CMLS, Cell. Mol. Life Sci. 61 (2004) 97–109
1420-682X/04/010097-13
DOI 10.1007/s00018-003-3334-y

CMLS Cellular and Molecular Life Sciences

Research Article

Horizontal gene transfer from Eukarya to Bacteria and domain shuffling: the α-amylase model

J.-L. Da Lage[a,*], G. Feller[b] and Š. Janeček[c]

[a] UPR 9034 Populations, Génétique et Évolution, C.N.R.S., 91198 Gif sur Yvette cedex (France), Fax: +33 1 69 07 04 21, e-mail: jldl@pge.cnrs-gif.fr
[b] Laboratoire de Biochimie, Institut de Chimie B6, Université de Liège, 4000 Liège-Sart Tilman (Belgium)
[c] Institute of Molecular Biology, Slovak Academy of Sciences, Dùbravska cesta 21, 84551 Bratislava (Slovakia)

Received 17 October 2003; accepted 6 November 2003

Abstract. α-Amylases are present in all kingdoms of the living world. Despite strong conservation of the tertiary structure, only a few amino acids are conserved in interkingdom comparisons. Animal α-amylases are characterized by several typical motifs and biochemical properties. A few cases of such α-amylases have been previously reported in some eubacterial species. We screened the bacterial genomes available in the sequence databases for new occurrences of animal-like α-amylases. Three novel cases were found, which belong to unrelated bacterial phyla: *Chloroflexus aurantiacus*, *Microbulbifer degradans*, and *Thermobifida fusca*. All the animal-like α-amylases in Bacteria probably result from repeated horizontal gene transfer from animals. The *M. degradans* genome also contains bacterial-type and plant-type α-amylases in addition to the animal-type one. Thus, this species exhibits α-amylases of animal, plant, and bacterial origins. Moreover, the similarities in the extra C-terminal domains (different from both the α-amylase domain C and the starch-binding domain), when present, also suggest interkingdom as well as intragenomic shuffling.

Key words. Horizontal gene transfer; α-amylase; C-terminal domain; *Microbulbifer degradans.*

α-Amylases (α-1,4-glucan-4-glucanohydrolases, EC 3.2.1.1) are ubiquitous enzymes synthesized by animals, microorganisms and plants that catalyze the hydrolysis of internal α-(1-4)-glycosidic bonds in starch, glycogen and related oligosaccharides. These enzymes display strong conservation of their overall conformation, as shown by the numerous X-ray structures available, but amino acid sequence identity remains below 10% between the main groups of organisms, mostly concentrated in a few, well-defined regions [1, 2]. However, within-kingdom comparisons show much higher similarity. In animals, for example, all the sequences known to date are alignable manually since they share at least 40% identity, including specific stretches that are absent from other non-animal α-amylases [3, 4].

A few Bacteria have been known for a number of years to have α-amylases that exhibit high sequence similarity with their animal counterparts, remarkably higher than with any other bacterial α-amylase. In addition, these sequences share the typical animal motifs [5]. This was first reported in the actinobacterium, *Streptomyces limosus* [6, 7]. Several other cases were reported from unrelated bacterial phyla. These Bacteria are actinomycetes – *Streptomyces*, *Thermomonospora* [8–10], firmicutes – *Bacillus* sp. no. 195- [11], and γ-proteobacteria – *Halomonas meridiana* [12], *Pseudoalteromonas haloplanktis* [13]. The α-amylase of the Gram-negative Antarctic bacterium *P. haloplanktis* (AHA) has been studied extensively

* Corresponding author.

[13, 14], so that its amino acid sequence, three-dimensional structure and biochemical properties were found to be typical of animal α-amylases [13–17]. The animal α-amylases are characterized by the strong conservation of sequence motifs bearing the catalytic, substrate-binding, and calcium-binding residues, and mainly the ligands of chloride which is an allosteric activator of these enzymes [2, 14, 18, 19].

In all organisms, α-amylase is made of three domains: domain A is the catalytic domain, shaped as a $(\beta/\alpha)_8$ barrel (TIM barrel). Domain B is a long loop between β_3 and α_3, and domain C is a C-terminal β sandwich (greek key) [20–22]. In addition, in a number of Bacteria, extra C-terminal domains are found, mainly carbohydrate-binding modules (CBMs), which are often, but not always, starch-binding domains of the CBM-20 family (CAZy database; http://afmb.cnrs-mrs.fr/CAZY/). These CBMs may be involved in substrate specificity, and they may be present in other glycoside hydrolases [23–25]. Surprisingly, the precursor of the bacterial α-amylase from *P. haloplanktis* possesses an extra domain at the C terminus (hereafter called the AHA C-terminal domain), involved in membrane anchoring and spanning [26]. This is of particular interest because this domain is unrelated to any CBM, and seems to have a very different function. On the other hand, most animal α-amylases lack C-terminal domains succeeding domain C. These observations prompted us to screen the increasing data of bacterial genomes in data banks, to search for other occurrences of animal-type α-amylases and to estimate their frequency.

