estimates are based on data sets in Table 3–2.

genes, with relative errors of 0.0354 and 0.0263 respectively between the initial and ML estimates of the α parameter. Conversely, both the Murphy and Madsen data sets have fewer genes (Table 3–2). The respective relative errors of the initial α estimates are 0.1871 and 0.1898. The relative error does not seem to be affected by the number of species since the animal mtDNA data set has the greatest number of species but the smallest relative error.

3.5.3 *Topology resolution under n–parameter and α–parameter methods*

Given that the α–parameter and n–parameter methods give potentially different ML topologies, even when there is little difference in the improved fit over the concatenated model, it is important to determine at what point the two methods provide congruent answers. Figure 3–5 shows the 'best' (or favoured) topology, along with the branchings that prove difficult to resolve (in the data set of Rokas et al., 2003).

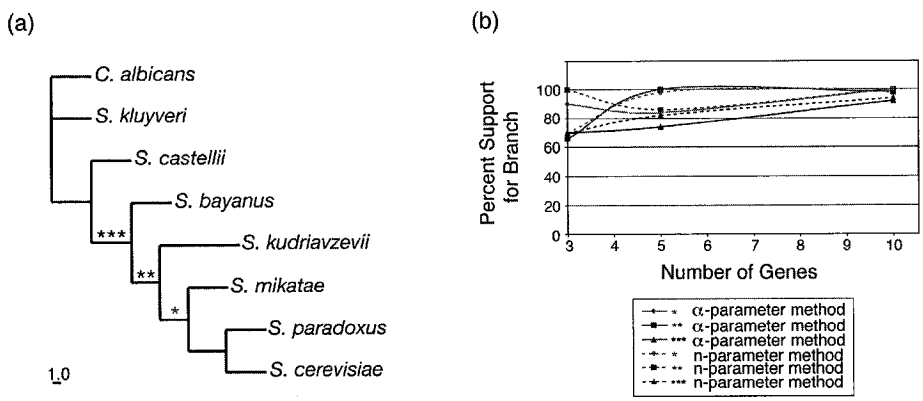


Figure 3–5: Bootstrap support for 8 fungal species under the n–parameter and α–parameter methods. (a) The correct tree topology. The branches which are difficult to resolve are labelled with *, **, and ***. (b) Bootstrap support for three branches for 3 genes, 5 genes and 10 genes sampled 50 times each from 106 genes in total.

The different methods of accounting for gene rates leads to different bootstrap support. Additionally, adding more genes leads to an increase in support as shown by Rokas *et al.* (Rokas et al., 2003). The paraphyletic branching of *S. bayanus* with *S. castellii* and the remainder of the in–group (Figure 3–5) was the most consistent in terms of improved bootstrap support with more genes, for both the

n–parameter and α–parameter methods. Conversely, the other two branches had inconsistent results across the two methods (Figure 3–5). For example, when using 3 genes, the *S. mikatae* branching starts with low support under the n–parameter method, then quickly reaches 90% support at 5 genes, and 100% support at 10 genes. For the α–parameter method the opposite is true: at 3 genes the *S. mikatae* branching has 90% bootstrap support, which drops to just above 80% support with 5 genes and increases to 100% support with 10 genes.

Evidently, in order to obtain a consistent ML tree between the two methods, more data are necessary. Thus, one explanation for the inconsistency in topologies found by the two methods (e.g. for the data in Table 3–2) is lack of data. This problem is exacerbated when more species are under analysis (as opposed to the 8 species used in this experiment). However, even with sufficient data, if model assumptions of the gene rates models are not valid (i.e. the model is misspecified with respect to true sequence evolution) than topology resolution artifacts can occur, even with sufficient data (see for instance, Philippe et al., 2005).

3.5.4 Correlation of gene rates with improved fit under n–parameter and α–parameter methods

The gene resampling experiment on the 106 gene Rokas data set not only provides information on how much data are necessary for both methods to provide congruent ML topologies, but it also demonstrates that some data sets have little to no improvement over the concatenated model (Figure 3–2).

Statistic	3 Genes		5 Genes		10 Genes	
	ρ	P–value	ρ	P–value	ρ	P–value
Max(GR)-Min(GR)	0.812	$4.008e^{-13}$	0.662	$8.407e^{-8}$	0.485	$1.185e^{-4}$
Max(GR)	0.464	$3.398e^{-4}$	0.244	$4.376e^{-2}$	0.144	0.1595
Min(GR)	-0.682	$2.537e^{-8}$	-0.723	$1.055e^{-9}$	-0.774	$2.111e^{-11}$

Table 3–4: Correlation of ΔAIC of the n–parameter method of accounting for gene rate heterogeneity with: the minimum gene rate (Min(GR)); maximum gene rate (Max(GR)); the difference between the minimum and maximum gene rates (Max(GR) - Min(GR)). Gene rates were estimated globally over the 106 genes from which the data sets were sampled. As the number of genes under analysis increases, so does the correlation of the ΔAIC with the minimum gene rate. Conversely, both the maximum gene rate and the difference between the two become less correlated with the ΔAIC. ΔAIC values are based upon accounting for gene rate heterogeneity using the n–parameter with one Gamma distribution for site rate heterogeneity.

To determine what leads to an improved model fit of the gene rates model over the concatenated model we calculated the correlation between the ΔAIC and three values: the rate of the slowest evolving gene in the data set; the rate of the fastest evolving gene in the data set; the difference between the rates of the fastest and slowest evolving genes in the data set. In order to compare gene rates properly across all resampled gene data sets, the gene rates were estimated over all 106 genes.

Correlation was tested only on the n–parameter and α–parameter methods with one–Gamma distribution for site rates. This allows for the influence of the gene rates on the ΔAIC to be tested without the influence of separate Gamma distributions for site rate heterogeneity for each gene. Results are given for the n–parameter method. Results based upon the α–parameter method were identical except in one data set where the concatenated model was preferred (Figure 3–2).

Results (Table 3–4) demonstrate that it is not only the number of genes in the analysis which affects the improvement of the rates incorporated model over the concatenated model. Both the minimum rate in the analysis and the difference between the maximum and minimum rates also have a strong effect. Correlation values show that with fewer genes both the difference between the maximum and minimum rates, and the maximum rate are positively correlated with ΔAIC. Conversely, the minimum rate is negatively correlated with ΔAIC (Table 3–4). However, as the number of genes increases, correlation of ΔAIC with the maximum gene rate decreases and becomes statistically insignificant (Table 3–4). Correlation of the ΔAIC with the difference between the maximum and minimum rates also decreases, as does the statistical significance. Interestingly, the negative correlation of the ΔAIC with the minimum gene rate increases, as does the significance of the correlation (Table 3–4). Thus, although the difference between maximum and minimum rate (i.e. the degree of rate heterogeneity) is important for improved fit, it is not as important as the minimum rate of the gene under analysis.

The results indicate that it is the minimum gene rate that is the primary variable that determines whether there is improved model fit when using a model that accounts for gene rate heterogeneity. Indeed, a slower global minimum rate indicates that a higher improvement in the fit of the model to the data are likely when accounting for gene rates. This is partially due to the fact that a slower global rate will likely lead to a slower relative rate in the data set under analysis, and thus greater gene rate heterogeneity. However, if gene rate heterogeneity

were the only factor influencing improved fit, we would expect to see that the correlation of improved fit with maximum rate would remain high (or at least significant) with more genes under analysis. This is because a faster global rate should also lead to greater gene rate heterogeneity. However, the maximum rate does not correlate with improved fit when there are more genes under analysis. Conversely, the minimum rate has a higher correlation with improved model fit when more genes are under analysis. Thus, the minimum rate of the gene has an effect upon the improved model fit, independent of the fact that a slower gene will likely lead to greater gene rate heterogeneity.

When the relative rates of the genes are used to test for correlation, the slowest evolving gene under analysis has an even more significant negative correlation with ΔAIC (-0.857, P–value of $1.064e^{-15}$ for data sets with 10 resampled genes). This correlation indicates that the DistR method can be run to test initial gene rates, and if there are very slow rates a much higher improved fit under the gene rates model can be expected.

Some analyses focus on eliminating fast sites/genes from phylogenetic analysis since these sites typically violate model assumptions, or lead to long branch attraction (LBA) (Brinkmann and Philippe, 1999; Hirt et al., 1999; Dacks et al., 2002; Brinkmann et al., 2005). It has also been noted that invariant sites can cause problems in phylogenetic reconstruction (Lockhart et al., 1996; Hirt et al., 1999; Dacks et al., 2002), leading to the removal of these sites from the analysis. This analysis indicates that properly accounting for the slow genes is quite important. Perhaps accounting for the slow genes correctly causes the invariant

sites to no longer violate model assumptions by shortening the branch lengths, and thus increasing the probability of no change over the branches. Conversely, given the low correlation of fast genes with improved model fit, fast sites which violate model assumptions (i.e. are saturated) probably continue to violate model assumptions.

Although correlations were tested over only one resampled data set, with few species, these results provide a preliminary indication that the more heterogeneous the data, the more likely an improvement will occur when accounting for the heterogeneity. This is especially true with few genes under analysis. However, as the number of genes increases this becomes less important than the evolutionary rate of the slowest gene.

3.5.5 Correlation of gene rate with site rate heterogeneity

Given that accounting for site rate heterogeneity separately for each gene leads to a much better model fit, the question arises of whether or not there is any correlation between the rate of evolution of a gene and the ML estimate of the α parameter accounting for site rate heterogeneity. Figure 3–6 shows the gene rate versus the ML estimate of the α parameter estimated over the data sets in Table 3–2. The positive correlation (Pearson's one–tailed correlation 0.432, $p = 1.887e^{-14}$) is significant.

Thus, it is not evident that accounting for gene rates, and site rates within a gene, is the best way to model the rate heterogeneity of all the sites. The rate of a site is here modelled based on both the site and gene rate heterogeneity.

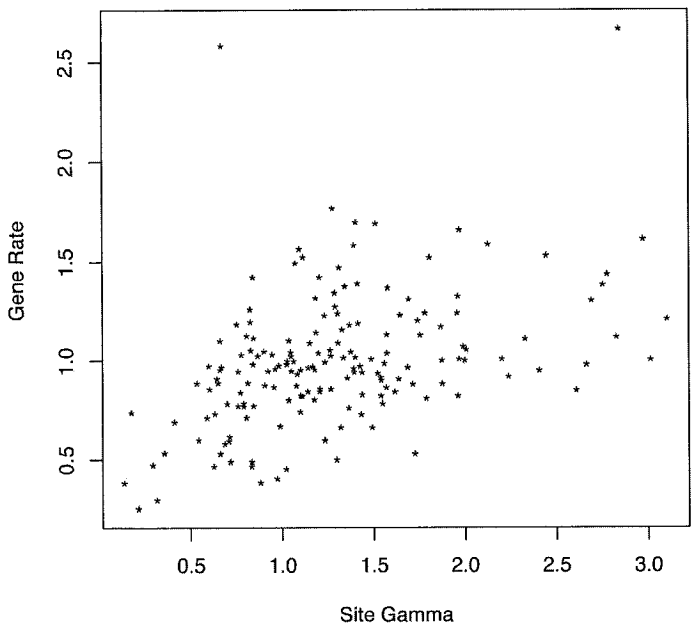


Figure 3–6: Correlation of the gene–Gamma site rate heterogeneity parameter with the maximum likelihood rate of the gene. Maximum likelihood gene rates were estimated using the JTT model of evolution with a site rate Gamma distribution for each gene and a proportion of invariant sites. Estimates are based on data sets in Table 3–2.

Yet there is only one absolute rate at which a given site evolves, ignoring rate heterogeneity through time. Clearly modelling this rate separately through site rate heterogeneity and gene rate heterogeneity is not completely correct. The correlation between the α parameter for site rate heterogeneity with the rate of evolution of the gene supports this conclusion. The gene rate parameter and the α parameter of the Gamma distribution are dependent. Thus, to a certain extent the different parameters are modelling the same information in the data, even though the parameters are estimated independently of one another. Thus, perhaps it is possible to use a model that will account for the correlation between the two, in order to find even better improvements of the model fit to the data.

3.6 Conclusions

In conclusion, given the current analysis, there is no reason to prefer the α–parameter method over the n–parameter method in phylogenetic inference. This is a promising result since it means that it is not necessary to use additional computation time to find a good fit of a model with gene rates to the data. However, these analyses also suggest that there is further work to be done in improving rate heterogeneity modelling in maximum likelihood methods. Since there is no guarantee of an improved model fit, even with an increasing number of genes, and there is high correlation between α estimates of site rate heterogeneity and gene rate estimates, clearly there are problems with current approaches.

3.7 Acknowledgements

We thank Trevor Bruen, Stéphane Guindon and Nicolas Lartillot for helpful comments on the manuscript. Thanks to Stéphane Guindon for kindly providing the source code of PHYML v2.2 for our use. Thanks to Joe Felsenstein for initially proposing the α–parameter approach to accounting for gene rate heterogeneity. Salary and support from the Canadian Institutes of Health Research (MOP 42475; BFL), the Canadian Institute for Advanced Research (CIAR; BFL), National Science and Engineering Research Council (NSERC grant 238975-01; DB), New Zealand Marsden Grant (DB), and supply of laboratory equipment and informatics infrastructure by Genome Quebec/Canada (BFL) is gratefully acknowledged. RBB is supported by an NSERC PGS-B scholarship, and Genome Quebec.

3.8 Appendix

3.8.1 Calculating the log–likelihood of a gene when integrating over gene rates

Below all calculations are for gene g. Let $LogL_{\alpha,k}$ be the log–likelihood of gene g, for category k of the probability density h over gene rates from (3.7). There are C categories which approximate distribution h. SF is a scale factor that is used to prevent over–flow and under–flow errors and it is the maximum of the log–likelihoods ($LogL_{\alpha,k}$) over all categories $k \in C$. $SLogL_{\alpha,k}$ is the $LogL_{\alpha,k}$ of gene g for category k, scaled by both the scale factor SF and the log of the probability of rate category $\hat{r}_k$. SL_g is the total scaled likelihood of gene g and $LogSL_g$ is the total scaled *log–likelihood* of gene g. Overall, the likelihood of gene g is computed as follows:

Compute $LogL_{\alpha,k} = Log(L_g)$ where $r_g = \hat{r}_k$ in (3.3) for each category C, using all sites i in gene g. This results in $LogL_{\alpha,1}, ..., LogL_{\alpha,C}$. Next calculate $SF = \max_k LogL_{\alpha,k}$ and $SLogL_{\alpha,k} = LogL_{\alpha,k} - SF - 1 + \log p(\hat{r}_k)$ for $k = 1, ..., C$. This scaling is performed in order prevent over– and under– flow errors. Thus,

$$SLogL_{\alpha,k} = \log\left(\frac{p(\hat{r}_k)\prod_{\text{site } i \in g}\sum_{j=1}^{S} p(\hat{R}_j)P(g_i|T,\hat{R}_j,\hat{r}_k)}{e^{SF+1}}\right) \tag{3.8}$$

from equation (3.7).

From (3.8) compute $e^{SLogL_{\alpha,k}}$ in order to calculate the scaled likelihood

$$\frac{p(\hat{r}_k)\prod_{\text{site } i \in g}\sum_{j=1}^{S} p(\hat{R}_j)P(g_i|T,\hat{R}_j,\hat{r}_k)}{e^{SF+1}}$$

for every category $k = 1, .., C$. Thus the total scaled likelihood is

$$SL_g = \sum_{k=1}^{C} \frac{p(\hat{r}_k) \prod_{\text{site } i \in g} \sum_{j=1}^{S} p(\hat{R}_j) P(g_i | T, \hat{R}_j, \hat{r}_k)}{e^{SF+1}} \tag{3.9}$$

The scaled log–likelihood for gene g is computed from (3.9) as

$$LogSL_g = \log(\sum_{k=1}^{C} e^{SLogL_{\alpha,k}}) - \log(e^{SF+1}) \tag{3.10}$$

Solve for the log–likelihood of gene g from equation (3.10) as

$$log(L_g) = LogSL_g + SF + 1$$

Note that the smallest scaled log–likelihood $LogSL_{\alpha,k}$ value possible that will not result in over– or under– flow is approximately -707 (where the smallest signed number than can be expressed with a double is $2.225074e^{-308}$). Thus when the scaled log–likelihood is less than -707, it is set to -707, essentially setting the likelihood for this rate category to 0. This means that for that particular gene rate category, the probability of the data, given the rate and other parameters, approached 0.

Chapter 4
Using evolutionary models to detect codon selection

4.1 Background

This chapter is unrelated to the previous two chapters. It focuses on the issue of rate heterogeneity, however in the context of detecting synonymous selection on sets of codons. Codons that undergo synonymous selection will only use one out of a set of synonymous codons at a codon site with greater frequency than expected. This is due to the fact that there is a selective (evolutionary) advantage to using a particular codon. Thus such codons should have slower rates of evolution. However, synonymous selection on a set of codons is currently detected using non–parametric methods that do not account for evolutionary properties of the data, such as rate heterogeneity of synonymous and non–synonymous mutations, gene rate heterogeneity and codon/nucleotide bias. Parametric methods exist that account for such properties of evolution, however they are only used to determine *sites* (where a codon common to many species defines a site) that are under positive or purifying selection.

In this chapter parametric codon models are used to test whether certain codons are under synonymous selection. It is possible to simulate data under

a model of neutral evolution, and compare the codon usage patterns found in the simulated data (under neutral evolution) to the patterns found in real (empirical) data. If the patterns differ significantly, then the codons might be under synonymous selective pressure.

4.2 Abstract

The forces that drive codon evolution, in particular why codons at certain sites are invariant (with no mutation over time), remain only partially explained and modelled at the informatics level. There are currently two main approaches to identify codon selection for a given set of genes, parametric and non–parametric. Non–parametric approaches involve inference of codon usage patterns, and correlation to experimentally determined features such as protein expression levels — a methodology that does not account for evolutionary forces acting upon the codons. Parametric approaches rely upon an explicit model of codon evolution. Such models are used to infer site–specific synonymous and non–synonymous substitution rates, which can used to identify individual sites that are under purifying selection. Yet this approach, which accounts for evolutionary forces acting upon codons, does not identify a set of codons that are under strong selective pressure, as non–parametric methods do. We propose to use such models to address the problem of detecting sets of codons under synonymous selection.

Two data sets are studied: nuclear genes of three *Saccharomyces* species that are known to undergo selection for translational efficiency; mitochondrial genes of

several *Reclinomonas* species that are highly A+T biased (as are mitochondrial genes of other eukaryotes and genes of many bacterial pathogens). Applying neutral models of evolution to detect synonymous codon selection in these data sets will answer two questions: (i) whether it is possible to detect known codons in the *Saccharomyces* genomes that are posited to be under synonymous translational selection (major codons); (ii) whether codon selection can also be identified in *Reclinomonas* mitochondrial genomes despite their high A+T bias. Results indicate that applying phylogenetic models of neutral evolution detects 11 *Saccharomyces* codons as under synonymous selection. Nine of these codons were previously identified as under selection for translational efficiency. Similarly, 10 *Reclinomonas* codons are identified as undergoing synonymous selection. This is especially interesting because *Reclinomonas* mitochondrial genomes have a much smaller effective number of codons given the high A+T bias of the genes.

4.3 Introduction

The ability to predict the degree of synonymous selection upon a set of codons is of paramount importance to understand protein evolution. For instance, elimination of unfavourable codons and the adjustment of codon bias is key to accurate over–expression of proteins in heterologous systems (e.g. Gustafsson et al., 2004; Sorensen and Mortensen, 2005). As codon selection is subject to evolutionary change, it is further important to consider variation among groups of

related species, as evidenced by the publication of a multitude of species–specific codon tables (Nakamura, 2005; Nakamura et al., 2000).

Current analyses of synonymous selection are performed using either a non–parametric or parametric approach. Non–parametric statistics (such as the number of effective codons (N_c) (Wright, 1990), the codon adaptation index (CAI) (Sharp and Li, 1987), and the codon bias index (CBI) (Bennetzen and Hall, 1982)) are calculated from various biological data sets without employing an evolutionary model. Parametric inference of sites under positive or purifying selection is performed using evolutionary codon models (Goldman and Yang, 1994; Muse and Gaut, 1994; Yang et al., 2000; Pond and Muse, 2005).

Although useful for summarizing properties of the data, non–parametric statistics are unable to directly predict the selection acting on specific codons over a set of genes. For example, the non–parametric codon adaptation index (CAI) is currently a popular non–parametric method to predict optimal codon usage. Using the CAI as a measure of gene expression level, it is possible to identify codons (called *major codons*) that have high usage in highly expressed genes (Sharp et al., 1988; Duret and Mouchiroud, 1999; Duret, 2000; Akashi, 2001). These *m*ajor codons are inferred to be under synonymous selection. Notably, these major codons *are* used more frequently in genes with both high mRNA abundance (Duret and Mouchiroud, 1999; Akashi, 2001) and high gene expression levels (Sharp and Li, 1987; Coghlan and Wolfe, 2000; Kliman et al., 2003). Coupled with evidence that codon usage is biased towards the more abundant cognate tRNAs (Moriyama and Powell, 1997; Beirne and Eyre-Walker, 2005)(Akashi, 2001), this is

posited as evidence of selection for translational efficiency, since tRNA abundance will have a greater effect upon highly expressed genes (Akashi, 2001).

However, selection for translational efficiency is only one of the potential biological processes affecting codon evolution. Other factors that will influence codon selection include the precision of the decoding apparatus, and constraints on species evolution such as long–term survival under sub–optimal conditions. Indeed, yeast is a fast–growing model species, in which translation that is optimized for protein expression levels might predominate. However, the large fraction of little known species that grow only marginally or under extreme environmental conditions may have other constraints upon codon evolution.

The CAI is currently the best statistical measure of gene expression in yeast compared to CBI, N_c, F_{op} and iterative methods that do not depend upon a set of highly expressed genes (Coghlan and Wolfe, 2000; Carbone et al., 2003). However, the CAI has only (a marginal) 0.65 correlation with yeast expression levels (Friber et al., 2004), and will not properly identify highly expressed genes when there is a high level of codon bias obscuring the signal (Peden, 1999). It can also not be applied meaningfully when reference gene expression levels are unknown for a given species (and cannot be inferred from known expression levels in a close relative). Furthermore, iterative methods (e.g. Merkl, 2003) are susceptible to codon bias, often selectively identifying the most highly biased codons, which may not correspond to codons under synonymous selective pressure. Finally, non–parametric approaches to codon bias do not account for the evolutionary rate of genes, although recent studies strongly suggest that the expression level of a

gene is highly correlated with evolutionary rate in *S. cerevisiae* (Drummond et al., 2005).

Phylogenetic parametric approaches have the advantage that they can account for an explicit model of codon evolution. Codon models offer the most realistic null hypothesis of codon evolution available, as they account for mutational bias, heterogeneity of non–synonymous and synonymous codon rates, and gene rate heterogeneity. All of these factors will influence the identification of *highly biased* codons with non–parametric methods. Unlike non–parametric approaches, codon models do not depend upon prior knowledge of any data from a given species to make inferences about codon selection, including biological data such as RNA or protein expression levels. However, using codon models does require access to sufficient comparative sequence information, and the building of multiple alignments of gene sequences.

Codon models are currently used to identify specific sites under purifying or positive selection. They are not used to identify whether a specific codon is under synonymous selective pressure across the genome. However, due to the fact that evolutionary forces acting upon a set of genes can be accurately modelled, codon models provide an ideal framework to identify which codons are under synonymous selection.

To identify codons that are under synonymous selection we first divide our data into two subsets: sites (codons) that are invariant (with no mutation) over a set of species, and sites (codons) that vary. Codons that are invariant (with no mutation) over a set of species, are either invariant due to chance, or due to some

selective pressure. It is possible to measure the Relative Synonymous Codon Usage (RSCU) for the two subsets of data. Here the RSCU is a normalized percentage usage of each codon for a particular amino acid. We are interested in sites where the RSCU is greater in invariant sites, than in sites that vary. In other words, if synonymous usage of a codon (as measured by RSCU) is higher in invariable sites, the candidate codon might be under synonymous selection. Codons that have this property in the real data are termed Increased RSCU in Invariant Sites (IRIS) (Figure 4–1).

To determine if codons are under synonymous selection we use parametric codon models to simulate data under a null hypothesis of neutral evolution. This simulated data will reflect the synonymous usage that is expected under the null hypothesis of no synonymous selection. If the synonymous usage for an IRIS codon in the real data differs significantly from the simulated data, the codon is under synonymous selective pressure according to that model. We define a codons as a Highly Selected Codon (HSC) if it is under synonymous selective pressure according to all models studied.

The goals of this paper are: (i) to identify IRIS codons and determine if they correspond to major codons in a well–known system (yeast); (ii) to identify HSCs and determine which parametric codon models best explain the difference in RSCU between invariant and variant sites observed in IRIS codons in real data; (iii) to determine if non–parametric methods can identify HSCs in highly biased data.

4.4 METHODS

4.4.1 A parametric approach to detecting synonymous selection – finding HSCs

Parametric codon models (Goldman and Yang, 1994; Yang et al., 2000; Pond and Muse, 2005) are used to detect whether a particular codon is under synonymous selection across a set of genes. Sites in an alignment are first subdivided into two subsets: sites (codons) that are invariant over time (with no mutation) and sites that are variant over time (with at least one mutation). Sites that are invariant are invariant either due to chance or due to synonymous selective pressure. Although variant sites may also be under synonymous selection, this cut–off is based upon the assumption that sites that do not mutate are more likely to be under synonymous selective pressure than sites with even one mutation. It is possible to use a less stringent cut–off, by dividing the data based upon the rate of evolution of the site (where the rate is estimated as the ratio of non–synonymous change to synonymous change). However, a conservative cut–off of invariant *versus* variant sites provides an ideal framework to determine whether codon models can identify codons that are under synonymous selection.

4.4.1.1 Defining the relationship between codon usage in invariant and variant sites

To quantify the relationship between codon usage in invariant and variant sites define $RSCU_{c,s}$ to be the RSCU of codon c calculated over sites s. The RSCU for a particular codon is calculated based upon the number of codons in a sequence

c and the number of amino acids a which the codon encodes. The ratio of the two is taken, and multiplied by the number of codons that code for the amino acid. For example, suppose there are 50 AAA codons in a set of sites under analysis. This codes for lysine which is encoded by 2 codons. Suppose also that there are 75 lysine residues in sites s. Then the $RSCU_{AAA,s}$ for codon AAA is calculated as $2\frac{50}{75} = \frac{4}{3}$. Likewise, the $RSCU_{AAG,s}$ of codon AAG, which also codes for lysine is calculated as $2\frac{25}{75} = \frac{2}{3}$.

Define $R_{c,v}$ as the Relative Synonymous Codon Usage for codon c computed over only *variant* codons, and define $R_{c,i}$ as the Relative Synonymous Codon Usage for codon c computed over only *invariant* codons. There are certain codons for which the RSCU will increase in invariant sites (relative to variant sites), and other codons for which the RSCU will decrease in invariant sites (relative to variant sites), even under a neutral model of evolution. The codons for which $R_{c,i} > R_{c,v}$ in the original data set of interest will hereafter be referred to as *increased RSCU in invariant sites (IRIS)* codons. Our test statistic measures $R_{c,i} - R_{c,v}$ in IRIS codons. We ignore non–IRIS codons because the RSCU in invariant sites is not over–represented when compared to the RSCU in variant sites. Thus, it is unlikely that non–IRIS codons are under synonymous selective pressure.

The method to determine if a codon is a (HSC) is as follows (Figure 4–1):

1. For each gene build a tree and infer maximum likelihood (ML) parameter estimates for the codon model of interest (Step 1, Figure 4–1);

2. Simulate 150 data sets for each gene based on the ML estimates obtained in Step 1 (Step 2, Figure 4–1);
3. For each of the 150 simulated data sets (one data set is a group of simulated genes), calculate $R_{c,i} - R_{c,v}$ for all IRIS codons c for variant sites v and invariant sites i (Step 3, Figure 4–1);
4. Compare $\hat{R}_{c,i} - \hat{R}_{c,v}$, the difference between invariant and variant sites RSCU for codon c as calculated on the real data set, to the distribution of 150 values for codon c obtained in Step 3 (Step 4, Figure 4–1); if this value is greater than 95 % of the simulated values obtained in Step 3, label codon c as a potential HSC;
5. If all codon models of interest label codon c as a potential HSC then codon c is a HSC;

4.4.1.2 Evolutionary models of interest

Two types of models were initially studied: codon models and DNA models where sites are divided into groups based upon codon position. However, only results on codon models are reported since they have much better explanatory power for the observed RSCU of variant and invariant sites. The codon models used consider both non–synonymous and synonymous rates of substitution across sites and a transition/transversion ratio (Yang et al., 2000; Yang and Nielsen, 2000; Pond and Muse, 2005). Furthermore, gene rate heterogeneity was accounted for by analyzing each gene separately.

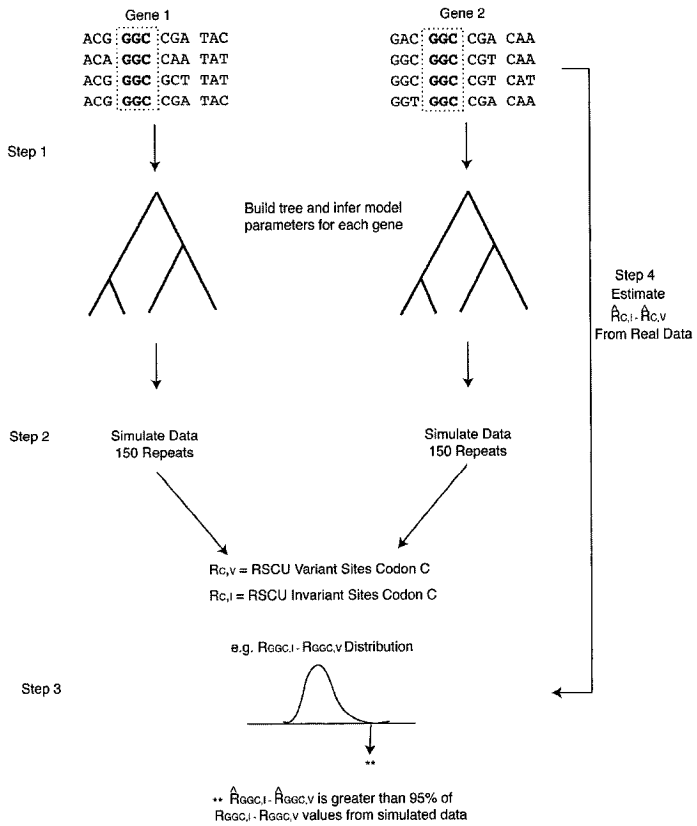


Figure 4–1: Diagram explaining how Highly Selected Codons (HSC) are detected using codon models. Gene trees and maximum likelihood (ML) parameters are inferred using phylogenetic codon models. Data are simulated (150 data sets) for each gene and set of parameters estimated. Therefore, there are collectively 150 simulated multi–gene data sets. Monte–Carlo P–values are calculated for the difference in RSCU for invariant versus variant sites ($R_{c,i} - R_{c,v}$ for codon c, where i denotes invariant sites and v denotes variant sites). This is done by estimating $R_{c,i} - R_{c,v}$ for each codon c over each simulated multi-gene data set. This gives a distribution of the difference in RSCU between invariant and variant sites expected under a neutral model of evolution. The estimated difference in RSCU between invariant and variant sites ($\hat{R}_{c,i} - \hat{R}_{c,v}$) is calculated for codon c based upon all genes in the real data set. If this value is greater than 95 percent of the values obtained under the simulated data for all models of interest then codon c is a Highly Selected Codon (HSC).

Estimating the ML parameters. All parameters were estimated separately for each gene using a modified version of PAML (Yang, 1997). The starting

tree was estimated in PAML using the DNA tree from PHYML, where each codon position is treated as a separate gene. This results in a starting tree with branch lengths estimated in terms of the expected number of *nucleotide substitutions per codon.* Parameters estimates were computed over a superset of species for which HSCs are determined in order to find accurate ML parameter estimates. This assumes that global and local evolutionary pressures are the same.

Accounting for site rate heterogeneity. Two different models were used for analysis:

1. A synonymous rate of 1.0 for each site, with a discrete Beta distribution to account for the non–synonymous rate heterogeneity;
2. Both non–synonymous and synonymous rates heterogeneity are accounted for using separate discrete Gamma distributions, with a mean synonymous rate of 1.0.

Estimating codon frequencies. Codon frequencies were estimated using two methods: by gene or over all genes. When calculating codon frequency separately for each gene, the product of nucleotide frequencies at each positions are used, since codons counts are not statistically significant. Thus, the equilibrium frequency of codon c is equal to $\pi_j^{(1)}\pi_j^{(2)}\pi_j^{(3)}$ where $\pi_j^{(l)}$ is the equilibrium frequency of nucleotide j at codon position l. When calculating codon frequencies based on all genes it is possible to use the codon counts over all genes and species to be analyzed.

