

Sequence conservation from human to prokaryotes of Surf1, a protein involved in cytochrome *c* oxidase assembly, deficient in Leigh syndrome

Alain Poyau, Karine Buchet, Catherine Godinot*

Centre de Génétique Moléculaire et Cellulaire, Centre National de la Recherche Scientifique, Université Claude Bernard Lyon I, 69622 Villeurbanne, cedex, France

Received 12 October 1999

Edited by Vladimir Skulachev

Abstract The human *SURF1* gene encoding a protein involved in cytochrome *c* oxidase (COX) assembly, is mutated in most patients presenting Leigh syndrome associated with COX deficiency. Proteins homologous to the human Surf1 have been identified in nine eukaryotes and six prokaryotes using database alignment tools, structure prediction and/or cDNA sequencing. Their sequence comparison revealed a remarkable Surf1 conservation during evolution and put forward at least four highly conserved domains that should be essential for Surf1 function. In *Paracoccus denitrificans*, the Surf1 homologue is found in the quinol oxidase operon, suggesting that Surf1 is associated with a primitive quinol oxidase which belongs to the same superfamily as cytochrome oxidase.

© 1999 Federation of European Biochemical Societies.

Key words: SURF1 (SURF-1); Cytochrome *c* oxidase; Leigh syndrome; Sequence alignment; Quinol oxidase

1. Introduction

Cytochrome *c* oxidase (COX, EC 1.9.3.1) is the terminal oxidase in the eukaryotic respiratory chain. In mammals, it is a 13-subunit complex embedded in the mitochondrial inner membrane. The three subunits coded by mitochondrial DNA constitute the catalytic core and are responsible for electron transfer and proton pumping while the nuclear-coded subunits are likely to be involved in enzyme regulation (see [1] for review). The mechanisms of biogenesis of the complex are not yet fully understood [2]. However, several human proteins homologous to yeast proteins essential for the COX assembly have been identified. One of them, the human Surf1, is homologous to the yeast *Saccharomyces cerevisiae* Shy1, which is a mitochondrial membrane protein somehow involved in the transfer of electrons between ubiquinol-cytochrome *c* reductase and COX [3].

Recently, the *SURF1* gene was identified as responsible for most cases of Leigh syndrome with a specific cytochrome oxidase deficiency (LS-COX⁻) [4,5]. The Leigh syndrome (OMIM 256000) is a severe infantile neurodegenerative disorder characterised by a subacute necrotising encephalopathy (see [6] for review), very often associated with COX deficiency [7]. Analysis of the *SURF1* gene in LS-COX⁻ patients re-

vealed several mutations that are predicted to result in a truncated protein [4,5,8,9]. In addition, we have recently identified two missense mutations of *SURF1* in two independent LS-COX⁻ patients ([9], SWISS-PROT Q15526). In our patients [9], as in other similar patients [7], COX subunits are barely detectable while their transcripts are normally expressed. This suggests that COX assembly is deficient and, hence, that Surf1 interferes in this process although its role remains unknown.

The human *SURF1* gene contains nine exons that code a 300 aa protein. It belongs to the Surfeit locus mapped at chromosome band 9q34 [10], including six housekeeping genes (*SURF1* to 6) separated by small intergenic regions. For example, *SURF1* and *SURF2* are controlled by a bidirectional promoter and their transcription start sites are distant by less than 100 bp [11]. This locus organisation is conserved in mice and chickens [10] but is not present in invertebrates [12].

In this paper, we have studied the conservation of the Surf1 protein during evolution. We show that proteins homologous to human Surf1 can be found in at least eight other eukaryotes and six prokaryotes. Several extremely conserved domains can be put forward and permit the prediction of Surf1 regions that should be essential for the protein structure and/or function.

