Monatshefte für Chemie 136, 1365–1381 (2005)
DOI 10.1007/s00706-005-0331-x

Monatshefte für Chemie
Chemical Monthly
Printed in Austria

What to Learn from a Comparative Genomic Sequence Analysis of *L*-Carnitine Dehydrogenase

Clemens Uanschou, Roswitha Frieht, and **Fritz Pittner***

Max F. Perutz Laboratories, Institute of Biochemistry and Molecular Cell Biology, University of Vienna and Ludwig-Boltzmann-Forschungsstelle für Biochemie, A-1030 Vienna, Austria

Received December 14, 2004; accepted January 26, 2005
Published online July 6, 2005

Summary. In contrast to eukaryotic cells certain eubacterial strains have acquired the ability to utilize *L*-carnitine (*R*-(−)-3-hydroxy-4-(trimethylamino)butyrate) as sole source of energy, carbon and nitrogen. The first step of the *L*-carnitine degradation to glycine betaine is catalysed by *L*-carnitine dehydrogenase (*L-C*DH, EC 1.1.1.108) and results in the formation of the dehydrocarnitine. During the oxidation of *L*-carnitine a simultaneous conversion of the cofactor NAD^+ to *NAD*H takes place. This catabolic reaction has always been of keen interest, because it can be exploited for spectroscopic *L*-carnitine determination in biological fluids – a quantification method, which is developed in our lab – as well as *L*-carnitine production.

Based on a cloned *L-C*DH sequence an expedition through the currently available prokaryotic genomic sequence space began to mine relevant information about bacterial *L*-carnitine metabolism hidden in the enormous amount of data stored in public sequence databases. Thus by means of homology-based and context-based protein function prediction is revealed that *L-C*DH exists in certain eubacterial genomes either as a protein of approximately 35 kDa or as a homologous fusion protein of approximately 54 kDa with an additional putative domain, which is predicted to possess a thioesterase activity. These two variants of the enzyme are found on one hand in the genome sequence of bacterial species, which were previously reported to decompose *L*-carnitine, and on the other hand in gram-positive bacteria, which were not known to express *L-C*DH. Furthermore we could not only discover that *L-C*DH is located in a conserved genetic entity, which genes are very likely involved in this *L*-carnitine catabolic pathway, but also pinpoint the exact genomic sequence position of several other enzymes, which play an essential role in the bacterial metabolism of *L*-carnitine precursors.

Keywords. *L*-Carnitine metabolism; Comparative genomics; Protein function prediction; Sequence database.

* Corresponding author. E-mail: fritz.pittner@univie.ac.at

Introduction

Carnitine Metabolism in Bacteria

According to current knowledge prokaryotes have developed three different strategies to make use of *L*-carnitine [1]. First of all *L*-carnitine can be cleaved into trimethylamine and malic acid (Fig. 1). This transformation is catalyzed by

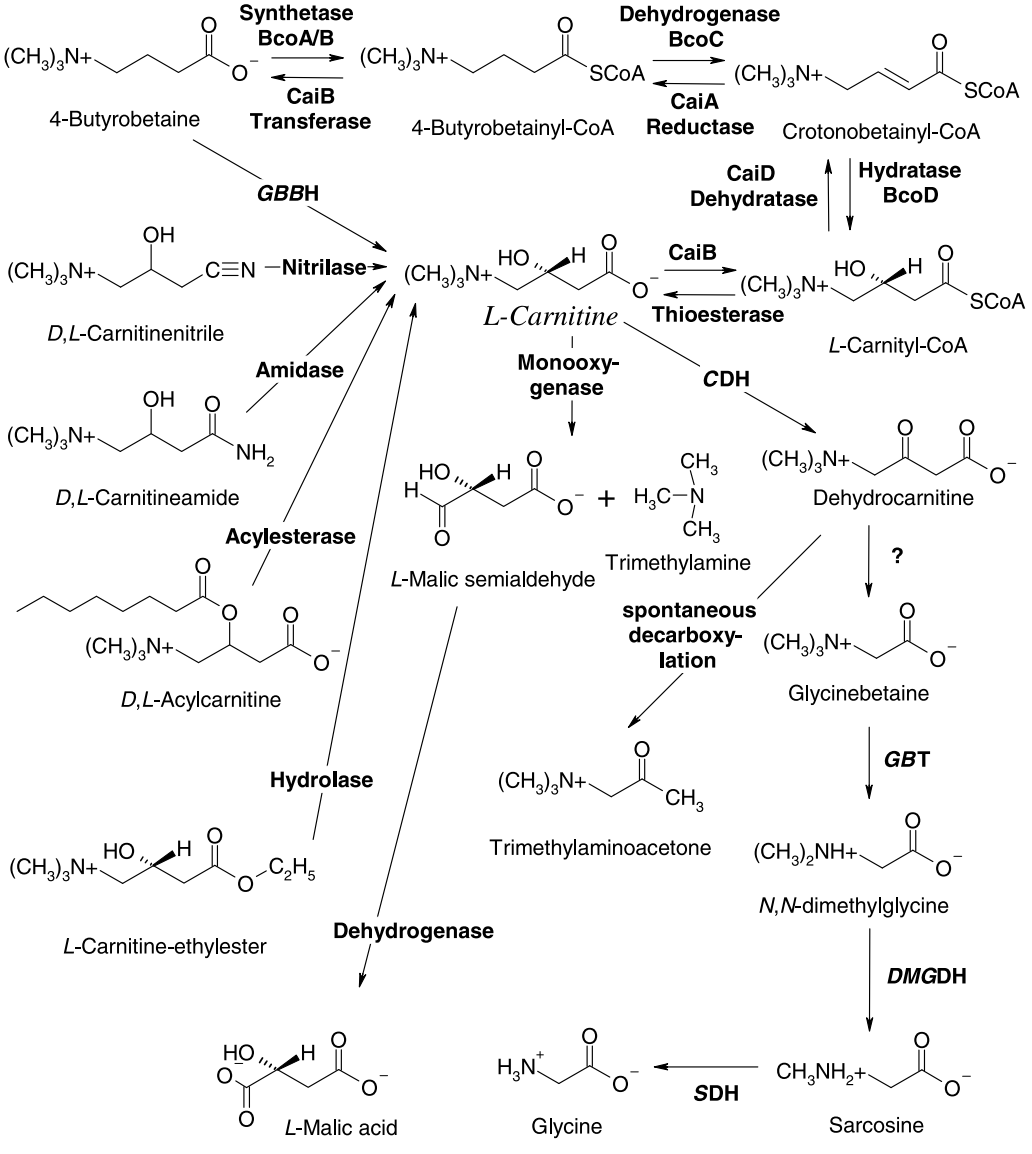


Fig. 1. *L*-Carnitine metabolism in eubacteria; explanation about the single reactions should be taken from the main text; abbreviations: bco. . .butyrobetaine-*L*-carnitine operon, cai. . .*L*-carnitine-butyrobetaine operon, CoA. . .coenzyme A, *DMG*DH. . .dimethylglycine-dehydrogenase, *GBB*H. . .gamma butyrobetaine hydroxylase, *GB*T. . .glycinebetaine-transmethylase, *L-C*DH. . .*L*-carnitine dehydrogenase, *S*DH. . .sarcosine dehydrogenase

enzymes not yet characterized, which are synthesized by strains of *Acinetobacter calcoaceticus* as well as *Serratia marcescens*. By means of this degradation pathway the above named bacterial species are capable of assimilating the carbon-chain of 3-hydroxy-4-(trimethylammonio)butyric acid betaine, whereas trimethylamine is not metabolically used by these strains and remains unchanged. Because eukaryotes can not catabolize *L*-carnitine, studies in mammals lead to the conclusion that microorganisms colonizing the intestinal tract convert *L*-carnitine, either as described above or in the way described as follows [1b].