4.4.1.3 Models of evolution: Simulating the data

In order to determine if evolutionary models account for the observed pattern of codon usage in invariant and variant sites a Monte–Carlo simulation study (parametric bootstrap) was performed. For both of the models studied (non–synonymous rates only and both non–synonymous and synonymous rates across sites), 150 sequences of length 4000 codons were simulated over the maximum likelihood tree from each gene in the data set using a modified version of PAML. Since two methods were used to estimate codon frequencies, both sets of codon frequencies were used when simulating data. Hence, there are four models that were fit to each data set, and four groups of 150 simulated genes as follows (Table 4–1).

1. Model NSG: Synonymous and non–synonymous rates across sites with codon frequencies estimated separately for each gene based upon codon counts;
2. Model NG: Non–synonymous rates across sites with codon frequencies estimated separately for each gene based upon codon counts;
3. Model NSD: Synonymous and non–synonymous rates across sites with codon frequencies estimated over the data set based upon codon counts;
4. Model ND: Non–synonymous rates across sites with codon frequencies estimated over the data set based upon codon counts;

Gene codon frequencies were computed using codon counts over the gene (models NG and NSG) as opposed to using nucleotide frequencies at each codon position, because results (not shown) demonstrated that this led to simulated data that

Model	rate heterogeneity		codon counts	
	non–synonymous	synonymous	global	gene
ND	yes	no	yes	no
NSD	yes	yes	yes	no
NG	yes	no	no	yes
NSG	yes	yes	no	yes

Table 4–1: The models under which data are simulated. Both synonymous and non–synonymous rates across sites are accounted for, as well as different methods of accounting for codon bias. ND – non–synonymous rates across sites, with codon bias calculated over all genes; NSD – non–synonymous and synonymous rates across sites, with codon bias calculated over all genes; NG – non–synonymous rates across sites, with codon bias calculated separately for each genes; NSG – non–synonymous and synonymous rates across sites, with codon bias calculated separately for each genes.

better approximated the difference is RSCU between invariant and variant sites found in the real data.

4.4.2 Codon usage bias statistics — non–parametric analyses of codon bias

The majority of techniques used to study codon bias focus on statistical tests of the data, such as the effective number of codons (EN_c, (Wright, 1990)), codon adaptation index (*CAI*, (Sharp and Li, 1987)), frequency of optimal codons (F_{op}, (Ikemura, 1981)), and codon bias index (*CBI*, (Bennetzen and Hall, 1982)). Assuming that these statistics are able to differentiate genes based upon expression level, sets of major codons can be identified which correspond to those codons with higher RSCU in highly expressed genes.

A more extensive technique involves the use of *correspondence analysis* a multivariate statistical technique that finds the largest axis of variation (principal

axis) to explain the data. In the case of codon usage bias the data consist of gene codon frequencies. If, for example, the principal axis correlates highly with GC bias then it is likely that GC bias best explains the observed codon usage bias (Peden, 1999). Correspondence analysis provides a method to determine what property of the data correlates most highly with the different codon usage of different genes. Often the CAI and CBI are correlated with the principal axis of greatest variation. When this correlation is high, it is assumed that the CAI /CBI (which are non–parametric measures of the expression level of the data) can explain the variation in codon usage across genes. Hence, translational selection is assumed to occur.

Because the CAI relies upon prior knowledge of expression levels of the genes, and both CBI and F_{op} rely upon a known set of optimal codons, it is not possible to calculate these statistics for the *Reclinomonas* data. However, it is possible to use correspondence analysis to determine which properties of the data best explain variation in gene codon usage. Correspondence analysis was performed using CodonW to determine if non–parametric methods can detect HSCs (Peden, 1999).

4.4.3 Data

Two different data sets are studied. We want to analyze a data set in which there is known selection of synonymous codons, and determine if such selection can be detected with phylogenetic approaches. We also want to analyze a data set in which there are strong mutational biases that are likely to obscure the observation of selection at the synonymous codon level under non–parametric

methods. A data set that fits well into the first category is the *Saccharomyces* data set from seven fungi (Rokas et al., 2003; Rokas and Carroll, 2005): *Saccharomyces cerevisiae*, *Candida glabrata*, *Saccharomyces bayanus*, *Saccharomyces kudriavzevii*, *Saccharomyces mikatae*, *Saccharomyces paradoxus*, and *Saccharomyces castellii*. This data set is a modified version of the data analyzed in Rokas *et al.* (Rokas et al., 2003), where *Candida glabrata* is used as an out–group, as opposed to *Candida albicans*. There are 106 nuclear genes of moderate A+T bias.

Reclinomonas mtDNA genes provide a data set that fits well into the second category. These genes have high A+T bias and there are a sufficient number of genes to obtain accurate inferences. *Reclinomonas* mtDNA contains the greatest number of genes out of all eukaryotes sequenced to date (Bevan and Lang, 2004, and references therein). Furthermore, this genome is representative of a data set in which the genes expression levels are unknown. Most non–parametric approaches will thus choose the most A+T biased codons as the codons undergoing synonymous selection due to the high A+T bias of the genome (Peden, 1999). Sixty–one genes from the mitochondria of four *Reclinomonas americana* strains (83, 84 and 94 and sp) were aligned to orthologous genes (if present) from *Rhodomonas salina*, *Seculamonas ecuadoriensis*, *Phytophorus infestans*, and *Ochromonas*. Protein alignments were performed using default settings of ClustalW (Thompson et al., 1994). Codon alignments were obtained by aligning the DNA sequence to the protein alignment using a custom–made program.

Although all species for a particular data set are used to infer parameter estimates in the codon models, detecting HSCs was performed only over subsets of

the species: in the *Saccharomyces* data set the species *Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, and *Saccharomyces mikatae*; in the *Reclinomonas* data set the four *Reclinomonas americana* species.

4.5 Results and Discussion

4.5.1 Identifying IRIS codons

To identify HSCs, codons with the IRIS property must first be identified (codons with Increased RSCU in Invariant Sites versus variant sites). Once the IRIS codons are known, it is possible to simulate data under a neutral model of evolution, and for each IRIS codon determine whether or not the observed difference in RSCU between invariant and variant sites is expected.

Figure 4–2 gives the invariant sites RSCU versus the variant sites RSCU for both the *Saccharomyces* (Figure 4–2a) and *Reclinomonas* (Figure 4–2b) data sets. It is clear for both data sets that the invariant and variant RSCU are correlated (*Reclinomonas* $\rho = 0.943$ P–value $< 2.2^{-16}$, *Saccharomyces* $\rho = 0.762$ P–value $= 1.345^{-13}$). Thus, codons that have high RSCU over variant sites are more likely to have high RSCU over invariant sites. This suggests that codon usage in invariant sites is not independent from codon usage in variant sites. *Reclinomonas* has a number of codons for which there is low RSCU in the variant sites and almost no synonymous usage in the invariant sites (Figure 4–2b). This indicates that there is no selective pressure to conserve codons with low RSCU in *Reclinomonas* (Figure 4–2a). Conversely, in *Saccharomyces* all codons have

moderate RSCU in invariant sites. This indicates that even codons with RSCU are conserved in yeast, whereas they are not in *Reclinomonas*.

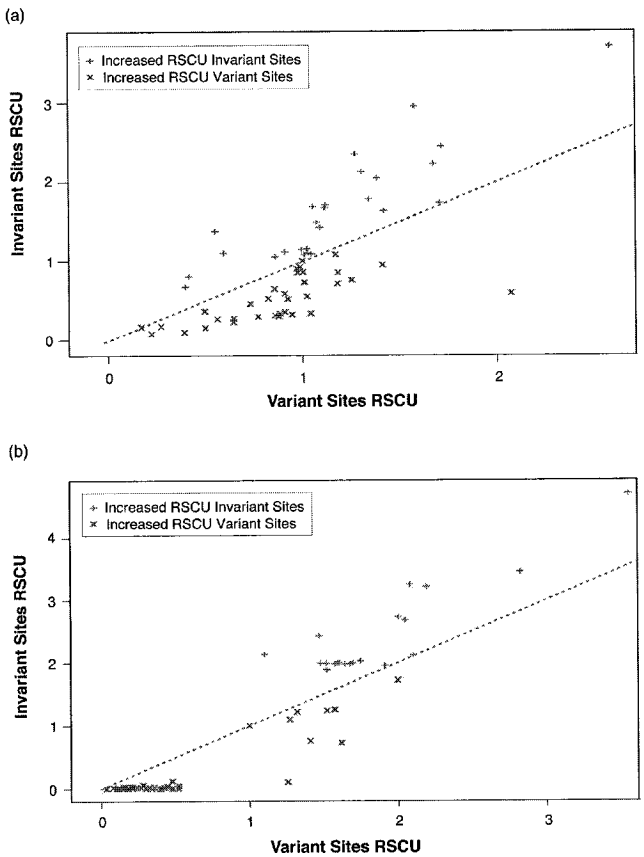


Figure 4–2: Invariant Sites RSCU versus Variant Sites RSCU for (a) *Saccharomyces* data set and (b) *Reclinomonas* data set. All 61 coding codons are shown. Points above the $x = y$ axis are IRIS codons. *Reclinomonas* has a number of codons for which there is no usage in invariant sites, likely due to the high A+T bias of the mtDNA. In general codons ending in A/T are conserved.

The difference in the number of IRIS codons in the two data sets occurs because the codon usage bias is less significant in *Saccharomyces* compared to *Reclinomonas*. The effective number of codons (N_c) as measured across all genes is approximately 32 for the *Reclinomonas* species and approximately 48 for the

Saccharomyces species. The N_c provides a measure of the extent of codon bias of the genome, with a lower value indicating a more highly biased genome. Since *Reclinomonas* has a low N_c and such extreme A+T and codon bias, there are fewer codons to which the IRIS statistic will apply. Conversely, *Saccharomyces* has much less A+T and a higher N_c. Hence, there are more codons to which the IRIS statistic will apply.

The codons for which there is increased RSCU in invariant sites largely overlap for both data sets, with the major difference in aliphatic amino acids (Figure 4–3). In both data sets, IRIS codons primarily end in A/T, except for Leu (TTG) in *Saccharomyces.* Table 4–3 gives a list of the IRIS codons for each species. Both species have a number of codons with only A/T (7 in *Saccharomyces*, 6 in *Reclinomonas*). However, in *Saccharomyces* none of these codons correspond to major codons (those codons with high RSCU in highly expressed genes where gene expression is measured using the CAI), which are given in bold (Table 4–3). There are only 11 major codons in *Saccharomyces* that correspond to IRIS codons. The remaining 10 major codons (Sharp et al., 1988) all end in C/G and have lower RSCU in invariant sites. Since the codon usage (RSCU) in invariant sites is lower than in variable sites, this suggests these codons are not over–represented in invariable sites (sites that are potentially under synonymous selective pressure). However, the genes under analysis from *Saccharomyces* are low to moderately expressed according to the CAI, and the major codons not identified as IRIS codons all have low usage in lowly expressed

Species	aliphatic	non–aromatic	basic	acidic	aromatic	sulfur	imino
R.am	GGT, GCA, GTT, TTA ATT	ACA, AGT, TCT	CAT, AAA AGA, CGT	GAA, CAA GAT, AAT	TTT, TAT	TGT	CCA, CCT
yeast	**GGT**, GCA CTT, TTA, **GCT**, **GTT** **TTG**, ATA ATT	ACA, **ACT** AGT, TCA **TCT**	CAT, AAA **AGA**, CGT	**GAA**, **CAA** GAT, AAT	TTT , TAT	**TGT**	**CCA**

Table 4–3: IRIS codons in *Saccharomyces* (yeast) and *Reclinomonas* (R.am). Bold codons in *Saccharomyces* denote those codons that have high RSCU in highly expressed genes and low RSCU in lowly expressed genes from (Sharp et al., 1988) (or major codons).

genes (Sharp et al., 1988). Perhaps if more highly expressed genes were present in the data set, some of these codons would also be identified as IRIS codons.

4.5.2 Identification of HSCs

HSCs are identified as IRIS codons that have an unexpected difference in RSCU between invariant and variant codons. Here, unexpected is defined based upon comparing the observed or empirical difference to the expected difference found in the simulated data. If the observed difference in RSCU between invariant and variant codons is greater than the expected difference found under 95 percent

of the simulated data, then the observed difference is significant at a P–value of 0.05 (Figure 4–1). Therefore, the probability of finding the observed difference *by chance* under a neutral model of evolution is less than 0.05 (or 5 percent). HSCs are IRIS codons for which none of the codon models studied (Table 4–1) can explain the observed difference in RSCU between invariant and variant sites (Figure 4–3a and b, P–value < 0.05).

Even though *Saccharomyces* has more IRIS codons (with an invariant RSCU increase) than *Reclinomonas*, approximately the same number of HSCs are identified in both data sets (Figure 4–3). In both species there is some biological pressure that is conserving these codons more than would be expected under a neutral model of evolution. Both data sets have HSC that correspond to hydroxyl/non–aromatic, sulfuric, basic and imino amino acid groups. However almost half of the HSC codons for *Saccharomyces* are aliphatic, whereas *Reclinomonas* has none. Rather, *Reclinomonas* prefers aromatic and acidic residues (plus two additional basic residues Lys AAA and His CAT). Selection on such codons might occur because many mitochondrial proteins are trans–membrane, and thus would have large regions that are hydrophobic, and surface domains that are hydrophilic.

4.5.2.1 Which codon models best explain the empirical data?

Four codon models are tested to determine which codons are HSCs. HSCs are defined based upon all four models detecting that a codon is under synonymous selective pressure. However, it is also possible to look at which model best explains the data – i.e. which model can explain best explain the RSCU for IRIS codons.

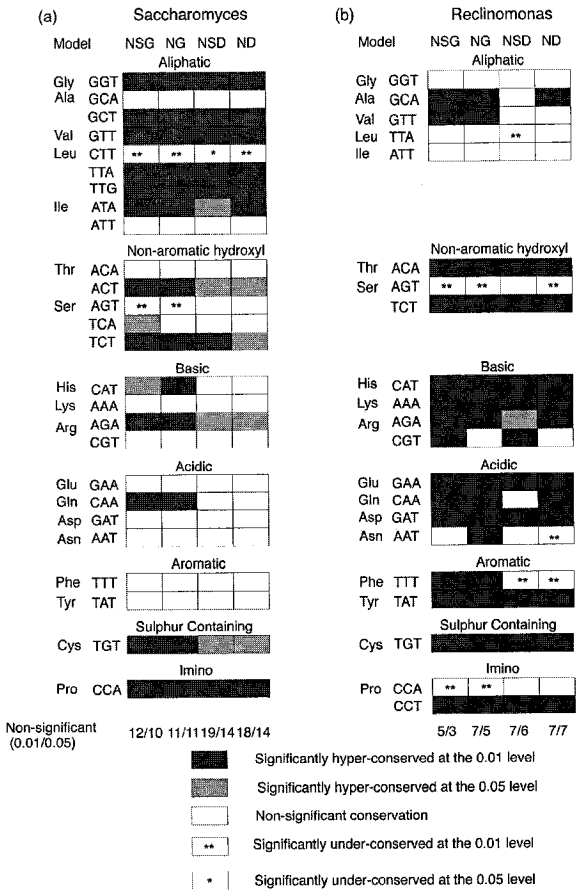


Figure 4–3: P–values that that difference in RSCU between invariant sites and variant sites for a particular codon is distributed according to a neutral model of evolution for (a) *Saccharomyces* and (b) *Reclinomonas*. For each codon c the RSCU of invariant sites ($R_{c,i}$) and variant sites ($R_{c,v}$) is calculated from 150 data sets simulated under four models of neutral evolution. This gives a distribution of values expected for $R_{c,i} - R_{c,v}$ as expected under the models of neutral evolution. The same statistic ($R_{c,i} - R_{v,i}$) is then calculated based upon the real data and compared to the values obtained from the simulated data. If the statistic on the real data is greater than 95 percent of the values obtained on the simulated data then the codon is high–lighted light grey. If the statistic on the real data is greater than 99 percent of the values obtained on the simulated data then the codon is high–lighted dark grey. Codons that are high–lighted light/dark grey under all four models of evolution are highly selected codons (HSC). If the statistic on real data is less than 95 or 99 percent of the values obtained under the simulated data then it is labelled * or ** respectively. The number of codons that dont have any synonymous selection are given for both the 0.01 and 0.05 significance levels. Please see Table 4–1 for details on models analyzed (NG, NSG, ND and NSD).

In *Reclinomonas* there are two codons (ATT and GGT) for which the observed difference in RSCU between invariant and variant sites is captured by each codon model (Figure 4–3b). Conversely, in *Saccharomyces* there are 10 codons (AAA, AAT, ACA, ATT, CGT, GAA, GAT, GCA, TAT, TTT) that are explained by all models of neutral evolution (Figure 4–3a). Although *Saccharomyces* has many more IRIS codons than *Reclinomonas*, many of these codons are explained by all neutral models of evolution studied.

On average it appears that estimating codon frequencies over all genes best models the IRIS codons (Figures 4–3a and b) in both *Reclinomonas* and *Saccharomyces*. For *Reclinomonas*, when a P–value of 0.01 is used to determine if an IRIS codon is under synonymous selection, three codon models can explain the difference in RSCU between variant and invariant sites found in seven IRIS codons (Figure 4–3b, models ND, NSD, NG). The remaining 14 codons are detected as under synonymous selection (although the codons that are under synonymous selection differ at times for the three models, Figure 4–3b). When a less stringent cut–off is used to classify a codon as under synonymous selection (P–value of 0.05), more codons are identified as under synonymous selection under each model except ND. Thus, each model was able to explain the observed RSCU difference between variant and invariant sites for fewer codons (Figure 4–3b), except model ND (which calculates codon bias across all genes and accounts only for non–synonymous site rates, Table 4–1). Furthermore, it appears that a model of non–synonymous rates over sites, with codon frequencies computed over all

genes in the data set, best accounts for the observed RSCU difference between variant and invariant sites found in *Reclinomonas*.

For the *Saccharomyces* data set, simulating according to codon frequencies estimated over all genes in the data set also lead to the detection of fewer IRIS codons as under synonymous selection (models NSD and ND, Table 4–1, Figure 4–3a). Therefore, accounting for codon bias over all genes, as opposed to a single gene at a time, better explains the observed difference in RSCU in invariant and variant sites for IRIS codons at a P–value of 0.05 (Figure 4–3a).

Collectively, the evidence in *Reclinomonas* and *Saccharomyces* suggests that there are perhaps global evolutionary pressures on a set of codons, as opposed to individual gene pressures as assumed in most non–parametric analyses.

Accounting for synonymous rates across sites did not change the degree explanatory power of the phylogenetic model greatly. However, these synonymous rates were calculated based upon all the species in the data sets (globally), not just the *Saccharomyces* and *Reclinomonas* species which were tested for HSCs. Analysis (not shown) indicates that the global mean synonymous substitution rate of a codon is approximately the same for all codons. Thus, the synonymous selection that occurs, appears to be only among the *Saccharomyces* and *Reclinomonas* (local) species that were tested for HSCs. This indicates that the synonymous selection detected by parametric analysis of codon selection finds codons that are conserved more than expected under neutral evolution at a local level only.

4.5.3 Major codons in Saccharomyces

When non–parametric analyses are performed on the *Saccharomyces* data, it is possible to identify major codons which agree with the results of (Sharp et al., 1988) (data not shown). In this well–known data set, the differences in codon bias of a gene can be explained by the expression level of the gene as measured by the CAI, where highly expressed genes (those with a high CAI) contain more major codons. Hence, one question that arises is whether HSCs correspond to major codons previously identified in yeast (Sharp et al., 1988).

When the HSCs in *Saccharomyces* are compared to codons with high RSCU in highly expressed genes of *Saccharomyces cerevisiae*, 9 of the 11 (from Table 4–3) have high RSCU in highly expressed genes (Sharp et al., 1988). Indeed, these nine codons have high RSCU in both highly and lowly expressed genes (Sharp et al., 1988). Conversely, the 15 codons that were not identified as HSC have low RSCU in highly expressed genes, and high RSCU in lowly expressed genes. This indicates that HSCs in *Saccharomyces* correspond in part to previously identified major codons that are more commonly found in highly expressed genes. Furthermore, these codons also have the highest RSCU out of all codons for a particular amino acid in lowly expressed genes, implying that HSCs have high RSCU over a set of genes, regardless of expression level. Perhaps the reason that major codons and HSCs correspond in *Saccharomyces* is not only because of tRNA fidelity, but because highly expressed genes in *Saccharomyces* tend to evolve more slowly (Drummond et al., 2005) and thus have more invariant sites.

The only two HSCs that are not major codons as identified by Sharp (Sharp et al., 1988) are ATA and TTA. This identification could be due to codon bias toward A+T, however none of the five other IRIS codons in *Saccharomyces* with only A/T at each codon position were identified as a HSC. Furthermore, in *Saccharomyces* both ATA and TTA are translated by single tRNAs that only translate that particular codon. There could be some biological constraint causing these codons to be conserved over short evolutionary distances more than expected under neutral models of evolution.

Two previously identified major codons from Table 4–3 that are not HSCs are CAA and GAA. Although CAA is highly selected according to two of the models (NSG and NG where codon frequencies are calculated by gene) the other models find no significant synonymous selection. Since neither CAA nor GAA is a HSC none of the major codons identified by Sharp et al. (Sharp et al., 1988) that have adenine in the second codon position are HSCs. This suggests that codons with adenine in the second position are not selected for by yeast, which is surprising in an A+T biased genome.

4.5.4 Non–parametric approaches to detecting codons under synonymous selection in Reclinomonas

Non–parametric methods are currently applied to data sets to detect codon bias, and infer which codons might be selected at a synonymous level by determining the most highly biased codons. It is important to apply such a non–parametric

approach to the *Reclinomonas* data to determine if these methods can also identify codons that are under synonymous selective pressure.

Non–parametric approaches to codon usage analysis attempt to determine if synonymous selection occurs in a particular group of genes based upon summary statistics of the data of interest. These summary statistics are correlated to the axis of greatest variation in the data. If this axis of variation correlates highly with a given summary statistic or property of the data, it is concluded that these properties explain the variation in the data. Table 4–4 gives the correlation values and significance of correlation for various properties of the *Reclinomonas* species that can be calculated directly from the data, with no prior information. A1 represents the principal axis in correspondence analysis (or axis of greatest variation). Pearson's correlation was calculated between the principal axis (A1) and the G+C content, gene rates (as calculated using the non–parametric DistR approach (Bevan et al., 2005)), fraction of codons conserved, and effective number of codons N_c (Table 4–4). In all *Reclinomonas* species there is strong correlation between the principal axis and both the G+C content and gene rate. Moderate correlation is found with the percentage of codons conserved between the three species.

These results indicate that the codon bias of individual genes in *Reclinomonas* is primarily explained through the rate of mutation of the gene, and the G+C content of genes. Correlation with the percentage of codons conserved is a further reflection of the mutational bias. Based on this analysis, accounting for gene rate heterogeneity and mutational bias in parametric models, should explain

	R.am. 83	*R.am. 84*	*R.am 94*	*R.am.sp*
A1–G+C	-0.779**	-0.740**	-0.763 ***	-0.713 **
A1–rates	0.720 **	0.716 **	0.741 **	0.698**
A1–PC	-0.489 **	-0.417 *	-0.538 **	-0.438*
A1–N_c	-0.078	-0.058	—0.034	0.053

Table 4–4: Pearson's correlation of Axis 1 (A1), which represents the greatest variation within the codon bias of the data with properties of the data for *Reclinomonas*. These properties include correlation of Axis 1 with G+C Content (A1–G+C), correlation of Axis 1 with gene rate (A1–rate), correlation of Axis 1 with percent conserved (A1–PC), and correlation of Axis 1 with N_c (A1–N_c, where N_c is the effective number of codons (Wright, 1990) as measured for each gene). A higher correlation value indicates that the property of interest explains a lot of the variation in codon usage observed over different genes. *P–value $< 5.0e-3$, **P–value $< 1.0e-8$, ***P–value $< 1.0e-30$.

the difference in RSCU between invariant and variant sites in all IRIS codons identified in *Reclinomonas*. However, parametric analysis identifies 10 IRIS codons as HSCs. Thus, in the case of the highly A+T biased *Reclinomonas* mtDNA genomes, non–parametric methods fail to identify that synonymous selection occurs in particular codons.

4.6 Conclusions

Both *Reclinomonas* and *Saccharomyces* have certain codons that undergo selection against synonymous mutation according to the codon models applied in this analysis. Using codon frequencies calculated over all genes versus gene codon frequencies does not significantly affect the results. This indicates that it is not clear that there are gene specific synonymous codon selective pressures.

In the *Saccharomyces* data set a number of the HSCs correspond to codons which have a greater RSCU in highly expressed genes (or major codons), and are

thought to be selected for at the translational level (Akashi, 2001). This indicates that codons with unusual usage patterns under non–parametric methods are also found to be selected for at the synonymous level under parametric codon models. However, there are also codons in the *Reclinomonas* data set which are detected to be under synonymous selective pressure. Because the results in *Saccharomyces* correlate well with well–known results regarding codons that are used in highly expressed genes, this indicates that HSCs in *Reclinomonas* have some other important biological function.

HSCs are not detected in *Reclinomonas* using non–parametric methods which attribute the observed codon usage bias to gene rate and mutational bias. This suggests that the strong A+T bias of *Reclinomonas* misleads non–parametric methods, which focus solely on summary statistics of codon usage bias so that unusual patterns of codon selection cannot be detected.

4.7 Acknowledgements

We thank Trevor Bruen and David Bryant for helpful comments on the manuscript. Salary and support from the Canadian Institutes of Health Research (MOP 42475; BFL), the Canadian Institute for Advanced Research (CIAR; BFL) and supply of laboratory equipment and informatics infrastructure by Genome Canada is gratefully acknowledged.

Chapter 5
Conclusions

Two main problems were addressed in this thesis: (i) Quickly estimating rate heterogeneity on a set of sequences and how best to incorporate such heterogeneity into phylogenetic methods; (ii) Identification of sets of codons under synonymous selection using parametric phylogenetic models that provide a null hypothesis of evolution.

5.1 Gene Rate Heterogeneity

The DistR algorithm was developed to quickly infer gene rates from pairwise distances between taxa in Chapter 2, (Bevan et al., 2005). Simulation demonstrates that the DistR approach is quite accurate when compared to ML gene rate estimates. Furthermore, when incorporated into PHYML in order to account for gene rate heterogeneity, better model fit was found, with a more accurate tree for a fungal mtDNA data set (Bevan et al., 2005).

This work was extended to study the question of how best to account for rate heterogeneity in phylogenetic analyses in Chapter 3. Rate heterogeneity was accounted for by either allowing each gene to have a ML gene rate (the n–parameter approach (Yang, 1996; Pupko et al., 2002b; Bevan et al., 2005)), or by

integrating over gene rates (the α–parameter approach (Felsenstein, 2004b)). The former approach requires less computation, and thus is faster. The later approach, although slower, has the advantage of requiring fewer parameters. Yet, results demonstrate that there is no reason to prefer the computationally more complex α–parameter method.

Furthermore, improved model fit according to the AIC was correlated with three properties of the data: slowest gene rate, fastest gene rate, and the difference between the two. As more data was present in the analysis (in terms of number of genes) the correlation with the slowest gene rate increased. Correlation with the other two properties decreased. This indicates that accounting for the rate of the slow sites properly is what primarily leads to improved model fit of a gene rate heterogeneity model over the concatenated model.

Chapters 2 and 3 demonstrate the importance of accounting for gene rate heterogeneity in phylogenetic analysis. However, the current analyses provide only a first step in how to account for heterogeneity in gene evolution when analyzing multi–gene data sets.

Analysis in Chapter 3 indicates that gene rate heterogeneity and site rate heterogeneity are correlated. This raises the question of partitioning data: if data could be partitioned so that sites with similar rates of evolution are clustered, then it is possible that only one rate of evolution would be necessary at each site, where the site rate would be equal to the cluster rate similarly to the *fixed rates* models of Yang (Yang, 1994).

Furthermore, there is no method to account for heterotachy (change in the rate of evolution for specific sites along a given lineage) in this work. Incorporating both heterotachy and rate heterogeneity into phylogenetic analysis would likely lead to even better model fit. Mixture models that allows for both gene rate heterogeneity and heterotachy in different clusters should prove ideal to address this problem, and have been widely used in phylogenetics to address related questions (Pagel and Meade, 2004; Lartillot and Philippe, 2004; Kolaczkowski and Thornton, 2004; Mayrose et al., 2005).

5.2 Codon Selection

In Chapter 4 the focus of the thesis changed to using phylogenetic models that account for rate heterogeneity, mutational bias etc. to study synonymous codon selection. Selection on a set of codons is currently determined using non–parametric approaches which summarize properties of the data. Well–known codon models that provide a null hypothesis of codon evolution exist, however they are used only to determine if a site is under positive or purifying selection (Yang and Nielsen, 2000; Yang et al., 2000; Pond and Muse, 2005). In this chapter codon models were used to simulate data under neutral evolution. The difference in Relative Synonymous Codon Usage in invariant and variant sites from the empirical data was then compared to the simulated data. In *Saccharomyces* this lead to the identification of 11 *Highly Selected Codons (HSCs)*, nine of which were identified as codons under synonymous selective pressure using non–parametric

approaches. However, in *Reclinomonas* mtDNA, non–parametric approaches were unable to identify any codons under synonymous selective pressure. Rather the codon usage bias across different genes was attributed to A+T bias and gene rate. However, the approach developed in this thesis identified 10 HSCs.

The analysis in Chapter 4 provides an ideal start to studying codon evolution using codon models. However, the techniques used are quite conservative. For instance, only codon sites that have no mutation over the species of interest are assumed to be potentially under synonymous selection. It is possible that sites with few mutations are also under synonymous selection. Thus, sites could be categorized not according to the number of observed mutations, but according to the rate of synonymous evolution. It might also be possible to incorporate a more formal definition of synonymous selection into codon model using mixture models (i.e. sites that are free to vary versus sites that are not free to vary) in order to infer if particular codons are under synonymous selection.

References

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control AC19:716–723.

Akashi, H. 2001. Gene expression and molecular evolution. Current Opinion in Genetics and Development 11:660–666.

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Duruflé, T. Gaasterland, P. Lopez, M. Müller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proceedings of the National Academy of Sciences 99:1414–1419.

Beirne, N. and A. Eyre-Walker. 2005. Variation in synonymous codon use and DNA polymorphism within the *d*rosophila genome. Journal of Evolutionary Biology 19:1–11.

Bennetzen, J. L. and B. D. Hall. 1982. Codon selection in yeast. Journal of Biological Chemistry 257:3026–3031.

Bevan, R. B. and B. F. Lang. 2004. Mitochondrial genome evolution: the origin of mitochondria and of eukaryotes. Topics in Current Genetics 8:1–15.

Bevan, R. B., B. F. Lang, and D. Bryant. 2005. Calculating the evolutionary rates of different genes: A fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. Systematic Biology 54:900–915.

Brinkmann, H. and H. Philippe. 1999. Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Molecular Biology and Evoluton 16:817–825.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. P. de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Systematic Biology 54:743–757.

Bull, J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. Systematic Biology 42:384–397.

Bullerwell, C. E., L. Forget, and B. F. Lang. 2003. Evolution of monoblepharidalean fungi based on complete mitochondrial genome sequences. Nucleic Acids Research 31:1614–1623.

Bulmer, M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Molecular Biology and Evolution 8:868–883.

Burnham, K. P. and D. R. Anderson. 2003. Model Selection and Multimodel Inference: A Practical Information–Theoretic Approach. Springer–Verlag, New York, NY.

Carbone, A., A. Zinovyev, and F. Kepes. 2003. Codon adaptation index as a measure of dominating codon bias. Bioinformatics 19:2005–2015.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17:540–552.

Churchill, G. A., A. von Haeseler, and W. C. Navidi. 1992. Sample size for a phylogenetic inference. Molecular Biology and Evolution 9:753–769.

Coghlan, A. and H. Wolfe, Kenneth. 2000. Relationship of codon bias to mRNA concentration and protein length in *s*accharomyces cerevisiae. Yeast 16:1131–1145.

Cranston, K. and B. Rannala. 2005. Closing the gap between rocks and clocks. Heredity 94:461–462.

Dacks, J. B., A. Marinets, W. F. Doolittle, T. Cavalier-Smith, and J. M. Logsdon Jr. 2002. Analyses of RNA polymerase II genes from free–living protists: Phylogeny, long branch attraction and the eukaryotic big bang. Molecular Biology and Evolution 19:830–840.