2. Materials and methods

2.1. 5'-RACE

Drosophila melanogaster RNA was kindly provided by Dr. B. Durand while *Rattus norvegicus* and *Arabidopsis thaliana* RNA were extracted from liver and leaves, respectively, with the 'Trizol Reagent' (Life Technologies) according to the manufacturer's protocol suggested for samples with high content of proteoglycans and/or polysaccharides. RNA (5 µg) was treated for 1 h with 3 U of DNase I (Life Technologies) that was then inactivated for 10 min at 65°C, in the presence of 2.5 mM EDTA. RNA (1 µg) was denatured for 10 min at 70°C in the presence of 10 pmol of reverse primer SP1 specific of each *SURF1* and chilled on ice. RNA was reverse transcribed at 42°C for 1 h with 200 U of Superscript II (Life Technologies) in reverse transcription buffer containing 0.5 mM of each dNTP, 10 mM dithiothreitol and 0.5 unit of RNasin. The reaction was stopped by incubation at 95°C for 5 min. *SURF1* cDNA was purified using the 'High pure PCR product purification kit' (Boehringer Mannheim), denatured for 3 min at 94°C and chilled on ice. A homopolymeric tail was appended to the cDNA 3'-end by incubation for 20 min at 37°C with 0.2 mM dATP and 10 U of terminal transferase (Boehringer Mannheim). The enzyme was inactivated for 10 min at 70°C. *SURF1* cDNA was amplified from tailed cDNA by a first PCR with the oligo dT-anchor primer (GACCACGCGTATCGATGTGCA-C(T)₁₆V) and a *SURF1* specific reverse primer SP2 in the presence of *Taq* buffer (50 mM KCl, 20 mM Tris-HCl pH 8.4), 0.3 µM of each primer, 200 µM of each dNTP, 1.5 mM MgCl₂ and 1.5 U of *Taq* DNA polymerase (Life Technologies). Thirty five cycles with 30 s at 94°C, 30 s at 55°C and 45 s at 72°C were carried out.

*Corresponding author. Fax: (33)-4-72 44 05 55.
E-mail: godinot@univ-lyon1.fr

Abbreviations: aa, amino acids; COX, cytochrome *c* oxidase; EST, expressed sequence tag; LS-COX⁻, Leigh syndrome associated with cytochrome *c* oxidase deficiency

2.2. cDNA sequencing

SURF1 cDNAs were re-amplified with two other successive PCR rounds from 1 µl of the previous PCR product, using the PCR anchor primer GACCACGCGTATCGATGTCGAC and nested reverse primers, firstly SP3 and secondly SP4. The final PCR products were purified on agarose gels with the 'Concert rapid gel extraction System' (Life Technologies). The DNA fragments (20 ng) were sequenced with the 'ABI Prism Dye Terminator Cycle Sequencing Ready Reaction Kit' (Perkin Elmer) on a 370A automated DNA sequencer (Applied Biosystems).

2.3. Computer analysis

Surf1 homologue sequences were obtained from several databanks using the BLAST2 (tblastn) program. For species in which a genomic but no complete cDNA sequence was available, the intron locations were predicted using the 'Splice site prediction program by Neural Network' [13]. Bacteria initiation codons were identified by the GeneMark.hmm program [14]. Multiple alignment was performed with the Clustal W program [15]. For each species, the secondary structure was predicted, using the NPS@ program (<http://pbil.ibcp.fr>) that retained a defined structure only when there is a consensus prediction between the PREDATOR program [16], the GOR IV program [17] and the Self Optimized Prediction Method [18].

2.4. Sequence accession numbers

Human *SURF1*: Z35093 (cDNA), AC002107 (9q34 chromosomal region) and Q15526 (protein in SWISS-PROT). *Mus musculus*: M14689; *R. norvegicus*: H33785, AI104283 (ESTs) and AF182952 (this study); *Fugu rubripes*: Y15171; *D. melanogaster*: A1404004 (EST), AC004351 (genomic DNA) and AF182954 (this study); *Caenorhabditis elegans*: C70766, C59428 (ESTs) and AC006871 (genomic DNA); *A. thaliana*: AB019230 and AF182953 (this study); *S. cerevisiae*: Z72897; *Schizosaccharomyces pombe*: AL096846 with protein CAB50922.1; *Rickettsia prowazekii*: AJ235273 with protein CAA15162.1; *Paracoccus denitrificans*: X78196; *Caulobacter crescentus*: contig gcc_630 (at <http://www.tigr.org>); *Pseudomonas aeruginosa*: contig 54 (at <http://www.pseudomonas.com/>); *Mycobacterium tuberculosis*: AL123456 with hypothetical protein MTCY427.16; *Mycobacterium leprae*: AL023635 with hypothetical protein MLCB1243.32C.