Strains of *E. coli*, *S. typhimurium*, *P. mirabilis*, and *P. vulgaris* are able to convert *L*-carnitine into γ-butyrobetaine (*GBB*) *via* crotonobetaine with the help of several proteins. Their genes are grouped in the so called caiTABCDE-operon, which was characterized in *E. coli* [3]. The above mentioned members of the family enterobacteriaceae and surely a lot of other microorganisms can not directly assimilate the carbon- and nitrogen-backbone of carnitine, in lieu the prokaryotes use the γ-trimethyl-β-hydroxybutyrobetaine as final electron-acceptor under anaerobic conditions, if other respiratory chains are absent. In order to replace, *e.g.*, a missing nitrate respiration, it has been demonstrated experimentally that the enzymes of the cai-operon depend on the existence of the cosubstrates crotonobetainyl- and γ-butyrobetainyl-CoA [4]. Additional studies have assigned functions to almost all gene products of the cai-operon and analyzed this metabolic pathway in detail [5–7]. As a result, it was discovered that the three proteins CaiA, CaiB, and CaiD are directly involved in the transformation of *L*-carnitine to γ-butyrobetaine: the gene caiD for instance codes an enoyl-CoA hydratase, which converts *L*-carnitinyl-CoA to crotonobetainyl-CoA in a reversible reaction [5]. The gene product of caiD is also supposed to have a carnitine racemase activity and is assisted by CaiB, a CoA transferase, its structure was determined recently [8]. The latter is particularly needed because the gene product of caiA alone is not sufficient to ensure crotonobetaine-reductase-activity [7] (Fig. 1).

The two previous mentioned pathways of *L*-carnitine catabolism found in the above named bacterial species both depend either on the addition of exogenic nitrogen or of nitrogen and carbon together to grow on a minimal culture medium with *L*-carnitine as the sole source of energy, carbon and nitrogen. So, these γ-proteobacteria are unable to metabolize the quaternary ammonium group, in contrast to prokaryotes, whose genomes contain the necessary information to code the *L*-carnitine dehydrogenase (*L-C*DH, EC 1.1.1.108 [9]). This enzyme catalyzes the first step in this carnitine degradation pathway and forms glycine betaine (*GB*) *via* dehydrocarnitine by using (an) unidentified enzyme(s). *Lindstedt et al.* supposed a CoA-depending dehydrocarnitine catabolism due to the fact that cell-free extracts of *Pseudomonas sp.* AK 1 decompose carnitine to *GB* if NAD^+, *ATP*, and CoA are added. If only NAD^+ is supplied, trimethylaminoacetone is observed [10]. This trimethylamine compound is generally obtained by spontaneous decarboxylation from unstable dehydrocarnitine under high and middle *pH* values and low ionic strength of the reaction medium [11]. Remarkably, the reversibility of this starting carnitine degradation reaction led to the development of several patents for *L*-carnitine production from dehydrocarnitine [2].

Several different proteobacterial species have been reported to produce *L-C*DH, namely *Agrobacterium sp.* [12, 13], *Alcaligenes sp.* [14, 15], *Pseudomonas putida*

[16], *P. aeruginosa* [17], and *Xanthomonas translucens* [18], or to use *L*-carnitine as the sole source of energy, carbon and nitrogen, namely *Burkholderia cepacia* [19], *Pseudomonas spp.* [19], and *Sinorhizobium meliloti* [20]. According to current knowledge the ability of a bacterial strain to survive on *L*-carnitine minimal culture medium can be set nearly equal with the occurrence of a gene for *L-C*DH. But due to the fact that some bacterial species degrade trimethylamine aerobically and anaerobically [21], a theoretical possibility exists that bacteria have developed another alternative *L*-carnitine degradation pathway, which enables them to grow on *L*-carnitine minimal culture medium. However such *L*-carnitine catabolism has never been observed. In addition to the above named gram-negative prokaryotes only one gram-positive bacterium, the *Brevibacterium linens* ATCC 19391 is supposed to produce *L*-carnitine dehydrogenase because this microorganism converts *L*-carnitine into *GB* under normal and raised osmotic pressure. However, this bacterium seems unable to make use of *GB* and therefore lacks the ability to grow on *L*-carnitine minimal culture medium [22].

To utilize the nitrogen a few bacterial species transform *GB* to glycine in a three step demethylation process starting with the formation of *N*,*N*-dimethylglycine (*DMG*) in *Pseudomonas aeruginosa* PAO1 [23] and *Sinorhizobium meliloti* [24] by the catalytical help of *glycinebetaine-transmethylase* (*GBT*, EC 2.1.1.5), that is also named *betaine-homocysteine-methyltransferase* (*BHMT*). The next conversion from *DMG* to sarcosine (*N*-methylglycine) is catalyzed by *dimethylglycine-dehydrogenase* (*DMGDH*, EC 1.5.99.2), which was found in *Sinorhizobium meliloti* [24]. Last but not least it was experimentally proved that *sarcosine dehydrogenase* (*SDH*, EC 1.5.99.1) transforms sarcosine to glycine. This enzyme was purified from *Pseudomonas putida* [25], *P. aeruginosa* [26], and postulated in *Sinorhizobium meliloti* [24] (Fig. 1). Moreover, it should be mentioned that the cyanobacterium *Aphanothece halophytica* [27] and species of the gram-positive genus *Arthrobacter* [28] are supposed to degrade *GB* in a similar demethylation process.

By working on the improvement of a *L*-carnitine quantification system for biological fluids, which has been developed in our lab and is based on the measurement of arising *NAD*H concentration caused by the *L-C*DH driven oxidation of carnitine to dehydrocarnitine by UV spectroscopy [29, 30], we were striven to identify new potential *L-C*DH producing microorganisms with more suitable enzyme characteristics to be able to determine *L*-carnitine more sensitively. To decrease the necessary sample volume a lower K_m value for *L*-carnitine on one hand and a better heat-resistancy of *L-C*DH on the other hand would be of need. The molecular weight of monomeric *L-C*DH was reported to range from 32 to 37 [16–18] and from 50 to 57 kDa [12, 31] and the enzyme occurs as homodimer under native conditions and possesses a K_m value for *L*-carnitine between 1 and 14 m*M* [32]. It also has to be mentioned that a *D*-carnitine dehydrogenase was isolated from *Agrobacterium* species, which shows no sequence homology with *L-C*DH [1a, 13].

Bacterial Osmoregulation Through Betaines – Another Purpose of L-Carnitine

L-Carnitine like glycine betaine is not only an exotic nutrient for only a few proteobacteria, but also serves as a *compatible solute* to support certain organisms in coping with hyperosmotic conditions [33]. *Compatible solutes* are low-molecular,

easy soluble compounds without a net charge at physiological *pH* and can be accumulated at high concentration within the cell without disturbing critical cellular processes or protein folding [34]. This allows cells to keep their turgor at constant level. For the eubacteria *Escherichia coli*, *Lactobacillus plantarum*, *Listeria monocytogenes* (citations included in Ref. [22]), *Pediococcus pentosaceus* [35], *Staphylococcus aureus* [36], and *Tetragenococcus halophila* [35], *L*-carnitine acts as osmoprotectant. These microorganisms lack the enzymatic equipment to convert this trimethylamine compound to *GB* when the osmotic stressful situation is over. However, it is well-known that glycine betaine is the preferred compatible solute in eubacteria and provides probably the highest level of osmotolerance [34b, 35]. So it has been supposed that the purpose of a *L-C*DH initiated *L*-carnitine degradation in *Brevibacterium linens* [22] and *Pseudomonas aeruginosa* [37] is the production of the more effective osmoprotectant *GB*. For the same reason choline (2-hydroxyethyltrimethylammoniumhydroxide) also seems to be oxidized in *P. aeruginosa* [37] to glycine betaine by choline dehydrogenase (EC 1.1.3.17) and betaine aldehyde dehydrogenase (EC 1.2.1.8) [24, 38].