Darwin, C. 1859. On the origin of species by means of natural selectioin. John Murray, Albemarle Street, London.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences 102:14338–14343.

Duret, L. 2000. tRNA gene number and codon usage in the *c.* elegans genome are co–adapted for optimal tranlation of highly expressed genes. Genetics 16:287–289.

Duret, L. and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *caenorhabditis, drosophila*, and *arabidopsis*. Proceedings of the National Academy of Sciences U.S.A. 96:4482–4487.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. Annals of Statistics 7:1–26.

Eulenstein, O., D. Chen, J. G. Burleigh, D. Fernández-Baca, and M. J. Sanderson. 2004. Performance of flip supertree construction with a heuristic algorithm. Systematic Biology 53:299–308.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology 27:401–410.

Felsenstein, J. 1981a. Evolutionary trees from dna sequences: A maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

Felsenstein, J. 1981b. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biological Jouranl of the Linnean Society 16:368–376.

Felsenstein, J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylgenies. Journal of Molecular Evolution 53:447–455.

Felsenstein, J. 2004a. Inferring Phylogenies. Sinauer Associates, Inc.

Felsenstein, J. 2004b. Inferring Phylogenies chap. 30, Pages 537–538. Sinauer Associates, Inc.

Felsenstein, J. 2004c. Inferring Phylogenies chap. 11, Pages 148–149. Sinauer Associates, Inc.

Felsenstein, J. 2004d. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle. URL: http://evolution.genetics.washington.edu/phylip.html.

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Systematic Zoology 20:406–416.

Fitch, W. M. 1977. On the problem of discovering the most parsimonious tree. The American Naturalist 111:223–257.

Foulds, L. R. and R. L. Graham. 1982. The steiner problem in phylogeny is NP–complete. Advances in Applied Mathematics 3:43–49.

Friber, M., P. von Rohr, and G. Gonnet. 2004. Limitations of codon adaptation index and other coding DNA–based feature for prediction of protein expression in *saccharomyces cerevisiae*. Yeast 21:1083–1093.

Gascuel, O. 1997. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Molecular Biology and Evolution 14:685–695.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2000. Bayesian Data Analysis. Chapman and Hall/CRC.

Gill, P., W. Murray, and M. Wright. 1982. Practical Optimization. Academic Press.

Goldman, N. and Z. Yang. 1994. A codon–based model of nucleotide substitution for protein–codon dna sequences. Molecular Biology and Evolution 11:725–746.

Gontcharov, A. A., B. Marin, and M. Melkonian. 2004. Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and *rbc*l sequence comparisons in the Zygnematophyceae (Streptophyta). Molecular Biology and Evolution 21:612–624.

Grassly, N. C., J. Adachi, and A. Rambaut. 1997. PSeq-Gen: An application for the monte carlo simulation of protein sequence evolution along phylogenetic trees. Computational Applications in Bioscience 13:559–560.

Guindon, S. and O. Gascuel. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52:696–704.

Gustafsson, C., S. Govindarajan, and J. Minshull. 2004. Codon bias and heterologous protein expression. Trends in Biotechnology 22:346–353.

Harlow, T. J., J. P. Gogarten, and M. A. Ragan. 2004. A hybrid clustering approach to recognition of protein families in 114 microbial genomes. BMC Bioinformatics 5:1–14.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22:160–174.

Hirt, R. P., J. M. Logsdon Jr., B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. 1999. Microsporidia are related to fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. Proceedings of the National Academy of Sciences 96:580–585.

Huelsenbeck, J. P., J. Bull, and C. W. Cunningham. 1996. Combining data in phylogenetic analysis. Tree 11:152–158.

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of bayesian inference of phylogeny. Systematic Biology 51:673–688.

Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Ikemura, T. 1981. Correlation between the abundance of escherichia coli transfer rnas and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the e. coli translational system. Journal of Molecular Biology 151:389–409.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Computer Applications in the

Biosciences (CABIOS) 8:275–282.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature 217:624–625.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press.

King, J. L. and T. H. Jukes. 1969. Non–Darwinian evolution. Science 164:788–798.

Kishino, H. and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the braching order in hominoidea. Journal of Molecular Evolution 29:170–179.

Kliman, R. M., N. Irving, and M. Santiago. 2003. Selection conflicts, gene expression, and codon usage trends in yeast. Journal of Molecular Evolution 57:98–109.

Kolaczkowski, B. and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature Pages 980–984.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger. 2002. The closest unicellular relatives of animals. Current Biology 12:1773–1778.

Lapointe, F. and G. Cucumel. 1997. The average consensus procedure: Combination of weighted trees containing identical or overlapping sets of taxa. Systematic Biology 46:306–312.

Larget, B. and D. L. Simon. 1999. Markov Chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution 16:750–759.

Lartillot, N. and H. Philippe. 2004. A bayesian mixture model for across–site heterogeneities in the amino–acid replacement process. Molecular Biology and Evolution 21:1095–1109.

Lockhart, P. J., A. W. D. Larkum, M. A. Steel, P. J. Waddell, and D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proceedings of the National Academy of Sciences 93:1930–1934.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610–614.

Massingham, T. and N. Goldman. 2005. Detecting amino acid sites under positive selection and purifying selection. Genetics 169:1753–1762.

Mayrose, I., N. Friedman, and T. Pupko. 2005. A "gamma" mixture model better accounts for among site rate heterogeneity. Bioinformatics 21:ii151–ii158.

Mayrose, I., D. Graur, N. Ben-Tal, and T. Pupko. 2004. Comparison of site-specific rate-inference methods: Empirical Bayesian methods are superior. Molecular Biology and Evolution 21:1781–1791.

Merkl, R. 2003. A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. Journal of Molecular Evolution 57:453–466.

Moriyama, E. N. and J. Powell. 1997. Codon selection in yeast usage bias and tRNA abundance in *d*rosophila. Journal of Molecular Evolution 45:514–523.

Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618.

Muse, S. V. and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Molecular Biology and Evolution 11:715–724.

Nakamura, Y. 2005. Codon usage database. http://www.kazusa.or.jp/codon/.

Nakamura, Y., T. Gojobori, and T. Ikemura. 2000. Codon usage tabularted from the international DNA sequence databases: status for the year 2000. Nucleic Acids Research 28:292.

Needleman, S. G. and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48:443–453.

Nei, M. and S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Systematic Biology 53:47–67.

Olsen, G. J., S. Pracht, and R. Overbeek. 1993. DNArates. URL: http://geta.life.uiuc.edu/ gary/programs/DNArates.html.

Pagel, M. and A. Meade. 2004. A phylogenetic mixture model for detecting pattern–heterogeneity in gene sequence or character–state data. Systematic Biology 53:571–581.

Peden, J. F. 1999. Analysis of codon usage. Ph.D. thesis University of Nottingham.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005. Heterotachy and long–branch attraction in phylogenetics. BMC Evolutionary Biology **5**.

Pond, S. K. and S. V. Muse. 2005. Site–to–site variation of synonymous substitution rates. Molecular Biology and Evolution .

Pupko, T., R. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. 2002a. Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18:S71–S77.

Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa. 2002b. Combining multiple data sets in a likelihood analysis: Which models are the best? Molecular Biology and Evolution 19:2294–2307.

Reeves, J. H. 1992. Heterogeneity in the substitution process of amino acid sties of proteins coded for by mitochondrial DNA. Journal of Molecular Evolution 35:17–31.

Rokas, A. and S. B. Carroll. 2005. More genes of more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Molecular Biology and Evolution 22:1337–1344.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome–scale approaches to resolving incongruence in molecular phylogeniesscale apporaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Ross, S. M. 2003. Introduction to Probability Models. Academic Press.

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

Sharp, P. and W.-H. Li. 1987. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research 15:1281–1295.

Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe, and F. Wright. 1988. Codon usage patterns in *escherichia coli, bacillus subtilis, saccharomyces cerevisiae, schizosaccharomyces pombe, drosophila melanogaster* and *homo sapiens*; a review of the considerable within–species diversity. Nucleic Acids Research 16:8207–8211.

Shimodaira, H. and M. Hasegawa. 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

Smith, T. F. and M. S. Waterman. 1981. Identification of common molecular subsequences. Journal of Molecular Biology 147:195–197.

Smyth, P. 2000. Model selection for probabilistic clustering using cross–validated likelihood. Statistics and Computing **10**:63–72.

Sorensen, H. P. and K. K. Mortensen. 2005. Advanced genetic strategies for recombinant protein expression in escherichia coli. Journal of Biotechnology 115:113–128.

Springer, M. S., M. J. Stanhope, O. Madsen, and W. W. de Jong. 2004. Molecules consolidate the placental mammal tree. Trends in Ecology and Evolution 19:430–438.

Sumida, M., Y. Kanamori, H. Kaneda, Y. Kato, M. Nishioka, M. Hasegawa, and H. Yonekawa. 2001. Complete nucleotide sequence and gene rearrangement of the mitochondrial genome of the japanese pond frog *Rana nigromaculata*. Genes and Genetic Systems 76:311–325.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Molecular Systematics chap. Phylogenetic Inference, Pages 407–514. Sinauer Associates.

Tanaka, J. S. and G. J. Huba. 1985. A fit index for covariance structure models under arbitrary GLS estimation. British J. Math. Statist. Psych. 38:197–201.

Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lectures in mathematics and in the life sciences 17:57–86.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acid Research 22:4673–4680.

Tomita, K., S. Yokobori, T. Oshima, T. Ueda, and K. Watanabe. 2002. The cephalopod *Loligo bleekeri* mitochondrial genome: Multiplied noncoding regions and transposition of tRNA genes. Journal of Molecular Evolution 54:486–500.

Voet, D., J. G. Voet, and C. W. Pratt. 1999. Fundamentals of Biochemistry. John Wiley and Sons, Inc.

Waddell, P. J. and M. A. Steel. 1997. General time–reversible distances with unequal rates across sites: Mixing γ and inverse gaussian distributions with invariant sites. Molecular Phylogenies and Evolution 8:398–414.

Whelan, S. and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a Maximum–Likelihoo approach. Molecular Biology and Evolution 18:691–699.

Wright, F. 1990. The 'effective number of codons' used in a gene. Gene 87:23–29.

Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10:1396–1401.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution 39:306–314.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. Journal of Molecular Evolution 42:587–596.

Yang, Z. 1997. Phylogenetic analysis by maximum likelihood (paml). URL: http://abacus.gene.ucl.ac.uk/software/paml.html.

Yang, Z. and R. Nielsen. 2000. Estimating synonymous and nonsynonymou substitution rates under realistic evolutionary models. Molecular Biology and Evolution 17:32–43.

Yang, Z., R. Nielsen, N. Goldman, and A.-M. K. Pedersen. 2000. Codon-substitution models for heterogenous selective pressure at amino acid sites. Genetics 155:431–449.

Appendix 2

Reprints

Syst. Biol. 54(6):900–915, 2005

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500354829

Calculating the Evolutionary Rates of Different Genes: A Fast, Accurate Estimator with Applications to Maximum Likelihood Phylogenetic Analysis

RACHEL B. BEVAN,[1] B. FRANZ LANG,[2] AND DAVID BRYANT[1]

[1]*McGill Centre for Bioinformatics, Duff Medical Building, 3775 University Street, Montréal, Quebec, H3A 2B4, Canada*
[2]*Program in Evolutionary Biology, Canadian Institute for Advanced Research; Centre Robert Cedergren, Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, Montréal, Québec, H3T 1J4, Canada;*
E-mail: rachel@mcb.mcgill.ca. (R.B.B.)

Abstract.—In phylogenetic analyses with combined multigene or multiprotein data sets, accounting for differing evolutionary dynamics at different loci is essential for accurate tree prediction. Existing maximum likelihood (ML) and Bayesian approaches are computationally intensive. We present an alternative approach that is orders of magnitude faster. The method, Distance Rates (DistR), estimates rates based upon distances derived from gene/protein sequence data. Simulation studies indicate that this technique is accurate compared with other methods and robust to missing sequence data. The DistR method was applied to a fungal mitochondrial data set, and the rate estimates compared well to those obtained using existing ML and Bayesian approaches. Inclusion of the protein rates estimated from the DistR method into the ML calculation of trees as a branch length multiplier resulted in a significantly improved fit as measured by the Akaike Information Criterion (AIC). Furthermore, bootstrap support for the ML topology was significantly greater when protein rates were used, and some evident errors in the concatenated ML tree topology (i.e., without protein rates) were corrected. [Bayesian credible intervals; DistR method; multigene phylogeny; PHYML; rate heterogeneity.]

It is widely recognized that the analysis of multiple unlinked genes is superior to single gene analyses for phylogenetic reconstruction. These unlinked genes may, however, be evolving according to very different rules. Heterogeneity of the evolutionary process must be accounted for in phylogenetic analyses (Bapteste et al., 2002; Bull et al., 1993; Huelsenbeck et al., 1996; Nylander et al., 2004; Pupko et al., 2002b; Yang, 1996). The concept of accounting for differing evolutionary pressures within phylogenetic analysis is not new (Yang, 1993). Site-specific rates of evolution can be computed for amino acids (e.g., Rate4Site, Mayrose et al., 2004; Pupko et al., 2002a) and DNA (e.g., DNArates, Olsen et al., 1993) using both Bayesian and maximum likelihood approaches.

Site rates within a gene are likely to be more correlated than rates for sites in different genes. To account for this, it can be assumed that each gene evolves at a different average rate and that these gene rates are drawn from some common distribution (Cranston and Ranala, 2005; Felsenstein, 2001, 2004a). Both Bayesian (Huelsenbeck and Ronquist, 2001) and maximum likelihood (Pupko et al., 2002b; Yang, 1996) methods exist to estimate gene rates (or more generally, locus rates) but these are computationally expensive.

We present a fast, accurate method to estimate the relative evolutionary rates of genes/proteins. For example, when run on a data set with 63 proteins over 123 taxa the algorithm takes less than a second. The method can be applied to protein or nucleotide data, though here we focus on protein sequences. The basic idea is to use pairwise estimates of evolutionary divergence (distances) to deduce the relative rates of different proteins, even when the proteins are not all present in all of the taxa. Although this approach does not give the ML estimates for the rates (Pupko et al., 2002b, Yang, 1996), it does provide an excellent approximation.

After computing rates they are incorporated as extra parameters into the ML tree search, resulting in improved fit as measured by the AIC. The rates estimated using the DistR procedure have been coded into PHYML version 2.2, available at http://atgc.lirmm.fr/phyml/ (Guindon and Gascuel, 2003). PHYML was used because incorporation of the rates was straightforward and because PHYML is an especially fast implementation of ML.

METHODS

The DistR Method

To begin with, the method will be explained through an example. Figure 1 represents three different protein alignments. Not all taxa are present in all three alignments. Suppose that the three proteins have rates r_1, r_2, and r_3. These rates will affect distances inferred from the alignments. Reversing the problem involves using the pairwise distances between species to estimate the different rates r_1, r_2, and r_3.

Figure 1 outlines two ways of obtaining distances from each protein. In the first method ML trees are constructed and the length of the path between two taxa in these trees is measured (referred to hereafter as *patristic ML distances*). In the second method distances are estimated directly from the alignments, as is customary in distance-based methods (referred to hereafter as *pairwise ML distances*). The end result from both methods is a distance matrix for each protein.

If the rate in one protein is twice the rate in a second protein, then the expected distance estimates from the first protein should be twice the expected distance estimates from the second protein. This should hold, approximately, for both pairwise ML distances and patristic ML distances. Equivalently, the distance estimate from the first protein, divided by two, should be approximately the distance estimate of the second protein.

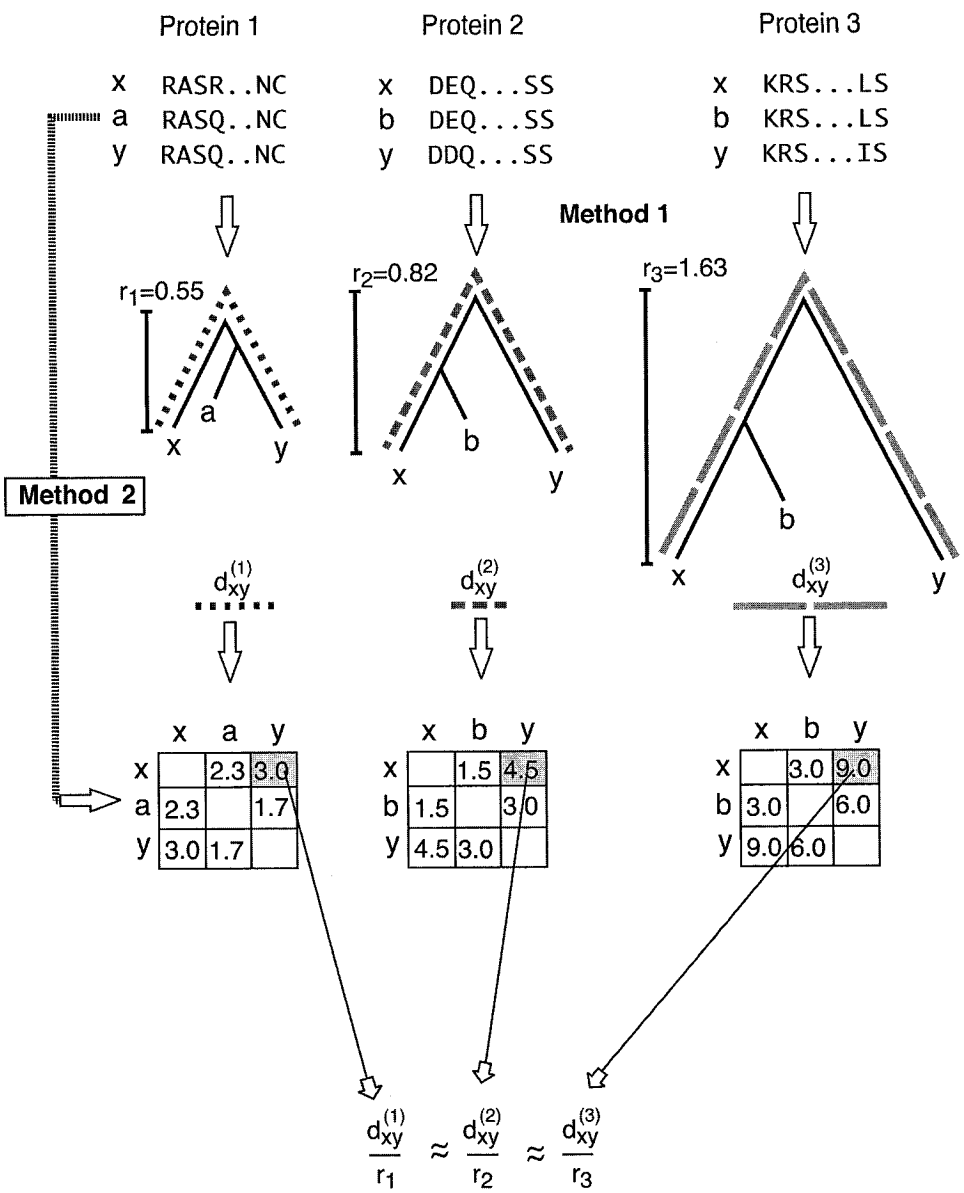


FIGURE 1. The general idea of the DistR estimation procedure. Beginning with individual protein alignments over a set of taxa (with missing data), distances between the species are estimated for each protein alignment. There are two choices of how to estimate the distances: directly from the alignment data (method 2); as the sum of the pairwise distances between taxa on a tree built from the alignment data (method 1). The result is a matrix of pairwise distances between taxa. The ratio of the pairwise distances to the rate of evolution of the protein should be approximately the same for all proteins.

In the example (Fig. 1), and later on, the distance between taxa x and y estimated from protein k is denoted $d_{xy}^{(k)}$, irrespective of whether it is a pairwise or patristic ML distance. Suppose that, for each k, the rate in protein k equals r_k. It follows that $\frac{d_{xy}^{(1)}}{r_1}$ will be approximately equal to $\frac{d_{xy}^{(2)}}{r_2}$ which in turn will be approximately equal to $\frac{d_{xy}^{(3)}}{r_3}$. This is denoted as

$$\frac{d_{xy}^{(1)}}{r_1} \approx \frac{d_{xy}^{(2)}}{r_2} \approx \frac{d_{xy}^{(3)}}{r_3}, \quad (1)$$

where "≈" means "approximately equal." In Figure 1, this gives $\frac{3.0}{0.55} \approx \frac{4.5}{0.82} \approx \frac{9.0}{1.63}$.

In a sense, the distance estimates obtained from each gene are normalized so that the scale is the same. Define this normalized distance or *consensus distance* between any two taxa as p_{xy}, with the assumption that

$$p_{xy} \approx \frac{d_{xy}^{(1)}}{r_1} \approx \frac{d_{xy}^{(2)}}{r_2} \approx \frac{d_{xy}^{(3)}}{r_3}.$$

Assume that rates r_1, r_2, and r_3 in Figure 1 are unknown, whereas the distances remain known. The above

approximate equality leads to

$$p_{xy} \approx \frac{3.0}{r_1} \approx \frac{4.5}{r_2} \approx \frac{9.0}{r_3}. \quad (2)$$

The unknowns p_{xy}, r_1, r_2, and r_3 can be solved for using a least squares approach.

The relation in Equation (2) provides a framework to solve for the *relative rates* r_1, r_2, and r_3, given estimates for the distances $d_{xy}^{(k)}$. This is the basic idea behind the method. The main issues are how to (a) handle the fact that the relations are only approximate; (b) deal with missing distances; (c) compute the rate estimates quickly. These issues are addressed in the following text and in Appendix 2.

To formalize the problem, suppose that there are n proteins (or genes, etc.) over m species. The distance between species x and y derived from protein k is denoted $d_{xy}^{(k)}$. The basic assumption made is that the ratio of the estimated distance between a pair of taxa for a given protein ($d_{xy}^{(k)}$ for protein k and taxa x,y), to the rate of the protein (r_k for protein k), is approximately equal across all proteins.

The rates $r_1, r_2, \ldots, r_n$ are unknown quantities to be estimated based upon the distance data from a given protein alignment. To do this, assume that there exists an unknown consensus distance p_{xy} such that

$$p_{xy} \approx \frac{d_{xy}^{(1)}}{r_1} \approx \frac{d_{xy}^{(2)}}{r_2} \approx \cdots \approx \frac{d_{xy}^{(n)}}{r_n},$$

where $n = 3$ for the example in Figure 1. All the consensus distances and rates can now be estimated using a least-squares approach.

In the least squares method it is possible to incorporate measures of uncertainty about the estimated distances $d_{xy}^{(k)}$. Distance estimates with low variance should contribute more to the analysis, whereas distance estimates with high variance (or infinite variance in the case of missing entries) should contribute little. Let $w_{xy}^{(k)} \geq 0$ be a measure of the uncertainty in the distance estimate between taxa x and y derived from protein k. If $d_{xy}^{(k)}$ is accurate, then $w_{xy}^{(k)}$ should be high. If there is less certainty about the accuracy of $d_{xy}^{(k)}$, then $w_{xy}^{(k)}$ should be low. This is achieved using the inverse of the variance of $d_{xy}^{(k)}$, that is, $w_{xy}^{(k)} = \frac{1}{\mathrm{Var}(d_{xy}^{(k)})}$. If protein k is not present in both x and y, then $w_{xy}^{(k)} = 0$. To measure the variance of the distance estimates the approximate formula of Bulmer (1991) is used in the implementation of DistR. Other variance estimators could also be used.

Under a weighted least-squares (WLS) framework the total discrepancy between the ratios $\frac{d_{xy}^{(n)}}{r_n}$ and the consensus distances p_{xy} is measured by

$$q(\mathbf{p}, \mathbf{r}) = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} \left(p_{xy} - \frac{d_{xy}^{(k)}}{r_k} \right)^2 \quad (3)$$

where $\mathbf{p}$ denotes the vector $[p_{12}, p_{13}, \ldots, p_{\frac{(m-1)m}{2}}]^T$ and $\mathbf{r}$ denotes the vector $[r_1, \ldots, r_n]^T$. This is similar to the minimization function used by Lapointe and Cucumel (1997) in the average consensus method. The main difference is that they assume one rate over all proteins, whereas this method includes different rates for each protein. Note that if taxa x and y are missing from a protein k then an estimate for $d_{xy}^{(k)}$ cannot be obtained. However, this is not a problem since the weight $w_{xy}^{(k)}$ will be zero in this case.

Estimating both rates and consensus distances using $q(\mathbf{p}, \mathbf{r})$ leads to the problem of *nonidentifiability*. In the absence of any error each estimated protein distance $d_{xy}^{(k)}$ is the product of the rate of the protein r_k and the consensus distance p_{xy}. Thus, a perfect fit to the equation is still achieved if all the rates are multiplied by some constant and all the consensus distances divided by the same constant. There is a problem of determining scale. Hence, Equation (3) does not have a well-defined minimum. To solve this problem a constraint

$$\sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} p_{xy} = \kappa \quad (4)$$

must be added to system, where κ is an arbitrary positive constant. The particular value of κ is irrelevant since changing κ merely causes all estimated rates to be multiplied by the same constant value. For this reason, it is possible to infer *relative rates* only. In DistR $\kappa = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} d_{xy}^{(k)}$, thus constraining the weighted estimated distances to be equal to the weighted consensus distances. This was empirically determined to minimize the variance of the DistR estimates.

Appendix 3 describes an extremely fast algorithm for minimizing the function $q(\mathbf{p}, \mathbf{r})$ subject to the constraint in Equation (4). The algorithm takes $O(nm^2 + n^3)$ time and $O(n^2 + m^2)$ memory. For example, when run on a data set with 63 proteins over 123 taxa, the algorithm takes less than a second. An implementation with source code is available at http://www.mcb.mcgill.ca/~rachel.

Experimental Studies

An extremely rapid method for estimating the relative rates of different genes has been proposed. The method is orders of magnitude faster than existing ML and Bayesian approaches. The most important question remaining is to what extent this increase in speed affects the accuracy of the estimates. In order to address this question, the accuracy of the new method was assessed using both simulated and empirical data.

In all the analyses PHYML (version 2.2) was used (Guindon and Gascuel, 2003) to compute ML distances and trees, with a JTT protein model, eight gamma categories plus invariant sites and the default (BIONJ) starting tree. The gamma shape parameter and proportion of invariant sites were estimated using default optimization routines in the program. When constructing ML trees from real data several bootstrap values were computed. As detailed below these values depend upon: whether patristic or pairwise ML distances were used in the DistR procedure; whether the rates were reestimated for each bootstrap replicate.

For both the simulated and empirical data, DistR estimates based upon patristic and ML distances were compared. This comparison was made in order to determine whether or not the additional computational effort required for estimating patristic ML distances is justified.

Experimental Studies—Simulated Data

The two key questions addressed through the simulation studies are:

- *Patristic versus pairwise ML distances.*—How accurate are the rate estimates using pairwise versus patristic ML distances?
- *Missing distances between taxa.*—How are DistR rate estimates affected when proteins are not present in all taxa?

To answer these questions protein alignments were simulated using Pseq-Gen (Grassly et al., 1997) with the JTT model of evolution. The initial tree and branch lengths were taken from an independent analysis of mitochondrial Atp8 proteins in 58 eukaryotes. Two types of simulations were carried out. The first, intended to address the first question, involved construction of 20 protein trees by randomly deleting taxa from the starting tree. In total there were four protein trees with 53 taxa, four with 48 taxa, four with 43 taxa, four with 38 taxa, and four with 33 taxa. For each tree a rate was sampled from a precomputed distribution of rates based on real data (data not shown), and protein alignments of length 100, 300, 500, and 1000 generated using Pseq-Gen (Grassly et al., 1997) (note that the average length of naturally occurring proteins is approximately 300 amino acids). The second analysis, intended to address the second question, increased the number of taxa deleted from the starting tree. In total there were seven trees with 25% of the taxa, seven with 50% of the taxa, and seven with 75% of the taxa. This resulted in 21 trees, 7 each with 16, 30, and 44 taxa, respectively. For each tree a rate was sampled from a precomputed distribution of rates based on real data (data not shown), and protein alignments of length 1000 generated using Pseq-Gen (Grassly et al., 1997). This experiment follows a protocol proposed by (Eulenstein et al., 2004). For both experiments, and for every set of parameters, 10 replicates of the experiment were performed. See Figure 2 for an overview of the simulations.

Statistics measured on the simulated data, including goodness-of-fit and mean squared error, are explained in detail in Appendix 1. These statistics were used to relate the accuracy of the DistR rate estimates to the known rates at which the proteins were simulated.

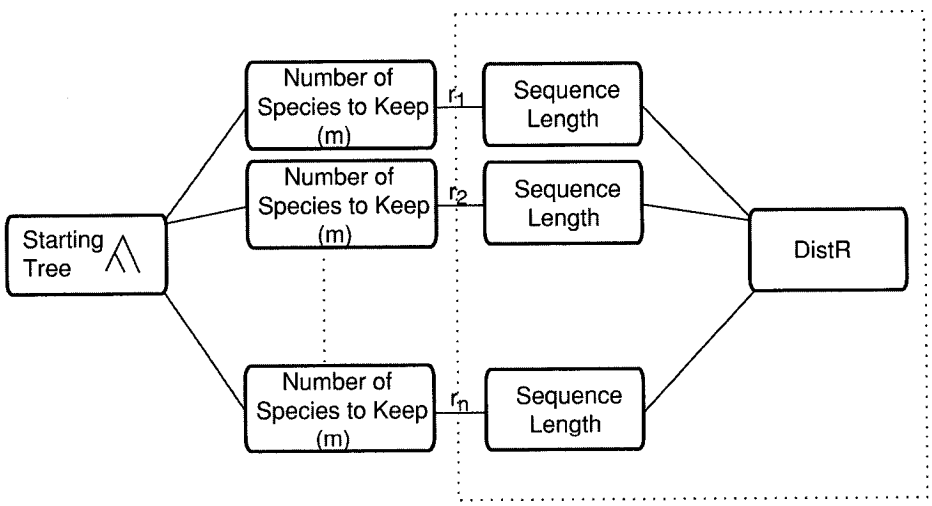


FIGURE 2. The general flow of the simulation studies. Two studies were performed, one with $n = 20$ and the other with $n = 21$ (where n is the number or proteins). The first study compared different methods of estimating distances using different alignment lengths. In the first study, 20 random subtrees from an original tree of 58 species were created, four each of size $m = 33$, $m = 38$, $m = 43$, $m = 48$, and $m = 53$ (where m is the size of the taxon set for a given protein). For each tree, a rate was sampled from a precomputed distribution of rates based on real data (data not shown). Protein alignments of length 100, 300, 500, and 1000 were simulated using Pseq-Gen (Grassly et al., 1997). A second analysis compared rate estimates with increasing amounts of data. Twenty-one random subtrees from the original tree of 58 species were created, 7 each of size $m = 16$, $m = 30$, and $m = 44$ (corresponding to approximately 25%, 50%, and 75% of the species [as in Eulenstein et al., 2004]). For each tree, a rate was sampled from a precomputed distribution of rates based on real data (data not shown). Alignments of length 1000 were generated. For both studies, 10 replicates were performed for each set of parameters.

Experimental Studies—Empirical Data

The data analyzed in this study consist of a set of 15 aligned mitochondrial protein sequences from 29 taxa. The taxon names and accession numbers are given in Table 1. Protein names and alignment accession numbers appear in Table 2. This multiprotein data set is of moderate size, and variants thereof have been used in numerous publications (e.g., Bullerwell et al., 2003; Lang et al., 2002; Sumida et al., 2001; Tomita et al., 2002). Furthermore, some of the species have high evolutionary rates and substitutional saturation of sites (i.e., *Smittium*), whereas others have very short branches in the resulting phylogenetic tree. Combined, these two properties can cause inaccurate grouping of the taxa due to long-branch attraction artifacts (Felsenstein, 1978).

TABLE 1. Empirical data analyzed. Names and accession numbers for protein sequences studied from Fungal species and outgroup. Fifteen proteins were downloaded for each species (if present in the species), the names of which are in Table 2.