3. Results and discussion

3.1. Search for Surf1 homologue sequences

Starting with a BLAST search, 14 homologues of human Surf1 have been identified in eight eukaryotes, including mammals, a fish, invertebrates, a plant or yeasts and in six prokaryotes (Fig. 1). For *M. musculus*, *F. rubripes*, *S. cerevisiae*, *S. pombe* and *R. prowazekii*, the proteins had been previously identified as human Surf1 homologues [3,11,19,20]. However, in *F. rubripes*, the amino acids coded by exons 1 and 2 that could not be identified, were unknown [19]. For *P. denitrificans*, *M. tuberculosis* and *M. leprae*, the predicted Surf1 homologues made up of 211, 271 and 270 aa, respectively, correspond to previously predicted ORFs of unknown function. In *P. denitrificans*, this protein is the ORF1 of the operon coding for a quinol oxidase [21]. For *C. crescentus* and *P. aeruginosa*, we put forward a Surf1 homologue by translation of genomic DNA contigs in which no ORF had been reported and by prediction of the initiation codons. In *C. crescentus*, we retrieved a degenerated GTG initiation codon and the protein is constituted of 245 aa, whereas in *P. aeruginosa*, the protein is 243 aa long. For *D. melanogaster*, *C. elegans* and *A. thaliana*, the BLAST2 program found genomic DNA sequences coding for proteins homologous to parts of Surf1. These Surf1 homologues were then extended by aligning these sequences with ESTs when available (invertebrates) and with other Surf1 homologues or by predicting the intron locations. In *C. elegans*, the 5 kb gene that we identified is made of six

exons coding for a predicted protein of 323 aa, homologous to Surf1. For *D. melanogaster* and *A. thaliana*, the N-terminal sequence could not be predicted because of its weak conservation between species. Likewise, for *R. norvegicus*, no EST was retrieved in the cDNA 5'-end.

For these last three species, the cDNA 5'-ends were retrieved by RACE experiments. After RNA isolation, each *SURF1* mRNA was reverse transcribed with a gene specific primer. After three successive PCR rounds with nested primers, we purified and sequenced cDNA fragments of 393 bp (*R. norvegicus*), 366 bp (*D. melanogaster*) and 359 bp (*A. thaliana*). These fragments correspond to the amino acids 1 to 105 (with a 15 bp 5'-UTR), 1 to 71 (with a 87 bp 5'-UTR) and 1 to 77 (with a 53 bp 5'-UTR), respectively. The resulting Surf1 proteins contain 306 aa, 300 aa and 354 aa, respectively. We can also deduce from the genomic DNA clones that the *D. melanogaster* and *A. thaliana* *SURF1* genes contain respectively 1200 bp with four exons and 2600 bp with six exons. For *R. norvegicus*, the absence of known genomic clones containing the *SURF1* homologue prevents the retrieval of the full-length gene. It should be noted that the *D. melanogaster* *SURF1* gene does not have the same structure as and is smaller than those of the other higher eukaryotes. Moreover, the 5'-UTR of *D. melanogaster* *SURF1* cDNA is longer (87 bp) than those of humans (14 bp), mice (4 bp) [11] and rats (15 bp) (this study). The small 5'-UTRs found in mammals were likely due to the compact structure of the Surf1 locus. These observations confirm those of Armes and Fried [12] which suggest that the Surf1 locus is not conserved in invertebrates.

3.2. Sequence alignment highlights the remarkable conservation of Surf1

Alignment made with Clustal W program of proteins homologous to the human Surf1 (300 aa) revealed striking similarities with the eight eukaryotic and six prokaryotic proteins (Fig. 1). All known eukaryotic Surf1 sequences are extended on the N-terminal end as compared to the bacterial species. This extension contains a typical mitochondrial targeting presequence. The secondary structure of all homologous sequences tentatively predicted using the consensus obtained between three different methods [16–18] shows that the location of several α -helix and β -sheet structures is conserved in almost all species. Table 1 shows the high percentage of identity or

Table 1
Percentage of identity and of similarity to the human Surf1 protein (300 aa)

Species	Identity (%)	Similarity (%)
Eukaryotes		
<i>M. musculus</i>	77.2	90.9
<i>R. norvegicus</i>	75.9	90.6
<i>F. rubripes</i>	68.7	90.4
<i>D. melanogaster</i>	41.1	71.2
<i>C. elegans</i>	35.5	64.8
<i>A. thaliana</i>	22.3	50.5
<i>S. cerevisiae</i>	23.2	46.1
<i>S. pombe</i>	27.6	57.0
Prokaryotes		
<i>R. prowazekii</i>	20.9	51.4
<i>P. denitrificans</i>	29.6	55.8
<i>C. crescentus</i>	31.5	55.8
<i>P. aeruginosa</i>	17.1	42.8
<i>M. tuberculosis</i>	12.8	38.7
<i>M. leprae</i>	13.1	39.5