Materials and methods

To detect sequences similar to animal α-amylases in bacterial genomes, we screened the completed or unfinished genomes (including Archaea) available through GenBank [27] (release April 2003) with the TBLASTN tool [28], using the *Drosophila melanogaster* α-amylase protein sequence (GenBank: X04569) [29] as a query. A cutoff expect value of e^{-70} was chosen. Positive results were, in turn, screened against GenBank using BLASTP, to find out the best hit with a true animal α-amylase. The AHA C-terminal domain (GenBank: X58627 [26], residues 472–669) was also searched in the same data banks. Percentages of pairwise identity between α-amylase protein sequences were estimated with the BLAST2 program [30], with no filtering and the BLOSUM62 matrix.

Alignments were done either with the MAP program [31] at the Baylor College of Medicine HGSC server (http://www.hgsc.bcm.tmc.edu/), or CLUSTALW [32]. For the global tree of α-amylases in living organisms, a set of α-amylases retrieved from GenBank [27] and SwissProt [33], representing the individual living kingdoms, was constructed (table 1) and their amino acid sequences were aligned by the CLUSTALW program as follows: (i) the best conserved regions {β_1, β_2, β_3, β_4, β_5, β_7 and β_8 of the catalytic $(\beta/\alpha)_8$ barrel and the region V in domain B [34]} were identified in each sequence; (ii) the segments preceding and succeeding the regions β_1 and β_8, respectively, were cut off; (iii) the shortened sequences were aligned by CLUSTALW; (iv) the identified conserved sequence regions were adjusted manually, if necessary; and (v) the remaining parts of the alignment (between the regions) were manually tuned where applicable. The alignment procedure is helped by the conservation of the three-dimensional (3D) structure in all the organisms: Since at least one α-amylase in each kingdom has been studied by crystallography, helices and sheets may be easily identified and superimposed. The whole alignment is available on the website http://imb.savba.sk/~janecek/Papers/HGT-Micde/. The global tree was built using this alignment by

Table 1. The α-amylases used in the present study for alignment and tree reconstruction.

Alpha-amylase	Abbreviation	GenPept*	Length†	C terminus
Bacteria				
Actinoplanes sp. SE50	Acpsp	CAC02970.1	1021	
Aeromonas hydrophila	Aemhy	AAA21936	464	
Bacillus amyloliquefaciens	Bacam	AAA22191.1	514	
Bacillus licheniformis	Bacli	CAA26981.1	512	
Bacillus stearothermophilus	Bacst	AAA22235.2	549	
Bacillus sp. No. 195	Bacsp	BAA22082.1	700	CBM-25
Bacillus subtilis	Bacsu	CAA23437.1	660	
Clostridium acetobutylicum	Cloac	AAD47072.1	760	
Chloroflexus aurantiacus	Chlau	ZP_00017646.1	575	CMB-20
Escherichia coli CFT 073	Escco	AAN82828.1	676	
Halomonas meridiana	Halme	CAB92963.1	457	
Micrococcus sp. 207	Micsp	CAA39321.1	1104	
Novosphingobium aromaticivorans	Nspar	ZP_00094545.1	617	
Pseudoalteromonas haloplanktis	Psaha	CAA41481.1	669	Extra C
Pseudomonas sp. KFCC10818	Psesp	AAA86835.1	563	
Salmonella typhimurium	Salty	AAL22523.1	675	

Table 1 (continued)

