The origin of mitochondria in light of a fluid prokaryotic chromosome model

Christian Esser, William Martin and Tal Dagan*

1. INTRODUCTION

There is consensus among biologists that mitochondria descend from free-living prokaryotes and that the organelle arose only once during evolution (Gray et al. 1999; Dolezal et al. 2006). There is considerably less agreement concerning the biochemical capabilities and phylogenetic affinity of the mitochondrial ancestor. Various eubacterial groups have been proposed as the ancestor of mitochondria. Even before the time of endosymbiosis; Magnetospirillum magnetotacticum.

The conventional approach to identify the mitochondrial ancestor is founded in the comparison of mitochondrially encoded genes with those in the genomes of free-living prokaryotes. By this means, early analyses of 16S rRNA suggested Agrobacterium tumefaciens to be the closest relative of the mitochondrion (Yang et al. 1985). More recent studies have attributed the mitochondrial ancestor to the Rickettsiales order, which mostly contains parasitic species with highly reduced genomes (Lang et al. 1999; Emelyanov 2003). Other studies have pointed specifically to Rickettsia prowazekii as the ancestral genome (Andersson et al. 1998), or a common ancestor of Rickettsia and Wolbachia (Wu et al. 2004).

1975), later renamed as upon the purple non-sulphur bacteria (John & Whatley 1998). But there is no consensus as to what constitutes an \( \alpha \)-proteobacterial gene. Is it a gene found in all or several \( \alpha \)-proteobacteria, or in only one? Here, we examine the proportion of \( \alpha \)-proteobacterial genes in \( \alpha \)-proteobacterial genomes by means of sequence comparisons. We find that each \( \alpha \)-proteobacterium harbours a particular collection of genes and that, depending upon the lineage examined, between 97 and 33% are \( \alpha \)-proteobacterial by the nearest-neighbour criterion. Our findings bear upon attempts to reconstruct the mitochondrial ancestor and upon inferences concerning the collection of genes that the mitochondrial ancestor possessed at the time that it became an endosymbiont.

**Keywords:** genomics; hydrogenosomes; lateral gene transfer; microbial evolution; endosymbiosis; Magnetococcus

2. MATERIAL AND METHODS

(a) Data

Prokaryotic and mitochondrial genomes were downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/; versions of April 2005; table S1 in electronic supplementary material). All 288 prokaryotic genomes were formatted into a single Blast (Altschul et al. 1990) database. Eighteen \( \alpha \)-proteobacterial and six mitochondrial genomes were used as queries, including only proteins longer than 50 amino acids.

(b) Nearest-neighbour inference

The nearest neighbour of each protein was inferred by a best Blast hit (BBH) approach and a phylogenetic tree approach using the neighbour-joining (Saitou & Nei 1987) and maximum-likelihood methods. Neither approach is infallible (Koski & Golding 2001; Penny et al. 2001); we used both approaches for comparison.

For BBH analysis, each protein in each query genome was blasted against the 288 genome database. The nearest neighbour was defined as the BBH above an E-value of \( 1 \times 10^{-20} \) that is neither the query protein nor stems from the same genus as the query protein. The taxonomic group of the nearest neighbours was specified as the phylum according to the NCBI taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/), or as the class in the case of proteobacteria.

In the NJ approach, the BBHs from each genus were selected. These were aligned with CLUSTALW (Thompson et al. 1994), protein distances were calculated with PROTDIST (Felsenstein 2005) using the JTT matrix, and used to reconstruct an NJ tree with NEIGHBOUR (Felsenstein 2005) using 100 bootstrap replicates. Maximum-likelihood trees were reconstructed using FASTML (Pupko et al. 2000). The nearest neighbour was defined as the operational taxonomic unit with smallest sum of branch lengths to the query protein that appears in more than or equal to 90% of the replicates.

3. RESULTS

If every gene contained within an \( \alpha \)-proteobacterial genome were of an \( \alpha \)-proteobacterial origin (i.e. most closely related to homologues in other \( \alpha \)-proteobacteria), then every nearest neighbour of every gene in each \( \alpha \)-proteobacterial genome would be found in another \( \alpha \)-proteobacterium. Our results (figure 1) indicate that \( \alpha \)-proteobacterial genomes are mosaic to varying degrees.

The highest proportion of \( \alpha \)-proteobacterial BBH nearest neighbours (92%) was found in Sinorhizobium meliloti, while the lowest proportion (64%) was found in Magnetospirillum magnetotacticum. An even lower proportion of \( \alpha \)-proteobacterial BBH nearest neighbours...
The BBH nearest-neighbour analysis of the mitochondrial genes results in a phyletic distribution that is similar to that of the α-proteobacterial genes (figure 1d). The majority of the nearest neighbours are α-proteobacterial (82% ± 7%), and additional frequent taxa are γ-proteobacteria and firmicutes (table S2 of electronic supplementary material).