Since phosphatidylcholine is a common constituent of the eukaryotic membranes, choline like *L*-carnitine is widespread in various environments. Therefore bacterial strains exhibit a selection advantage if their genomes include the necessary information to code for these catabolizing enzymes. The gram-negative mammal pathogen, *Pseudomonas aeruginosa* for instance, is known to meet these three advantages (*L*-carnitine, choline, *GB* degradation capabilities) and is therefore perfectly adapted to hostile environment, which exists in the respiratory system of cystic fibrosis patients. If this bacterium destroys the host's bronchopulmonary cell membranes, the local concentration of *L*-carnitine outside the cells will likely increase [37]. Strains of the *Burkholderia cepacia* group and *P. aeruginosa* have been isolated from the sputum of cystic fibrosis patients [39], a fact that indicates the presumable occurrence of similar adaptation mechanisms in these related species. In comparison to mammals the *L*-carnitine levels in higher plants are low, nevertheless mentionable amounts have been reported for those plants with high concentrations of fatty acids as wheat seeds and alfalfa seedlings [40]. This observation was the spur for *Goldmann* to investigate the ability of *Sinorhizobium meliloti* – a soil bacterium, famous for its symbiotic interaction with the feeding plant alfalfa – to grow on *L*-carnitine as the sole source of energy, carbon and nitrogen [20]. In a following publication the same working group experimentally proved that glycine betaine is the accumulated intermediate compound, which they had noticed in their first study as an unknown *L*-carnitine degradation product. Therefore they concluded that the *L*-carnitine catabolism in *Sinorhizobium meliloti* has to be very similar to the one in various *Pseudomonas* species [41]. These two examples help to answer to the question why bacteria have developed metabolic pathways to convert *L*-carnitine to *GB* and represent a parasitic and a symbiotic case of prokaryotic–eukaryotic interaction promoted by *L*-carnitine.

Protein Function Prediction by Means of Similarity- and Context-based Approaches

The exact function of two proteins can not be doubtlessly predicted by comparing their sequence similarity only, because small changes – like an amino acid

substitution in the reactive center – can already alter the protein-function and substrate specificity effectively [42]. In fact, even proteins with completely equal sequences can exhibit different functions depending on the environmental working conditions under which they fulfill their duty [43]. Another latent problem of homology-based protein function prediction, which is based on the assumption that similar sequences will have similar functions, is the fact that it is only possible to assign functions for approximately 60% of the predicted open reading frames (ORFs) in the lately sequenced genomes, because of missing homologous proteins, whose functions have been verified experimentally. Additionally 10–30% of predicted protein functions are supposed to be incorrectly annotated because of generous BLAST expectation value thresholds [45]. Therefore it would be desirable to include additional information from sequence data about the biological context of an investigated protein in order to annotate functions of given protein sequences more accurately.

In the last decade an enormous effort has been made to decipher the nucleotide order of several hundred genomes, starting with the publication of the first completely sequenced bacterial genome from *Haemophilus influenzae* in 1995 [46]. Until October 2004 approximately 340 microbial genomes, which comprise together one and a half billion nucleotides, were made public at *NCBI's* website to allow homology-based search operations. To get an overview a nearly complete list of current and finished sequencing projects can be accessed at "http://www.genomesonline.org" [47]. Now that this huge amount of accumulated data is waiting to be mined, a promising nonhomology-based method, namely the genomic context-based protein function prediction [44], should be very helpful, especially since contents of DNA database are increasing continuously. What makes this approach extraordinary is that it uses three different types of genomic associations to infer functional links between proteins: a) gene fusion, b) conservation of genetic neighborhood, c) phylogenetic profiles.

Ad a) If different proteins have fused homologues in distant related organisms, they exhibit associated functions, so that it will be very likely that they physically interact directly or indirectly (in a complex) with each other or that they are part of the same metabolic pathway. The latter is by far most frequently observed, but fused proteins do not always catalyze subsequent steps. However, the amount of genomic sequence data grows exponentially, so that it is easier to predict fused genes, seeing that the number of observed fusion events increases with the number of genome sequences.

Ad b) Genomes and their operons are randomly rearranged. Therefore it is possible that the gene order and content as well as regulatory mechanisms of operons vary even in closely related species. However some small gene clusters seem to be conserved in evolution over a wide range of species. Their genes tend to be part of the same operon due to the fact that they have to be maintained and regulated together in order to allow their products to functionally interact.

Ad c) If groups of scattered genes from different genomes have the same phylogenetic profile it is likely that they are functionally connected.

In conclusion it should be noted that similarity-based methods are used to assign functions to proteins, whereas context-based methods are intended to reveal functional interaction between proteins. Therefore the latter should be seen

as complementary to the former approach and furthermore they can often help to elucidate biochemical pathways or functional networks in cooperation. Remarkably, chromosomal proximity represents the most powerful approach of the three [48].

Results and Discussions

Homology Search for L-CDH

Unfortunately an ORF annotated as *L-C*DH was not found in any of the numerous annotated genomes, which are stored in several sequence databases, although a complete *L-C*DH nucleotide sequence is situated in the patent section of the GenBank (www.ncbi.nlm.nih.gov), that was cloned and determined from *Alcaligenes sp.*

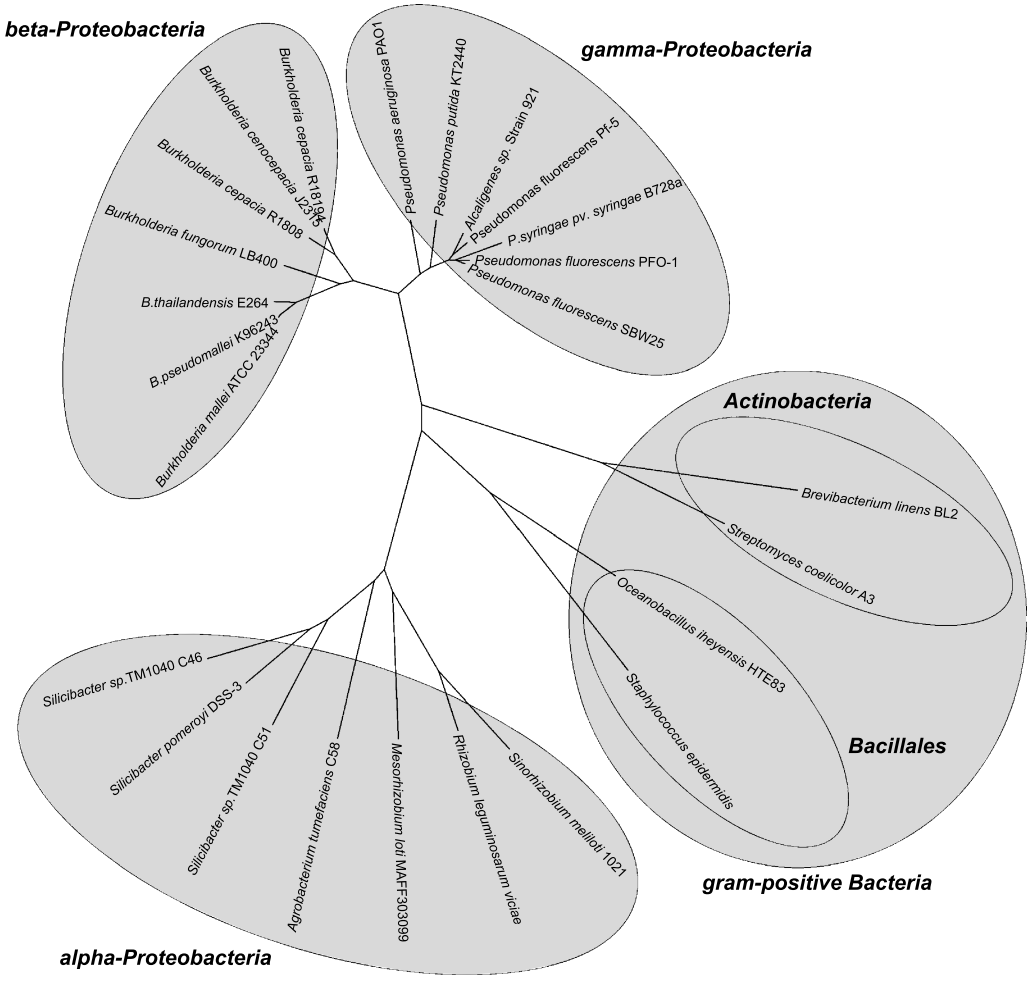


Fig. 2. This phylogenetic tree derived from the predicted *L-C*DH sequences was generated using *Neighbor Joining* method [73]; phylogenetic analysis was conducted using MEGA version 2.1 [74]; taxonomical information can be accessed at "http://www.ncbi.nlm.nih.gov/Taxonomy/"