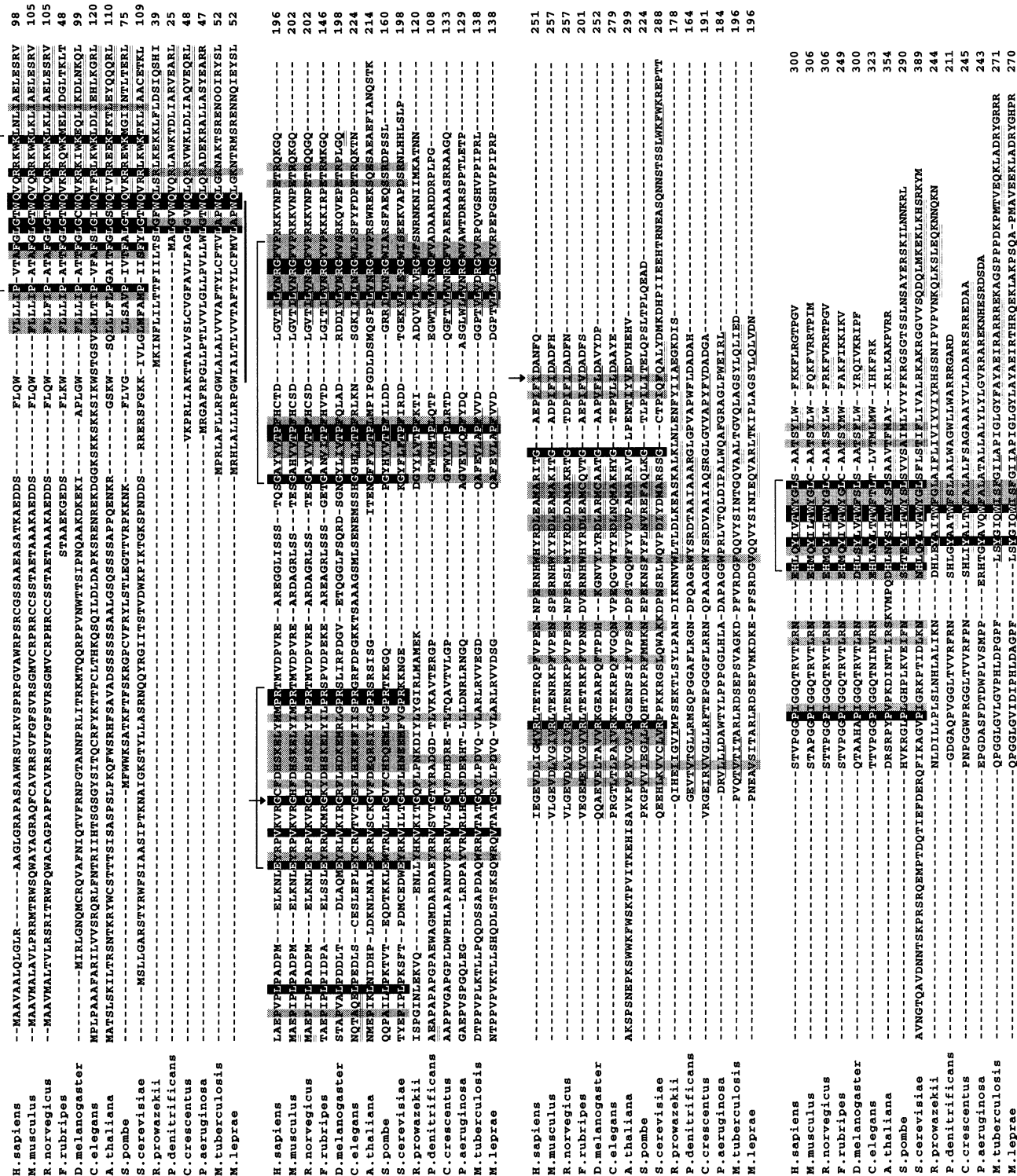


Fig. 1. Multiple alignment of the Surf1 protein. The black and grey vertical bars represent identical and conserved amino acids, respectively, either in eukaryotes or in all species. For each species, the secondary structure was predicted using the NPS@ program (see Section 2). Only the β -sheets and the α -helix which were predicted at a similar location in almost all studied species are underlined with one or two lines, respectively. Two hydrophobic transmembrane domains predicted in all species are indicated with a bold line below the sequence. The brackets above the sequence show the four domains containing highly conserved sequences. The two missense mutations found in our patients [9] are indicated by arrows. The dotted lines correspond to gaps introduced to improve the alignment. The numbers on the right indicate the amino acid number of each protein. For *F. rubripes*, the first amino acid is not the real first residue because the predicted sequence is still incomplete.