The vast majority (93% ± 4%) of the BBH nearest neighbours among proteins encoded within the 18 α-proteobacterial genomes sampled here reside within genomes of the proteobacterial phylum (figure 1). The α-proteobacterial genomes sampled encode many open reading frames, sequences that have no known homologues (17 953 in all α-proteobacteria), and whose history cannot presently be addressed by sequence comparisons. In the present study, the BBH approach detected, on average, no nearest neighbour for 27% of the proteins in each α-proteobacterial genome (figure S1 in electronic supplementary material). Our results were found to be independent of the tree size or the sampled species (figures S2–S4 in electronic supplementary material).

The distribution of NJ nearest neighbours of α-proteobacterial genes is similar to the results of the BBH method (figure 1c). The majority of the genes (87% ± 13%) have an α-proteobacterial nearest neighbour, while other frequent nearest neighbours are γ-proteobacterial and actinobacterial genes (table S3 of electronic supplementary material).

The BBH nearest-neighbour analysis of the mitochondrial genes results in a phyletic distribution that is similar to that of the α-proteobacterial genes (figure 1d). The majority of the nearest neighbours are α-proteobacterial (82% ± 7%), and additional frequent taxa are γ-proteobacteria and firmicutes (table S2 of electronic supplementary material).

The vast majority (93% ± 4%) of the BBH nearest neighbours among proteins encoded within the 18 α-proteobacterial genomes sampled here reside within genomes of the proteobacterial phylum (figure 1). The α-proteobacterial genomes sampled encode many open reading frames, sequences that have no known homologues (17 953 in all α-proteobacteria), and whose history cannot presently be addressed by sequence comparisons. In the present study, the BBH approach detected, on average, no nearest neighbour for 27% of the proteins in each α-proteobacterial genome (figure S1 in electronic supplementary material). Our results were found to be independent of the tree size or the sampled species (figures S2–S4 in electronic supplementary material).

Using the ML reconstruction, the proportion of α-proteobacterial nearest neighbours was lower by approximately 10% (figure S5 in electronic supplementary material).

Figure 1. Distribution across taxonomic groups for (a) α-proteobacterial nearest neighbours by BBH, (b) prokaryotic genes, (c) α-proteobacterial nearest neighbours by NJ and (d) mitochondrial nearest neighbours by BBH. The number of query proteins used for the analysis is shown to the left of the species name (y-axis).
4. DISCUSSION

On the basis of sequence similarity to α-proteobacterial homologues, it has been estimated that 630 eukaryotic genes trace to α-proteobacteria (Gabaldon & Huynen 2003). But there are thousands of eukaryotic nuclear genes that are clearly eubacterial, but not specifically α-proteobacterial, in terms of their patterns of sequence similarity (Esser et al. 2004; Rivera & Lake 2004; Embley & Martin 2006). Finding a eukaryotic gene that branches with a group other than α-proteobacteria is often taken as evidence for an origin from that group (for example, Baughn & Malamy 2002), the methodological problems of deep phylogenetic trees notwithstanding (Susko et al. 2006). Hence, we wished to know how many of the α-proteobacterial genes pass the test of being α-proteobacterial by the nearest-neighbour criterion.

The answer, based upon the current sample, ranges from approximately 97% for Sinorhizobium to...
approximately 33% for Magnetococcus sp. The mitochondrial genomes studied (figure 1d) did not differ in terms of the nearest-neighbour composition from α-proteobacterial genomes.

Prokaryotic gene content is shaped not only by inheritance, but also by gene loss and LGT (Doolittle 2004; Kunin et al. 2005; Lerat et al. 2005). But this realization is only slowly being assimilated into thinking on the mitochondrial origin and eukaryotic gene origins (Esser et al. 2004). Our findings indicate that modern α-proteobacterial genomes represent transient collections of genes that stem from diverse sources. By inference, the ancestor of mitochondria had a mosaic genome as well; hence, a criterion that is often used to infer whether a eukaryotic nuclear gene of eubacterial origin stems from the mitochondrion or not—namely branching with an α-proteobacterial gene (Kurland & Andersson 2000)—is probably too strict, because it tacitly assumes a static model of bacterial chromosome evolution in which LGT and gene loss do not exist, either now or in the past. Incorporating a fluid bacterial chromosome model into endosymbiotic theory generates the prediction that nuclear genes can accurately recover events that occurred more than 1.5 billion years ago (Embley & Martin 2006)—but that they should not necessarily belong to the known set of contemporary α-proteobacterial genes, regardless of how one were to define it.