strain 921 more than thirteen years ago [14]. It seems that this GenBank entry of the accession number *E05045* is not included in the annotation process of genome sequences, which tries to assign functions to identified open reading frames by means of homology search. To look for *L-C*DH homologues in the available genomic sequence space a BLAST search [49] was performed using this translated *L-C*DH-sequence from *Alcaligenes sp.* strain 921 as query. The investigation resulted in a manageable number of promising hits in several prokaryotic genomes that formed phylogenetic clusters according to the taxonomical relationships between the microorganisms in which they were detected (Fig. 2).

To confirm the reliability of *E05045* a multiple sequence alignment (MSA) of the predicted *L-C*DH sequences including short sequence fragments of experimentally determined N-terminal *L-C*DH-sequences was computed. At least this MSA confirms the correctness of the predicted *L-C*DH for those species for which an *Edman* degradation has been performed (Fig. 3).

Several publications examined the correlations between sequence similarity and functional similarity, to assess the significance of results obtained by BLAST. The authors concluded that at levels above 40% sequence identity and BLAST expectation values below 10^{-50} protein functions as well as all four EC numbers of compared enzymes are almost always conserved as long as the sequence length is

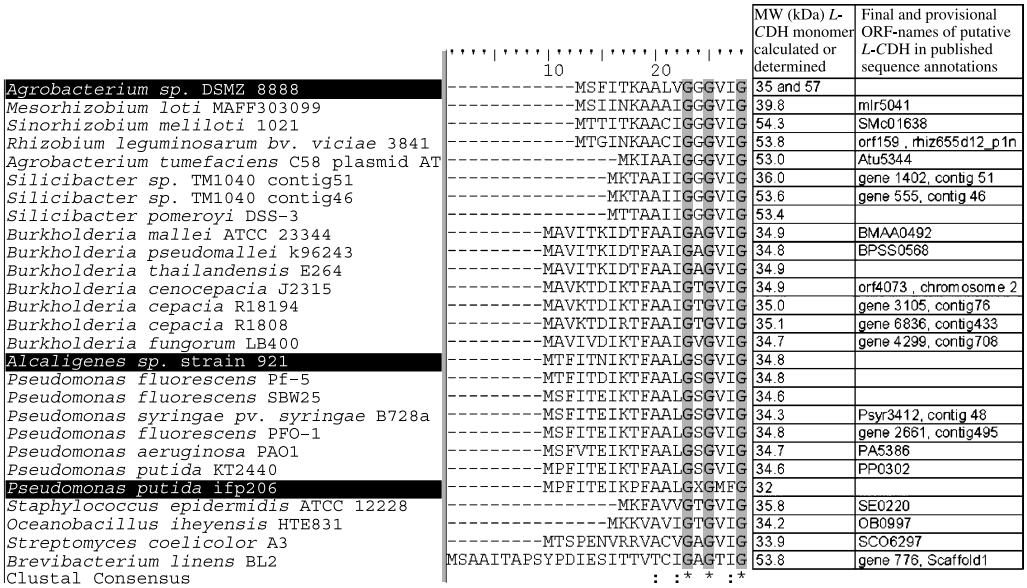


Fig. 3. Multiple sequence alignment of the N-terminal ends of several predicted *L*-carnitine dehydrogenases; the unmarked sequences are identified by homology search, the black marked bacterial strain names were either experimentally obtained by *Edman* degradation of a short N-terminal part of the purified enzyme or determined by sequencing after cloning [12, 14, 16]; the typical GXGXXG motif (marked grey) of the *NAD*-binding domain is the characteristic consensus sequence of all *NAD* depending dehydrogenases [75]; the asterisk at the bottom of the alignment indicates the equivalence of the amino acid residue on the respective position, whereas the colon indicates only strong conservation; more detailed information can be found in the documentation of the computer program *clustalx*, which was used to calculate the MSA [76]

sufficient [43]. Thus, only BLAST hits with an appropriate score above or below the previously mentioned thresholds were taken into consideration. The bacterial species containing the predicted *L-C*DH gene in their genome sequence were compared to microorganisms that were reported to express *L-C*DH or convert *L*-carnitine to *GB* or grow on culture medium with *L*-carnitine as the sole source of energy, carbon and nitrogen. On the basis of the BLAST results we defined four groups of phylogenetically clustering bacteria and these will be discussed corresponding to their increasing BLAST expectation values.

Identification of *L-C*DH genes with a molecular weight of 35 kDa in the genomic sequences of the γ-proteobacterial genus *Pseudomonas*: Considering the experimental confirmation for the *L-C*DH production in this bacterial genus [16, 17], this was not unexpected. But the extreme high sequence identities of more than 90% between the translated *L-C*DH ORF in the entry *E05045* and the predicted *L-C*DH genes of all three sequenced strains of *Pseudomonas fluorescens* give reason to doubt the correctness of the taxonomic classification of *Alcaligenes sp.* strain 921. The same applies to *Pseudomonas sp.* YS-240 because this strain was reported to express a *L-C*DH with a molecular mass of 55 kDa [31]. It is striking that we only observed *Pseudomonas* species with a L-*C*DH of 35 kDa in our genomic sequence analysis. So it is plausible that strain 921 belongs to the genus *Pseudomonas* and YS-240 does not. This assumption is understandable if we are aware of the fact that the genera *Pseudomonas* [50–52] and *Alcaligenes* [53] have constantly been and will be under taxonomical revision. Today the genus *Pseudomonas* is coarsely split into five groups of phylogenetically clustering strains [54]: *P. aeruginosa*, an opportunistic animal pathogen, *P. putida* and *P. fluorescens*, which are in general rhizosphere-colonizing and non-pathogenic, *P. syringae*, a plant pathogen, and last but not least the non-fluorescing *P. stutzeri*, which seems to be the only *Pseudomonas* core-species without *L*-carnitine degradation capability [19] and furthermore without a sequenced genome.

In contrast to the species of *Pseudomonas* all the seven available fully sequenced strains of the β-proteobacterial genus *Burkholderia* contained *L-C*DH homologous ORFs. The sequence identity between the *L-C*DH genes of *Alcaligenes sp.* 921 or *Pseudomonas* species and members of the *Burkholderia* cluster are exceptionally high (more than 70% and the *E*-value is below 10^{-135}). These results may be explained by the fact that bacterial species were transferred of the genus *Pseudomonas* to the genus *Burkholderia* nearly 12 years ago [51] and therefore seems to be closely related. Today the genus *Burkholderia* contains more than 30 species including plant pathogens, rhizosphere-colonizing, nitrogen-fixing soil bacteria, and animal pathogens [55]. The access to the genomes of the major representatives of these three groups is very helpful, but only the species *Burkholderia cepacia* was experimentally tested positive to growth on *L*-carnitine minimal culture medium [19].