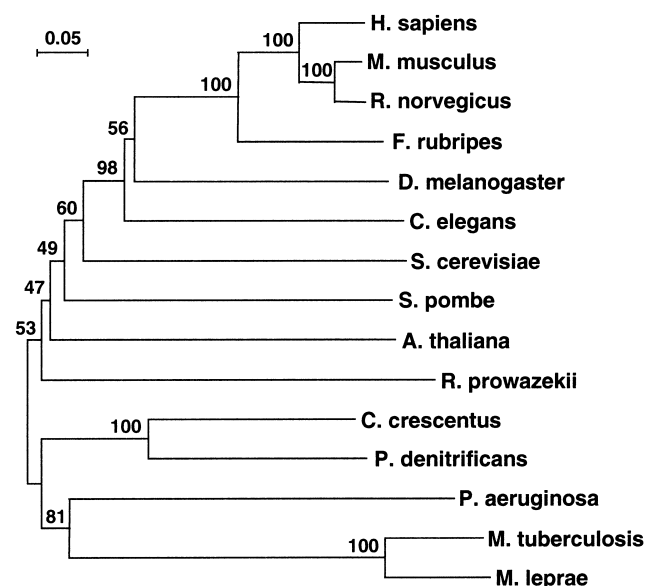


Fig. 2. Dendrogram obtained from the total human Surf1 protein aligned with eukaryotes and bacterial homologues by the PHYLO-WIN program with the neighbour joining method [28]. Bootstrap values are indicated at the nodes (1000 replicates).

similarity of all species to the human Surf1. Furthermore, many amino acids are conserved: 28 were identical in the nine eukaryotes and 48 were conservatively substituted. When the comparison included eukaryotes and prokaryotes, 11 aa were strictly identical: Leu⁷⁶, Trp⁷⁹, Gln⁸⁰, Val¹²⁰, Gly¹²⁴, Pro¹⁶⁵, Leu¹⁷⁶, Arg¹⁷⁹, Gly¹⁸⁰, Tyr²⁷⁴ and Trp²⁷⁸ (numbering of the human sequence) and 19 other amino acids were conservatively substituted. The conservation was particularly striking on four hot spots of the sequence. The first similar region is located between the human Pro⁷⁰ and Lys⁸⁷, at the junction between a first conserved hydrophobic domain and an α -helix made of about 50% charged amino acids found in all species except mycobacteria. This hydrophobic domain is large enough to represent a transmembrane domain (bold line in Fig. 1) except in the case of *P. denitrificans* for which the hydrophobic N-terminal domain is only made of 6 aa.

Another cluster of very similar amino acids is located between Glu¹¹⁶ and Arg¹³⁷. The third region, between Gly¹⁵⁹ and Val¹⁸², includes two clusters with amino acids identical or conservatively substituted in all species and two β -sheets always predicted in this region. These regions should have an important role for the structure and/or function of the Surf1 protein. Indeed, in the region between Glu¹¹⁶ and Arg¹³⁷ surrounding the strictly conserved Gly¹²⁴, a change of this Gly into a Glu gives an inactive Surf1 in one LS-COX⁻ patient [9]. It is likely that some mutations inducing LS-COX⁻ will also be found in the region located between Gly¹⁵⁹ and Val¹⁸².

The C-terminal domain is also well conserved between His²⁷¹ and Leu²⁸¹ and, at the same level, a stretch of mostly hydrophobic residues is very likely to constitute a transmembrane domain (bold line) present in all species, aside a β -sheet and/or an α -helix. The essential role of this C-terminal domain was previously stressed by the studies of Zhu et al. [4] who have shown that the SURF1 cDNA of an LS-COX⁻ patient exhibiting a deletion of the last 10 aa and the modification of the nine previous ones predicted a non-functional

truncated protein. In addition, many pathogenic mutations of LS-COX⁻ patients are clustered in this region of exons 8 and 9 [4,5].