Species	GenBank accession number
Ascomycota	
Aspergillus nidulans	CAA33481, AAA99207, AAA31737, CAA25707, AAA31736, CAA23994, X15442, P15956, CAA23995, CAA33116, X00790, X15441, X06960, J01387, X01507
Candida albicans	AF285261
Candida glabrata	CGL511533
Hypocrea jecorina	AF447590
Penicillium marneffei	NC_005256
Pichia canadensis	NC_001762
Podospora anserina	X55026
Saccharomyces cerevisiae	AJ_011856
Schizosaccharomyces japonicus	NC_004332
Schizosaccharomyces octosporus	AF275271
Schizosaccharomyces pombe	X54421
Torrubiella confragosa	AF487277
Yarrowia lipolytica	AJ307410
Basidiomycota	
Cryptococcus neoformans	NC_004336
Schizophyllum commune	AF402141
Cantharellus cibarius[a]	
Choanoflagellida	
Monosiga brevicollis	AF538053
Chytridiomycota	
Allomyces macrogynus	U41288
Harpochytrium94	NC_004760
Harpochytrium105	NC_004623
Hyaloraphidium curvatum	AF402142
Monoblepharella	AY182007
Rhizophydium136	NC_003053
Spizellomyces punctatus	AF402142
Metazoa	
Homo sapiens	NC_001807
Metridium senile	AF000023
Zygomycota	
Smittium culisetae	AY8632133
Mortierella verticillata	AY863211
Rhizopus oryzae	AY863212

[a] Downloaded from http://megasun.bch.umontreal.ca/People/lang/FMGP/proteins.html.

Alignments were performed using the default settings of ClustalW (Thompson et al., 1994). Highly variable sites or those with many gaps were eliminated using Gblocks (Castresana, 2000) with the following settings: number of sequences for a flank position equal to half the number of species plus one; number of contiguous nonconserved positions equal to 10; minimum length of a block four; half the species allowed gaps. All other parameters were set to default.

The key questions addressed using real protein data are:

- *Comparison of DistR estimates to ML estimates.*—How do DistR rate estimates compare to those obtained using the ML based method COMBINE (Pupko et al., 2002b)?
- *Comparison of DistR estimates to Bayesian estimates.*—How do DistR rate estimates compare to those obtained by MrBayes (Huelsenbeck and Ronquist, 2001) under a Bayesian approach?
- *Patristic versus pairwise ML distances.*—How do rate estimates from pairwise ML distances and rate estimates from patristic ML distances compare when applied to real data?
- *Inclusion of DistR estimates into the phylogenetic tree search of PHYML.*—What is the affect of including DistR estimates in an ML tree search? Is there a significantly improved fit? Are improved phylogenetic estimates obtained?

Comparison of DistR estimates to ML estimates.—Note that when comparing DistR rates to those computed using COMBINE (Pupko et al., 2002b), the number of taxa and proteins had to be restricted, because COMBINE can currently only handle data sets for which all taxa are present in all proteins. Two different starting trees were included in the analysis: the ML tree from PHYML based upon the concatenated data set and the ML tree from PHYML when protein rates were incorporated. Rates were estimated under three different models: global amino acid frequencies with one gamma distribution; local amino acid frequencies (for each protein partition) with one gamma distribution; local amino acid frequencies with one gamma distribution for each partition.

Comparison of DistR estimates to Bayesian estimates.—Bayesian estimation of the posterior distribution of the protein rates was performed using MrBayes version 3.0 (Huelsenbeck and Ronquist, 2001). Default priors were used with the JTT model of evolution plus one gamma distribution (eight categories), one parameter for the proportion of invariant sites, and one set of branch lengths for the entire data set. This is the same model that is used for the PHYML + protein rates analysis of the data. Two runs of four chains with 300,000 iterations were performed; the burn-in used was 30,000. A further analysis of the data was performed without protein rates (using the same model) in order to compare to the concatenated PHYML analysis. Four chains were run for 150,000 iterations, with a burn-in of 15,000. Convergence of the chains was determined empirically.

TABLE 2. DistR estimates for empirical data based on pairwise and patristic ML distance estimates. Mean rate estimates and variances for rate estimates are based upon bootstrap replicates over the fungal data set. Rates are normalized so that the average rate is one. Acc. no. = accession number for the alignment in EMBL. AL = alignment length. Patristic refers to rates estimated based on distances from maximum likelihood trees. Pairwise refers to rates estimated based on maximum likelihood distances.

				Patristic		Pairwise	
Protein	Acc. no.	No. of species	AL	Mean	Variance $\times 10^{-3}$	Mean	Variance $\times 10^{-3}$
Atp8	ALIGN_000885	28	32	1.08	8.68	1.15	11.8
Atp9	ALIGN_000886	26	73	0.55	5.12	0.55	4.35
Rps3	ALIGN_000900	11	77	2.02	41.1	2.33	31.5
Nad3	ALIGN_000893	24	79	1.13	8.82	1.15	10.1
Nad4	ALIGN_000894	24	424	1.14	3.52	1.10	2.76
Nad4L	ALIGN_000895	23	85	0.87	5.91	0.91	6.45
Nad6	ALIGN_000897	24	96	1.05	7.21	1.10	7.80
Atp6	ALIGN_000884	29	203	1.07	3.76	1.03	4.07
Cox2	ALIGN_000889	29	220	0.75	3.81	0.71	2.98
Cox3	ALIGN_000890	29	245	1.05	4.75	0.86	3.24
Nad1	ALIGN_000891	24	294	0.89	2.61	0.84	2.30
Nad2	ALIGN_000892	23	313	1.21	2.16	1.29	2.69
Cob	ALIGN_000887	29	375	0.67	1.17	0.61	1.04
Cox1	ALIGN_000888	29	487	0.53	1.76	0.46	.749
Nad5	ALIGN_000896	24	520	1.01	2.79	0.89	1.94

Inclusion of DistR estimates into the phylogenetic tree search of PHYML.—DistR rates were incorporated into the ML framework of PHYML following the proportional approach (Pupko et al., 2002b; Yang, 1996); however, optimization over the rates was not performed. ML trees over the entire data set were calculated in four different ways using this modified version of PHYML. In the first analysis, the proteins were simply concatenated (equivalent to a rate of one for each protein). In the second analysis, the estimated protein rates from the real data set (based on patristic ML distances) were used for each bootstrap replicate when computing the likelihood. In the third and fourth analyses, protein rates were estimated for each bootstrap replicate using patristic and pairwise ML distances respectively. These rates were incorporated into the likelihood computation for each bootstrap replicate. Consensus trees were computed using the CONSENSE program available in the PHYLIP package (Felsenstein, 2004b).

RESULTS AND DISCUSSION

Simulated Data

Patristic versus pairwise ML distances.—The first simulation study demonstrates two important results: pairwise ML distances provide equally good distance estimates as patristic ML distances to the DistR method (Fig. 3); if the fit of the initial pairwise/patristic ML distances to the data is accurate then the DistR estimates will be accurate (Figs. 3 and 4). The first result is important since pairwise ML distances are very fast to compute. The second result indicates that error in the rate estimates stems principally from error in the distance estimates, rather than the DistR method itself.

The numerical results from the first experiment are summarized in Figure 3. The proteins are sorted in order of increasing rate, and the histogram indicates the mean squared error (MSE) over the 10 different replicates (see Appendix 1 for the exact formula used to compute MSE). Mean rate estimates are labelled to the right of each MSE bar, with the rate at which the data was simulated on the left. Results are presented only for alignments of length 100 and 1000. The results for alignments of length 300 and 500 fall in-between these two extremes. Note that the MSE increases in proportion to the rate, so results are presented on two scales.

The mean estimates for the different methods were quite close to the real rates at which the data were simulated, regardless of the alignment length, procedure used to estimate the distances, or rate at which the data was simulated (Fig. 3). However, it is clear from the mean squared error that the DistR estimates based on shorter alignments have larger error (or greater variation), despite the fact that the mean rate estimate is often almost as accurate as that for longer alignments. Furthermore, the mean squared error tends to increase with higher rates. This is likely because the error is often in the third significant digit; for slower rates this will lead to a smaller MSE. Overall there is negligible difference between the mean and MSE statistics for a given alignment length (comparing DistR estimates based on patristic versus pairwise ML distances).

Results also indicate that errors in the rate estimates are due to errors in the original distances rather than approximations introduced in the DistR method. For each protein and alignment length the absolute error between the mean rate estimates and the real rate at which the alignments were simulated was compared to the goodness-of-fit between the estimated and true distances (Fig. 4). This fit can be measured since the data are simulated under a known model at a particular rate. Alignments of length 100 and 300 only were examined, since the errors become negligible for longer alignments. The fit was measured using the goodness-of-fit statistic of Tanaka et al. (Tanaka and Huba, 1985), which is determined from the sum of squares error between true and estimated distances, normalized by the sum of the true distances squared. The exact formula for goodness-of-fit is

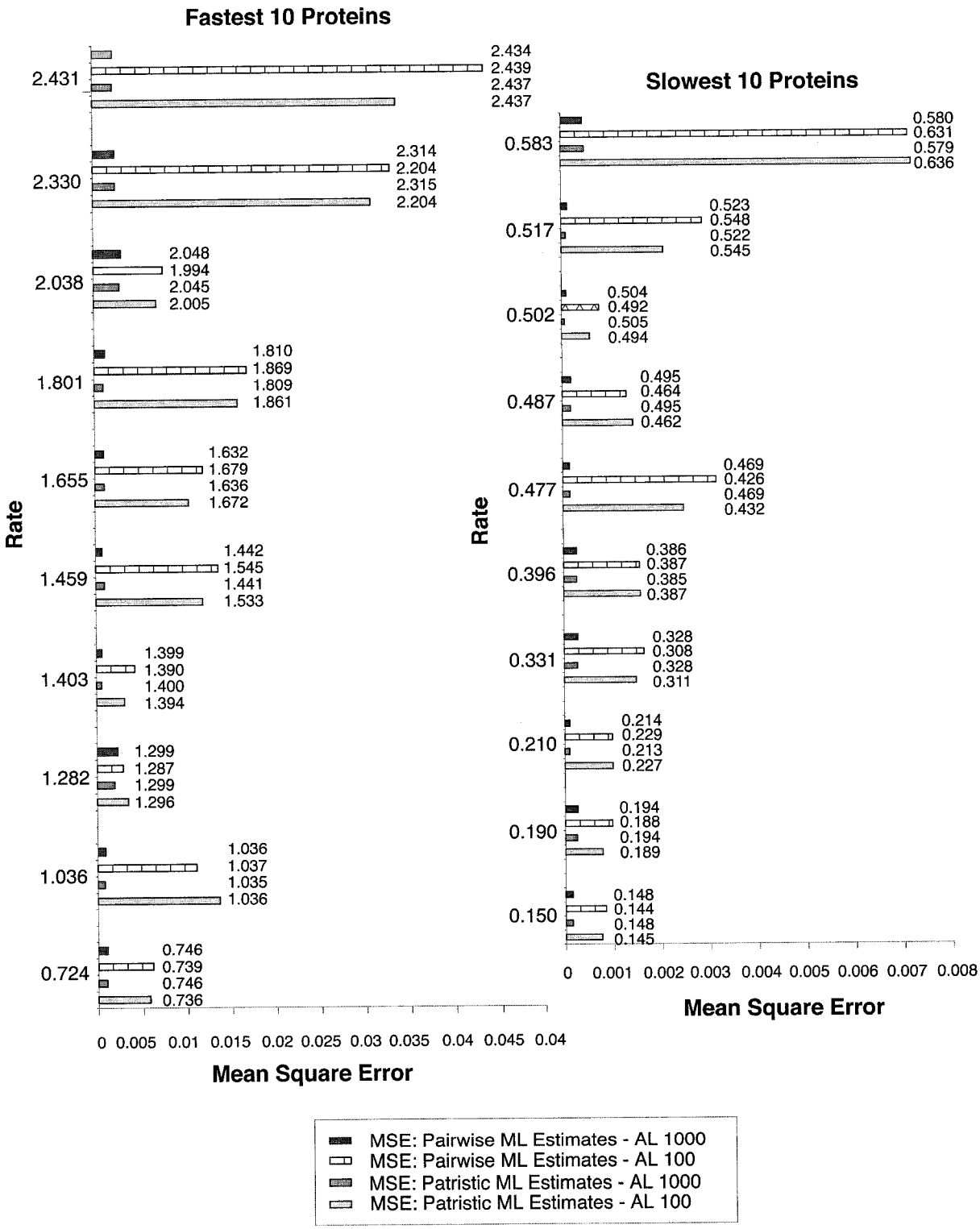


FIGURE 3. Mean squared error for different methods of distance estimation and different alignment lengths. The rates at which the data were simulated are labeled on the left-hand side of the graph. The mean rate estimate for a given distance estimation method, alignment length, and rate is given on the right of the MSE bar. AL = alignment length. The 10 fastest proteins are in the left-hand column. The number of species in each protein (from fastest to slowest) are Protein 1: 53 species; Protein 2: 38 species; Protein 3: 33 species; Protein 4: 53 species; Protein 5: 38 species; Protein 6: 48 species; Protein 7: 53 species; Protein 8: 48 species; Protein 9: 43 species; Protein 10: 33 species. The 10 slowest proteins are in the right-hand column. The number of species in each protein (from fastest to slowest) are Protein 1: 33 species; Protein 2: 48 species; Protein 3: 43 species; Protein 4: 43 species; Protein 5: 48 species; Protein 6: 33 species; Protein 7: 43 species; Protein 8: 53 species; Protein 9: 38 species; Protein 10: 38 species. All rates are normalized so that the average rate is one over all 20 proteins. The total number of taxa in the data set is 58.

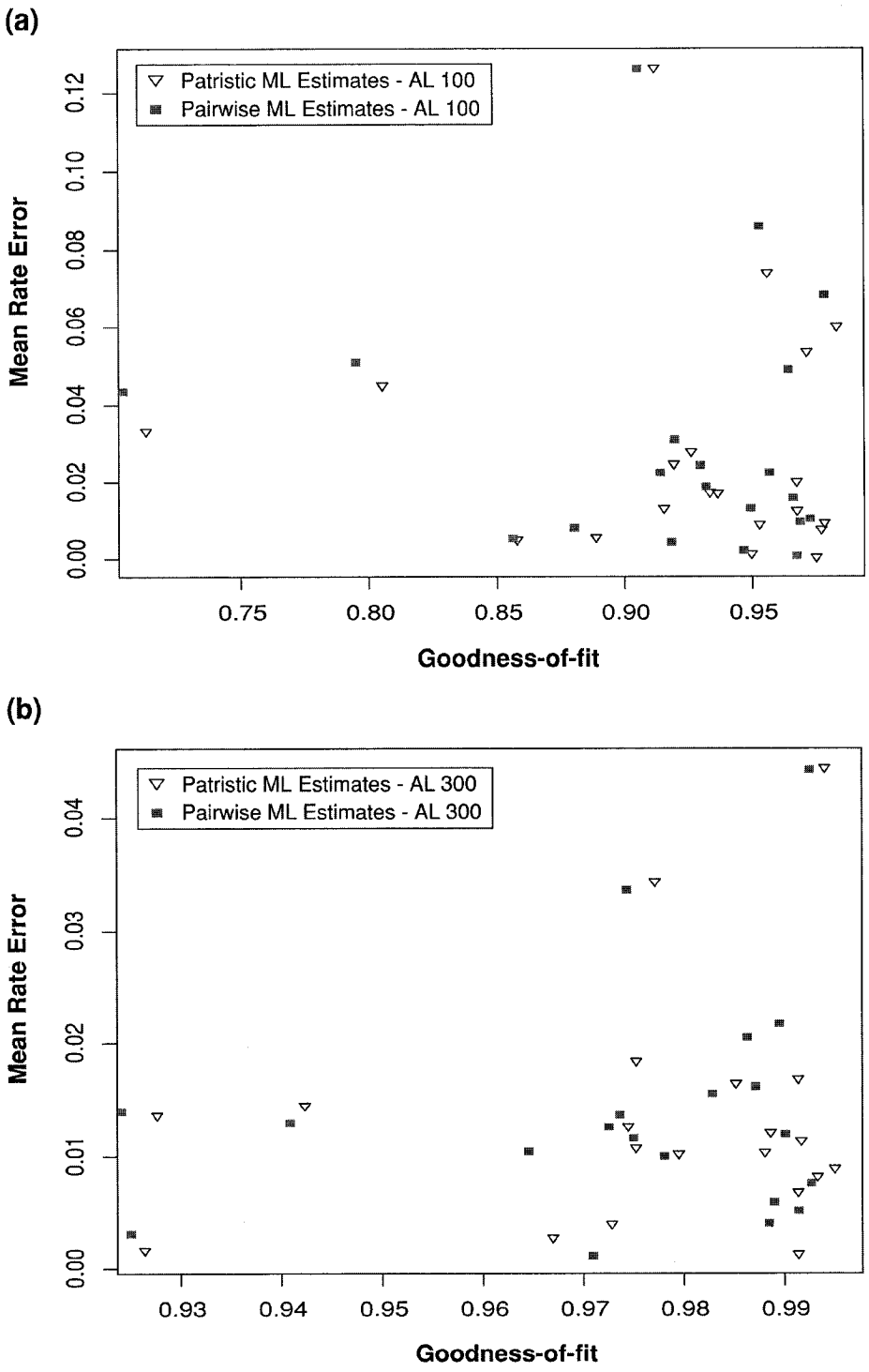


FIGURE 4. Average error of DistR rate estimates compared to goodness-of-fit of distances based upon patristic and pairwise ML distance estimates. (a) DistR rate estimates were based upon simulated proteins of length 100. (b) DistR rate estimates were based upon simulated proteins of length 300. A higher value for goodness-of-fit means that the fit of the estimated distances to the original distances is better.

presented in Appendix 1. The statistic has a maximum of one, which indicates a perfect fit.

It is expected that with longer alignments the goodness-of-fit will increase, indicating that the fit of the model to the data is better. This is clearly the case as seen when comparing goodness-of-fit for alignments of length 100 (Fig. 4a) to that for alignments of length 300 (Fig. 4b). The fit is further improved, and relative error reduced, with alignments of length 500 and longer (data not shown). The decrease in the goodness-of-fit (indicating a worse fit) seen with short alignment lengths indicates that the error of the method is dependent upon

the error of the distance estimates and is not a property of the estimation procedure itself.

Interestingly, the error in rate estimation is in some cases less when based upon pairwise ML distances, rather than patristic ML distances. Given that the multiple sequence alignments are short (100 and 300 amino acid residues) and include many species (at least 33 in each protein alignment), there are many trees that will fit the data equally well. Thus, there is high variation in building a ML tree to fit the original tree on which the data were simulated. Hence, estimating a ML tree with few data will likely lead to an incorrect topology. This will result in a worse fit between the original tree and the tree estimated from the alignment data. This is not true for pairwise ML distances, which do not account for topology.

Missing distances between taxa.—In the previous experiment, less than half of the taxa were missing in each protein, and 20 proteins were used to estimate rates. The effects of more extreme missing taxa were also tested, where no distance estimates were present between some pairs of taxa. To achieve this, up to 75% of the taxa were removed from the starting tree. Additionally, many fewer proteins were used for DistR estimation. Results indicate that the DistR method is robust to missing taxa, though having many missing taxa led to the expected increase in variance of the rate estimates.

Figure 5 summarizes the error in rate estimates for two simulated data sets. In the first example (Fig. 5a) there are four protein trees, each with 16 taxa ($\approx 28\%$ of the total taxon set). In the second example (Fig. 5b) there are eight protein trees. Seven of these have 16 taxa and the other has 30 taxa. The proteins are ordered from fastest to slowest rate in both Figure 5a and Figure 5b. Mean rate estimates are shown on the right of the MSE, and the rate at which the protein simulated (averaged to equal one) is given on the left. Simulated proteins in Figure 5a are labeled from I to IV. The same simulated proteins in Figure 5b are likewise labeled.

Once again it is evident that pairwise ML distances and patristic ML distances give almost identical average relative rate estimates (to within two or three decimal places). Furthermore, the missing data has little effect on mean rate estimates, but does have a large effect on the variance. For instance, comparing the MSE for the first protein in Figure 5a to that of the second protein in Figure 5b (it is the same simulated protein), it is clear that although the mean rate estimate is approximately as accurate with more taxa (Fig. 5b), the MSE is clearly smaller when more distances between a pair of taxa are included in the analysis. Thus it is evident that more data in terms of pairwise distances between taxa (over multiple proteins) will reduce the error of the DistR estimate.

Calculation of the relative rates within groups of the same number of species was also performed (i.e., proteins with 16 species, proteins with 30 species, and proteins with 44 species). For each subset of proteins mean rate estimates based on pairwise ML distances were slightly worse or identical to those based on patristic ML distances (data not shown). In addition, the variances were greater in general for rates estimated based on pairwise ML distances. The major difference between the three analysis was that the variance of the rate estimates was lower when more species were included in the analysis. Furthermore, the mean rate estimates were slightly more accurate for the data sets over larger taxon groups (data not shown).

Accuracy in spite of missing taxa demonstrates that the rate estimation procedure is consistent (assuming that the initial distance estimates are accurate), regardless of the number of proteins under analysis. This is because rates are not computed relative to the distance estimates of one protein. Rather, they are constrained by all the distance estimates. Thus, if one set of distance estimates is extremely biased with respect to the remainder of the distances they will not have a strong effect on the final rate estimates.

Empirical Data

Comparison of DistR estimates to ML estimates.—Rates were calculated in a ML framework using only those proteins that are present over the entire species set (Atp6, Cob, Cox1, Cox2, and Cox3) due to a constraint of the program COMBINE (Pupko et al., 2002b). Table 3 shows the time for rate estimation and rate estimates based on different models under the ML framework in comparison to DistR estimates based on pairwise and patristic ML distances. Two sets of ML estimates are given for each model. The first based upon the concatenated tree, and the second on the DistR incorporated ML tree. DistR estimates are computed far more rapidly and are still accurate in comparison to ML estimates. In comparison to the six ML estimates, the DistR rates based on patristic ML distances are slight overestimates for Cob and Cox1, and slight underestimates for Cox2 and Cox3. The estimate for Atp6 is an average of the 6 ML estimates (Table 3). Notably, the patristic DistR estimates for Cob and Cox1 are closest to the ML estimates based on

TABLE 3. Comparison of ML rate estimates to DistR estimates. Comparison of relative rate estimates and estimation time from COMBINE and DistR for five proteins (Atp6, Cob, Cox1, Cox2, and Cox3) from the fungal data set. For each model, rates based upon the maximum likelihood concatenated tree from PHYML are given on the first line, and rates based upon the maximum likelihood tree incorporating DistR rates (computed in PHYML) are given on the second. All estimates were normalized so that the average rate is one. GF = global amino acid frequencies; LF = local amino acid frequencies (calculated for each protein); 1-GAM = one gamma distribution estimated for the entire data set; 5-GAM = one gamma distribution for each protein; DistR Pat = DistR estimation using patristic ML distances; DistR Pair = DistR estimation using pairwise ML distances.

Method	Time	Atp6	Cob	Cox1	Cox2	Cox3
GF + 1-GAM	776s	1.24	0.81	0.62	0.99	1.34
		1.25	0.81	0.63	0.99	1.33
LF + 1-GAM	842s	1.35	0.80	0.61	0.94	1.31
		1.36	0.80	0.62	0.93	1.30
LF + 5-GAM	648s	1.36	0.79	0.59	0.94	1.31
		1.39	0.78	0.61	0.92	1.30
DistR Pat	0.116s	1.32	0.83	0.66	0.91	1.29
DistR Pair	0.122s	1.40	0.83	0.64	0.96	1.18

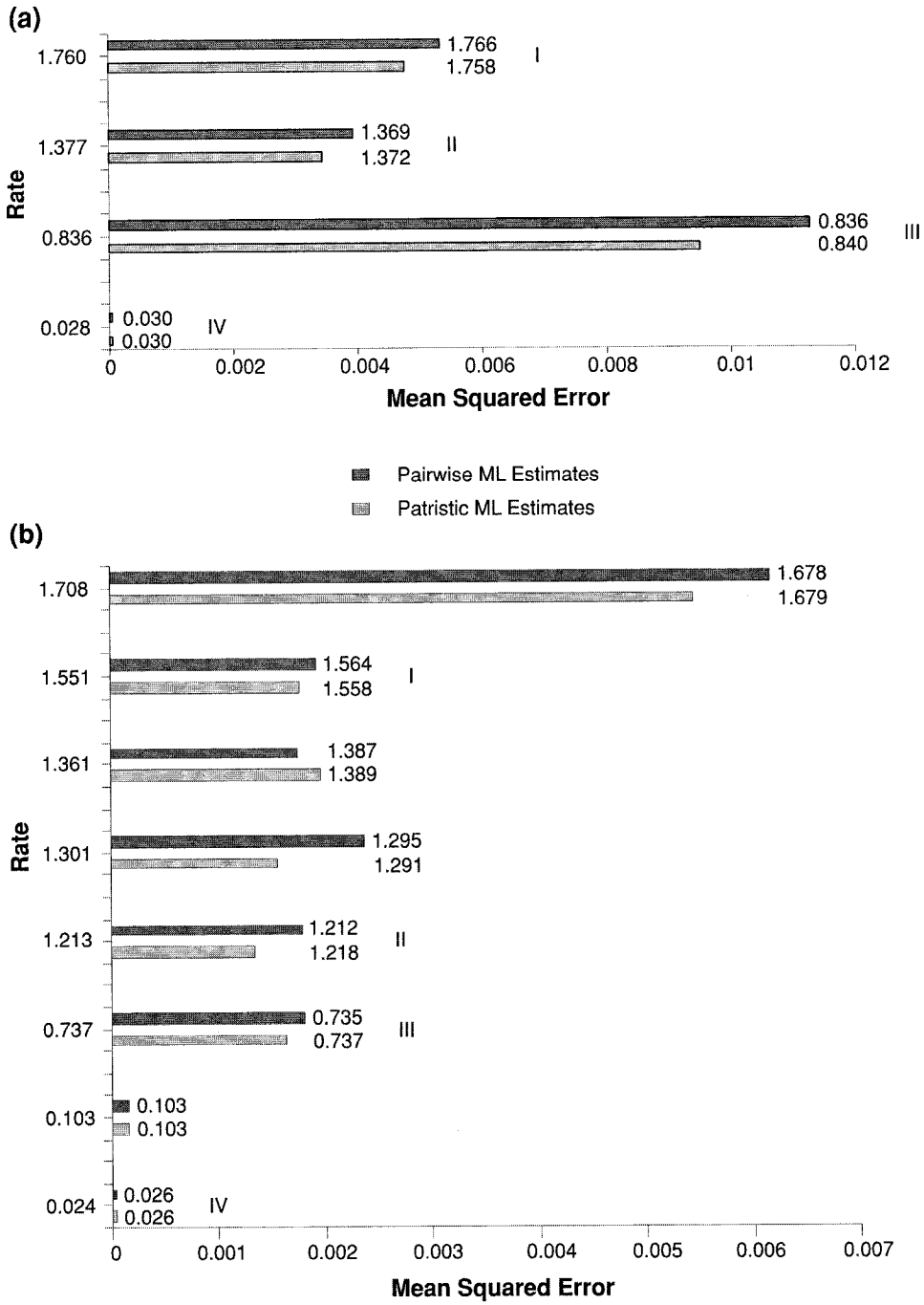


FIGURE 5. Mean squared error for different methods and different amounts of distance data. The rates at which the data were simulated are labelled on the left-hand side of the graph in both (a) and (b). Mean rate estimates for both distance estimation methods are labelled on the right of the MSE bars for each protein. All rates are normalized so that the average rate is one in both (a) and (b) and are sorted from fastest to slowest. Proteins that are the same in both (a) and (b) are labelled. (a) Rate estimates based upon a data set consisting of four proteins with 16 taxa each. (b) Rate estimates based upon a data set consisting of eight proteins; seven with 16 taxa and one with 30 taxa.

the rate-incorporated tree using global amino acid frequencies plus the one-gamma-distribution model. Conversely, the DistR estimates for Cox2 and Cox3 are closest to the ML estimates based on the same tree, using local amino acid frequencies and the five-gamma-distribution model. The DistR estimates based on pairwise ML distances are quite close to those based on patristic ML distances, except for Atp6 and Cox3. Atp6 has a much higher rate—quite close to the ML estimate for the LF + 5-GAM model where the estimates were based on the rate-incorporated ML tree. However, the Cox3 estimate is quite low compared to all ML estimates; Cox3 had a higher variation in rate estimation over all proteins (Table 3), a case where perhaps the lack of topological

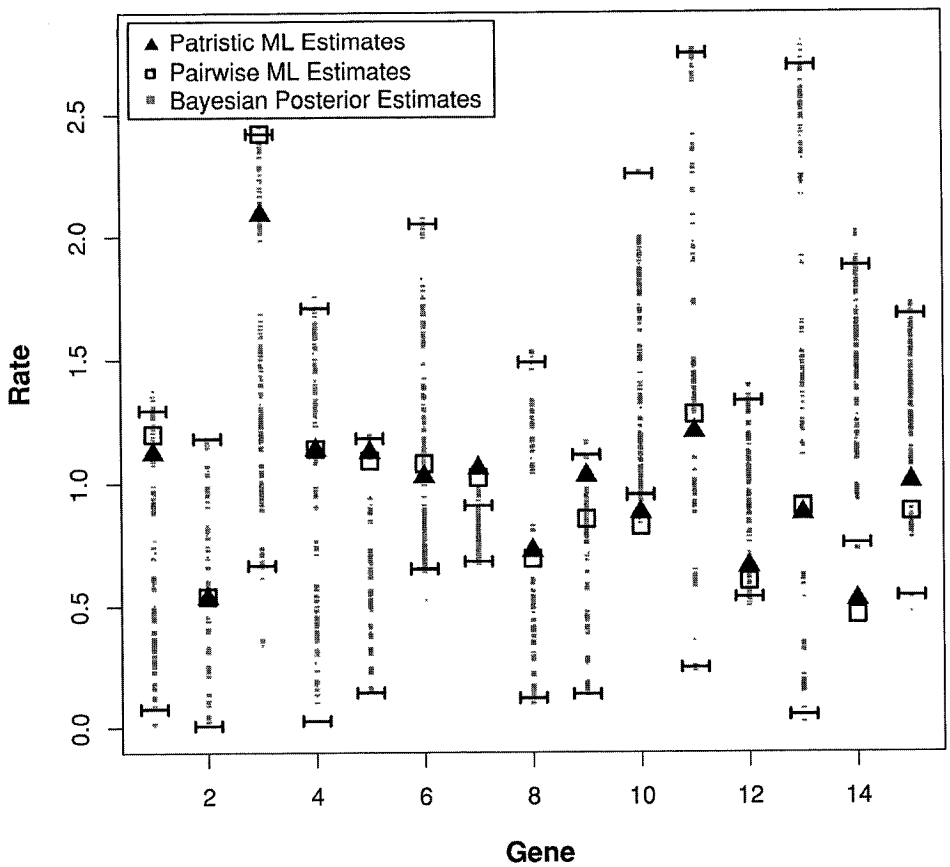


FIGURE 6. Distribution of rates from the MrBayes proportional model analysis compared to DistR estimates. Bars at either end represent the 95% credible interval. The DistR estimate based upon patristic ML distances is marked by a solid triangle. The DistR estimate based upon pairwise ML distances is marked by a square. The posterior rate estimates of MrBayes are given by a solid square. DistR estimates are normalized so that the average rate is one (as in MrBayes). Proteins are ordered from shortest to longest as follows: Atp8, Atp9, Rps3, Nad3, Nad4, Nad4L, Nad6, Atp6, Cox2, Cox3, Nad1, Nad2, Cob, Nad4, Cox1, and Nad5.

information decreases the accuracy of the DistR estimate. Clearly this is not an issue for most proteins, but can be an issue for some. Overall it appears that the DistR estimates are model independent regardless of distance estimation procedure and provide excellent first approximations to the ML estimates.

Comparison of DistR estimates to Bayesian estimates.— The posterior distribution of rates from MrBayes is shown in Figure 6. For all but three of the proteins the DistR estimates fall within the 95% posterior credible interval for the protein rate. Each of Nad6, Cox1, and Cox3 have DistR estimates that do not fall between the 95% posterior credible interval. Both Cox1 and Cox3 have average sequence lengths, and 29 taxa each. Nad6 is shorter at less than 100 amino acids, with only 24 species. In the case of Nad6 perhaps the short sequences length contributes to uncertainty in the DistR estimates. However, it is unlikely that the Bayesian posterior distributions of the rates are accurate. This conclusion is based upon the fact that the four chains were mixing quite poorly in both runs even after 300,000 iterations (data not shown). Sampling from the posterior distribution is unlikely to be correct since the chain might be oversampling from areas of low likelihood. Comparison of the tree of the highest likelihood from this analysis to the tree of highest likelihood based on the concatenated data indicates that MrBayes was in a suboptimal topological space when sampling rate estimates (using the Bayesian information criterion, data not shown). Furthermore, the DistR ML tree is a significantly better fit of the model to the data based on the AIC (Felsenstein, 2004a) when compared to the likelihood of the MrBayes rate incorporated tree as computed in PHYML. Thus, although the posterior distribution of the rates appears reasonable, the chain seems to be having difficulty sampling through topology space.