The third cluster of *L-C*DH BLAST hits (46–51%, *E* 10^{-75}–10^{-83}) comprises genera from α-proteobacteria namely *Agrobacterium*, *Mesorhizobium*, *Rhizobium*, *Sinorhizobium*, and *Silicibacter*. Considering the experimental results described in the introduction [12, 13, 20], we expected to detect species of the family *Rhizobiaceae*, whereas *Mesorhizobium loti* and especially the members of the marine *Roseobacter clade Silicibacter pomeroyi* and *Silicibacter sp.* TM1040 are unknown

quantities. Although *Roseobacters* are known to be cosmopolitan in the marine environment, *Silicibacter sp.* TM1040 for instance was isolated from a culture of the dinoflagellate *Pfiesteria piscicida* indicating a close association between these organisms [56]. This biological interaction between a eukaryote and a prokaryote should again help to figure out why certain marine bacteria obviously produce *L-C*DH. Moreover it is interesting to see that we could detect two similar probable *L-C*DH sequences with a calculated molecular weight of 36 and 53.6 kDa in the genome of *Silicibacter sp.* TM1040. *Hanschmann* analogously supposed that *Agrobacterium sp.* DSMZ 8888 expresses two variants of *L-C*DH in presence of *L*-carnitine as the only diet, but he purified only the bigger enzyme (57 kDa) [12]. We also isolated a *L-C*DH from *Agrobacterium sp.* DSMZ 8888, using a different purification method to *Hanschmann*: We enriched the smaller *L-C*DH with a molecular mass of 35 kDa. On the whole, the predicted *L-C*DH sequences appear quite homogeneous except for the sequence of *Mesorhizobium loti* MAFF303099. In comparison with the others this small variant of *L*-carnitine dehydrogenase exhibits an unusual higher molecular weight of 40 kDa as well as a considerable higher calculated hydrophobicity and isoelectric point (data not shown).

Finally the predicted *L-C*DH genes of the last cluster are exclusively located in the genomes of the gram-positive bacteria, namely the cheese colonizer *Brevibacterium linens*, the deep-sea isolate *Oceanobacillus iheyensis*, the opportunistic human pathogen *Staphylococcus epidermidis*, and the soil-dwelling *Streptomyces coelicolor* (45–55%, E 10^{-74}–10^{-93}). This result confirms the assumption of a French workgroup that *Brevibacterium linens* produces *L-C*DH [22] and makes it very likely that other gram-positive bacteria do the same. It seems doubtful that gram-positive bacteria are generally able to grow on *L*-carnitine minimal culture medium, because otherwise it would have already been observed in the countless growth tests. In our experiments we could not notice an according behaviour of *Oceanobacillus iheyensis* HTE381 (DSMZ 14371) either. Therefore it is more likely that these genera will exclusively use *L-C*DH, in order to obtain *GB* for the purpose of osmoprotection. *B. linens* BL2 and *O. iheyensis* HTE381 [57] for example resist salinities of 20%.

Results from Context-based Function Prediction: L-CDH Fusion Protein and Conserved Genetic Neighborhood

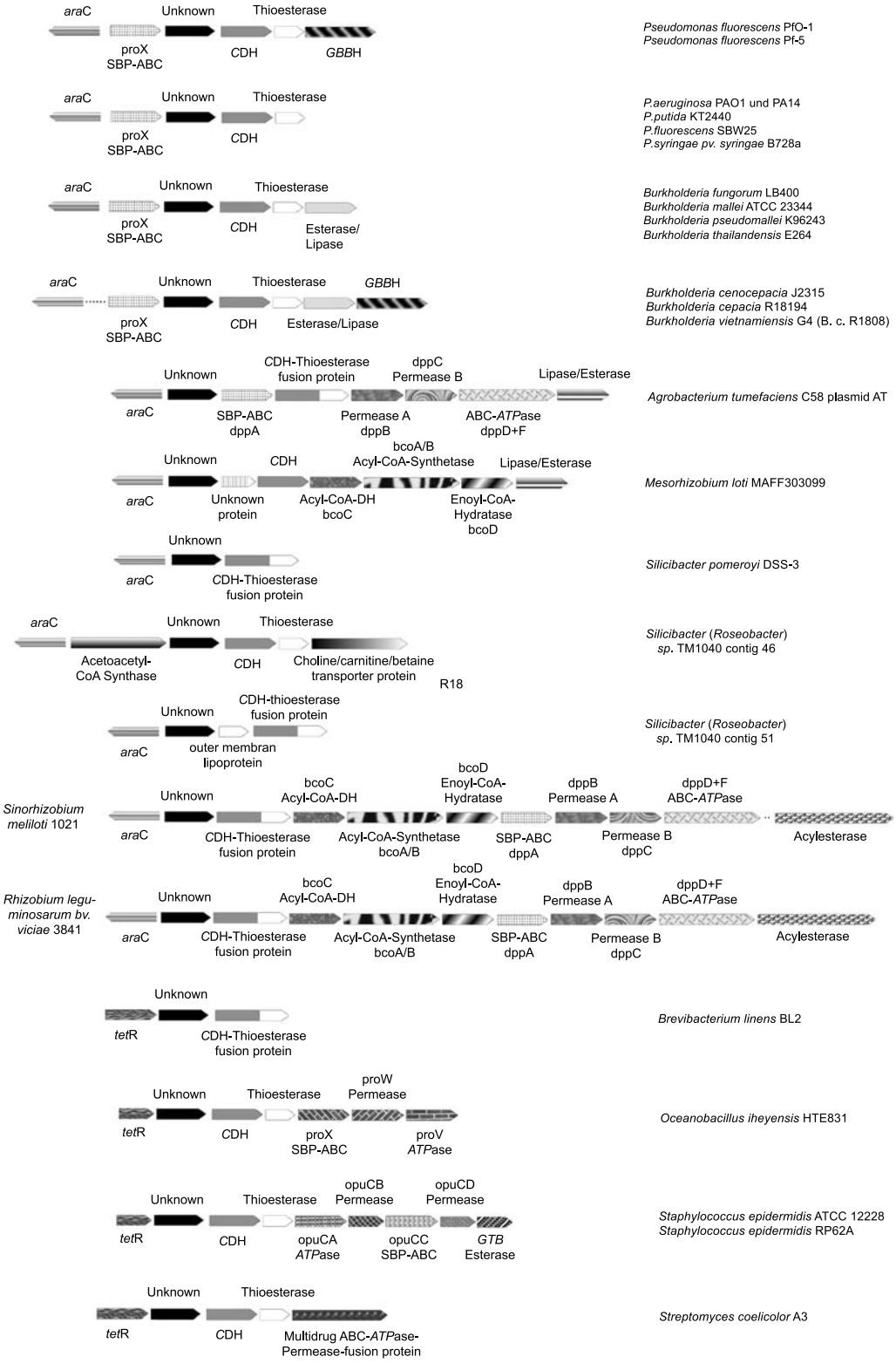
As previously mentioned the database entry E05045 includes a nearly identical copy of the *L-C*DH sequence of *Pseudomonas fluorescens* Pf-5. Therefore we decided to take a closer look at the entire sequence of *Alcaligenes sp.* strain 921 to find out if vector sequence from cloning procedure is present. On the examination it became quite clear that the database entry represents an uncontaminated sequence piece of the *Alcaligenes* genome (as a whole 1540 bases long) including two small parts of the ending of a first ORF and the beginning of a second ORF in addition to the *L-C*DH-gene. Both ORFs neighbour on the *L-C*DH sequence of *P.f.* Pf-5 and *Alcaligenes sp.* strain 921. Due to the fact that the entry E05045 only comprises small fragments of these two ORFs with a sequence length of 67 and 17 amino acids and that they are identical to those in strain Pf-5 except for one residue, we continued our investigation with the sequence of *P. fluorescens*. The ORF