Another region surrounding Ile²⁴⁶ could also be important for Surf1 structure or function, even if amino acids are not as well conserved in this region. At the position corresponding to Ile²⁴⁶ (mutated into a Thr in LS-COX⁻ patient L1 in [9]), a hydrophobic amino acid is conserved at least in all eukaryotes and the predicted secondary structure was always a β -sheet surrounded by two coiled regions. The same secondary structure could be predicted in all superior eukaryotic Surf1 proteins and in some bacteria. Replacing Ile²⁴⁶ by a Thr in patient L1 changed the conserved hydrophobic character of this amino acid and prevented the prediction of the β -sheet. Probably the α -helix-coil- β -sheet structure that could be found at this level was more essential for the Surf1 protein than the amino acid conservation.

3.3. Phylogenetic analysis of Surf1

Fig. 2 shows the dendrogram built to compare the protein evolution in the 15 species where Surf1 was found. Although this tree should not be taken to represent the precise evolution of distances, the reconstituted phylogeny is informative. *R. prowazekii* was proposed to be the bacteria closest to the mitochondria ancestor [20] and Surf1 is a mitochondrial protein. Therefore, it is logical to find the *R. prowazekii* Surf1 closer to eukaryote Surf1 than that of other bacteria.

Our data show that the *P. denitrificans* gene homologous to human SURF1 is the ORF1 predicted at the 3'-end of the quinol oxidase operon (cytochrome *ba*₃) [21]. The protein coded by this ORF1 is highly homologous to human Surf1 since it has 29.6% identical and 55.8% similar aa (Table 1). Surf1 is the only protein in the operon in addition to the four subunits constituting the quinol oxidase. Furthermore, the *P. denitrificans* COX aa₃ is also coded by an operon [22] containing at least two ORFs corresponding to proteins homologous to those involved in the human COX assembly: COX10 [23] and COX11 [24]. Therefore, the eukaryotic genes involved in COX assembly should then originate from both the bacterial quinol oxidase *ba*₃ and COX aa₃ operons since the quinol oxidase and COX belong to the same oxidase superfamily [25]. A Surf1 homologue could not be identified in two bacteria for which the entire genomes have been sequenced: *Escherichia coli* and the Gram⁺ *Bacillus subtilis*. *E. coli* has no COX but a quinol oxidase, while *B. subtilis* has a COX and a quinol oxidase. A gene duplication event which gave COX and quinol oxidase has happened during evolution of Gram⁺ bacteria, the duplicated ancestral gene being either a COX [26] or a quinol oxidase [27]. In addition, several lateral gene transfer events of quinol oxidase may have occurred during evolution. In any case, since Surf1 is present neither in *E. coli* nor in *B. subtilis*, it was not mandatory in primitive COX or quinol oxidases but it is found in the mitochondria bacterial ancestor *R. prowazekii* which has no quinol oxidase but several COX genes. Surf1 might have appeared during one of the gene transfers.

In conclusion, the remarkable conservation of Surf1 in eukaryotes as well as in relatively recent prokaryotes should provide new opportunities either to make site-directed mutagenesis or to use various species as models to understand the Surf1 function in COX assembly, which hopefully, could lead one day to a therapy for LS-COX⁻ patients.

Acknowledgements: We would like to thank C. Donne-Goussé, Dr C. Hänni and Dr M. Gouy for fruitful discussions and for their help in the sequence comparison analyses. Thanks are also due to Dr B. Durand for her generous gift of *Drosophila* RNA. This work was supported by the 'Région Rhône-Alpes', by the 'Association Française contre les Myopathies' (grant and support to A.P.) and by the 'Fondation pour la Recherche Médicale' (support to K.B.).