Thus, it appears that the proportional model under MrBayes, when used without different parameters for each partition (as in Nylander et al., 2004), does not search tree space as well as PHYML with the rate multipliers included. Perhaps this is due to an incorrect prior on the rate parameters used. If this is the problem the DistR method can certainly be used to find a distribution of the rates of proteins, which could be used as the prior on these parameters. The discrepancy could also be due to the different search heuristics used in MrBayes. Given the computational complexity of the search, it might be

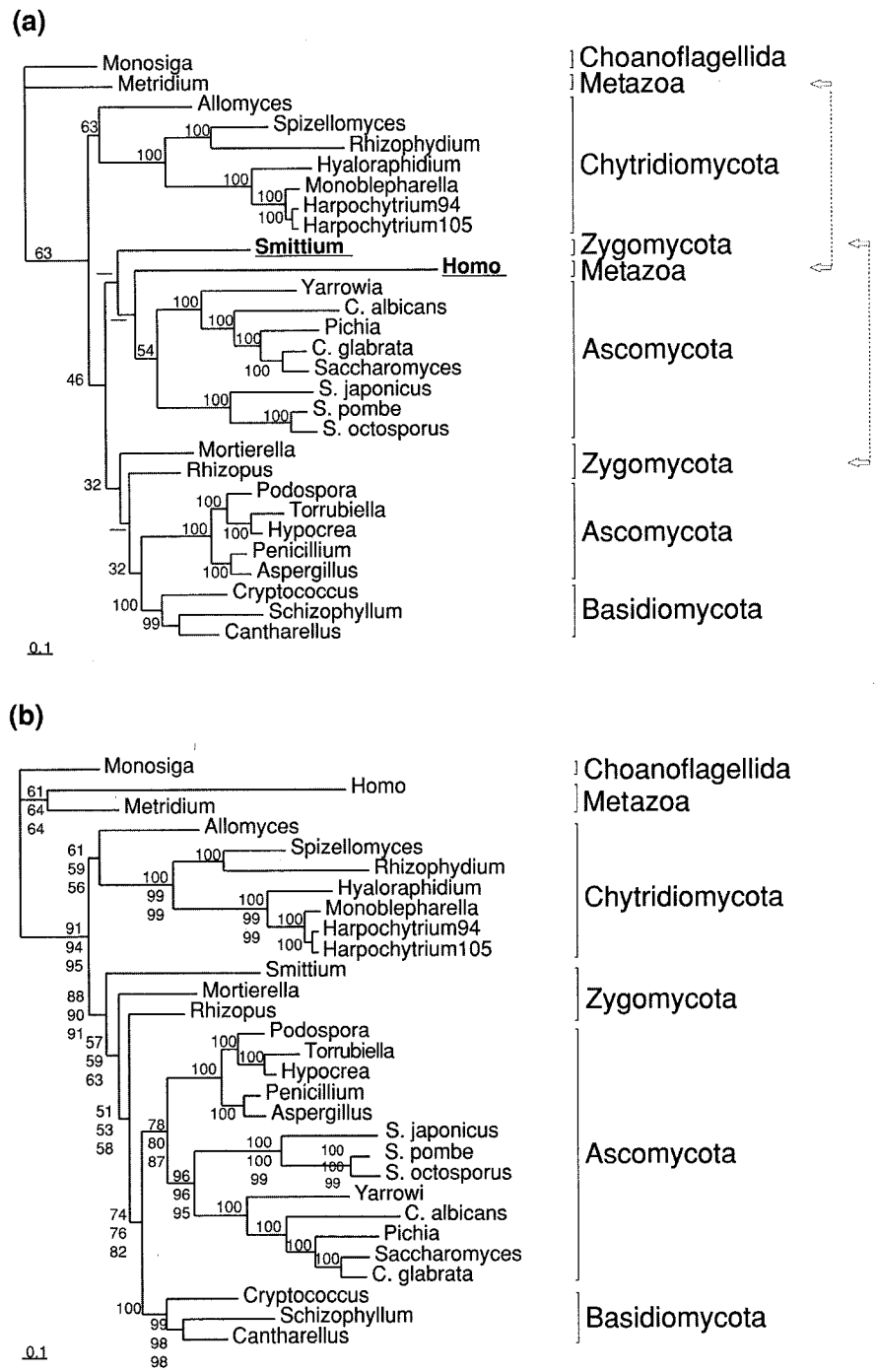


FIGURE 7. (a) Phylogenetic analysis based upon the mitochondrial data set. The topology shown was inferred using PHYML without DistR protein rates, using the JTT model of protein evolution, with eight gamma categories, and ML estimation of the alpha parameter of the gamma distribution and the proportion of invariant sites. It was constructed using the concatenated "unambiguously" aligned proteins. Bootstrap support for this topology was computed based upon 100 replicates. The percentage of support for each clade is given at the root of the clade. In cases where the consensus tree differed from the maximum likelihood topology a "–" is written. (b) Phylogenetic analysis based upon mitochondrial data set. The topology shown was inferred using PHYML with DistR protein rates, using the JTT model of protein evolution, with eight gamma categories, and ML estimation of the alpha parameter of the gamma distribution and the proportion of invariant sites. It was constructed using the concatenated unambiguously aligned proteins and protein rate estimates. The percentage of support for each clade is given. Bootstrap support for this topology was computed based upon 100 replicates, using three different methods. The top numbers give the percentage of support based upon using the patristic ML distance DistR estimates from the real data as rate values in computing the ML tree for each bootstrap replicate. The middle numbers give the percentage of support based upon reestimating DistR estimates for each bootstrap replicate using patristic ML distances. The bottom numbers give the percentage of support based upon reestimating DistR estimates for each bootstrap replicate using pairwise ML distances. When bootstrap support was the same for each method of incorporating rates it is given only once.

difficult for the program to search for the best rate parameters while also searching for the best topology.

Patristic versus pairwise ML distances.—The relative protein rates of the real data are unknown. However the variance of the rate estimates using both patristic and pairwise ML distances can be compared, a smaller estimate being preferable. Contrary to expectations, but confirming the simulation studies, rate estimates from pairwise ML distances had smaller variance than rate estimates from patristic ML distances.

Variances of the rate values computed were estimated by nonparametric bootstrap of the protein alignments, and reestimation of the distances and DistR rates for each bootstrap data set. The mean and variance of the DistR estimates for pairwise and patristic ML distances show some interesting trends (Table 2). In general, the average rate estimates were similar, with the notable exception of Atp8, Cox3, and Rps3 (and to a lesser extent Nad2, Nad5, and Nad6). Ten of the 15 protein rates derived from patristic ML distances had greater variance than their counterparts derived from pairwise ML distances. (Table 2). These results support the conclusion that introducing topology into the distance estimation procedure is not likely to lead to better distances estimates for the DistR procedure when so many taxa are involved and the alignments are short. This is a consequence of the large number of distinct trees that can fit a short alignment equally well.

Inclusion of DistR estimates into phylogenetic tree search of PHYML.—The experimental results when DistR estimates are incorporated into the ML tree search demonstrate the importance of accounting for different evolutionary pressures in phylogenetic inference.

Bootstrap support values for the ML tree using concatenated data are presented in Figure 7a. The bootstrap support for some of the clades was quite weak. Incorporating DistR estimates based upon both patristic and pairwise ML distances into the tree search led to the same ML tree, presented in 7b. Overall, bootstrap support was improved in most clades when DistR estimates were incorporated into the tree search.

The topology of the ML concatenation-based tree does not separate Zygomycota and Ascomycota as distinct clades, which is not surprising because the Zygomycota are traditionally difficult to place. Furthermore, the outgroup is incorrect since it should also contain *Homo sapiens* (which groups incorrectly with the zygomycete *Smittium* and the Ascomycota). This long-branch-attraction problem is due to the highly derived *Smittium* and *Homo* sequences. Using DistR estimates improves the bootstrap support in certain clades, and corrects the most evident topological problems, notably that Zygomycota more accurately group together (although as an unresolved paraphyletic group). Indeed, almost every branch that does not show 100% bootstrap support with the concatenated data have improved support when using protein rates. The only branching where support somewhat lessened from the concatenated to the protein-rate-based trees (and with using individual bootstrap rates) was the branching of *Allomyces* (a species that is difficult to place whatever the method or data set) with the remainder of the Chytridiomycota (Figs. 7a and b). Bootstrap support is strongest when using protein rates based upon pairwise ML distances, where the rate estimates were recomputed for each bootstrap replicate. This is perhaps because the variation in the pairwise ML distance rate estimates was smaller than, or on the same order of magnitude as, the rate estimates based on patristic ML distances.

Both the Kishino-Hasegawa (KH) test and Akaike Information Criterion (AIC) support the ML topology with protein rates as a better fit for the model to the data than the concatenated topology. Under the KH test (Kishino and Hasegawa, 1989, Shimodaira and Hasegawa, 2001); the concatenated topology was significantly worse than the DistR topology ($P < 0.0001$) when the topology was computed with rate estimates calculated based on both patristic and pairwise ML distances. The AIC provides a statistical measurement of the significance of the change in log-likelihood when using two different models to fit the data. The measure compensates for the increase in the number of parameters in the rates model. When DistR estimates based on pairwise ML distances are used, the AIC is 1043.65182 greater than the AIC for a single rate, concatenated analysis. When patristic ML distances are used for rate estimation, the increase in AIC over the concatenated analysis is 1068.7542. Both increases in AIC are very substantial, indicating that important information in the data that is disregarded by traditional concatenated analysis is captured by modeling protein rates.

CONCLUSION

A fast and accurate method to calculate the rates of partitioned data sets is presented. Although the analyses performed here are based upon protein sequence data, using nucleotide sequences should prove as effective. The error in the method is largely due to incorrect initial distance estimates for the proteins, which tend to be worse with smaller or poorly conserved sequences. Using pairwise ML distances for DistR estimation is just as accurate as using patristic ML distances. The estimates are accurate when compared to ML estimates and Bayesian posterior credible intervals for the rates. Incorporating the DistR estimates into PHYML leads to statistically better likelihood and topology.

ACKNOWLEDEGEMENTS

We thank Scott Bunnell, Alain Vandal, Tad Pupko, Tim Collins, and Olivier Gascuel for helpful comments on the manuscript. Thanks to Stéphane Guindon for kindly providing the source code of PHYML v2.2 for our use. Salary and support from the Canadian Institutes of Health Research (MOP 42475; BFL), the Canadian Institute for Advanced Research (CIAR; BFL), National Science and Engineering Research Council (NSERC grant 238975-01; DB), Fonds de recherche sur la nature et les technologies (FQRNT grant 2003-NC-81840; DB), and supply of laboratory equipment and informatics infrastructure by Genome Canada are gratefully acknowledged. RBB is supported by an NSERC PGS-B scholarship.

REFERENCES

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Duruflé, T. Gaasterland, P. Lopez, M. Müller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Nat. Acad. Sci. 99:1414–1419.

Bull, J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Bio. 42:384–397.

Bullerwell, C. E., L. Forget, and B. F. Lang. 2003. Evolution of monoblepharidalean fungi based on complete mitochondrial genome sequences. Nucleic Acids Res. 31:1614–1623.

Bulmer, M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol. Biol. Evol. 8:868–883.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Cranston, K., and B. Rannala. 2005. Closing the gap between rocks and clocks. Heredity 94:461–462.

Eulenstein, O., D. Chen, J. G. Burleigh, D. Fernández-Baca, and M. J. Sanderson. 2004. Performance of flip supertree construction with a heuristic algorithm. Syst. Biol. 53:299–308.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylgenies. J. Mol. Evol. 53:447–455.

Felsenstein, J. 2004a. Inferring phylogenies, pages 148–149. Sinauer Associates, Sunderland, Massachusetts.

Felsenstein, J. 2004b. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle. URL: http://evolution.genetics.washington.edu/phylip.html

Gill, P., W. Murray, and M. Wright. 1982. Practical optimization. Academic Press.

Grassly, N. C., J. Adachi, and A. Rambaut. 1997. PSeq-Gen: An application for the monte carlo simulation of protein sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:559–560.

Guindon, S., and O. Gascuel. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Huelsenbeck, J. P., J. Bull, and C. W. Cunningham. 1996. Combining data in phylogenetic analysis. Tree 11:152–158.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the braching order in hominoidea. J. Mol. Evol. 29:170–179.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger. 2002. The closest unicellular relatives of animals. Curr. Biol. 12:1773–1778.

Lapointe, F., and G. Cucumel. 1997. The average consensus procedure: Combination of weighted trees containing identical or overlapping sets of taxa. Syst. Biol. 46:306–312.

Mayrose, I., D. Graur, N. Ben-Tal, and T. Pupko. 2004. Comparison of site-specific rate-inference methods: Empirical Bayesian methods are superior. Mol. Biol. Evol. 21:1781–1791.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–67.

Olsen, G. J., S. Pracht, and R. Overbeek. 1993. DNArates. URL: http://geta.life.uiuc.edu/gary/programs/DNArates.html.

Pupko, T., R. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. 2002a. Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18:S71–S77.

Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa. 2002b. Combining multiple data sets in a likelihood analysis: Which models are the best? Mol. Biol. Evol. 19:2294–2307.

Shimodaira, H., and M. Hasegawa. 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

Sumida, M., Y. Kanamori, H. Kaneda, Y. Kato, M. Nishioka, M. Hasegawa, and H. Yonekawa. 2001. Complete nucleotide sequence and gene rearrangement of the mitochondrial genome of the japanese pond frog *Rana nigromaculata*. Genes Genet. Systems 76:311–325.

Tanaka, J. S., and G. J. Huba. 1985. A fit index for covariance structure models under arbitrary GLS estimation. Br. J. Math. Statist. Psych. 38:197–201.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acid Res. 22:4673–4680.

Tomita, K., S. Yokobori, T. Oshima, T. Ueda, and K. Watanabe. 2002. The cephalopod *Loligo bleekeri* mitochondrial genome: Multiplied noncoding regions and transposition of tRNA genes. J. Mol. Evol. 54:486–500.

Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

First submitted 24 November 2004; reviews returned 18 March 2005; final acceptance 24 May 2005
Associate Editor: Tim Collins

APPENDIX 1

FORMULA FOR MEAN SQUARED ERROR AND GOODNESS-OF-FIT

Mean squared error is used to describe the accuracy of rate estimates. Because only relative rates can be computed rates are normalized so that the average rate over all proteins is one. Let $\tilde{r}$ denote the true rate (that is, the rate used in simulations), and let $\hat{r_1}, \ldots, \hat{r_{10}}$ be the rates estimated in the 10 replicates of the experiment. The mean squared error (MSE) is defined as

$$\frac{1}{10}\sum_{i=1}^{10}(\tilde{r} - \hat{r_i})^2.$$

Goodness-of-fit is used to measure the fit of the distance estimates to the distances in the tree used for simulation. There is a slight problem with scales since Pseq-Gen treats branch lengths as the expected number of substitutions per 100 sites while PHYML treats branch lengths as the expected number of substitutions per site. Let $\tilde{d}_{xy}^{(k)}$ be the distance between x and y in the tree used to simulate protein k, let r_k denote the rate used when simulating protein k, and let $\hat{d}_{xy}^{(k)}$ be the distance estimated by PHYML.

Given the differences in scale the goodness-of-fit measure used was

$$1.0 - \frac{\sum_{xy}\left(r_k\tilde{d}_{xy}^{(k)} - 100\hat{d}_{xy}^{(k)}\right)^2}{\sum_{xy}\left(r_k\tilde{d}_{xy}^{(k)}\right)^2}.$$

Note that the goodness-of-fit is at most one, and equals one if and only if there is a perfect fit.

APPENDIX 2

FAST ALGORITHM FOR LEAST-SQUARES ESTIMATION

This appendix shows how to quickly determine the vectors $\mathbf{p}$ and $\mathbf{r}$ that minimize the function $q(\mathbf{p}, \mathbf{r})$ in Equation (3)

$$q(\mathbf{p}, \mathbf{r}) = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} \left(p_{xy} - \frac{d_{xy}^{(k)}}{r_k} \right)^2$$

subject to the constraint that $h(\mathbf{p}) = \kappa$, where

$$h(\mathbf{p}) = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} p_{xy}$$

and κ is an arbitrary, positive constant. In the implementation of DistR

$$\kappa = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} d_{xy}^{(k)}$$

which corresponds to the assumption that the unknown consensus distances are roughly centered on the average of the observed distances. This value can be computed in $O(nm^2)$ time for n proteins and m taxa. Any other positive constant will work, as the only effect is to change the scale of the rate estimates.

To simplify the mathematics substitute $s_k = \frac{1}{r_k}$ for each $k = 1, \ldots, n$. Let $\mathbf{s}$ denote the vector $[s_1, \ldots, s_n]^T$. Minimizing $q(\mathbf{p}, \mathbf{r})$ is then equivalent to minimizing

$$f(\mathbf{p}, \mathbf{s}) = \sum_{k=1}^{n} \sum_{x,y} w_{xy}^{(k)} \left(p_{xy} - s_k d_{xy}^{(k)} \right)^2. \tag{5}$$

Recall from calculus that the minimum of a one dimensional function can be found by determining where the first derivative is equal to zero. This condition extends to multidimensional functions with constraints. Refer to Gill et al. (1982) for an excellent introduction to the optimization tools used here.

If $(\mathbf{p}, \mathbf{s})$ together minimize the function f, subject to the condition that $h(\mathbf{p}) = \kappa$, then there exists a real number λ such that

$$\begin{aligned} \frac{\partial f(\mathbf{p}, \mathbf{s})}{\partial p_{xy}} - \lambda \frac{\partial h(\mathbf{p})}{\partial p_{xy}} &= 0 \quad \text{for all taxa } x, y \\ \frac{\partial f(\mathbf{p}, \mathbf{s})}{\partial s_k} &= 0 \quad \text{for all proteins } k \\ h(\mathbf{p}) &= \kappa. \end{aligned} \tag{6}$$

In general, (6) is only a necessary condition for reaching the minimum, and not a sufficient condition. However, in this case the matrix formed from the second derivatives of $f(\mathbf{p}, \mathbf{s})$ is *positive definite*, so that the function f is convex (Gill et al., 1982). It follows that if $(\mathbf{p}, \mathbf{s})$ and λ satisfy (6) then $(\mathbf{p}, \mathbf{s})$ gives the global minimum.

It is possible to derive the partial derivatives of the functions f and h explicitly. To help with notation define the quantities:

$$\alpha_k = \sum_{xy} 2 w_{xy}^{(k)} \left(d_{xy}^{(k)} \right)^2 \quad \text{for all proteins } k;$$

$$\beta_{xy} = 2 \sum_{k=1}^{n} w_{xy}^{(k)} \quad \text{for all taxa } x, y;$$

$$\beta_{xy,k} = -2 w_{xy}^{(k)} d_{xy}^{(k)} \quad \text{for all proteins } k \text{ and taxa } x, y.$$

The partial derivative of f with respect to s_k, for some protein k, is

$$\begin{aligned} \frac{\partial f(\mathbf{p}, \mathbf{s})}{\partial s_k} &= \sum_{xy} -2 w_{xy}^{(k)} \left(p_{xy} - d_{xy}^{(k)} s_k \right) d_{xy}^{(k)} \\ &= \alpha_k s_k + \sum_{xy} \beta_{xy,k} p_{xy}. \end{aligned}$$

The partial derivatives of f and h with respect to p_{xy}, for some taxa x, y, are

$$\begin{aligned} \frac{\partial f(\mathbf{p}, \mathbf{s})}{\partial p_{xy}} &= \sum_{k=1}^{n} 2 w_{xy}^{(k)} \left(p_{xy} - d_{xy}^{(k)} s_k \right) \\ &= \sum_{k=1}^{n} \beta_{xy,k} s_k + \beta_{xy} p_{xy} \\ \frac{\partial h(\mathbf{p})}{\partial p_{xy}} &= \sum_{k=1}^{n} w_{xy}^{(k)} \\ &= \beta_{xy}/2. \end{aligned}$$

Note from the partial derivatives that the conditions in Equation (6) are linear equations involving the entries of $\mathbf{p}$, $\mathbf{s}$, and λ. As such, the next step is to rewrite 6 in terms of matrix algebra. Given that there are n proteins and m taxa define the following: let D be the $n \times n$ matrix with $\alpha_1, \alpha_2, \ldots, \alpha_n$ down the diagonal and zeros off the diagonal; let C be the $\frac{m(m-1)}{2} \times \frac{m(m-1)}{2}$ matrix with $\beta_{12}, \beta_{13}, \ldots, \beta_{(m-1)m}$ down the diagonal and zeros off the diagonal; let B be the $\frac{m(m-1)}{2} \times n$ matrix with rows indexed by unique pairs of taxa, columns indexed by proteins, and the entry corresponding to row xy and column k equal to $\beta_{xy,k}$; let $\mathbf{v}$ be the $\frac{m(m-1)}{2}$ dimensional vector $\mathbf{v} = \frac{1}{2}[\beta_{12}, \beta_{13}, \ldots, \beta_{(m-1)m}]^T$.

The conditions in Equation (6) can now be rewritten as

$$D\mathbf{s} + B^T \mathbf{p} = 0 \tag{7}$$

$$B\mathbf{s} + C\mathbf{p} + \mathbf{v}\lambda = 0 \tag{8}$$

$$\mathbf{v}^T \mathbf{p} = \kappa. \tag{9}$$

Define

$$\mathbf{u} = B^T C^{-1}\mathbf{v}$$
$$\omega = \mathbf{v}^T C^{-1}\mathbf{v}.$$

Solving for $\mathbf{p}$ in (8) gives:

$$\mathbf{p} = C^{-1}(-B\mathbf{s} - \mathbf{v}\lambda) \tag{10}$$

Substituting this into (9) and solving for λ gives:

$$\lambda = \frac{\kappa + \mathbf{v}^T C^{-1} B\mathbf{s}}{-\mathbf{v}^T C^{-1}\mathbf{v}}.$$
$$= \frac{\kappa + \mathbf{u}^T\mathbf{s}}{-\omega}.$$

Replacing λ with the above equation in (10) provides a solution for $\mathbf{p}$ in terms of the above defined matrices, vectors and $\mathbf{s}$ (i.e., there are no longer any unknowns except for $\mathbf{p}$ and $\mathbf{s}$):

$$\mathbf{p} = C^{-1}\left(-B\mathbf{s} + \mathbf{v}\frac{\kappa + \mathbf{u}^T\mathbf{s}}{\omega}\right) \tag{11}$$
$$= C^{-1}\left(\frac{\mathbf{v}\mathbf{u}^T}{\omega} - B\right)\mathbf{s} + \frac{\kappa}{\omega}C^{-1}\mathbf{v}. \tag{12}$$

Finally, substitute (12) into (7) to get

$$0 = D\mathbf{s} + B^T\mathbf{p}$$
$$= \left(D + \frac{\mathbf{u}\mathbf{u}^T}{\omega} - B^T C^{-1} B\right)\mathbf{s} + \frac{\kappa}{\omega}\mathbf{u}.$$

Let

$$M = \left(D + \frac{\mathbf{u}\mathbf{u}^T}{\omega} - B^T C^{-1} B\right).$$

Then, $\mathbf{s}$ is found by solving the equation:

$$M\mathbf{s} = -\frac{\kappa}{\omega}\mathbf{u}. \tag{13}$$

Consensus distances $\mathbf{p}$ are obtained by substituting $\mathbf{s}$ into Equation (12).

The entire computation is summarized in Appendix 3. The running time of the algorithm is $O(nm^2 + n^3)$ which is time optimal. The algorithm uses $O(n^2 + m^2)$ memory in addition to the $O(nm^2)$ required to store the distance estimates $d_{xy}^{(k)}$.

There are two complications that can arise in the above calculations. Firstly, it could be the case that for a particular pair of taxa x, y there is no single protein that contains both x and y. This means that β_{xy} is undefined, so that C is no longer invertible. This problem is easily solved. If there is no protein with both x and y then the line in (6) involving the partial derivative with respect to p_{xy} is satisfied trivially. Therefore, the row and column of C, the row of B, and entry of $\mathbf{v}$ indexed by the pair x, y can be removed. The reduced problem can be solved as before, although no estimate for p_{xy} is obtained. Row removal is handled in the pseudocode for the algorithm given in Appendix 3 by using constraints in the summations.

The second complication is that the optimization problem might have more than one solution, in which case the matrix M in (13) will not be invertible. This indicates that more information is required to estimate the relative rates, as would arise, for example, in a concatenation of two protein alignments over entirely different sets of taxa.

APPENDIX 3
THE DISTR ALGORITHM

Algorithm DISTR(d, w)
Input: Distance estimates $d_{xy}^{(k)}$ for each pair of taxa and each protein k.
Weights $w_{xy}^{(k)}$ for each distance estimate.
Missing distances have weight zero.
Output: Rate estimates r. Consensus distances p.

$\kappa = \sum_{k=1}^{n}\sum_{xy} w_{xy}^{(k)} d_{xy}^{(k)}$
for k from 1 to n do
  $\alpha_k \leftarrow \sum_{xy} 2w_{xy}^{(k)}(d_{xy}^{(k)})^2$
  for all taxa x, y do
    $\alpha_{k,xy} \leftarrow -2w_{xy}^{(k)} d_{xy}^{(k)}$
    $\beta_{xy,k} \leftarrow -2w_{xy}^{(k)} d_{xy}^{(k)}$
for all taxa x, y do
  $\beta_{xy} \leftarrow 2\sum_{k=1}^{n} w_{xy}^{(k)}$
$\omega \leftarrow \frac{1}{4}\sum_{xy}\beta_{xy}$
for k from 1 to n do
  $u_k \leftarrow \sum_{xy}\beta_{xy,k}$
for k from 1 to n do
  $z_k \leftarrow -\frac{\kappa}{\omega}u_k$.
  for l from 1 to n do
    $M_{kl} \leftarrow -\sum_{xy:\beta_{xy}\neq 0}\frac{\beta_{xy,k}\beta_{xy,l}}{\beta_{xy}} + \frac{1}{\omega}u_k u_l$
    if $k = l$ then $M_{kl} \leftarrow M_{kl} + \alpha_k$
if M is nonsingular then output "Insufficient data to estimate rates"
solve $M\mathrm{s} = -\frac{\kappa}{\omega}\mathrm{u}$ to obtain s
for all taxa x, y such that $\beta_{xy} \neq 0$ do
  $p_{xy} \leftarrow \sum_k\left(\frac{u_k}{2\omega} - \frac{\beta_{xy,k}}{\beta_{xy}}\right)s_k + \frac{\kappa}{2\omega}$
for k from 1 to n do
  $r_k \leftarrow \frac{1}{s_k}$
output r and p.

Mitochondrial genome evolution: the origin of mitochondria and of eukaryotes

Rachel B. Bevan and B. Franz Lang

Abstract

Mitochondria, the energy-producing organelles of eukaryotic cells, evolved from an endosymbiotic α-Proteobacterium more than one billion years ago. These organelles contain their own genetic system, a remnant of the endosymbiont's genome that varies considerably in size, genome architecture, and coding capacity throughout eukaryotes. The five to ~100 genes contained in mitochondrial DNA (mtDNA) code for mitochondrial components involved in up to five mitochondrial processes: respiration/oxidative phosphorylation and translation (invariantly), as well as transcription, RNA maturation, and protein import. These mtDNA-encoded proteins have provided an invaluable alternative to nuclear gene sequences as a source for molecular phylogenetics, by both elucidating and confirming relationships among eukaryotes. However, only a small fraction of the mitochondrial proteome is encoded by the mitochondrion. Indeed, nuclear genes code for much of the proteome. It is likely that most of these genes migrated from the mitochondrion to the nucleus over the course of eukaryotic evolution. In some cases, however, it is clear that genes were recruited to the mitochondrion from the nucleus or other undefined sources. New insights into early mitochondrial genome evolution come from both the investigation of primitive (minimally derived) eukaryotes and the comparison of mitochondria to intracellular bacterial symbionts. Defining more precisely both the α-proteobacterial ancestry of the mitochondrial genome and the contribution of the endosymbiotic event to the nuclear genome will be essential for a full understanding of the origin and evolution of the eukaryotic cell as a whole.

1 Introduction

It has become widely accepted that the mitochondrion derives from an endosymbiotic α-Proteobacterium. Furthermore, there is accumulating evidence that this powerhouse of the eukaryotic cell, which carries out ATP production through oxidative phosphorylation, was acquired only once. However, the phylogenetic reconstruction of such a primordial event, both with respect to the precise timing and the nature of the endosymbiotic partners, is a non-trivial problem (Philippe and Forterre 1999). Not surprisingly, numerous mutually exclusive hypotheses on

Topics in Current Genetics, Vol. 8
C. Köhler, M.F. Bauer (Eds.): Mitochondrial Function and Biogenetics
DOI 10.1007/b96830 / Published online: 9 March 2004

the origin of eukaryotes exist that are based on conjecture, rather than on evidence from molecular phylogenetics. These hypotheses differ as to the nature of the partners involved in this endosymbiosis and the driving forces and timing of their fusion (reviewed in Lang et al. 1999). One of the central issues in this on-going debate is whether or not amitochondriate eukaryotes ever existed.

Although mitochondria are almost omnipresent throughout eukaryotic lineages, and are in most species essential for survival, a few eukaryotes lack both functional mitochondria and a mitochondrial genome. However, among these 'amitochondriate' eukaryotes several contain derived mitochondria (e.g. hydrogenosomes) that generate ATP anaerobically (Embley et al. 2003). Still, others contain what appear to be remnant mitochondrial organelles; 'mitosome' or 'crypton' (Mai et al. 1999; Tovar et al. 1999, 2003; Williams et al. 2002) of currently unknown function. Significantly, these organelles express a few nucleus-encoded proteins typically targeted to and functioning in mitochondria (Roger and Silberman 2002). Together, these observations favor the view that the last common ancestor of extant eukaryotes was a mitochondrion-containing organism. This implies that primitively amitochondriate eukaryotes (eukaryotes whose ancestors never had mitochondria) may not exist and by extension may never have existed. Alternatively, strictly anaerobic eukaryotes might have existed only during the early evolutionary periods with anoxic conditions that would have favored their survival.

Some of these issues have been discussed in a number of comprehensive reviews on animal (Boore 1999), fungal (Paquin et al. 1997; Bullerwell et al. 2003c; Hauser 2003; Kennell and Cohen 2003; Leigh et al. 2003), plant (Wolstenholme 1992; Giege and Brennicke 2001), and protist (Gray et al. 1998, 1999; Lang et al. 1999; Burger et al. 2003b) mitochondrial genomes. Yet, due to the numerous new publications and significant amounts of recently available data, an updated general review on mitochondria appears timely. Accordingly, this review comments upon: (i) various new aspects of mitochondrial genome structure and evolution; (ii) gene migration between mitochondrial and nuclear genomes; (iii) how well mitochondrial data support phylogenetic hypotheses; (iv) how recent hypotheses on the origins of eukaryotes have changed our perception about the importance of mitochondria in defining eukaryotic evolution. To conclude, we will provide a short update on the diversity of mitochondrial genomes across eukaryotes.

2 Evolution of mitochondrial genomes and genes: anything is possible

There is astounding diversity throughout extant eukaryotes with respect to mitochondrial genome size, genome architecture, gene content, and gene expression. In fact, many of these differences appear to evolve within relatively short evolutionary periods (for a recent review see Burger et al. 2003b) and go far beyond the spectrum of variation seen in chloroplast or nuclear genomes. The notable and often bewildering deviations in genome architecture and in gene expression, contrast

markedly with the conservatism of biological functions encoded by mtDNA. To provide a better understanding of the molecular and genetic basis for the rapid changes of mitochondrial genomes, the following provides an overview of what is known about mitochondrial genome architecture and evolution.

2.1 The perplexing diversity of mitochondrial genome architecture

The previously held view that mitochondrial genomes are circular molecules has been contradicted by molecular evidence that a number of circular-mapping mtDNAs (and possibly also eubacterial DNAs) consist of linear, multimeric head-to-tail concatamers (Bendich 1993, 1996, 2001; Lecrenier and Foury 2000). The end structures of the molecules include terminal repeats of varying lengths, terminally attached proteins and single-stranded DNA termini closed covalently (reviewed in Nosek and Tomaska 2003a, 2003b). In contrast, linear-mapping, monomeric DNA molecules have been found in numerous unrelated organisms, including ciliates (Suyama et al. 1985; Pritchard et al. 1990; Burger et al. 2000), chlorophycean green algae (*Chlamydomonas* and relatives (Gray and Boer 1988; Vahrenholz et al. 1993; Fan and Lee 2002)), oomycetes (Martin 1995), chlorarachniophytes (Gilson et al. 1995), several cnidarian animals (Bridge et al. 1992), and fungi (Kovac et al. 1984; Fukuhara et al. 1993; Nosek et al. 1995; Forget et al. 2002). The mtDNA of one organism in particular illustrates the perplexing diversity of mitochondrial genomes: the mtDNA of *Amoebidium parasiticum* (Lang et al. 2002; Burger et al. 2003a) is organized into several hundred distinct linear chromosomes (more details on genome structure and content can be found in section 3.4). This is in sharp contrast to most other mitochondrial genomes that are made up of only one type of chromosome.