located downstream of the *L-C*DH gene in the strain Pf-5 possesses a predicated thioesterase activity (white arrow in Fig. 4) and the other ORF located upstream contains a highly conserved sequence coding an unknown protein function (black arrow in Fig. 4). Furthermore, it turned out that this gene constellation is conserved in the genomes of nearly all probable *L-C*DH expressing bacteria except for three minor deviations in the genomes of members of the α-proteobacteria (*cf.* Fig. 4). This case of gene order conservation strongly indicates a functional correlation between the gene products of these three ORFs and the probable involvement of the protein ''Unknown'' in the still unsolved dehydrocarnitine degradation process. This prokaryotic protein of unknown function belongs to the members of a PFAM-protein-family [58] named DUF849 (PF05853). Moreover, the occurrence of a fusion protein consisting of *L-C*DH and thioesterase leads to an identical conclusion for these two enzymes. Considering the dehydrocarnitine degradation mechanism proposed by *Lindstedt* – previously mentioned in the introduction [10] – and the published metabolic pathway from *GBB* to *L*-carnitine discovered by *Lonza* [59] it seems obvious that the predicted thioesterase activity is responsible for splitting off CoA. Considering these results it becomes quite clear that sequence analysis can not replace experimental work, but it can give us a hint where it would be most promising for making an effort.

Prediction of L-Carnitine Precursor Degradation Enzymes

The carnitine metabolism in microorganisms has been under intense observation for the last decades due to the fact that some bacteria and fungi possess the ability to synthesize the commercially interesting quaternary ammonium compound from diverse precursors, namely γ-butyrobetaine (*GBB*), crotonobetaine, 3-dehydrocarnitine, acyl-carnitine, carnitine-amide and carnitine-nitrile [2] (Fig. 1). Under good conditions it should be possible to pinpoint the exact genomic positions of the genes of the enzymes that are responsible for the *L*-carnitine precursors degradation by means of sequence analysis, provided that the sequences of homologous proteins were previously determined. According to this the numerous public accessible genome sequences of several prokaryotes were investigated.

First the gene of γ-butyrobetaine hydroxylase (*GBB*H, EC 1.14.11.1) was discovered in the genome annotation of *Pseudomonas fluorescens* PFO-1 and later also in the *Burkholderia cepacia* group in the immediate vicinity of the predicted *L-C*DH-ORF. This *GBB*H catalyses the single-stage conversion of *GBB* to *L*-carnitine in bacteria (Fig. 1) and was sequenced of *Pseudomonas sp.* AK 1 [60]. An alternative multistage way of *GBB* degradation to *L*-carnitine was used by the Swiss company *Lonza* to develop the most successful and efficient method for *L*-carnitine production. (The industrial process is described in detail in this issue.) The sequences of the involved enzymes (BcoA/B, BcoC, BcoD) have not been published, but it is known that their genes are directly located next to the ORFs of *L-C*DH on one end and next to a not further characterized transporter protein at the other end in the genome of the strain HK4 (DSMZ 2903), which belongs to a genus related to *Agrobacterium* or *Rhizobium* [59]. Due to the fact that we have identified *L-C*DH by similarity search in several bacteria with a fitting taxonomic profile, we were eager to examine the genetic neighborhood of these candidates. Actually we


araC
Unknown
Thioesterase
proX SBP-ABC
CDH
GBBH
Pseudomonas fluorescens PfO-1
Pseudomonas fluorescens Pf-5
araC
Unknown
Thioesterase
proX SBP-ABC
CDH
P.aeruginosa PAO1 und PA14
P.putida KT2440
P.fluorescens SBW25
P.syringae pv. syringae B728a
araC
Unknown
Thioesterase
proX SBP-ABC
CDH
Esterase/ Lipase
Burkholderia fungorum LB400
Burkholderia mallei ATCC 23344
Burkholderia pseudomallei K96243
Burkholderia thailandensis E264
araC
Unknown
Thioesterase
GBBH
proX SBP-ABC
CDH
Esterase/Lipase
Burkholderia cenocepacia J2315
Burkholderia cepacia R18194
Burkholderia vietnamiensis G4 (B. c. R1808)
Unknown
CDH-Thioesterase fusion protein
dppC Permease B
Lipase/Esterase
araC
SBP-ABC dppA
Permease A dppB
ABC-ATPase dppD+F
Agrobacterium tumefaciens C58 plasmid AT
Unknown
CDH
bcoA/B Acyl-CoA-Synthetase
Lipase/Esterase
araC
Unknown protein
Acyl-CoA-DH bcoC
Enoyl-CoA-Hydratase bcoD
Mesorhizobium loti MAFF303099
Unknown
araC
CDH-Thioesterase fusion protein
Silicibacter pomeroyi DSS-3
araC
Unknown
Thioesterase
Acetoacetyl-CoA Synthase
CDH
Choline/carnitine/betaine transporter protein
R18
Silicibacter (Roseobacter) sp. TM1040 contig 46
Unknown
CDH-thioesterase fusion protein
araC
outer membran lipoprotein
Silicibacter (Roseobacter) sp. TM1040 contig 51
Sinorhizobium meliloti 1021
Unknown
bcoC Acyl-CoA-DH
bcoD Enoyl-CoA-Hydratase
dppB Permease A
dppD+F ABC-ATPase
araC
CDH-Thioesterase fusion protein
Acyl-CoA-Synthetase bcoA/B
SBP-ABC dppA
Permease B dppC
Acylesterase
Rhizobium leguminosarum bv. viciae 3841
Unknown
bcoC Acyl-CoA-DH
bcoD Enoyl-CoA-Hydratase
dppB Permease A
dppD+F ABC-ATPase
araC
CDH-Thioesterase fusion protein
Acyl-CoA-Synthetase bcoA/B
SBP-ABC dppA
Permease B dppC
Acylesterase
Unknown
tetR
CDH-Thioesterase fusion protein
Brevibacterium linens BL2
Unknown
Thioesterase
proW Permease
tetR
CDH
proX SBP-ABC
proV ATPase
Oceanobacillus iheyensis HTE831
Unknown
Thioesterase
opuCB Permease
opuCD Permease
tetR
CDH
opuCA ATPase
opuCC SBP-ABC
GTB Esterase
Staphylococcus epidermidis ATCC 12228
Staphylococcus epidermidis RP62A
Unknown
Thioesterase
tetR
CDH
Multidrug ABC-ATPase-Permease-fusion protein
Streptomyces coelicolor A3

found a group of neighbouring ORFs in the genomes of *Mesorhizobium loti* MAFF303099, *Rhizobium leguminosarum bv. viciae* 3841 and *Sinorhizobium meliloti* 1021, which were arranged in the same order as the genes in the bco-operon with identical functions [59]. The enzymes' ORFs were predicted as acyl-CoA-dehydrogenase, acyl-CoA-synthetase, and enoyl-CoA-hydratase in the sequence annotations. These function assignments are not very specific and indicate an involvement of these enzymes in a fatty acids β-oxidation like reaction. This assumption is true, because their real functions are a γ-butyrobetaine-CoA dehydrogenase (BcoC), a γ-butyrobetaine/crotonobetaine-CoA synthetase (BcoA/B), and a γ-butyrobetaine-CoA hydratase (BcoD) (*cf.* Figs. 1 and 4). To emphasize the trustworthiness of our assertion, we compared the corresponding enzymes from the *E. coli* KT12 cai-operon with those of the predicated bco-operon and thus revealed that caiD possesses a significant sequence similarity to the probable bcoD gene in the α-proteobacterial species (50%, E 10^{-60}). This observation confirms the postulation that the whole reduction reaction of crotonobetaine to γ-butyrobetaine in *E. coli* is irreversible in contrast to the reversible transferase reaction catalyzed by CaiB and the reversible hydration of crotonobetainyl-CoA to *L*-carnitinyl-CoA catalyzed by CaiD [5].