References

- [1] Poyau, A. and Godinot, C. (1999) in: *Mitochondrial Diseases: Models and Methods* (Lestienne, P., Ed.), pp. 115–127, Springer Verlag, Berlin.
- [2] Nijtmans, L.G., Taanman, J.W., Muijsers, A.O., Speijer, D. and Van den Bogert, C. (1998) *Eur. J. Biochem.* 254, 389–394.
- [3] Mashkevich, G., Repetto, B., Glerum, D.M., Jin, C. and Tzagoloff, A. (1997) *J. Biol. Chem.* 272, 14356–14364.
- [4] Zhu, Z., Yao, J., Johns, T., Fu, K., De Bie, I., Macmillan, C., Cuthbert, A.P., Newbold, R.F., Wang, J.C., Chevrette, M., Brown, G.B., Brown, R.M. and Shoubridge, E.A. (1998) *Nat. Genet.* 20, 337–343.
- [5] Tiranti, V., Hoertnagel, K., Carrozzo, R., Galimberti, C., Munaro, M., Granatiero, M., Zelante, L., Gasparini, P., Marzella, R., Rocchi, M., Bayona-Bafaluy, P., Enriquez, J.A., Uziel, G., Bertini, E., Dionisi-Vici, C., Franco, B., Meitinger, T. and Zeviani, M. (1998) *Am. J. Hum. Genet.* 63, 1609–1621.
- [6] Rahman, S., Blok, R.B., Dahl, H.H., Danks, D.M., Kirby, D.M., Chow, C.W., Christodoulou, J. and Thorburn, D.R. (1996) *Ann. Neurol.* 39, 343–351.
- [7] Lombes, A., Nakase, H., Tritschler, H.J., Kadenbach, B., Bonilla, E., De Vivo, D.C., Schon, E.A. and Di Mauro, S. (1991) *Neurology* 41, 491–498.
- [8] Tiranti, V., Jaksch, M., Hofmann, S., Galimberti, C., Hoertnagel, K., Lulli, L., Freisinger, P., Bindoff, L., Gerbitz, K.D., Comi, G.P., Uziel, G., Zeviani, M. and Meitinger, T. (1999) *Ann. Neurol.* 46, 161–166.
- [9] Poyau, A., Buchet, K., Bouzidi, M.F., Zabot, M.T. and Godinot, C. (1999) *Biochimie* 81, S184.
- [10] Duhig, T., Ruhrberg, C., Mor, O. and Fried, M. (1998) *Genomics* 52, 72–78.
- [11] Lennard, A., Gaston, K. and Fried, M. (1994) *DNA Cell Biol.* 13, 1117–1126.
- [12] Armes, N. and Fried, M. (1996) *Mol. Cell. Biol.* 16, 5591–5596.
- [13] Brunak, S., Engelbrecht, J. and Knudsen, S. (1991) *J. Mol. Biol.* 220, 49–65.
- [14] Lukashin, A. and Borodovsky, M. (1995) *Nucleic Acids Res.* 26, 1107–1115.
- [15] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) *Nucleic Acids Res.* 22, 4673–4680.
- [16] Frishman, D. and Argos, P. (1996) *Protein Eng.* 9, 133–142.
- [17] Garnier, J., Gibrat, J.F. and Robson, B. (1996) *Methods Enzymol.* 266, 540–553.
- [18] Geourjon, C. and Deléage, G. (1994) *Protein Eng.* 7, 157–164.
- [19] Armes, N., Gilley, J. and Fried, M. (1997) *Genome Res.* 7, 1138–1152.
- [20] Andersson, S.G.E., Zomorodipour, A., Andersson, J.O., Sichert-Pontén, T., Alsmark, U.C.M., Podowski, R.F., Näslund, A.K., Eriksson, A.S., Winkler, H.H. and Kurland, C.G. (1998) *Nature* 396, 133–140.
- [21] Richter, O.M.H., Tao, J.S., Turba, A. and Ludwig, B. (1994) *J. Biol. Chem.* 269, 23079–23086.
- [22] Raitio, M., Jalli, T. and Saraste, M. (1987) *EMBO J.* 6, 2825–2833.
- [23] Nobrega, M.P., Nobrega, F.G. and Tzagoloff, A. (1990) *J. Biol. Chem.* 265, 14220–14226.
- [24] Tzagoloff, A., Capitanio, N., Nobrega, M.P. and Gatti, D. (1990) *EMBO J.* 9, 2759–2764.
- [25] Saraste, M. (1990) *Q. Rev. Biophys.* 23, 331–366.
- [26] Castresana, J., Lübken, M., Saraste, M. and Higgins, D.G. (1994) *EMBO J.* 13, 2516–2525.
- [27] Musser, S.M. and Chan, S.I. (1998) *J. Mol. Evol.* 46, 508–520.
- [28] Galtier, N., Gouy, M. and Gauthier, C. (1996) *Comput. Appl. Biosci.* 12, 543–548.