The transition from a circular to a linearly-mapped mtDNA conformation has been observed at short phylogenetic distances, as seen for instance in yeasts (Nosek and Tomaska 2003a), golden algae (Coleman et al. 1991; Chesnick et al. 2000), and chytridiomycete fungi (Forget et al. 2002). One hypothesis for the origin of linear genome structures is that the insertion of linear plasmids into mtDNA triggers the conversion of genome conformation from circular to linear (for a review of the well characterized fungal plasmids see Kennell and Cohen 2003). In fact, this phenomenon has been observed in maize cytoplasmic male sterility (Schardl et al. 1984).

Another interesting facet of mtDNA is its wide size variation from approximately six to several hundred kbp, and in some notable exceptions even larger (see Table 1, Fig. 1, 2). This is mostly attributable to accumulation and loss of introns, mobile elements, A+T-rich intergenic spacers, and repeat sequences. In particular, the length of non-coding mtDNA may vary extensively, even within a single genus. For example, the size variation of fission yeast mtDNAs (between ~17 and ~80 kbp; see Table 1; (Bullerwell et al. 2003b)) is for the most part due to large intergenic regions that contain multiple repeats. In addition, length variation due to introns (group I and group II) of various size (0.15 kbp to 4 kbp) and number

Organismal Group	Genome Size and Structure [1]	Protein Genes [2]	RNA Genes [3]	Translation Code [4]	Introns [5]
Metazoa (animals)					
Caenorhabditis elegans	13.8 (circular)	12	24	**UGA(W), AUA (M)** [6]	-
Homo sapiens (human)	16.6 (circular)	13	24	**UGA(W), AUA (M)** [6]	-
Metridium senile (sea anemone)	17.4 (circular)	13	4	**UGA(W)**	2
Fungi					
Hyaloraphidium curvatum (chytrid)	**30.0 (linear)**	14	9	standard	1
Pichia canadensis (Hansenula wingei) [13]	27.7	15	28	**UGA(W)**	2
Podospora anserina	94.2	14	27	**UGA(W)**	33
Rhizophydium sp.	68.8	14	9	**UAG(L)**	**37**
Saccharomyces cerevisiae (baker's yeast)	85.8 (circular)	8	27	**UGA(W), AUA(M), CUN(T)**	13
Schizosaccharomyces pombe (fission yeast)	19.4 (circular)	8	28	standard [7]	3
Schizosaccharomyces octosporus	44.2 (circular)	8	27	standard	6
Schizosaccharomyces japonicus var. jap.	≥ 80 (circular)	7	27	standard	2
Spizellomyces punctatus (chytrid)	58.8; 1.4; 1.1 **(3 circular DNAs)**	14	10	**UAG(L)**	12
Plants					
Marchantia polymorpha (liver wort)	186.6 (circular)	38	30	standard	32
Arabidopsis thaliana (thale cress)	**366.9 (circular)**	31	21	standard	23
Photosynthetic protists					
Chara vulgaris (charophyte green alga) [9]	67.7 (circular)			standard	27
Chlamydomonas reinhardtii (green alga)	15.8 (linear)	7	5	standard	-
Chondrus crispus (red alga) [10]	25.9 (circular)	19	27	**UGA(W)**	1
Prototheca wickerhamii (green alga) [8]	55.3 (circular)	31	29	standard	5
Ochromonas danica (golden alga)	41.0 (linear)	30	27	standard	-
Porphyra purpurea (red alga)	63.7 (circular)	22	26	**UGA(W)**	2
Pylaiella littoralis (brown alga) [11]	58.5 (circular)	32	27	standard	7
Non-photosynthetic protists					
Acanthamoeba castellanii (amoeba)	41.6 (circular)	34	18	**UGA(W)**	3
Amoebidium parasiticum (Holozoa)	≥ 200 **(several hundred linear)**	≥17	≥27	**UGA(W)**	≥23
Dictyostelium discoideum (slime mold) [12]	55.6 (circular)	33	20	standard	5
Monosiga brevicollis (choanoflagellate)	76.6 (circular)	26	27	**UGA(W)**	4
Phytophthora infestans (oomycete)	38.0 (circular)	35	27	standard	-
Plasmodium falciparum [14]	**6.0 (circular)**	3	2	**UGA(W)**	-
Reclinomonas americana (jakobid flagellate)	69.0 (circular)	66	31	standard	1

can be substantial. Indeed, 75% of the mitochondrial genome in *Podospora anserina*, an ascomycete fungus, is accounted for by introns (Table 1; Cummings et al. 1990).

Somewhat counter-intuitively, there is little correlation between mtDNA size and gene content (Table 1, Fig. 1, 2). The average coding capacity of the mitochondrial genome across eukaryotes is approximately 40-50 genes, with extremes of only five in *Plasmodium* (Wilson and Williamson 1997) and nearly 100 in jakobid flagellates (Lang et al. 1997). Despite this large difference in gene number, mitochondrial genes are only involved in at most five basic processes: protein import, RNA maturation, transcription and invariantly, respiration/oxidative phosphorylation and translation. Even in the gene-rich jakobid mtDNAs, only a small fraction of genes are implicated in processes other than translation and respiration/oxidative phosphorylation.

Table 1 (overleaf): explanations
[1] Size in kbp, rounded values; 'circular' stands for 'circular mapping', i.e. the major portion of these mtDNAs occur as long linear concatamers, not as monomeric circles. 'linear' stands for 'monomeric linear'. Extreme genome sizes and unusual genome architecture marked in bold.
[2] The number of identified genes (not including ORFs) is indicated. The basic set of protein-coding genes typically found in animals and fungi are, *cob* (apocytochome b), *cox1,2,3* (cytochrome oxidase subunits), *atp6,8,9* (ATPase subunits), and *nad1,2,3,4,4L,5,6* (NADH dehydrogenase subunits). The mtDNA of the coral *Sarcophyton glaucum* contains an additional gene with similarity to bacterial *mutS*, (Pont-Kingdon et al. 1998), the nematode *C. elegans* lacks *atp8* (Okimoto et al. 1992), and protists, fungi and plants usually contain additional hypothetical protein genes (ORFs).
[3] Genes for *rns*, *rnl* (small and large subunit rRNAs) occur in all mtDNAs, whereas genes coding for 5S rRNA (*rrn5*), and RNase P RNA (*rnpB*), might be absent. The number of mtDNA-encoded tRNAs varies. Duplicated genes are counted only once.
[4] Deviations from the standard bacterial translation code are indicated in bold.
[5] Total number of introns, the two mitochondrial intron classes 'I' and 'II' are not distinguished.
[6] Further codon reassignments include the use of AGA and AGG as stop codons, and additional translation initiation codons other than AUG and GUG.
[7] In *S. pombe*, one UGA(Trp) is present in *rps3*, and two in intronic ORFs.
[8] *Prototheca* (Wolff et al. 1994) belongs to green algae, but has secondarily lost its capacity for photosynthesis.
[9] (Turmel et al. 2003).
[10] (Leblanc et al. 1995).
[11] (Oudot-Le Secq et al. 2001).
[12] (Ogawa et al. 2000)
[13] (Sekito et al. 1995)
[14] Referred to as 'linear' in the literature (Feagin et al. 1992), but belongs to the class of 'circular-mapping' multimeric head-to-tail concatemers.

2.2 Unusual mitochondrial gene structure and gene expression

In terms of gene structure, mitochondria exhibit most unusual deviations, including gene fusions, genes-in-pieces, and gene reductions. Examples of the latter case include truncated tRNAs (lacking one or more of the helical arms) that are found in mitochondria of several animal lineages (e.g. Wolstenholme et al. 1987), and the severely reduced and structurally streamlined mitochondrial rRNAs of most animals (Boore 1999) and some protists (Feagin et al. 1997; Gray et al. 1998).

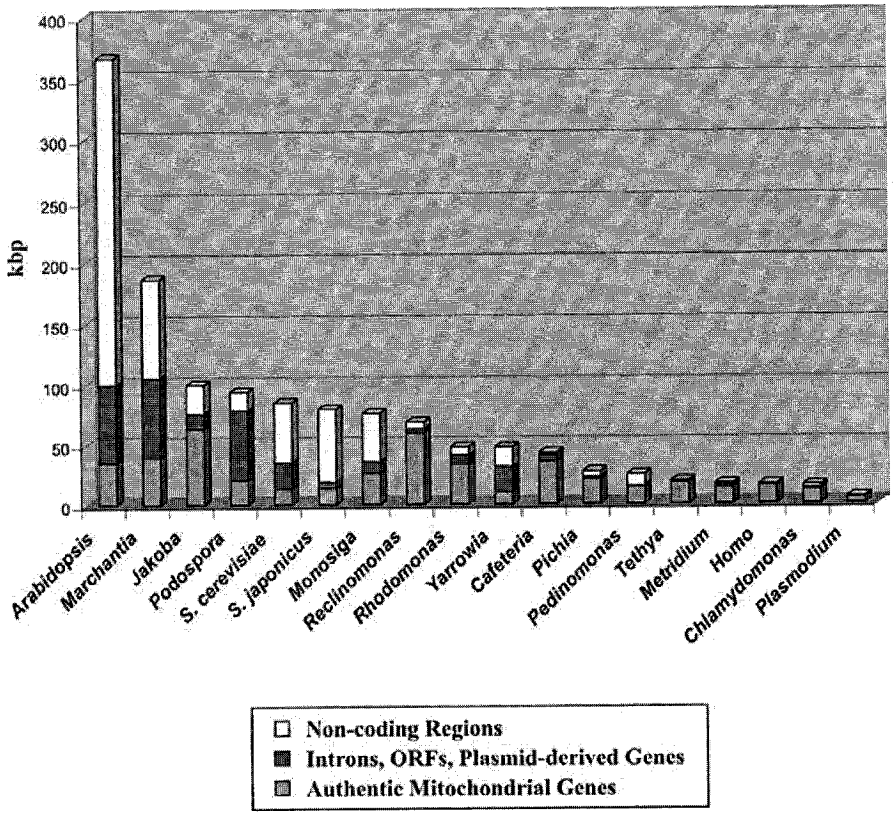


Fig. 1. Mitochondrial genome size and coding content across eukaryotes. Species are ordered by genome size, from left to right. Intergenic regions (yellow); introns, intronic, and other ORFs, phage-like reverse transcriptases and DNA polymerases (maroon); length of coding regions of authentic mitochondrial genes (violet). Species are: *Arabidopsis thaliana* (flowering plant, angiosperm); *Marchantia polymorpha* (liverwort, bryophyte); *Jakoba libera* (jakobid flagellate, (G. Burger and B. Franz Lang, unpublished)); *Saccharomyces cerevisiae*, *Schizosaccharomyces japonicus*; (ascomycete fungi); *Monosiga brevicollis* (choanozoan flagellate); *Reclinomonas americana* (jakobid flagellate); *Rhodomonas salina* (cryptophyte alga); *Yarrowia lipolytica* (ascomycete fungus); *Cafeteria roenbergensis* (stramenopile flagellate); *Pichia canadensis* (ascomycete fungus); *Pedinomonas minor* (green alga, chlorophyte); *Tethya sp.* (demosponge, (D. Lavrov and B. Franz Lang, unpublished)); *Metridium senile* (cnidarian animal); *Homo sapiens* (vertebrate animal); *Chlamydomonas reinhardtii* (green alga, chlorophyte); *Plasmodium falciparum* (apicomplexan protist). If not further specified, data have been retrieved from the Organelle Genome Database, GOBASE (http://megasun.bch.umontreal.ca/gobase/).

The structural constraints of the unusual animal mitochondrial tRNAs have been modeled recently into a three-dimensional L form (Steinberg et al. 1997). As well, protein-coding genes are occasionally shortened at their C-terminus compared to

genes in more primitive organisms (e.g. animal mitochondrial *nad5* and *atp8*, and several genes of the fungus *Harpochytrium* (Bullerwell et al. 2003a).

One of the best known examples of gene fusion is that of the *cox1* and *cox2* genes in *Acanthamoeba castellanii* (Burger et al. 1995) and *Dictyostelium discoideum* (Ogawa 2000 MGG). In both cases, a single open reading frame is formed because the genes are immediately adjacent, without an intervening translation termination codon. In *Acanthamoeba*, a bi-cistronic mRNA is produced, which may or may not be translated into a fusion protein (Lonergan and Gray 1996). In another example, the *cox2* gene of the brown alga *Pylaiella* has a large (3018 nucleotides) N-terminal extension of its open reading frame (Oudot-Le Secq et al. 2001).

Genes-in-pieces are often found on both strands of a genome interspersed with other genes, and may be broken into as many as 20 modules (e.g. Gillespie et al. 1999). Genes-in-pieces were first described for rRNA genes in the ciliate protozoon *Tetrahymena pyriformis* (Schnare et al. 1986; Heinonen et al. 1987) and the green alga *Chlamydomonas reinhardtii* (Boer and Gray 1988). The discrete rRNA transcripts are held together *via* base pairing of complementary sequence stretches (Boer and Gray 1988). Similarly, mature transcripts may be assembled from discrete, group II intron-containing protein-coding genes. In these cases, base-pairing of the intron sequences brings together the exons that are subsequently joined by trans-splicing. The well-studied examples of *nad1, 2, 3* and *nad5* genes in plants and in the prasinophyte green alga *Mesostigma viride* may involve two- as well as three-molecule interactions (e.g. Chapdelaine and Bonen 1991, 1993; Knoop et al. 1997; Malek and Knoop 1998; Morawala-Patell et al. 1998; Giege and Brennicke 2001; Turmel et al. 2002b). In contrast, the intron-less *nad1* genes in *Tetrahymena pyriformis* and *Paramecium aurelia* are probably translated from the two mRNA pieces into separate protein fragments (Edqvist et al. 2000). Finally, one of the most intriguing cases of genes-in-pieces is that of *cox2* in the green alga *Scenedesmus obliquus.* In this particular case (unlike in chlamydomonad algae in which the two pieces are encoded in the nucleus (Perez-Martinez et al. 2001)), the N-terminus is encoded in the mitochondrion (Nedelcu et al. 2000) and the C-terminus is thought to be encoded in the nucleus (Perez-Martinez et al. 2001). A similar situation has been documented in angiosperms. In this group of land plants, cases have been identified where an intact *rpl2* gene is present in either the mitochondrion or the nucleus, a split *rpl2* gene exists with the N-terminus encoded in the mitochondrion and the C-terminus in the nucleus; and finally, a split gene exists with both parts encoded in the nucleus (Adams et al. 2001a).

2.3 Past and current gene loss and gene emigration from the mitochondrial to the nuclear genome

A primary force that has shaped the evolution of mitochondrial genomes is the (usually irreversible) loss of genes from the mitochondrial genome, which may occur through three major processes. One is the removal of the selective pressure

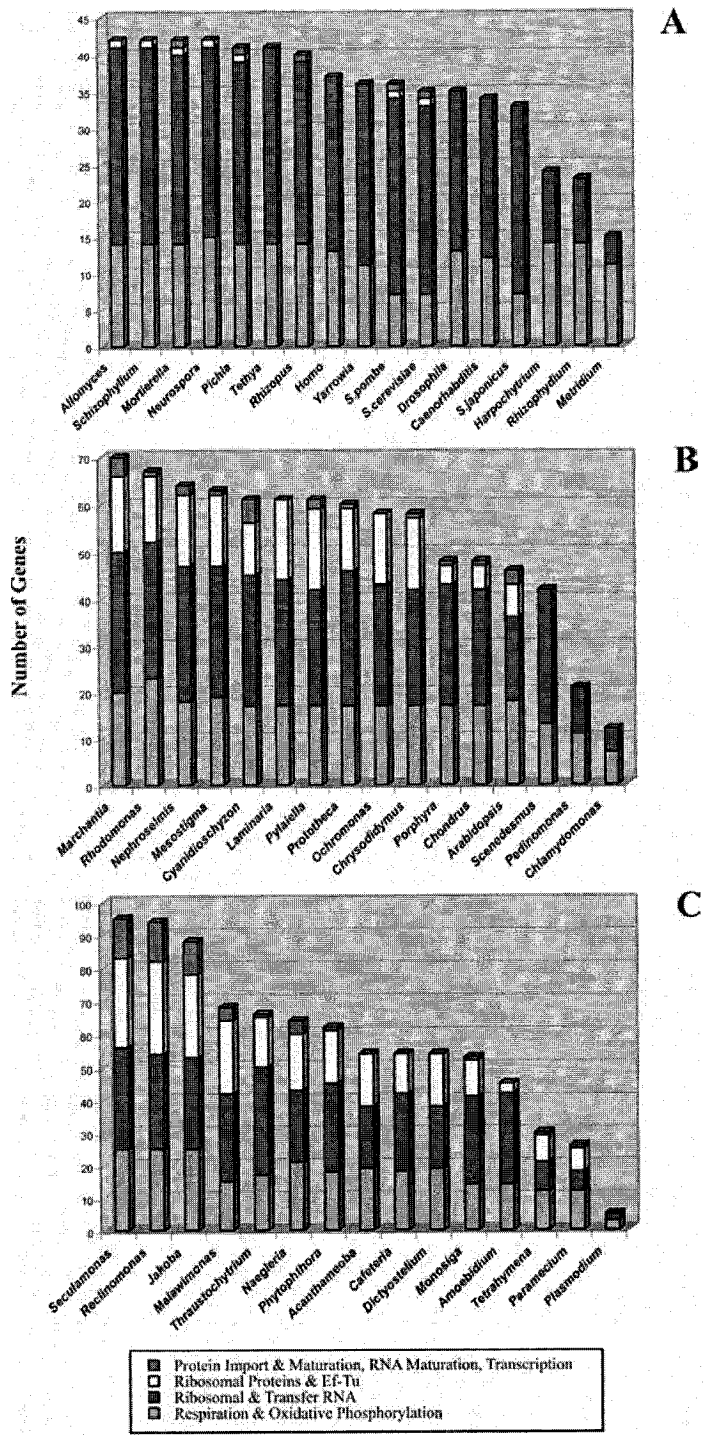

A
B
C
Number of Genes
Allomyces
Schizophyllum
Mortierella
Neurospora
Pichia
Tetrya
Rhizopus
Homo
Yarrowia
S.pombe
S.cerevisiae
Drosophila
Caenorhabditis
S.japonicus
Harpochytrium
Rhizophydium
Matridium
Marchantia
Rhodomonas
Nephroselmis
Mesostigma
Cyanidioschyzon
Laminaria
Pylaiella
Prototheca
Ochromonas
Chrysodidymus
Porphyra
Chondrus
Arabidopsis
Scenedesmus
Pedinomonas
Chlamydomonas
Seculamonas
Reclinomonas
Jakoba
Malawimonas
Thraustochytrium
Naegleria
Phytophthora
Acanthamoeba
Cafeteria
Dictyostelium
Monosiga
Amoebidium
Tetrahymena
Paramecium
Plasmodium
Protein Import & Maturation, RNA Maturation, Transcription
Ribosomal Proteins & Ef-Tu
Ribosomal & Transfer RNA
Respiration & Oxidative Phosphorylation

Fig. 2. (overleaf) Mitochondrial gene classes and their representation across eukaryotes. Species are ordered by the total number of genes, from left to right. Genes included in the function classes are Respiration & Oxidative Phosphorylation (violet): *atp1,3,4,6,8,9*, *cob*, *cox1-3*, *nad1-4,4L,6-11*, *sdh2-4*; Ribosomal and Transfer RNAs (maroon): *rnl*, *rns*, *rrn5*, *trnA*, *trnC*, etc.; Ribosomal Proteins and EF-Tu (yellow): *rps1-4,7,8,10-14,19*, *rpl1,5,6,10,14,16,18-20,27-31,32,34,36*, *tufA*; RNA Maturation, Protein Import & Maturation and Transcription (blue-green): *rnpB*, *secY*, *tatC*, *yejR (ccmF)*, *yejU (ccmC)*, *yejV (ccmB)*, *yejW (ccmA)*, *cox11*, *rpoA-D*. **Part A**, Fungi + Animals: *Allomyces macrogynus* (chytridiomycete); *Schizophyllum commune* (basidiomycete); *Mortierella verticillata* (zygomycete); *Pichia canadensis* (ascomycete); *Tethya sp.*, (demosponge); *Rhizopus stolonifer* (zygomycete); *Homo sapiens*; *Schizosaccharomyces pombe*; *Saccharomyces cerevisiae*; *Schizosaccharomyces japonicus* (ascomycetes); *Harpochytrium94*; *Rhizophydium sp.* (chytridiomycetes); *Metridium senile* (cnidarian animal); *Yarrowia lipolytica* (ascomycete). **Part B**, Plants + Algae: *Marchantia polymorpha* (liverwort, bryophyte); *Nephroselmis olivcea* (green alga); *Prototheca wickerhamii* (green alga); *Arabidopsis thaliana* (flowering plant, angiosperm); *Scenedesmus obliquuus* (green alga); *Chlamydomonas reinhardtii* (green alga); *Pedinomonas minor* (green alga, chlorophyte); *Cyanidioschyzon merolae*; *Porphyra purpurea*; *Chondrus crispus* (red algae); *Pylaiella littoralis*; *Ochromonas danica*; *Chrysodidymus synuroideus* (golden-brown algae); *Rhodomonas salina* (cryptophyte alga). **Part C**, Protists: *Seculamonas ecuadoriensis*; *Reclinomonas americana*; *Jakoba libera* (jakobid flagellates); *Malawimonas californiana* (malawimonad flagellate); *Thraustochytrium aureum* (stramenopile); *Naegleria gruberi* (heterolobosean amoeba); *Phytophthora infestans* (stramenopile); *Acanthameoba castellanii* (rhizopod amoeba); *Cafeteria roenbergensis* (stramenopile flagellate); *Dictyostelium discoideus* (slime mold); *Monosiga brevicollis* (choanoflagellate); *Amoebidium parasiticum* (ichtyosporean protist); *Tetrahymena*; *Paramecium aurelia* (ciliates); *Plasmodium falciparum* (apicomplexan protist).

on genes that are not needed by a specialized organelle (e.g. amino acid biosynthesis genes). Another process involves the replacement of the function of a mitochondrial gene by a nuclear gene after loss from the mitochondrion. An example of this is the loss of *rps8* in *Arabidopsis thaliana* and subsequent replacement by nuclear *rps15A* (Adams et al. 2002a). The final method of mitochondrial gene elimination is transfer to the nucleus. The most compelling evidence for this last process is seen in plants, in which such transfers are on-going. There is ample evidence of the intermediate stages of gene transfer, including all variants of gene presence, either in the nucleus, or in the mitochondrion, or both, as already discussed above for *rpl2* in connection with genes-in-pieces.

Numerous recent studies have been performed on angiosperms (well over 250 species) to detect evidence of organellar gene transfer events. Examples of such studies include: *sdh3* and *sdh4* (Adams et al. 2001b), 5' and 3' *rpl2* (Adams et al. 2001a), *cox2* (Adams et al. 1999), and a general study of 40 common mitochondrial genes (Adams et al. 2002b). Other analyses of unusual transfer events have also been performed including *infA* loss from the chloroplast (Millen et al. 2001) and *rbcL* transfer from the chloroplast to the mitochondrion (Cummings et al. 2003).

In each study, hybridization was used to detect presence or absence of the gene(s) of interest in both the organelle of interest and the nucleus. Sequencing

was used either to directly detect (i.e. *infA* (Millen et al. 2001)) or confirm (i.e. *cox2* (Adams et al. 1999)) gene loss in various plants. In each of these studies, loss events were mapped to a phylogenetic tree and nuclear mitochondrial targeting sequences were compared, in order to infer the pattern and number of gene losses to the nucleus. In one extraordinary case, an ancient gene loss for both mitochondrial *rps2* and *rps11* is predicted at the base of core eudicots, with hypothesized regains in a few cases (this will be discussed in further detail in section 2.4). Figure 3 depicts the angiosperm phylogenetic tree with loss and acquisition events from the various studies mapped to the tree.

However, despite the numerous well-studied examples of gene transfer to the nucleus, a good estimate of the number and identity of transferred genes is currently lacking. This is due to the difficulty in demonstrating, with confidence, that given nuclear genes are phylogenetically derived from the genome of the mitochondrial endosymbiont. Confronted with the fact that only a small minority of yeast mitochondrion-targeted proteins can be phylogenetically traced with confidence, either to an α–proteobacterial or to an alternative genomic source (Karlberg et al. 2000), BLAST analyses (or lack of sequence similarity to any known gene) have been used instead to assess the evolutionary origin of most of these genes. We share the concern of others (Koski and Golding 2001) that conclusions about the evolutionary history of genes, if based on such similarity measures, come with a serious degree of uncertainty, if not systematic error. In fact, yeast, with its highly accelerated rate of gene evolution and its highly reduced nuclear genome, is not an ideal organism for such estimates.

Due to redundancy of metabolic genes that were present in the ancestral eukaryotic host, such genes were likely among the first to be eliminated from mitochondrial genomes. From the remaining genes that are essential to core mitochondrial functions - oxidative phosphorylation, DNA replication, transcription and translation - a hierarchical pattern of gene loss or transfer to the nucleus (e.g. *nad* and *rps* genes) has been observed. This pattern has been seen in many phylogenetically diverse lineages, including algae, fungi, plants, and animals (Lang et al. 1999). For example, the genes encoding the ribosomal protein, Rps1, and the NADH dehydrogenase subunit, Nad8, are usually lost first, whereas *rps3*, *nad1-6,* plus *nad4L* are the last to go. In several taxa, including budding and fission yeasts, all *nad* genes have been lost from mtDNA, due to the complete elimination of the proton-pumping, rotenone-sensitive mitochondrial NADH dehydrogenase complex (Friedrich et al. 1995; Friedrich and Weiss 1997; Rasmusson et al. 1999). In these cases, electron transfer from NADH to ubiquinone is mediated by an alternative, single-polypeptide enzyme with FAD as sole prosthetic group (Yagi 1991).

The strongest support for hierarchical gene loss comes from the gene content of the most reduced mitochondrial genomes. Regardless of phylogenetic placement, their gene content is quite similar (see also chapter on mitochondrial genome comparisons) (Lang et al. 1999). Several, non-exclusive hypotheses can explain the ability for certain genes to be physically transferred to the nucleus. Factors possibly limiting the ability of gene transfer include: hydrophobicity (von Heijne 1986), preventing transfer through the mitochondrial membrane; mechanism of adoption of the functional structure of the protein depending on other mitochon-

drial proteins; and gene expression based upon the redox potential of the mitochondrial membrane (Forsberg et al. 2001).

2.4 Immigration of genes to mitochondria, and horizontal gene transfer among plant mitochondrial genomes

There is little evidence to support gain of genes by mitochondria. The few known cases include: gain of genes for DNA and RNA polymerase from fungal, protist, or plant mitochondrial plasmids (e.g. Wahleithner and Wolstenholme 1988; Court and Bertrand 1993; Grace et al. 1994; Hermanns and Osiewacz 1994; Takano et al. 1994) of a *mutS*-related gene in coral mtDNAs (Pont-Kingdon et al. 1998) of chloroplast and nuclear sequences in various plants (e.g. Dietrich et al. 1996; Glover et al. 2001; Cummings et al. 2003) and regain of the mitochondrial *rps2* and *rps11* in a few core eudicots (Fig. 3) after an ancient loss (Bergthorsson et al. 2003). Indeed, that latter case is one of the few known examples of the regain of a gene after it has been lost, and it is hypothesized to be due to horizontal transfer from unrelated non-eudicot plant species (Bergthorsson et al. 2003).

In the case of *rps2* and *rps11*, evidence for regain of the genes through horizontal gene transfer comes from a phylogenetic analysis including numerous plant species. Unexpectedly, the *rps2* sequence gained in Actinidia (order Ericales, Fig. 3) groups strongly with monocot *rps2* sequences, and the *rps11* gain in *Lonicea* and other Caprifoliaceae (order Dipsacles, Fig. 3) groups with the order Ranunculales, also with strong support. In the case of *rps11* gain in *Betula* (order Fagales, Fig. 3), the phylogenetic position of the acquired sequence was not resolved, but sequence divergence levels were low, suggesting horizontal gene transfer among mitochondria as the underlying mechanism for regain of the gene. Reverse gene transfer from the nucleus to the mitochondrion can be excluded because a substantially higher nuclear nucleotide substitution rate would have led to much more highly divergent sequences than were observed. Sequencing of the genes supports the possibility that some of the horizontally transferred genes are active, whereas others are clearly pseudogenes (Bergthorsson et al. 2003). A most surprising example for activity of genes without disabling mutations is a chimeric *rps11* gene in *Sanguinaria canadensis* (not included in Fig. 3), of half monocot and half eudicot origin, which is expressed and RNA-edited (Bergthorsson et al. 2003). The authors conclude that such horizontal transfer events are perhaps much more widespread in plants than initially supposed.

Another unique transfer event involves that of the chloroplast gene *rbcL* to mitochondria. Twenty angiosperms were tested for chloroplast-to-mitochondrion transfer of *rbcL* sequences, with results indicating five to six separate intraorganellar transfers events (Cummings et al. 2003). However, sequence comparisons show frame shift mutations, truncations and a great number of non-

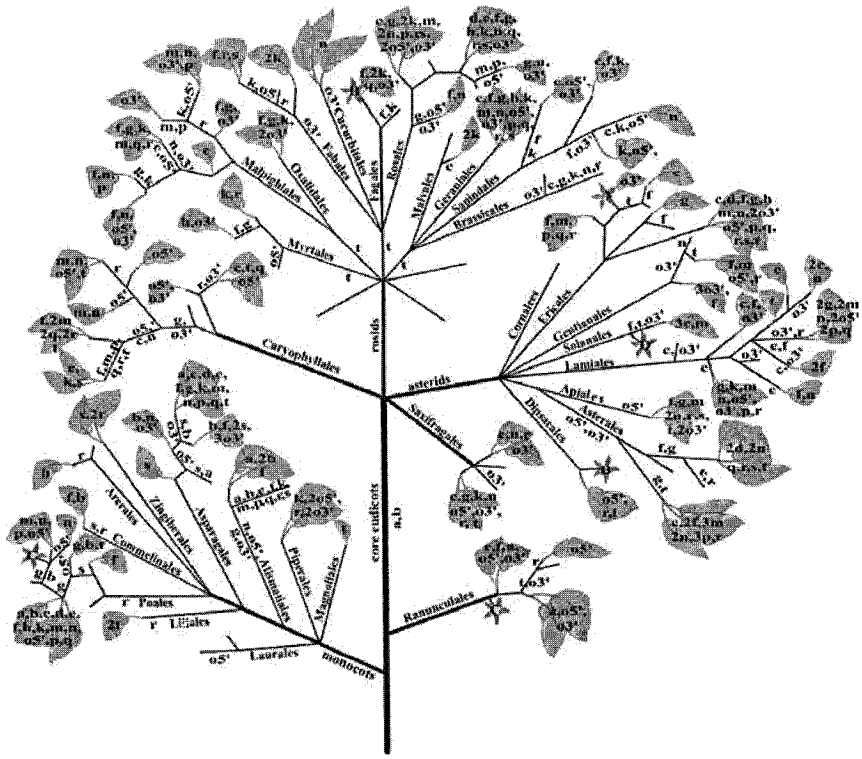


Fig. 3. Summary of gene transfer events according to studies on ~280 angiosperms. The major angiosperm orders for which gene loss from the mitochondria was detected are depicted based upon the phylogenetic tree of (Soltis et al. 1999). Additionally, unique gene regain events are also shown. For details of the data used see Adams et al. 1999, 2000, 2001a, 2001b and Millen et al. 2001. Transfer events marked along branches represent those that encompassed a clade of species. Transfer events in leaves represent either single events in individual species, or multiple events in several species (e.g. '2g' indicates that 2 separate transfers of gene g occurred in 2 different species in the clade). The following is a legend of gene transfers from the mitochondria to the nucleus: a – *rps2*, b – *rps11*, c – *rps1*, d – *rps3*, e – *rps4*, f – *rps7*, g – *rps10*, h – *rps12*, k – *rps13*, m – *rps14*, n – *rps19*, o3' – *rpl2* 3', o5' – *rpl2* 5', p – *rpl5*, q – *rpl16*, r – *sdh3*, s – *sdh4*, t – *infA*. The following is a legend of gene regain (horizontal transfer from other distantly related species): A – *rps2* gain, B – *rsp11* gain. Also of note are the presence of full *rpl2* in the mitochondria of a single species (labeled C on the tree) and the fact that gene *infA* was found missing in a species but the same species was not tested for any of the other genes as listed above (labeled D on the tree).

synonymous mutations in all mitochondrial *rbcL* sequences, which have no apparent mitochondrial function. The combined evidence indicates that they are most likely pseudogenes.