Another interesting detail is that we found an ORF with a predicted acylesterase (COG2936 [61]) activity next to the probable bco-operon in the genomes of *S. meliloti* and *R. leguminosarum*. These genes code for a protein of 75 kDa size and show a sequence similarity of 40% (E 10^{-137}) with a gene found in the fungi *Gibberella zeae* PH-1, also known as *Fusarium graminearum*, a close relative of *Fusarium oxysporum sp. lini*, which selectively hydrolyzes *L*-octanoylcarnitine [62]. So it is very likely that these bacterial genes have an acyl-*L*-carnitine esterase activity, just like a protein found in *Alcaligenes sp.* strain 981 [15]. As stated before we are not sure that the last mentioned strains really belongs to the genus *Alcaligenes*, because the taxonomic classification was performed by a phenotypical characterization many years ago. However, strain 981 (don't mix up with strain 921) expresses the larger variant of *L-C*DH with a molecular weight of approximately 51 ± 6 kDa and a K_m-value for *L*-carnitine of 9.3 in addition to an acylcarnitine esterase with a molecular mass of 63 ± 7 kDa and a substrate specificity to acetyl-, propionyl- and octanoyl-*L*-carnitine. This enzyme has been used to determine acetyl-*L*-carnitine solely or for the determination of acetyl-*L*-carnitine and *L*-carnitine simultaneously [63]. Three different esterases (except the acylesterase), which are located near the predicted *L-C*DHs, leap to the eye by inspecting the

Fig. 4. Genetic neighborhood of predicted *L-C*DHs; the protein functions in this figure have either been taken from genome annotations or have been identified by manually performed sequence similarity search and biological context evaluation; the arrow length is nearly proportional to sequence length, and arrows with identical design indicate significant sequence homology; the arrow orientation is depending on transcription direction and used acronyms are briefly explained in the following abbreviation section; abbreviations: ABC. . .*ATP*-binding cassette, *ara*C. . .Arabinose control gene [77], bco. . .butyrobetaine – *L*-carnitine operon [59], CDH. . .*L*-carnitine dehydrogenase, CoA. . .coenzyme A, dpp. . .dipeptide ABC transporter system [78], *GBB*H. . .γ-butyrobetaine hydroxylase, *GTB*. . .glyceroltributyrate, opu. . .osmoprotectant uptake – betaine ABC transporter system [67], pro. . .glycine betaine/proline betaine ABC transporter system [68], SBP. . .substrate binding protein, *tet*R. . .Tetracycline repressor

genes in Fig. 4. These esterases could be either responsible for the decomposition of *L*-carnitine-ethylester, *L*-carnitine-*n*-propylester [64, 65], or may be correlated to the betaine-ester-hydrolases recently described [66].

Betaine ABC Transporters

Numerous ABC-transport systems and single-component transporters for *L*-carnitine and glycine betaine have been described in bacteria [67, 68]. Their affinity for the different trimethylammonium compounds varies. The fact that different types of this kind of transporters with this particular specificity were found near the ORFs of predicted *L-C*DH genes confirms the probable occurrence of a *L*-carnitine conversion to *GB*, especially in gram-positive bacteria (Fig. 4). In *Brevibacterium linens* the *L*-carnitine uptake system seems not to be induced by *L*-carnitine, since the chloramphenicol addition under hyperosmotic conditions could not prevent the accumulation of *L*-carnitine in contrast to the production of *GB* [22]. Similar tests with *Pseudomonas aeruginosa* resulted in an opposite behavior of the γ-proteobacterium, because the influx of *L*-carnitine in the cell is inhibited due to the elimination of the *ATP* driven uptake [69]. These experimental observations are concordant with the occurrence of a predicted betaine uptake system component (proX) in the genetic neighborhood of the *L-C*DH gene in *Pseudomonas aeruginosa* and the absence of a corresponding ORF in the genome of *Brevibacterium linens* (*cf.* Fig. 4).

Methods

Completely annotated and finished genome sequences as well as drafts can be downloaded from the following websites: "http://www.ncbi.nlm.nih.gov", "http://www.sanger.ac.uk/Projects/Microbes/", and "http://genome.jgi-psf.org/microbial/index.html". Preliminary sequence data from *Pseudomonas fluorescens* Pf-5 and *Silicibacter pomeroyi* DSS-3 were obtained from The Institute for Genomic Research through the website at "http://www.tigr.org". Sequence similarity searches were performed either with the help of the online BLAST-server at the *NCBI*-website ("www.ncbi.nlm.nih.gov/BLAST/") or with locally installed stand-alone binaries of BLAST downloaded from "ftp://ftp.ncbi.nih.gov/blast/executables/". ORF predictions and annotations have been obtained from the websites at "http://pedant.gsf.de" [70] or "http://genome.ornl.gov/microbial/". In case that an annotated genomic sequence was missing its ORFs have been manually determined by means of the computer programs *Glimmer2* [71] or *GeneMarkS* [72]. Their success rate is reported to be more than 97% for the correct identification of prokaryotic ORFs. Context-based prediction has been performed manually and the results were compared with the results of the online database *String* [48].

Abbreviations

ATCC. . .American type culture collection, BLAST. . .basic local alignment search tool, *L-C*DH. . .*L*-carnitine dehydrogenase, COG. . .cluster of orthologous groups of

proteins, DSMZ...Deutsche Sammlung von Mikroorganismen und Zellkulturen, EC...enzyme commission, *GB*...glycine betaine, *GBB*(H)...γ-butyrobetaine (hydroxylase), kDa...kiloDalton, K_m...*Michaelis* constant, MSA...multiple sequence alignment, MW...molecular weight, *NAD*H...nicotinamide-adenine dinucleotide, ORF...open reading frame, PFAM...protein families database of alignments and hidden *Markov* models.

Acknowledgements

We are grateful to *Gisela Pittner* and *Irene Maier* for critical reading of the manuscript. Sequencing of *Pseudomonas fluorescens* Pf-5 was accomplished with support from USDA. Information about funding, performing institutions, and status of ongoing and finished genome sequencing projects can be obtained on the website of the Genomes Online Database (GOLD): ''http://www.genomesonline.org'' [47].

