Evolution of b-type cytochromes in prokaryotes

Vassiliki Lila Koumandou¹, Sophia Kossida²

Supplementary figure legends:

Figure S1: Phylogenetic reconstruction of all cytochrome-b sequences used in this study, except K00241.

The tree shown is based on PhyML analysis of 472 sequences from all orthology groups (except K00241) and all species. Numerical values at the nodes of the tree indicate statistical support based on 100 bootstraps. For species name abbreviations, and full details of accession numbers for all protein sequences used, refer to Table S1. All sequences are marked with a letter from "A" to "L" at the end to distinguish different orthology groups (see Table 1), and are also color-coded accordingly, as shown in the insert. Most sequences from the "A" (red) and the "B" group (blue) cluster together; "C" group sequences (cyan) cluster with the divergent "B" sequences; "D" and "F" sequences each form a small tight cluster (green, and purple, respectively) with good bootstrap support for the clustering of these two groups; "E" sequences (orange) are split into two clusters, one containing bacterial sequences, and the other containing archaeal sequences which group with archaeal "I" sequences (khaki); all but 2 "G" sequences (brown) cluster together; all but 11 "H" sequences (magenta) end up in one of two clusters; hypothetical "I" (khaki) sequences are dispersed, although most group with hypothetical "E" and "J" sequences (as might be expected); all "J" sequences (light blue) end up in one of three clusters interspersed with "I" sequences, except for two which end up in the
"G" cluster; "K" sequences (yellow) are dispersed and quite divergent, while most "L" sequences (grey) cluster together.

**Figure S2: Phylogenetic reconstruction of cytochrome-b of the $b_{6f}$ and the bc complex, including sequences with a cytochrome-b domain which are not part of an orthology group.**

The tree shown is the best Bayesian topology based on 205 sequences total, including cytochrome $b_6$ of the cytochrome $b_{6f}$ complex fused to subunit IV (A: K02635/7), cytochrome-b of the bc complex (B: K00410/2, C: K03887/8, D: K03891), narC (F: K15879), as well as sequences with a cytochrome-b domain which are not part of an orthology group (E & I). Numerical values at the nodes of the tree indicate statistical support (posterior probability) based on 100 bootstraps. For species name abbreviations, and full details of accession numbers for all protein sequences used, refer to Table S1. All sequences are marked with a letter at the end to distinguish different orthology groups (see Table 1), and are also color-coded accordingly, as shown in the insert. Some of the most prominent groupings discussed in the text are annotated.

**Figure S3: Phylogenetic reconstruction of the cytochrome-b subunit of ubiquinol-cytochrome c reductase (bc complex), formate dehydrogenase, Ni/Fe dehydrogenase, and the cytochrome b556 subunit of succinate dehydrogenase.**

A: The tree shown is the best Bayesian topology of 205 sequences total, including cytochrome-b of the bc complex (B: K00410/2), formate dehydrogenase (H: K00127), and Ni/Fe hydrogenase (L: K03620). Numerical values at the nodes of the tree indicate statistical support (posterior probability) based on 100 bootstraps. All sequences are marked with a letter at the end to distinguish different orthology groups (see Table 1), and are also color-
coded accordingly, as shown in the insert. The tree is the same as that presented in Figure 2, shown in star format for easier comparison with the tree in panel B.

B: The tree shown is the best Bayesian topology of 360 sequences total, including all sequences as in panel A (marked and color-coded as above), as well as 155 sdhC sequences (K00241). The six groups of K00241 sequences are annotated I-VI as in Figure 3.

Figure S4: Multiple sequence alignment of eukaryotic and prokaryotic cytochrome b561.

The alignment, including 22 eukaryotic cytochrome b561 (K08360) and 61 prokaryotic cytochrome b561 (K12262) sequences (dataset as in Figure 4), was generated using Muscle, and visualized in Jalview. The four helices (pink) predicted by PSIpred for the E. coli sequence, along with the 6 helices based on the structure of the Arabidopsis protein, as reported in Figure S3 of (Lu et al., 2014), are shown above the alignment. Red squares, blue dots, green triangles and pink diamonds correspond to histidine residues coordinating the iron atoms, amino acids that play important roles in substrate recognition on the cytoplasmic side and the non-cytoplasmic side, and amino acids that use their side chains to form H-bonds with the heme groups, respectively, as in Figure S3 of (Lu et al., 2014). Magenta bars correspond to previously reported substrate-binding motifs, ALLVYR and SLHSW, again as annotated in Figure S3 of (Lu et al., 2014). Conserved residues are highlighted in the alignment (blue: >80% agreement, mid blue: >60% agreement, light blue: >40% agreement; only the residues that agree with the consensus residue for each column are colored). The conservation annotation histogram below the alignment reflects conservation of the physicochemical properties, and marks absolutely conserved residues (score 11) with a yellow asterisk '*', and columns where physicochemical properties are conserved (score 10) with a yellow '+'; less conserved positions are shown in darker colors with decreasing score. The quality annotation histogram reflects the likelihood of observing a mutation in any
particular column of the alignment based on the BLOSUM62 matrix scores (for each column,
the sum of the ratios of the two BLOSUM62 scores for a mutation pair, and each residue's
conserved BLOSUM62 score, are normalized and plotted on a scale of 0 to 1) The consensus
histogram reflects the percentage of the modal residue per column, and the consensus
sequence logo is shown below the sequence for conserved regions ('+' denotes non-conserved
residues and '-' denotes gap residues).

**Figure S5:** Phylogenetic reconstruction of eukaryotic and prokaryotic cytochrome b561,
including sequences containing a cytochrome-b domain but not belonging to an
orthology group.

The tree shown is based on PhyML analysis of 223 sequences total, including eukaryotic
cytochrome b561 (K08360), prokaryotic cytochrome b561 (K12262 and K15879) sequences,
prokaryotic sequences annotated as cyt-b561 but which are not part of an orthology group,
and prokaryotic sequences containing a cytochrome-b domain but not belonging to an
orthology group (see Table 1). K12262 sequences are marked with a "G" at the end, K15879
sequences with an "F", and the ones not belonging to an orthology group with an "E", "I" or
"J" (see Table 1 for details). Sequences are also color-coded as indicated in the insert (E:
orange, F: purple, G: brown, I: khaki, J: light blue). For species name abbreviations, and full
details of accession numbers for all protein sequences used, refer to Table S1.
A: K02635/7 cytochrome b6
B: K00410/2 ubiquinol cytochrome c reductase
C: K03887/8 menaquinol cytochrome c reductase
D: K03991 ubiquinol cytochrome c reductase
E: cytochrome-b-domain-containing proteins
F: K15879 cytochrome b-561
G: K12262 cytochrome b-561 of E. coli
H: K00127 formate dehydrogenase
I: cytochrome-b/b6-domain-containing proteins
J: cytochrome b-561-like
K: cytochrome b5
L: K03620 Ni/Fe hydrogenase
A: K02635/7 cytochrome b6
B: K00410/2 ubiquinol cytochrome c reductase
C: K03887/8 menaquinol cytochrome c reductase
D: K03891 ubiquinol cytochrome c reductase
E: cytochrome-b-domain-containing proteins
F: K15879 cytochrome b-561
I: cytochrome-b/b6-domain-containing proteins
B: K00410/2 ubiquinol cytochrome c reductase
H: K00127 formate dehydrogenase
L: K03620 Ni/Fe dehydrogenase
black: succinate dehydrogenase