The mitochondrial genes *rps13* and *rps8* provide two other interesting examples in which loss of a mitochondrial gene is not accompanied by the usual trans-

fer to the nucleus, but instead by substitution with duplicated nuclear genes. In the case of *rps13* of cotton and two legumes (*M. trunculata* and *L. japonicus*) that have no mtDNA-encoded *rps13*, nuclear genes with mitochondrial targeting sequences have been identified (Adams et al. 2002a). According to phylogenetic analyses, these genes have been derived through a duplication of a chloroplast-targeted nuclear *rps13* that subsequently acquired a mitochondrial targeting sequence, an event that took place early in rosid evolution. A parallel study reveals that angiosperms do not contain a mtDNA-encoded *rps8* gene. However, closely related copies of *rps15A*, the cytosolic counterpart of *rps8*, are present in *Arabidopsis*, tomato, barley, and *M. trunculata*, all with specific mitochondrial targeting sequences (Adams et al. 2002a). Substitution of mitochondrial *rps8* by cytosolic *rps15A* is hypothesized to have occurred in the common ancestor of angiosperms and gymnosperms.

2.5 Possible mechanisms of gene migration

One of the central questions is why, and to what extent, transfer of organellar genes to the nucleus might be advantageous for the cell. In the aforementioned study of *cox2* in legumes (Adams et al. 1999), five species were found with dual, intact and transcribed, nuclear and mitochondrial *cox2* genes, indicating that either of the genes might potentially be silenced. In a phylogenetic context, the results of the study led to the inference of four and five separate inactivations of mitochondrial and nuclear *cox2* genes, respectively. From these results, it is unclear if there is any selective advantage for a mitochondrial *versus* nuclear location of *cox2* in these plants. In fact, in mitochondria throughout eukaryotes, the *cox2* gene is almost universally present.

Henze and Martin (2001) discuss alternatives for transfer of genes from organelles to the nucleus, whether as RNA, cDNA, or genomic DNA. Current views strongly favor cDNAs as the transport vehicle based upon a number of examples from plants, because RNA editing, non-standard translation codes, and (in a few genes) the presence of organelle-specific introns, render the genes difficult to express in the nucleus. However, direct transfer of genomic DNA might well occur in non-plant species. Support for the possibility of direct transfer of mtDNA to the nucleus is suggested by the presence of a complete, continuous piece of mtDNA (620 kbp, with repeated sections) in the nuclear genome of *Arabidopsis* (Lin et al. 1999; Stupar et al. 2001), and by substantial numbers of mtDNA insertions in the nuclear genomes of primates (Woischnik and Moraes 2002). However, conclusive evidence for a functional gene transfer mechanism *via* genomic DNA is currently lacking.

Successful gene translocation to the nucleus requires a number of conditions: (i) the coding material must be transferred to the nucleus, either by direct transfer of mtDNA, or *via* a cDNA intermediate; (ii) that export of the nucleic acid from the mitochondria takes place either through specific transport, or through escape after membrane disruption (Brennicke et al. 1993); (iii) sites undergoing RNA editing have to carry corresponding changes in the transferred gene, and removal of

introns has to occur, so that the nuclear gene will be functional. These constraints makes gene transfer *via* an RNA intermediate that is reverse transcribed into cDNA most likely; (iv) codons, altered according to mitochondrion-specific genetic codes, have to be changed when transferred to the nucleus. In fact, it is likely that species with a non-standard mitochondrial translation code no longer (or only in very rare instances) transfer genes from the mitochondrion to the nucleus (Brennicke et al. 1993). (v) Import of the gene material into the nucleus must occur to allow integration into the nuclear genome, which is possible in general because nucleic acids constitutively travel through nuclear pores (Brennicke et al. 1993); (vi) the gene has to be activated for its mitochondrial function by recombinative attachment of an amino-terminal mitochondrial targeting sequence. Experiments in both yeast and *E. coli* demonstrate that a large number of such potential presequences are available in these genomes (Baker and Schatz 1987), indicating that the recruitment of target sequences would be only a minor barrier against gene transfer to the nucleus. Finally, (vii) transcription has to occur from a promoter of adequate strength and regulation.

3 Mitochondrial genome diversity in major eukaryotic groups

Mitochondria have a wide variety of structural features, including size variation, circular- and linearly-mapping genomes, different chromosome counts, and differences in gene and intron content and in gene expression mechanisms. In view of a number of publications that have appeared since these topics have been reviewed in detail, the following section discusses features of protist, animal, fungal, and plant mitochondrial genomes, selected to highlight these attributes. For the investigation of more detailed sequence-related features, we propose use of the mitochondrial genome database, GOBASE (Korab-Laskowska et al. 1998; O'Brien et al. 2003), which contains up-to-date, expert-curated mitochondrial sequences, and which in their original form are also available through GenBank/NCBI.

3.1 Animals

At the time of writing, completely sequenced mtDNAs were available from more than 300 animals. These mitochondrial genomes are typically small (14 to 17 kbp in size; Table 1), although there are instances of larger genomes in a few species (e.g. in bark weevils and scallops), due to either gene duplications or to accumulation of sequence repeats in the control region (e.g. Boyce et al. 1989; Rigaa et al. 1995). In most instances, genes are compactly arranged on a single, circular (mono- or dimeric, supercoiled) mtDNA, with transcription on both strands (for more details and exceptions see Boore 1999). The majority of animal mitochondrial genomes contain the same set of 37 genes (Fig. 2A), including *cox1,2,3*, *cob*, the NADH dehydrogenase units *nad1-6* and *nad4L*, *atp6* and *atp8*, as well as

genes for the small and large rRNA subunits (*rnl* and *rns*), and up to 22 tRNAs (two for both leucine and serine, but a single one for both methionine and isoleucine). An additional 2-3 tRNA genes are encoded in demosponges, mitochondrial genomes that are the least derived among all Metazoa (Lavrov and Lang, unpublished). The structures of both animal rRNAs and tRNAs are very reduced and have little similarity with bacterial and other mitochondrial counterparts (e.g. see the extreme case of nematodes Okimoto et al. 1992). Furthermore, animal, like fungal, mtDNAs exhibit none of the operon-like bacterial gene clustering that is found in mtDNAs of protists, fungi and plants (see following sections; for a review see Lang et al. 1999).

Animal mitochondrial gene order is remarkably stable within the major taxonomic lineages, unlike in most other eukaryotes. However, there appear to be sufficient differences in mitochondrial gene order to provide the relevant information (i.e. synapomorphic changes) for resolving the phylogeny of deep metazoan branches (Boore and Brown 1998; Lavrov et al. 2002).

3.2 Fungi

Although fungi and animals share a common origin to the exclusion of other eukaryotes, and have a similar basic set of mitochondrial genes, fungi are much more variable in terms of size, structure, and content of genes and introns. In the following section, we will describe selected examples of mtDNAs representing both chytridiomycete ('lower') and ascomycete fungi (for more detailed recent reviews on fungal mtDNAs see: Paquin et al. 1997; Bullerwell et al. 2003c; Hauser 2003; Kennell and Cohen 2003; Leigh et al. 2003)

Currently, mtDNAs are available from 23 fungal taxa, including representatives of the four major divisions of this kingdom, Ascomycota, Basidiomycota, Zygomycota, and Chytridiomycota (Lang 2003). The chytridiomycete fungus *Hyaloraphidium curvatum* has a mitochondrial genome of approximately 30 kbp (Table 1). The genome is organized into linear monomers with inverted repeats at the termini (Forget et al. 2002). Inability to completely sequence the ends of the genome suggests that there are protein complexes or closed hairpin structures at the termini; however, further examination gave no indication of single-stranded closed loops (Forget et al. 2002). Genes in the *H. curvatum* mtDNA total 26 including only seven that code for tRNAs, and only the *cob* gene is interrupted by a group I intron. Despite a strong codon bias, the seven mtDNA-encoded tRNAs are largely insufficient to translate all codons occurring in protein coding genes. Consequently, the majority of tRNAs must be nucleus-encoded and imported. Mitochondrial genome size and the intron count varies significantly in the monoblepharidalean relatives of *H. curvatum*, from 19.5 kbp (*Harpochytrium94*; no introns) to 60.4 kbp (*Monoblepharella15*; eight group I introns, seven having intron ORFs (Bullerwell et al. 2003a).

Unlike *H. curvatum* and its monoblepharidalean relatives, the chytridiomycete *Allomyces macrogynus* (Blastocladiales) has a full set of tRNA genes (25) that are sufficient to recognize all codons occurring in its mitochondrial protein-coding

genes (Fig. 2A). Its circular-mapping genome is almost twice the size of *H. curvatum* mtDNA (Table 1; Paquin and Lang 1996). The intron content of *A. macrogynus* is even higher than in *Monoblepharella* (26 introns of group I and two of group II), accounting for 37% of the mtDNA sequence (Paquin and Lang 1996). In distinction to all other chytridiomycetes, the *A. macrogynus* mtDNA encodes the ribosomal protein gene *rps3*, the first ribosomal protein gene identified by sequence similarity in a fungal mtDNA, testifying to the ancestral nature of *A. macrogynus* (for identification of further *rps3* genes see Bullerwell et al. 2000).

Quite similar to the situation in chytridiomycetes, ascomycete mtDNAs are highly variable even at relatively close phylogenetic distance, as for instance in the three fission yeasts *S. pombe*, *S. octosporus* and *S. japonicus var. japonicus* (Bullerwell et al. 2003b). *S. japonicus var. japonicus* has the largest genome at over 80 kbp, *S. octosporus* is next at about 44.2 kbp and *S. pombe* the smallest at about 19.4 kbp. Despite the large size differences, *S. pombe* and *S. octosporus* have virtually identical gene content, including the only recently described *rnpB* gene that codes for the RNA subunit of mitochondrial RNase P (Seif et al. 2003). In comparison, although *S. japonicus var. japonicus* has the largest mtDNA, it lacks both *rnpB* and *rps3* (Table 1, Fig. 2A). The standard genetic code is used in all protein-coding genes in all three species, with the exception of *rps3* in *S. pombe* and intronic ORFs in each species. In these rare exceptions, TGA is expected to code for tryptophan (Bullerwell et al. 2003b).

One of the major differences between the three species is the amount of noncoding sequence in the genome. These highly A+T rich regions (at 77.6%, 80.7%, and 82.0% in *S. pombe*, *S. octosporus* and S. *japonicus var. japonicus*, respectively) account for 11.1% of the genome of *S. pombe*, 49.4% of the genome in *S. octosporus* and an exceptional 75.6% of the genome in *S. japonicus var. japonicus* (Fig. 1). Each species has few introns: two group I and one group II intron in *S. pombe*; four group I and two group II introns in *S. octosporus*; and two group I introns in *S. japonicus var. japonicus*.

3.3 Plants and green algae

Regarding gene content, the mitochondrial genomes of plants, green algae, and other protists resemble each other more closely than do those of either animals or fungi (Lang et al. 1999). Yet plant and protist mtDNAs are very different in terms of size (land plants have the largest known mtDNAs, with sizes ranging from 180 to 2,400 kbp (Ward et al. 1981; Palmer et al. 1992)) due to the presence of large intergenic regions, ORFs, pseudogenes, introns and foreign DNA. In angiosperms, mtDNA undergoes frequent genome rearrangements and formation of subgenomic circles through recombination at direct repeats (Backert et al. 1997). In spite of this high genome variability, plant genes have exceptionally low evolutionary rates compared to their animal and fungal counterparts and, even more surprisingly, compared to the nuclear genomes of plants.

Several mtDNAs of the green algal relatives of plants have been characterized within the last few years, falling into the two sister lineages of green algae, the

Chlorophyta (comprising the Prasinophyceae, Ulvophyceae, Trebouxiophyceae, and Chlorophyceae), and the Streptophyta, (comprising the Charophyceae and land plants as a monophyletic group). While the prasinophytes *Nephroselmis olivacea* and *Mesostigma viride* have the most ancestral, gene-rich green algal mtDNAs of those sequenced, *Pedinomonas minor* has a small mtDNA with only a few genes (similar to *Chlamydomonas* sp. mtDNAs), a non-standard translation code (UGA, tryptophan), a 'genes-in-pieces' LSU rRNA gene, and a disproportionally large A+T-rich repeat region (Turmel et al. 1999, 2002b). Overall, *Pedinomonas* and the chlamydomonads share virtually all characteristics of a reduced-derived genome (note that the terms 'ancestral', 'primitive', and 'reduced-derived' do not have a phylogenetic meaning, but rather compare how much mtDNAs resemble bacterial genomes). Conversely, *Nephroselmis* and *Mesostigma* have retained many prokaryotic features, which classify these mitochondrial genomes together with those of *Prototheca wickerhamii* (another chlorophyte (Wolff et al. 1994)) and two recently sequenced charophyte mtDNAs as little derived (Turmel et al. 2002a, 2002b, 2003).

As seen in this detailed analysis of plant and green algal mitochondrial genomes there is a pattern of progressive loss of genes and of ancestral features from primitive green algae, to primitive land plants (such as *Marchantia polymorpha* (Oda et al. 1992)), to flowering plants (Fig. 2B).

3.4 Protists

Protist are a highly heterogeneous group of several dozen phyla of eukaryotes that are negatively defined as not belonging to animals, fungi, or plants. Because of this definition, the closest relatives of animals and plants are found within the protists, exemplified by choanoflagellates in the case of animals, and charophyte algae in the case of plants. As protists comprise the majority of biological diversity, the evolutionary relationships among most protist phyla currently remain either unknown or contentious.

Many protist mitochondrial genomes are more bacteria-like than those of animals and fungi (Fig. 2C; for a more general overview see Lang et al. 1999 and Gray et al. 2001). In this respect, the most striking mtDNAs are those of jakobids, which resemble a bacterial genome in miniature (see section 5.1.2 for more details). At the other end of the spectrum, the malaria parasite *Plasmodium falciparum* has the smallest mtDNA with only three protein-coding and two high fragmented rRNA genes (Fig. 2C) (Feagin et al. 1997).

A highly unusual mitochondrial genome architecture has recently been identified in *Amoebidium parasiticum*, an ichthysporean protist, which together with the choanoflagellate *Monosiga brevicollis* has been recognized as a close specific relative of multicellular animals (Lang et al. 2002; Burger et al. 2003a). The mitochondrial genome of *A. parasiticum* consists of several hundred linear chromosomes ranging in size from 0.3 to 8.3 kbp, with virtually identical, short inverted terminal repeats. The orientation of the repeats correlates with the direction of

transcription, possibly indicating that they are involved in the transcription process (Burger et al. 2003a).

4 Value of mitochondrial sequences for phylogenetic inference

The resolution and predictive power of phylogenies using molecular sequence data is often unprecedented, when compared to morphological, biochemical and ultrastructural characters. Sequence-based phylogenies are not only less limited by the number of informative characters to consider, but are based on relatively well-understood models of sequence evolution that are mathematically tractable. Despite the high number of sequence positions in single genes or proteins, statistically well supported inferences of deep (old) divergences usually require the use of sequences from multiple, well conserved genes.

Sets of multiple protein sequences encoded by mtDNAs are now available for most eukaryotic groups, and have proven their value for phylogenetic analyses in many instances. An advantage to mitochondrial sequence use is the virtual lack of lateral gene transfer in mitochondria (but see the discussion on lateral exchange among flowering plant in section 2.4). It is now widely accepted that mitochondria originated only once, from within the α-Proteobacteria (no secondary endosymbiotic events have been discovered). Thus, the mitochondrial phylogeny reflects the phylogeny of (mitochondriate) eukaryotes, and the α-Proteobacteria can be used as a relatively close, unambiguous outgroup for these analyses.

However, while mitochondrial protein-based phylogenies often predict eukaryotic phylogenies with high statistical support (Fig. 4) (e.g. Burger et al. 1999; Forget et al. 2002; Lang et al. 2002; Bullerwell et al. 2003a), this dataset does not have sufficient phylogenetic signal for the prediction of numerous deep protist and animal phylogenies. This lack of support results from the unusually high evolutionary rate of animal mitochondrial genes, which obliterates most of the phylogenetic signal. In contrast, rearrangements in animal mitochondrial genomes are exceptionally rare events. Therefore, mitochondrial gene order data may be used for inferring deep evolutionary relationships within the animals (e.g. Boore et al. 1995; Boore and Brown 1998; Sankoff et al. 2000; Lavrov et al. 2002). A gene order approach is obviously restricted to cases where sufficient shared-derived characters (synapomorphies) are available. For this reason, this approach is not applicable to many animal groups including mammals, where mitochondrial gene organization is nearly identical.

5 Origin and evolution of mitochondria and of the eukaryotes

It has become increasingly clear that the nuclear genome of eukaryotes is an evolutionary chimera that incorporates substantial fractions of genetic material from other sources. Indeed, there are many parts of the genome that are most similar in sequence to organisms as unrelated as Proteobacteria, Archaebacteria, and (in photosynthetic species) Cyanobacteria. However, it has mostly remained a question of belief, rather than of phylogenetic evidence, whether the entire proteobacterial fraction of genes in eukaryotic genomes stems from the α-proteobacterial ancestor of mitochondria, whether additional cellular fusions that pre-date the mitochondrial endosymbiosis must be invoked, or whether many more punctual, lateral gene transfers from closely associated or ingested (food) bacteria have occurred. To resolve these questions, additional genomic information will be required that more precisely defines: (i) the nature of the α-proteobacterial endosymbiont that gave rise to the mitochondrion and its closest extant relatives; (ii) the nuclear genomes of a number of minimally-derived eukaryotes (protists); (iii) the mitochondrial genomes that have evolved in concert with these protist nuclear genomes.

5.1 Eubacterial ancestry of mitochondria

5.1.1 Rickettsia prowazekii, one of the closest living relatives of mitochondria

The study of the genome sequences of the pathogenic α-Proteobacteria *Rickettsia prowazekii* and its relatives such as *Ehrlichia, Anaplasma, and Wolbachia* (all belonging to the Rickettsiales) is of particular importance, since they are among the closest (if not the closest) living relatives of the mitochondrion (e.g. Gray 1993; Andersson et al. 1998; Ogata et al. 2000, 2001; Burger and Lang 2003). The approximately 1.1 mbp genome of *R. prowazekii* was the first α-proteobacterial genome to be completely sequenced. It is circular mapping, codes for only ~834 proteins, and compared to free-living α-Proteobacteria has a very high overall A+T content of 70.9%. Approximately 24% of the genome is non-coding, including 0.9% that consists of pseudogenes and 0.2% that consists of repeats. It is hypothesized that much of this non-coding region was once coding, and is slowly being eliminated from the genome (Andersson et al. 1998). Out of the identified genes 24.8% have no similarity to any other predicted or known gene, 12.5% have similarity to those predicted in other species, and 62.7% are known genes or pseudogenes with identified function (Andersson et al. 1998).

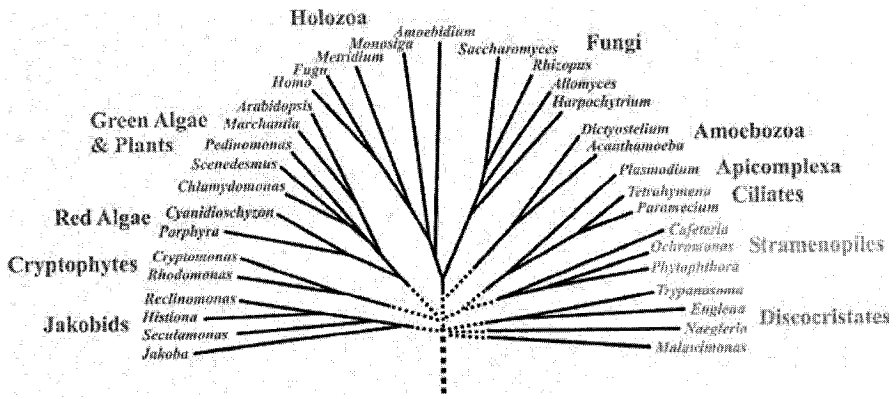


Fig. 4. Schematic tree of phylogenetic relationships among eukaryotes. Major organismal groups are color-coded; for taxonomic nomenclature see (Lang et al. 2002; Baldauf 2003). Solid lines indicate phylogenetic branching orders that are statistically well supported in analyses with concatenated mitochondrial protein sequences (and for most part, consistent with nuclear sequence – based phylogenies). Dotted lines indicate uncertain branching order based on conjecture, at the base of the tree, in excess of approximately 1 billion years.

5.1.2 Jakobid flagellates have the most gene-rich and eubacteria-like mtDNAs

Since minimally-derived eukaryotes are key to a better understanding of evolutionary events that took place in early eukaryotic evolution, it is necessary to develop criteria to identify those species that 'time forgot' (Palmer 1997). From the standpoint of mitochondria, jakobid flagellates are clearly in this group of organisms, as they do contain the most gene-rich and the most eubacteria-like mtDNAs among eukaryotes. One such jakobid, *R. americana,* carries a total of 98 genes, now also including an only recently identified tmRNA-like gene (Keiler et al. 2000; Jacob et al. 2004), and all the mitochondrial genes found in other sequenced mtDNAs (Lang et al. 1997) with the exception of *mutS* of corals (Pont-Kingdon et al. 1995). Eighteen protein-coding genes of known function and the tmRNA-like gene are unique among published mitochondrial sequences. Among the most surprising of these novel genes are: four (*rpoA-D*) encoding a eubacteria-like RNA polymerase with a regulatory sigma factor (*rpoD*); a gene encoding the RNA subunit of RNase P with all the structural hallmarks of its bacterial homologs; two genes (*tatA* and *tatC* (Jacob et al. 2004)) involved in the twin-arginine protein translocation pathway (Wu et al. 2000); and one (*secY*) involved in the Sec-dependant protein transport pathway. Other members of the jakobids such as *Histiona*, *Jakoba*, and *Seculamonas* have similarly gene-rich mitochondrial genomes (B. Franz Lang, unpublished). Together with *Reclinomonas*, these species are currently the subjects of extensive cDNA sequencing in order to explore the coding capacity of their nuclear genomes and to define the phylogenetic position of jakobids in the global eukaryotic tree.

5.2 Serial endosymbiosis, or metabolic syntrophy model of mitochondrial origin?

Analyses of the genetic material contained within mitochondria and chloroplasts have clearly shown that their closest contemporary relatives are α-Proteobacteria and Cyanobacteria respectively (e.g. Gray and Spencer 1996a). These results lend support to the hypothesis that the two organelles originated as bacterial endosymbionts, as originally postulated in the Serial Endosymbiosis Theory (Taylor 1974). Although the endosymbiont theory can be traced back almost a century (Schimper 1883; Altmann 1890) it was widely accepted only in its most recent forms (Gillham 1974; Margulis 1975, 1988; Cavalier-Smith 1987; Gray 1989, 1992, 1998; Gray and Spencer 1996b). It is even proposed to account for different organelles including undulipodia (9+2 flagella) (Margulis 1988) and peroxisomes (de Duve 1969, 1982, 1996). The SET has various formulations (de Duve 1996; Akhmanova et al. 1998), in which the host was either an archaebacterium or a nucleus-containing eukaryote. However, the central points to all formulations of SET are: (i) the step-wise association of a host with bacterial symbionts, in which chloroplast acquisition followed that of the mitochondrion (de Duve 1996); (ii) that the host's metabolism was "characteristic of the eukaryotic nucleocytoplasm" (Margulis 1981), being both heterotrophic and anaerobic (Fig. 5). A further notion inherent in the hypotheses is that the host cell was the primary source of the nuclear genome of the primitive eukaryote. Thus, genes in the nucleus that are related to mitochondrial function would have resulted from mitochondrion-to-nucleus transfer.

However, recent data show that although the eukaryotic nuclear genome contains genes of both archaebacterial and eubacterial ancestry, the eubacterial component of the genome is much larger than would be expected if it were due only to transfer from the mitochondria (Golding and Gupta 1995; Feng et al. 1997). Furthermore, there are some genes of posited eubacterial ancestry that are not directly involved in mitochondrial biogenesis and function (Markos et al. 1993; Keeling and Doolittle 1997; Hashimoto et al. 1998; Karlberg et al. 2000). It appears that, in general, genes of archaebacterial origins are informational, whereas those of eubacterial origin are operational (Rivera et al. 1998). Together, this new evidence calls for modifications of the hypothesis of mitochondrial origins that will account for the chimeric nature of eukaryotic nuclear genomes.

Various models involving fusion of eubacterial and archaebacterial partners in the creation of the nuclear genome have been proposed (Zillig et al. 1989). These models invoke a major eubacterial contribution to the nuclear genome during its initial formation and assume a subsequent endosymbiotic acquisition of mitochondria. The more recently formulated hydrogen hypothesis provides a somewhat different explanation for the mosaic nature of the eukaryotic nucleus: the fusion of a Proteobacterium that was able to respire, with a strictly hydrogen-dependent Archaebacterium, based on metabolic syntrophy (Fig. 5). Indeed, associations between these two types of prokaryotes arefrequently observed in nature, and according to the hydrogen hypothesis, environmental association of the two

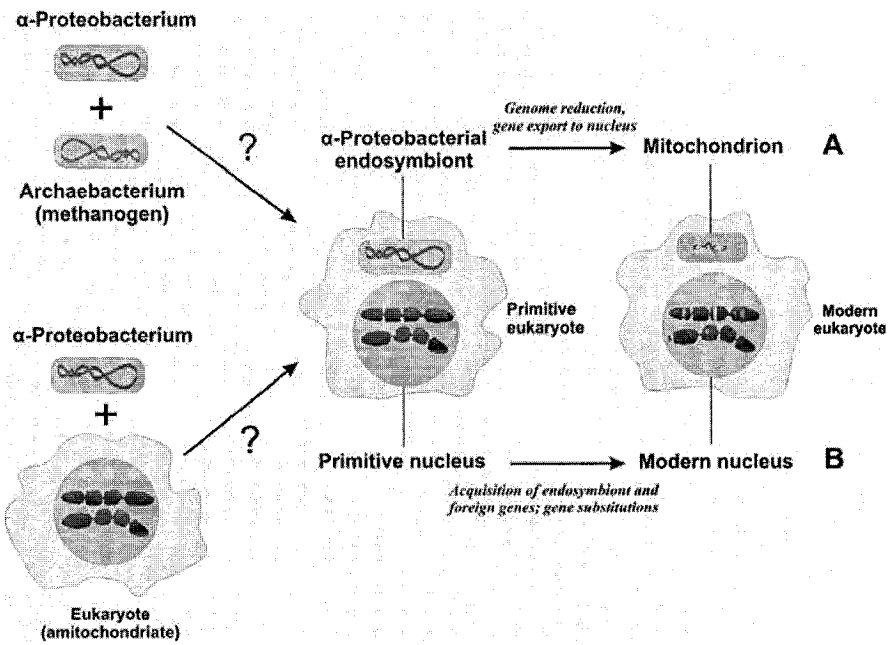


Fig. 5. Alternative hypotheses describing the origin of eukaryotes and of mitochondria. According to the 'Serial Endosymbiosis Theory', an α-proteobacterial endosymbiont is captured by a nucleus-containing eukaryotic host resembling extant amitochondriate protists. This theory does not endeavor to explain the origin of the eukaryotes through cell fusions. In contrast, the 'Hydrogen Hypothesis' of eukaryotic origins proposes the fusion of either an α-Proteobacterium with a methanogenic Archaebacterium to result either directly in a mitochondrion-containing cell; or alternatively (not shown in the figure), a fusion of an (α- or γ-) Proteobacterium with a methanogenic Archaebacterium, to create an amitochondriate eukaryote. According to accumulating evidence, however, most if not all extant amitochondriate eukaryotes might have secondarily lost their mitochondrion (converted to a hydrogenosome, 'mitosome' or 'crypton'); it is open to discussion if an amitochondriate eukaryote ever existed.

species led to a cellular integration of these two organisms (Martin and Müller 1998). Not only does this hypothesis account for the chimeric nature of the nuclear genome, but it also allows for a simultaneous origin of the eukaryotic cell and its mitochondrion, under the assumption that the proteobacterial partner was the α-Proteobacterium that gave rise to the mitochondrion. A similar hypothesis invokes the same principle of metabolic syntrophy, but involves a δ-proteobacterial symbiont instead (Moreira and Lopez-Garcia 1998).

Phylogenetic support is currently lacking to favor one or the other of these variant hypotheses. However, recent data have become available that might lend support to the hydrogen hypothesis. These data question the existence of early diverging amitochondriate organisms (collectively referred to as 'Archezoa' (Cavalier-Smith 1983; Martin and Müller 1998)) according to various phylogenetic analyses

(Vossbrinck et al. 1987; Sogin et al. 1989; Sogin 1991; Cavalier-Smith 1993; Leipe et al. 1993). The existence of such amitochondriate organisms, deemed to have never contained mitochondria, has been used as support for the hypothesis that the host in the endosymbiotic event was a primitive eukaryote that both had a nucleus and the ability to phagocytose bacterial cells. Yet, genes typical of mitochondrial function have recently been found in the nuclear genomes of almost all Archezoa (Bui et al. 1996; Germot et al. 1996, 1997; Horner et al. 1996; Hirt et al. 1997; Embley and Hirt 1998; Roger et al. 1998), some of which have can be traced phylogenetically to the rickettsial subdivision of the α-Proteobacteria. In several cases, such genes of (arguably) mitochondrial origin have been found to be targeted to hydrogenosomes (organelles that anaerobically produce ATP with hydrogen as a byproduct of the process), or to other organelles of currently unknown function ('mitosome' or 'crypton' (Mai et al. 1999; Tovar et al. 1999, 2003; Williams et al. 2002)). This suggests a common evolutionary origin for the hydrogenosome, mitosome, crypton, and the mitochondrion, and that extant amitochondriate eukaryotes once had mitochondria, which they lost only secondarily.

The lack of evidence for extant, primarily amitochondriate eukaryotes might be interpreted in favor of fusion theories such as the hydrogen hypothesis. However, such strictly anaerobic eukaryotes (the postulated host cells of the Serial Endosymbiosis Theory) might have existed during an early evolutionary period with still-anoxic conditions that would have favored their survival.

6 Concluding remarks

Questions about mitochondrial and eukaryotic evolution are most effectively approached by analyses of complete genomes from a phylogenetically comprehensive set of eukaryotes, i.e. through comparative genomics, and phylogenetic inferences using genomic data ('phylogenomics'). The comparative genomics approach to mitochondrial genome evolution has demonstrated many benefits, including the identification of previously unrecognized genes and genetic elements, inferences about unusual modes of gene expression, and a more precise portrayal of the evolutionary history of mitochondria, and by inference, of the eukaryotic cell. In fact, the data gathered has provided the strongest evidence so far for a monophyletic origin of mitochondria.

In addition, the quest for the most primitive (least derived) mitochondrial genomes has led to the identification of the jakobid flagellates, such as *Reclinomonas americana*. All known members of this lineage (including *Reclinomonas*, *Histiona*, *Seculamonas*, and *Jakoba*) have the most gene-rich and bacteria-like mtDNAs of all eukaryotes, as initially identified in *R. americana*. Although we feel that the search for even more ancestral mitochondrial genomes might have reached the upper limit with the jakobid mtDNAs sequenced to date, the ongoing exploration of the nuclear genomes of these species will likely provide surprises, including clues about the extent and nature of α-proteobacterial genes that were transferred into the eukaryotic nucleus. In order to fully understand the origin and

history of all other eukaryotic groups it is as necessary to study the genomes of the most early diverging and most primitive members of these groups. Species that might provide the most information include choanoflagellates and sponges for the animal lineage, chytrids for fungi, and charophyte and other primitive green algae for plants.

Furthermore, identification and study of minimally diverged, free-living, as well as intracellular/symbiotic, α-proteobacterial relatives of mitochondria will be important to shed light on the reduction process underlying the transition from the eubacterial to the proto-mitochondrial genome. *Rickettsia* are one such group of bacteria since they share with mitochondria many 'stunning examples of highly derived genomes'. While examples of less-derived, more mitochondria-like obligatory intracellular symbionts are still lacking, a number of genome sequences of free-living α-Proteobacteria are currently becoming available, including species that undergo facultative, close associations with eukaryotic host cells. Likewise, tracking the evolution of mitochondrial genes continues to be an effective means of tracking the evolution of the eukaryotic cell as a whole. Fortunately, examples of established lateral gene transfer of mtDNA-encoded genes are rare, with the notable exception of exchanges of gene material among flowering plants. Thus, although many mtDNA genomes have been currently sequenced, determination of a wider variety of protist mtDNA sequences and refinement of outgroup data should allow for the rigorous reconstruction of a eukaryotic phylogeny.