References

[1] For reviews see: a) Kleber HP (1997) FEMS Microbiol Lett **147**: 1; b) Rebouche CJ, Seim H (1998) Annu Rev Nutr **18**: 39
[2] For reviews see: a) Naidu GSN, Lee IY, Lee EG, Kang GH, Park YH (2000) Bioprocess Eng **23**: 627; b) Jung H, Jung K, Kleber HP (1993) Adv Biochem Eng Biotech **50**: 21
[3] Eichler K, Bourgis F, Buchet A, Kleber HP, Mandrand-Berthelot MA (1994) Mol Microbiol **13**: 775
[4] Elssner T, Hennig L, Frauendorf H, Haferburg D, Kleber HP (2000) Biochemistry **39**: 10761
[5] Elssner T, Engemann C, Baumgart K, Kleber HP (2001) Biochemistry **40**: 11140
[6] Jung H, Buchholz M, Clausen J, Nietschke M, Revermann A, Schmid R, Jung K (2002) J Biol Chem **277**: 39251
[7] Preusser A, Wagner U, Elssner T, Kleber HP (1999) Biochim Biophys Acta **1431**: 166
[8] Stenmark P, Gurmu D, Nordlund P (2004) Biochemistry **43**: 13996
[9] Information about EC numbers is available at http://www.chem.qmul.ac.uk/iubmb/enzyme/
[10] Lindstedt G, Lindstedt S, Midtvedt T, Tofft M (1970) J Bacteriol **101**: 1094
[11] Vandecasteele JP (1980) Appl Environ Microbiol **39**: 327
[12] Hanschmann H, Ehricht R, Kleber HP (1996) Biochim Biophys Acta **1290**: 177
[13] Setyahadi S, Harada E, Mori N, Kitamoto Y (1998) J Mol Catal B-Enzym **4**: 205
[14] Houriyou K, Takahashi M, Mizoguchi J, Imamura S (1991) Japan patent 1993161492-A/1
[15] Ueda S, Takahashi M (1992) Europe patent 0480676
[16] Goulas P (1988) Biochim Biophys Acta **957**: 335
[17] Aurich H, Kleber HP, Sorger H, Tauchert H (1968) Eur J Biochem **6**: 196
[18] Mori N, Kasugai T, Kitamoto Y, Ichikawa Y (1988) Agric Biol Chem **52**: 249
[19] Grimont PAD, Vancanneyt M, Lefevre M, Vandemeulebroecke K, Vauterin L, Brosch R, Kersters K, Grimont F (1996) System Appl Microbiol **19**: 510
[20] Goldmann A, Boivin C, Fleury V, Message B, Lecoeur L, Maille M, Tepfer D (1991) Mol Plant Microbe Interact **4**: 571
[21] Kim S-G, Bae H-S, Lee S-T (2001) Arch Microbiol **176**: 271
[22] Jebbar M, Champion C, Blanco C, Bonnassie S (1998) Res Microbiol **149**: 211
[23] Serra AL, Mariscotti JF, Barra JL, Lucchesi GI, Domenech CE, Lisa AT (2002) J Bacteriol **184**: 4301
[24] Smith LT, Pocard JA, Bernard T, Le Rudulier D (1988) J Bacteriol **170**: 3142
[25] Oka I, Yoshimoto T, Rikitake K, Ogushi S, Tsuru D (1979) Agric Biol Chem **43**: 1197
[26] Bater AJ, Venables WA (1977) Biochim Biophys Acta **468**: 209

[27] Incharoensakd A, Waditee R (2000) Curr Microbiol **41**: 227
[28] Meskys R, Harris RJ, Casaite V, Basran J, Scrutton NS (2001) Eur J Biochem **268**: 3390
[29] Schalkhammer T, Pittner F, Lepek M, Mayer C, Vater C, Dietrich-Ostry C, Lohninger A, Piller A (1999) US patent US2003/162241A1
[30] Manjón A, Obón JM, Iborra JL (2000) Anal Biochem **281**: 176
[31] Mori N, Mitsuzumi H, Kitamoto Y (1994) J Ferment Bioeng **78**: 337
[32] Mori N, Mitsuzumi H, Kitamoto Y (1993) J Ferment Bioeng **76**: 395
[33] Peluso G, Barbarsi A, Savica V, Reda E, Nicolai R, Benatti P, Calvani M (2000) J Cell Biochem **80**: 1
[34] For reviews see: a) Csonka LN, Hanson AD (1991) Annu Rev Microbiol **45**: 569; b) Poolman B, Glaasker E (1998) Mol Microbiol **29**: 397
[35] Baliarda A, Robert H, Jebbar M, Blanco C, Deschamps A, Le Marrec C (2003) Int J Food Microbiol **84**: 13
[36] Vilhelmsson O, Miller KJ (2002) J Food Prot **65**: 1008
[37] Lucchesi GI, Lisa TA, Casale CH, Domenech CE (1995) Curr Microbiol **30**: 55
[38] Velasco-García R, Mújica-Jiménez C, Mendoza-Hernández G, Muñoz-Clares RA (1999) J Bacteriol **181**: 1292
[39] Kolak M, Karpati F, Monstein HJ, Jonasson J (2003) Int J Med Microbiol **293**: 309
[40] Panter RA, Mudd JB (1969) FEBS Lett **5**: 169
[41] Burnet MW, Goldmann A, Message B, Drong R, El Amrani A, Loreau O, Slightom J, Tepfer D (2000) Gene **244**: 151
[42] Wu G, Fiser A, ter Kuile B, Sali A, Muller M (1999) Proc Natl Acad Sci USA **96**: 6285
[43] For review see: Whisstock JC, Lesk AM (2003) Q Rev Biophys **36**: 307
[44] For review see: Gabaldón T, Huynen MA (2004) Cell Mol Life Sci **61**: 930
[45] Devos D, Valencia A (2001) Trends Genet **17**: 429
[46] Fleischmann RD, Adams MD et al (1995) Science **269**: 496
[47] Bernal A, Ear U, Kyrpides N (2001) Nucleic Acids Res **29**: 126
[48] Von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snell B (2003) Nucleic Acids Res **31**: 258
[49] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res **25**: 3389
[50] Bossis E, Lemanceau P, Latour X, Gardan L (2000) Agronomie **20**: 51
[51] Yabuuchi E, Oyaizu H, Yano I, Hotta H, Hashimoto Y, Ezaki T, Arakawa M (1992) Microbiol Immunol **36**: 1251 and erratum in Microbiol Immunol **37**: 335 (1993)
[52] Regenhardt D, Heuer H, Helm S, Fernandez C, Strömpl E, Moore ERB, Timmis KN (2002) Environ Microbiol **4**: 912
[53] http://www.dsmz.de/download/bactnom/bactname.pdf
[54] Yamamoto S, Kasai H, Arnold DL, Jackson RW, Vivian A, Harayama S (2000) Microbiology **146**: 2385
[55] For review see: Coenye T, Vandamme P (2003) Environ Microbiol **5**: 719
[56] Alavi M, Miller T, Erlandson K, Schneider R, Belas R (2001) Environ Microbiol **3**: 380
[57] Takami H, Takaki Y, Uchiyama I (2002) Nucleic Acids Res **30**: 3927
[58] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe Kl, Marshall M, Sonnhammer ELL (2002) Nucleic Acids Res **30**: 276
[59] Zimmermann T, Werlen J (1998) US patent 5759824
[60] Rüetschi U, Nordin I, Odelhög B, Jörnvall H, Lindstedt S (1993) Eur J Biochem **213**: 1075
[61] Tatusov RL et al (2003) BMC Bioinformatics **4**: 41
[62] Aragozzini F, Manzoni M, Cavazzoni V, Craveri R (1986) Biotechnol Lett **8**: 95
[63] Tomita K, Sakurada S, Minami S (2001) J Pharm Biomed Anal **24**: 1147
[64] Strack E, Aurich H, Grüner E (1964) Z Allg Mikrobiol **4**: 154

[65] Kiyoshi N, Yukie O, Akiko M, Tetsuo O, Tatsuya O (1987) Japan patent 62-134092
[66] Tscherry B, Bornscheuer U, Musidlowska A, Werlen J, Zimmermann T (2002) Patent WO 02/061094 A2
[67] Angelidis A, Smith GM (2003) Appl Environ Microbiol **69**: 1013
[68] Schiefner A, Breed J, Bösser L, Kneip S, Gade J, Holtmann G, Diederichs K, Welte W, Bremer E (2004) J Biol Chem **279**: 5588
[69] Kleber HP, Aurich H (1967) Biochem Biophys Res Comm **26**: 255
[70] Frishman D et al (2003) Nucleic Acids Res **31**: 207
[71] Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Nucleic Acids Res **27**: 4636
[72] Besemer J, Lomsadze A, Borodovsky M (2001) Nucleic Acids Res **29**: 2607
[73] Saitou N, Nei M (1987) Mol Biol Evol **4**: 406
[74] Kumar S, Tamura K, Jakobsen IB, Nei M (2001) Bioinformatics **17**: 1244
[75] Lesk AM (1995) Curr Opin in Struct Biol **5**: 775
[76] Thompson JG (1997) Nucleic Acids Res **24**: 4876
[77] Martin RG, Rosner JL (2001) Curr Opin Microbiol **4**: 132
[78] Nickitenko AV, Trakhanov S, Quiocho FA (1995) Biochemistry **34**: 16585