Analysis of mitochondrial genome sequences has revealed that these organelles represent a microcosm of *nature's most advanced evolutionary laboratories*, confronting scientists with many examples of novel genetic mechanisms to discover. In many instances, principles first discovered in mitochondria have subsequently been recognized in bacterial or nuclear genomes (e.g. a variety of RNA editing mechanisms, autocatalytic intron RNAs, genes-in-pieces, trans-splicing, a number of deviations from the 'standard' translation code, quartet translation initiation, etc.). Thus, an understanding of mitochondrial systems has implications far beyond organelle biology.

It appears timely to begin exploring mitochondrial transcriptomes in much the same way as comparisons have been performed at the mtDNA level. This approach will reveal: whether besides informational and structural, regulatory RNA species are also synthesized; whether genetic information is more widely altered by co- or post-transcriptional processes such as RNA editing; and the activity, or lack thereof, of identified genes (see e.g. the examples of gene transfer from plant mtDNAs to the nucleus, and the exchange of mtDNAs among plants, where numerous silent genes have been identified). The last case is of more far-reaching importance, since even identification of a gene plus its transcription does not imply that a functional product of the gene exists. As proteomics technologies become more sensitive and accessible, the demonstration of gene activities and molecular interactions also at the protein level should become standard.

Acknowledgements

This work was supported by funds from the 'Canadian Institute of Health Research' (CIHR), Genome Quebec, and Genome Canada. B. Franz Lang is Imasco Fellow in the program of Evolutionary Biology of the Canadian Institute for Advanced Research (CIAR), whom we thank for salary and interaction support.

References

Adams KL, Daley DO, Qiu YL, Whelan J, Palmer JD (2000) Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. Nature 408:354-357

Adams KL, Daley DO, Whelan J, Palmer JD (2002a) Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts. Plant Cell 14:931-943

Adams KL, Ong HC, Palmer JD (2001a) Mitochondrial gene transfer in pieces: fission of the ribosomal protein gene *rpl2* and partial or complete gene transfer to the nucleus. Mol Biol Evol 18:2289-2297

Adams KL, Qiu YL, Stoutemyer M, Palmer JD (2002b) Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci USA 99:9905-9912

Adams KL, Rosenblueth M, Qiu YL, Palmer JD (2001b) Multiple losses and transfers to the nucleus of two mitochondrial succinate dehydrogenase genes during angiosperm evolution. Genetics 158:1289-1300

Adams KL, Song K, Roessler PG, Nugent JM, Doyle JL, Doyle JJ, Palmer JD (1999) Intracellular gene transfer in action: dual transcription and multiple silencings of nuclear and mitochondrial *cox2* genes in legumes. Proc Natl Acad Sci USA 96:13863-13868

Akhmanova A, Voncken F, van Alen T, van Hoek A, Boxma B, Vogels G, Veenhuis M, Hackstein JH (1998) A hydrogenosome with a genome. Nature 396:527-528

Altmann R (1890) Die Elementarorganismen und ihre Beziehungen zu den Zellen. Viet, Leipzig

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133-140

Backert S, Lurz R, Oyarzabal OA, Borner T (1997) High content, size and distribution of single-stranded DNA in the mitochondria of *Chenopodium album* (L). Plant Mol Biol 33:1037-1050

Baker A, Schatz G (1987) Sequences from a prokaryotic genome or the mouse dihydrofolate reductase gene can restore the import of a truncated precursor protein into yeast mitochondria. Proc Natl Acad Sci USA 84:3117-3121

Baldauf SL (2003) The deep roots of eukaryotes. Science 300:1703-1706

Bendich AJ (1993) Reaching for the ring: the study of mitochondrial genome structure. Curr Genet 24:279-290

Bendich AJ (1996) Structural analysis of mitochondrial DNA molecules from fungi and plants using moving pictures and pulsed-field gel electrophoresis. J Mol Biol 255:564-588

Bendich AJ (2001) The form of chromosomal DNA molecules in bacterial cells. Biochimie 83:177-186

Bergthorsson U, Adams KL, Thomason B, Palmer JD (2003) Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 424:197-201

Boer PH, Gray MW (1988) Scrambled ribosomal RNA gene pieces in *Chlamydomonas reinhardtii* mitochondrial DNA. Cell 55:399-411

Bonen L (1993) Trans-splicing of pre-mRNA in plants, animals, and protists. Faseb J 7:40-46

Boore JL (1999) Animal mitochondrial genomes. Nucleic Acids Res 27:1767-1780

Boore JL, Brown WM (1998) Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Curr Opin Genet Dev 8:668-674

Boore JL, Collins TM, Stanton D, Daehler LL, Brown WM (1995) Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature 376:163-165

Boyce TM, Zwick ME, Aquadro CF (1989) Mitochondrial DNA in the bark weevils: size, structure and heteroplasmy. Genetics 123:825-836

Brennicke A, Grohmann L, Hiesel R, Knoop V, Schuster W (1993) The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. FEBS Lett 325:140-145

Bridge D, Cunningham CW, Schierwater B, DeSalle R, Buss LW (1992) Class-level relationships in the phylum *Cnidaria*: evidence from mitochondrial genome structure. Proc Natl Acad Sci USA 89:8750-8753

Bui ET, Bradley PJ, Johnson PJ (1996) A common evolutionary origin for mitochondria and hydrogenosomes. Proc Natl Acad Sci USA 93:9651-9656

Bullerwell CE, Burger G, Lang BF (2000) A novel motif for identifying *rps3* homologs in fungal mitochondrial genomes. Trends Biochem Sci 25:363-365

Bullerwell CE, Forget L, Lang BF (2003a) Evolution of monoblepharidalean fungi based on complete mitochondrial genome sequences. Nucleic Acids Res 31:1614-1623

Bullerwell CE, Leigh J, Forget L, Lang BF (2003b) A comparison of three fission yeast mitochondrial genomes. Nucleic Acids Res 31:759-768

Bullerwell CE, Leigh J, Seif E, Longcore JE, Lang BF (2003c) Evolution of the fungi and their mitochondrial genomes. In: Arora D, Khachatourians GG (eds) Applied Mycology and Biotechnology. Elsevier Science, Amsterdam, pp 133-160

Burger G, Forget L, Zhu Y, Gray MW, Lang BF (2003a) Unique mitochondrial genome architecture in unicellular relatives of animals. Proc Natl Acad Sci USA 100:892-897

Burger G, Gray MW, Lang BF (2003b) Mitochondrial genomes - anything goes. Trends Genet 19:709-716

Burger G, Lang BF (2003) Parallels in genome evolution in mitochondria and bacterial symbionts. IUBMB Life 55:205-212

Burger G, Plante I, Lonergan KM, Gray MW (1995) The mitochondrial DNA of the amoeboid protozoon, *Acanthamoeba castellanii*: complete sequence, gene content and genome organization. J Mol Biol 245:522-537

Burger G, Saint-Louis D, Gray MW, Lang BF (1999) Complete sequence of the mitochondrial DNA of the red alga *Porphyra purpurea*. Cyanobacterial introns and shared ancestry of red and green algae. Plant Cell 11:1675-1694

Burger G, Zhu Y, Littlejohn TG, Greenwood SJ, Schnare MN, Lang BF, Gray MW (2000) Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison with *Paramecium aurelia* mitochondrial DNA. J Mol Biol 297:365-380

Cavalier-Smith T (1983) Members of the Archezoa include the Microsporidia, Metamonada, and Parabasalia. In: Schwemmler W, Schenk HEA (eds) Endocytobiology II. De Gruyter, Berlin, pp 1027-1034

Cavalier-Smith T (1987) The simultaneous symbiotic origin of mitochondria, chloroplasts, and microbodies. Ann N Y Acad Sci 503:55-71

Cavalier-Smith T (1993) Kingdom protozoa and its 18 phyla. Microbiol Rev 57:953-994

Chapdelaine Y, Bonen L (1991) The wheat mitochondrial gene for subunit I of the NADH dehydrogenase complex: a trans-splicing model for this gene-in-pieces. Cell 65:465-472

Chesnick JM, Goff M, Graham J, Ocampo C, Lang BF, Seif E, Burger G (2000) The mitochondrial genome of the stramenopile alga *Chrysodidymus synuroideus*. Complete sequence, gene content and genome organization. Nucleic Acids Res 28:2512-2518

Coleman AW, Thompson WF, Goff LJ (1991) Identification of the mitochondrial genome in the chrysophyte alga *Ochromonas danica*. J Protozool 38:129-135

Court DA, Bertrand H (1993) Expression of the open reading frames of a senescence-inducing, linear mitochondrial plasmid of *Neurospora crassa*. Plasmid 30:51-66

Cummings DJ, McNally KL, Domenico JM, Matsuura ET (1990) The complete DNA sequence of the mitochondrial genome of *Podospora anserina*. Curr Genet 17:375-402

Cummings MP, Nugent JM, Olmstead RG, Palmer JD (2003) Phylogenetic analysis reveals five independent transfers of the chloroplast gene *rbcL* to the mitochondrial genome in angiosperms. Curr Genet 43:131-138

de Duve C (1969) Evolution of the peroxisome. Ann N Y Acad Sci 168:369-381

de Duve C (1982) Peroxisomes and related particles in historical perspective. Ann N Y Acad Sci 368:1-4

de Duve C (1996) The birth of complex cells. Sci Am 274:38-45

Dietrich A, Small I, Cosset A, Weil JH, Marechal-Drouard L (1996) Editing and import: strategies for providing plant mitochondria with a complete set of functional transfer RNAs. Biochimie 78:518-529

Edqvist J, Burger G, Gray MW (2000) Expression of mitochondrial protein-coding genes in *Tetrahymena pyriformis*. J Mol Biol 297:381-393

Embley TM, Hirt RP (1998) Early branching eukaryotes? Curr Opin Genet Dev 8:624-629

Embley TM, van der Giezen M, Horner DS, Dyal PL, Foster P (2003) Mitochondria and hydrogenosomes are two forms of the same fundamental organelle. Philos Trans R Soc Lond B Biol Sci 358:191-201

Fan J, Lee RW (2002) Mitochondrial genome of the colorless green alga *Polytomella parva*: two linear DNA molecules with homologous inverted repeat Termini. Mol Biol Evol 19:999-1007

Feagin JE, Mericle BL, Werner E, Morris M (1997) Identification of additional rRNA fragments encoded by the *Plasmodium falciparum* 6 kb element. Nucleic Acids Res 25:438-446

Feagin JE, Werner E, Gardner MJ, Williamson DH, Wilson RJ (1992) Homologies between the contiguous and fragmented rRNAs of the two Plasmodium falciparum extrachromosomal DNAs are limited to core sequences. Nucleic Acids Res 20:879-887

Feng DF, Cho G, Doolittle RF (1997) Determining divergence times with a protein clock: update and reevaluation. Proc Natl Acad Sci USA 94:13028-13033

Forget L, Ustinova J, Wang Z, Huss VA, Lang BF (2002) *Hyaloraphidium curvatum:* a linear mitochondrial genome, tRNA editing, and an evolutionary link to lower fungi. Mol Biol Evol 19:310-319

Forsberg J, Rosenquist M, Fraysse L, Allen JF (2001) Redox signalling in chloroplasts and mitochondria: genomic and biochemical evidence for two-component regulatory systems in bioenergetic organelles. Biochem Soc Trans 29:403-407

Friedrich T, Steinmuller K, Weiss H (1995) The proton-pumping respiratory complex I of bacteria and mitochondria and its homologue in chloroplasts. FEBS Lett 367:107-111

Friedrich T, Weiss H (1997) Modular evolution of the respiratory NADH:ubiquinone oxidoreductase and the origin of its modules. J Theor Biol 187:529-540

Fukuhara H, Sor F, Drissi R, Dinouel N, Miyakawa I, Rousset S, Viola AM (1993) Linear mitochondrial DNAs of yeasts: frequency of occurrence and general features. Mol Cell Biol 13:2309-2314

Germot A, Philippe H, Le Guyader H (1996) Presence of a mitochondrial-type 70-kDa heat shock protein in *Trichomonas vaginalis* suggests a very early mitochondrial endosymbiosis in eukaryotes. Proc Natl Acad Sci USA 93:14614-14617

Germot A, Philippe H, Le Guyader H (1997) Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in *Nosema locustae*. Mol Biochem Parasitol 87:159-168

Giege P, Brennicke A (2001) From gene to protein in higher plant mitochondria. C R Acad Sci III 324:209-217

Gillespie DE, Salazar NA, Rehkopf DH, Feagin JE (1999) The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum* have short A tails. Nucleic Acids Res 27:2416-2422

Gillham NW (1974) Genetic analysis of the chloroplast and mitochondrial genomes. Annu Rev Genet 8:347-391

Gilson P, Waller R, McFadden G (1995) Preliminary characterisation of chlorarachniophyte mitochondrial DNA. J Eukaryot Microbiol 42:696-701

Glover KE, Spencer DF, Gray MW (2001) Identification and structural characterization of nucleus-encoded transfer RNAs imported into wheat mitochondria. J Biol Chem 276:639-648

Golding GB, Gupta RS (1995) Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol Biol Evol 12:1-6

Grace KS, Allen JO, Newton KJ (1994) R-type plasmids in mitochondria from a single source of *Zea luxurians teosinte*. Curr Genet 25:258-264

Gray M, Spencer D (1996a) Organellar evolution. In: Collins M (ed) Evolution of Microbial Life. Cambridge University Press, pp 109-126

Gray MW (1989) The evolutionary origins of organelles. Trends Genet 5:294-299

Gray MW (1992) The endosymbiont hypothesis revisited. Int Rev Cytol 141:233-357

Gray MW (1993) Origin and evolution of organelle genomes. Curr Opin Genet Dev 3:884-890

Gray MW, Boer PH (1988) Organization and expression of algal (*Chlamydomonas reinhardtii*) mitochondrial DNA. Philos Trans R Soc Lond B Biol Sci 319:135-147

Gray MW, Burger G, Lang BF (1999) Mitochondrial evolution. Science 283:1476-1481

Gray MW, Burger G, Lang BF (2001) The origin and early evolution of mitochondria. Genome Biol 2:Reviews1018

Gray MW, Lang BF, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Brossard N, Delage E, Littlejohn TG, Plante I, Rioux P, Saint-Louis D, Zhu Y, Burger

G (1998) Genome structure and gene content in protist mitochondrial DNAs. Nucleic Acids Res 26:865-878

Gray MW, Spencer DF (1996b) Organellar evolution. In: Collins M (ed) Evolution of Microbial Life. Cambridge University Press, pp 109-126

Hashimoto T, Sanchez LB, Shirakura T, Müller M, Hasegawa M (1998) Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. Proc Natl Acad Sci USA 95:6860-6865

Hauser G (2003) Fungal mitochondrial genomes, plasmids, and introns. In: Arora D, Khachatourians GG (eds) Applied Mycology and Biotechnology. Elsevier Science, Amsterdam, in press

Heinonen TY, Schnare MN, Young PG, Gray MW (1987) Rearranged coding segments, separated by a transfer RNA gene, specify the two parts of a discontinuous large subunit ribosomal RNA in *Tetrahymena pyriformis* mitochondria. J Biol Chem 262:2879-2887

Henze K, Martin W (2001) How do mitochondrial genes get into the nucleus? Trends Genet 17:383-387

Hermanns J, Osiewacz HD (1994) Three mitochondrial unassigned open reading frames of *Podospora anserina* represent remnants of a viral-type RNA polymerase gene. Curr Genet 25:150-157

Hirt RP, Healy B, Vossbrinck CR, Canning EU, Embley TM (1997) A mitochondrial Hsp70 orthologue in *Vairimorpha necatrix*: molecular evidence that microsporidia once contained mitochondria. Curr Biol 7:995-998

Horner DS, Hirt RP, Kilvington S, Lloyd D, Embley TM (1996) Molecular data suggest an early acquisition of the mitochondrion endosymbiont. Proc R Soc Lond B Biol Sci 263:1053-1059

Jacob Y, Seif E, Paquet P-O, Lang BF (2004) Loss of the mRNA-like region in the mitochondrial tmRNAs of jakobids. RNA: in press

Karlberg O, Canback B, Kurland CG, Andersson SG (2000) The dual origin of the yeast mitochondrial proteome. Yeast 17:170-187

Keeling PJ, Doolittle WF (1997) Evidence that eukaryotic triosephosphate isomerase is of alpha-proteobacterial origin. Proc Natl Acad Sci USA 94:1270-1275

Keiler KC, Shapiro L, Williams KP (2000) tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: A two-piece tmRNA functions in *Caulobacter*. Proc Natl Acad Sci USA 97:7778-7783

Kennell JC, Cohen SM (2003) Fungal mitochondria: genomes, genetic elements, and gene expression. In: Arora D (ed) Handbook of Fungal Biotechnology, 2nd edn. Marcel Dekker Inc., New York, in press

Knoop V, Altwasser M, Brennicke A (1997) A tripartite group II intron in mitochondria of an angiosperm plant. Mol Gen Genet 255:269-276

Korab-Laskowska M, Rioux P, Brossard N, Littlejohn TG, Gray MW, Lang BF, Burger G (1998) The organelle genome database project (GOBASE). Nucleic Acids Res 26:138-144

Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. J Mol Evol 52:540-542

Kovac L, Lazowska J, Slonimski PP (1984) A yeast with linear molecules of mitochondrial DNA. Mol Gen Genet 197:420-424

Lang BF (2003) The fungal mitochondrial genome project. URL - http://megasun.bch.umontreal.ca/People/lang/FMGP/

Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Gray MW (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387:493-497

Lang BF, Gray MW, Burger G (1999) Mitochondrial genome evolution and the origin of eukaryotes. Annu Rev Genet 33:351-397

Lang BF, O'Kelly C, Nerad T, Gray MW, Burger G (2002) The closest unicellular relatives of animals. Curr Biol 12:1773-1778

Lavrov DV, Boore JL, Brown WM (2002) Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: duplication and nonrandom loss. Mol Biol Evol 19:163-169

Leblanc C, Boyen C, Richard O, Bonnard G, Grienenberger JM, Kloareg B (1995) Complete sequence of the mitochondrial DNA of the rhodophyte *Chondrus crispus* (Gigartinales). Gene content and genome organization. J Mol Biol 250:484-495

Lecrenier N, Foury F (2000) New features of mitochondrial DNA replication system in yeast and man. Gene 246:37-48

Leigh J, Seif E, Rodriguez N, Jacob Y, Lang BF (2003) Fungal evolution meets fungal genomics. In: Arora D (ed) Handbook of Fungal Biotechnology, 2. edn. Marcel Dekker Inc., New York, pp 145-161

Leipe DD, Gunderson JH, Nerad TA, Sogin ML (1993) Small subunit ribosomal RNA+ of *Hexamita inflata* and the quest for the first branch in the eukaryotic tree. Mol Biochem Parasitol 59:41-48

Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, Feldblyum TV, Buell CR, Ketchum KA, Lee J, Ronning CM, Koo HL, Moffat KS, Cronin LA, Shen M, Pai G, Van Aken S, Umayam L, Tallon LJ, Gill JE, Venter JC, et al. (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402:761-768

Lonergan KM, Gray MW (1996) Expression of a continuous open reading frame encoding subunits 1 and 2 of cytochrome c oxidase in the mitochondrial DNA of *Acanthamoeba castellanii*. J Mol Biol 257:1019-1030

Mai Z, Ghosh S, Frisardi M, Rosenthal B, Rogers R, Samuelson J (1999) Hsp60 is targeted to a cryptic mitochondrion-derived organelle ("crypton") in the microaerophilic protozoan parasite *Entamoeba histolytica*. Mol Cell Biol 19:2198-2205

Malek O, Knoop V (1998) Trans-splicing group II introns in plant mitochondria: the complete set of cis-arranged homologs in ferns, fern allies, and a hornwort. RNA 4:1599-1609

Margulis L (1975) Symbiotic theory of the origin of eukaryotic organelles; criteria for proof. Symp Soc Exp Biol:21-38

Margulis L (1981) Symbiosis in Cell Evolution. Freeman, San Francisco

Margulis L (1988) Serial endosymbiotic theory (SET): undulipodia, mitosis and their microtubule systems preceded mitochondria. Endocyt Cell Res 5:133-162

Markos A, Miretsky A, Müller M (1993) A glyceraldehyde-3-phosphate dehydrogenase with eubacterial features in the amitochondriate eukaryote, *Trichomonas vaginalis*. J Mol Evol 37:631-643

Martin FN (1995) Linear mitochondrial genome organization in vivo in the genus *Pythium*. Curr Genet 28:225-234

Martin W, Müller M (1998) The hydrogen hypothesis for the first eukaryote. Nature 392:37-41

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ, Jermiin LS, Wolfe KH (2001) Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13:645-658

Morawala-Patell V, Gualberto JM, Lamattina L, Grienenberger JM, Bonnard G (1998) Cis- and trans-splicing and RNA editing are required for the expression of *nad2* in wheat mitochondria. Mol Gen Genet 258:503-511

Moreira D, Lopez-Garcia P (1998) Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. J Mol Evol 47:517-530

Nedelcu AM, Lee RW, Lemieux C, Gray MW, Burger G (2000) The complete mitochondrial DNA sequence of *Scenedesmus obliquus* reflects an intermediate stage in the evolution of the green algal mitochondrial genome. Genome Res 10:819-831

Nosek J, Dinouel N, Kovac L, Fukuhara H (1995) Linear mitochondrial DNAs from yeasts: telomeres with large tandem repetitions. Mol Gen Genet 247:61-72

Nosek J, Tomaska L (2003a) Mitochondrial genome diversity: evolution of the molecular architecture and replication strategy. Curr Genet 44:73-84

Nosek J, Tomaska L (2003b) Mitochondrial telomeres: alternative solutions to the end-replication problem. In: Parwaresch GKaR (ed) Telomeres and Telomerases: Cancer and Biology, in press

O'Brien EA, Badidi E, Barbasiewicz A, deSousa C, Lang BF, Burger G (2003) GOBASE--a database of mitochondrial and chloroplast information. Nucleic Acids Res 31:176-178

Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, Akashi K, Kanegae T, Ogura Y, Kohchi T, et al. (1992) Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA. A primitive form of plant mitochondrial genome. J Mol Biol 223:1-7

Ogata H, Audic S, Barbe V, Artiguenave F, Fournier PE, Raoult D, Claverie JM (2000) Selfish DNA in protein-coding genes of *Rickettsia*. Science 290:347-350

Ogata H, Audic S, Renesto-Audiffren P, Fournier PE, Barbe V, Samson D, Roux V, Cossart P, Weissenbach J, Claverie JM, Raoult D (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 293:2093-2098

Ogawa S, Yoshino R, Angata K, Iwamoto M, Pi M, Kuroe K, Matsuo K, Morio T, Urushihara H, Yanagisawa K, Tanaka Y (2000) The mitochondrial DNA of *Dictyostelium discoideum*: complete sequence, gene content and genome organization. Mol Gen Genet 263:514-519

Okimoto R, Macfarlane JL, Clary DO, Wolstenholme DR (1992) The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. Genetics 130:471-498

Oudot-Le Secq MP, Fontaine JM, Rousvoal S, Kloareg B, Loiseaux-De Goer S (2001) The complete sequence of a brown algal mitochondrial genome, the ectocarpale *Pylaiella littoralis* (L) Kjellm. J Mol Evol 53:80-88

Palmer JD (1997) Genome evolution. The mitochondrion that time forgot. Nature 387:454-455

Palmer JD, Soltis D, Soltis P (1992) Large size and complex structure of mitochondrial DNA in two nonflowering land plants. Curr Genet 21:125-129

Paquin B, Laforest MJ, Forget L, Roewer I, Wang Z, Longcore J, Lang BF (1997) The fungal mitochondrial genome project: evolution of fungal mitochondrial genomes and their gene expression. Curr Genet 31:380-395

Paquin B, Lang BF (1996) The mitochondrial DNA of *Allomyces macrogynus*: the complete genomic sequence from an ancestral fungus. J Mol Biol 255:688-701

Perez-Martinez X, Antaramian A, Vazquez-Acevedo M, Funes S, Tolkunova E, d'Alayer J, Claros MG, Davidson E, King MP, Gonzalez-Halphen D (2001) Subunit II of cytochrome c oxidase in chlamydomonad algae is a heterodimer encoded by two independent nuclear genes. J Biol Chem 276:11302-11309

Philippe H, Forterre P (1999) The rooting of the universal tree of life is not reliable. J Mol Evol 49:509-523

Pont-Kingdon G, Okada NA, Macfarlane JL, Beagley CT, Watkins-Sims CD, Cavalier-Smith T, Clark-Walker GD, Wolstenholme DR (1998) Mitochondrial DNA of the coral *Sarcophyton glaucum* contains a gene for a homologue of bacterial *MutS*: a possible case of gene transfer from the nucleus to the mitochondrion. J Mol Evol 46:419-431

Pont-Kingdon GA, Okada NA, Macfarlane JL, Beagley CT, Wolstenholme DR, Cavalier-Smith T, Clark-Walker GD (1995) A coral mitochondrial *mutS* gene. Nature 375:109-111

Pritchard AE, Seilhamer JJ, Mahalingam R, Sable CL, Venuti SE, Cummings DJ (1990) Nucleotide sequence of the mitochondrial genome of *Paramecium*. Nucleic Acids Res 18:173-180

Rasmusson AG, Svensson AS, Knoop V, Grohmann L, Brennicke A (1999) Homologues of yeast and bacterial rotenone-insensitive NADH dehydrogenases in higher eukaryotes: two enzymes are present in potato mitochondria. Plant J 20:79-87

Rigaa A, Monnerot M, Sellos D (1995) Molecular cloning and complete nucleotide sequence of the repeated unit and flanking gene of the scallop *Pecten maximus* mitochondrial DNA: putative replication origin features. J Mol Evol 41:189-195

Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci USA 95:6239-6244

Roger AJ, Silberman JD (2002) Cell evolution: mitochondria in hiding. Nature 418:827-829

Roger AJ, Svard SG, Tovar J, Clark CG, Smith MW, Gillin FD, Sogin ML (1998) A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. Proc Natl Acad Sci USA 95:229-234

Sankoff D, Bryant D, Deneault M, Lang BF, Burger G (2000) Early eukaryote evolution based on mitochondrial gene order breakpoints. J Comput Biol 7:521-535

Schardl CL, Lonsdale DM, Pring DR, Rose KR (1984) Linearization of maize mitochondrial chromosomes by recombination with linear episomes. Nature (London) 310:292-296

Schimper AFW (1883) Über die Entwicklung der Chlorophyllkörper und Farbkörper. Bot Z 41:105-114

Schnare MN, Heinonen TY, Young PG, Gray MW (1986) A discontinuous small subunit ribosomal RNA in *Tetrahymena pyriformis* mitochondria. J Biol Chem 261:5187-5193

Seif ER, Forget L, Martin NC, Lang BF (2003) Mitochondrial RNase P RNAs in ascomycete fungi: Lineage-specific variations in RNA secondary structure. RNA 9:1073-1083

Sekito T, Okamoto K, Kitano H, Yoshida K (1995) The complete mitochondrial DNA sequence of *Hansenula wingei* reveals new characteristics of yeast mitochondria. Curr Genet 28:39-53

Sogin ML (1991) Early evolution and the origin of eukaryotes. Curr Opin Genet Dev 1:457-463

Sogin ML, Gunderson JH, Elwood HJ, Alonso RA, Peattie DA (1989) Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia.* Science 243:75-77

Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402-404

Steinberg S, Leclerc F, Cedergren R (1997) Structural rules and conformational compensations in the tRNA L-form. J Mol Biol 266:269-282

Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, Jiang J (2001) Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: implication of potential sequencing errors caused by large-unit repeats. Proc Natl Acad Sci USA 98:5099-5103

Suyama Y, Fukuhara H, Sor F (1985) A fine restriction map of the linear mitochondrial DNA of *Tetrahymena pyriformis*: genome size, map locations of rRNA and tRNA genes, terminal inversion repeat, and restriction site polymorphism. Curr Genet 9:479-493

Takano H, Kawano S, Kuroiwa T (1994) Complex terminal structure of a linear mitochondrial plasmid from *Physarum polycephalum*: three terminal inverted repeats and an ORF encoding DNA polymerase. Curr Genet 25:252-257

Taylor FJR (1974) II. Implications and extensions of the serial endosymbiosis theory of the origin of eukaryotes. Taxon J1 - t 23:229-258

Tovar J, Fischer A, Clark CG (1999) The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite *Entamoeba histolytica.* Mol Microbiol 32:1013-1021

Tovar J, Leon-Avila G, Sanchez LB, Sutak R, Tachezy J, Van Der Giezen M, Hernandez M, Muller M, Lucocq JM (2003) Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation. Nature 426:172-176

Turmel M, Lemieux C, Burger G, Lang BF, Otis C, Plante I, Gray MW (1999) The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor.* Two radically different evolutionary patterns within green algae. Plant Cell 11:1717-1730

Turmel M, Otis C, Lemieux C (2002a) The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc Natl Acad Sci USA 99:11275-11280

Turmel M, Otis C, Lemieux C (2002b) The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. Mol Biol Evol 19:24-38

Turmel M, Otis C, Lemieux C (2003) The mitochondrial genome of *Chara vulgaris*: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. Plant Cell 15:1888-1903

Vahrenholz C, Riemen G, Pratje E, Dujon B, Michaelis G (1993) Mitochondrial DNA of *Chlamydomonas reinhardtii*: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication. Curr Genet 24:241-247

von Heijne G (1986) Why mitochondria need a genome. FEBS Lett 198:1-4

Vossbrinck CR, Maddox JV, Friedman S, Debrunner-Vossbrinck BA, Woese CR (1987) Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. Nature 326:411-414

Wahleithner JA, Wolstenholme DR (1988) Origin and direction of replication in mitochondrial plasmid DNAs of broad bean, *Vicia faba*. Curr Genet 14:163-170

Ward BL, Anderson RS, Bendich AJ (1981) The mitochondrial genome is large and variable in a family of plants (cucurbitaceae). Cell 25:793-803

Williams BA, Hirt RP, Lucocq JM, Embley TM (2002) A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. Nature 418:865-869

Wilson RJ, Williamson DH (1997) Extrachromosomal DNA in the Apicomplexa. Microbiol Mol Biol Rev 61:1-16

Wolff G, Plante I, Lang BF, Kück U, Burger G (1994) Complete sequence of the mitochondrial DNA of the chlorophyte alga *Prototheca wickerhamii*. Gene content and genome organization. J Mol Biol 237:75-86

Wolstenholme DR (1992) Animal mitochondrial DNA: structure and evolution. Int Rev Cytol 141:173-216

Wolstenholme DR, Macfarlane JL, Okimoto R, Clary DO, Wahleithner JA (1987) Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms. Proc Natl Acad Sci USA 84:1324-1328

Wu LF, Ize B, Chanal A, Quentin Y, Fichant G (2000) Bacterial twin-arginine signal peptide-dependent protein translocation pathway: evolution and mechanism. J Mol Microbiol Biotechnol 2:179-189

Yagi T (1991) Bacterial NADH-quinone oxidoreductases. J Bioenerg Biomembr 23:211-225

Zillig W, Klenk H-P, Palm P, Leffers H, Pühler G, Gropp F, Garrett RA (1989) Did eukaryotes originate by a fusion event? Endocyt Cell Res 6:1-25

Glossary

Archaebacteria (Archea): one of the three domains of life besides eukaryotes and eubacteria.

Basally/early diverging group: lineage that emerged early in the evolution of a clade (i.e. a deep branch in a tree).

Character: any heritable attribute of organisms that varies among species, and that can be used for phylogenetic inference.

Derived: taxa or characters that have evolved far away from the primitive (original, ancient, ancestral) state. In molecular phylogeny, short tree branches indicate little derived taxa.

Homology: similarity due to common evolutionary origin, i.e. derivation from the same ancestral character.

Monophyletic group: a monophyletic group (clade, lineage) has a unique origin in a single ancestral species, and includes the ancestor and all of its descendants.

Outgroup: taxon (taxa) used to root a phylogenetic tree, and that is (are) considered to lie outside of the group of interest.

Phylogeny: evolutionary relationships among organisms. Molecular phylogeny is based on DNA and protein sequences (or other molecular characters).

Proteobacteria: group of Gram-negative Eubacteria that are subdivided into α,β,γ.

RNA Editing: the programmed modification of primary transcript sequence, including substitutions, as well as insertions and deletions.

Synapomorphy: a derived character or character state, shared among a group of species to the exclusion of others.

Tree: line graph representing the evolutionary history of a set of taxa, connecting contemporary taxa *via* internal branches to hypothetical ancestors.

Bevan, Rachel B.
Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, Montréal Québec, H3T 1J4, Canada and McGill Centre for Bioinformatics, Duff Medical Building, 3775 University Street, Montreal, Quebec, H3A 2B4, Canada

Lang, B. Franz
Program in Evolutionary Biology, Canadian Institute for Advanced Research and Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, Montréal Québec, H3T 1J4, Canada
Franz.Lang@Umontreal.ca