Alpha-amylase	Abbreviation	GenPept*	Length†	C terminus
Shigella flexneri	Shifl	AAN45063.1	676	
Streptococcus bovis	Stcbo	AAA97431.1	485	
Streptococcus mutans	Stcmu	AAC35010.1	486	
Streptomyces coelicolor	Stmco	CAB88153.1	506	
Streptomyces lividans TK-24	Stmld	CAA49759.1	919	
Streptomyces limosus	Stmli	AAA88554.1	566	CBM-20
Streptomyces thermoviolaceus	Stmth	AAA26697.1	460	
Thermoactinomyces vulgaris	Thavu	CAA49465.1	482	
Thermomonospora curvata	Thscu	CAA41881.1	605	CBM-20
Thermotoga maritima	Thtma	CAA72194.1	553	
Vibrio cholerae	Vibch	AAF96758.1	690	
Xanthomonas campestris	Xamca	AAA27591.1	475	
Yersinia pestis	Yerpe	AAM87640.1	687	
Archaea				
Pyrococcus furiosus	Pycfu	AAB67705.1	460	
Pyrococcus sp. KOD1	Pycsp	BAA21130.1	461	
Thermococcus hydrothermalis	Thchy	AAC97877.1	457	
Thermococcus sp. AEPII 1a	Thcsp-AEP	AAM48113.1	461	
Thermococcus sp. Rt3	Thcsp-Rt3	AAB87860.1	469	
Fungi and yeasts:				
Aspergillus kawachii	Aspka	BAA22993.1	640	CBM-20
Aspergillus niger	Aspni	P56271	484	
Aspergillus oryzae	Aspor	AAA32708.1	499	
Cryptococcus sp. S-2	Crcsp	BAA12010.1	631	CBM-20
Lipomyces kononenkoae (Amy1)	Limko-1	AAC49622.1	570	
Saccharomycopsis fibuligera	Samfi	CAA29233.1	494	
Schwanniomyces occidentalis	Schoc	CAA34162.1	512	
Plants				
Arabidopsis thaliana (mouse-ear cress)	Arath	AAM64582.1	423	
Avena fatua (oat)	Avefa	CAA09323.1	434	
Hordeum vulgare (barley – high pI)	Horvu-H	AAA98790.1	427	
Hordeum vulgare (barley – low pI)	Horvu-L	AAA32929.1	438	
Ipomoea nil (morning glory)	Iponi	BAC02435.1	424	
Malus domestica (apple)	Maldo	AAF63239.1	413	
Musa acuminata (banana)	Musac	AAO11776.1	416	
Oryza sativa (rice)	Orysa	AAA33885.1	434	
Phaseolus vulgaris (kidney bean)	Phavu	BAA33879.1	420	
Solanum tuberosum (potato)	Soltu	AAA91884.1	407	
Triticum aestivum (wheat)	Triae	AAA34259.1	413	
Vigna mungo (black gram)	Vigmu	CAA51734.1	421	
Zea mays (maize)	Zeama	AAA50161.1	439	
Animals				
Aedes aegypti (yellow fever mosquito)	Aedae	AAB60934.1	486	
Apis mellifera (honey bee)	Apime	AAM20738.1	493	
Caenorhabditis elegans (nematode)	Caeel	CAB02856.1	713	extra C
Corbicula fluminea (asian clam)	Corfl	AF468016	699	extra C
Drosophila melanogaster (fruit fly)	Drome	CAA28238.1	494	
Euroglyphus maynei (mite)	Eurma	AAD38943.1	521	
Gallus gallus (chicken)	Galga	AAC60246.1	512	
Homo sapiens (human, saliva)	Homsa	AAA52279.1	511	
Litopenaeus vannamei (white shrimp)	Penva	CAA54524.1	512	
Pseudopleuronectes americanus (flounder)	Pspam	AAF65827.1	512	
Rattus norvegicus (rat, liver)	Ratno	BAB39466.1	521	
Sus scrofa (pig, pancreas)	Sussc	AAF02828.1	511	
Tenebrio molitor (yellow meal worm)	Tenmo	P56634	471	
Three different *Microbulbifer* proteins				
Microbulbifer degradans (bacterial-like)	Mibde-B	ZP_00065699.1	566	
Microbulbifer degradans (plant-like)	Mibde-P	ZP_00065690.1	643	extra C
Microbulbifer degradans (animal-like)	Mibde-A	ZP_00066069.1	563	CBM-20

* The accession numbers are the GenPept numbers form GenBank, except for the SwissProt accession number of the enzymes from *Aspergillus niger* and *Tenebrio molitor*.

† The length concerns the entire length of the precursor.

the neighbor-joining method [35] and drawn with TreeView [36].

Results

The global tree (fig. 1) shows that bacterial α-amylases are scattered in several clusters, according to their overall sequence similarities. Some of these clusters branch with eukaryote kingdoms. In figure 2, a selection of α-amylases of species from the different kingdoms of the living world are aligned, along with Bacteria clustered to them. Kingdom-specific stretches are highlighted. In this study, beside the overall sequence similarity, which is estimated by both percentage of identity and expect value, these stretches are considered as signatures of α-amylase types in Bacteria (animal-type, plant-type and so on). In the next section, we focus on bacterial α-amylases with high similarity to animals, which were at the origin of our investigations.

Animal-like α-amylases in Bacteria and the case of *Microbulbifer degradans*

The results of the BLAST searches with the *Drosophila* sequence were first inspected for the presence of typical animal motifs. The most remarkable animal motifs searched are highlighted in red in figure 2. An additional motif (CEHRW), more downstream, is not shown. The results of the search are summarized in figure 3. We found new animal-like α-amylases in three species: the actinomycete *Thermobifida fusca* (which is very closely related to the known *Thermomonospora curvata*), the ther-

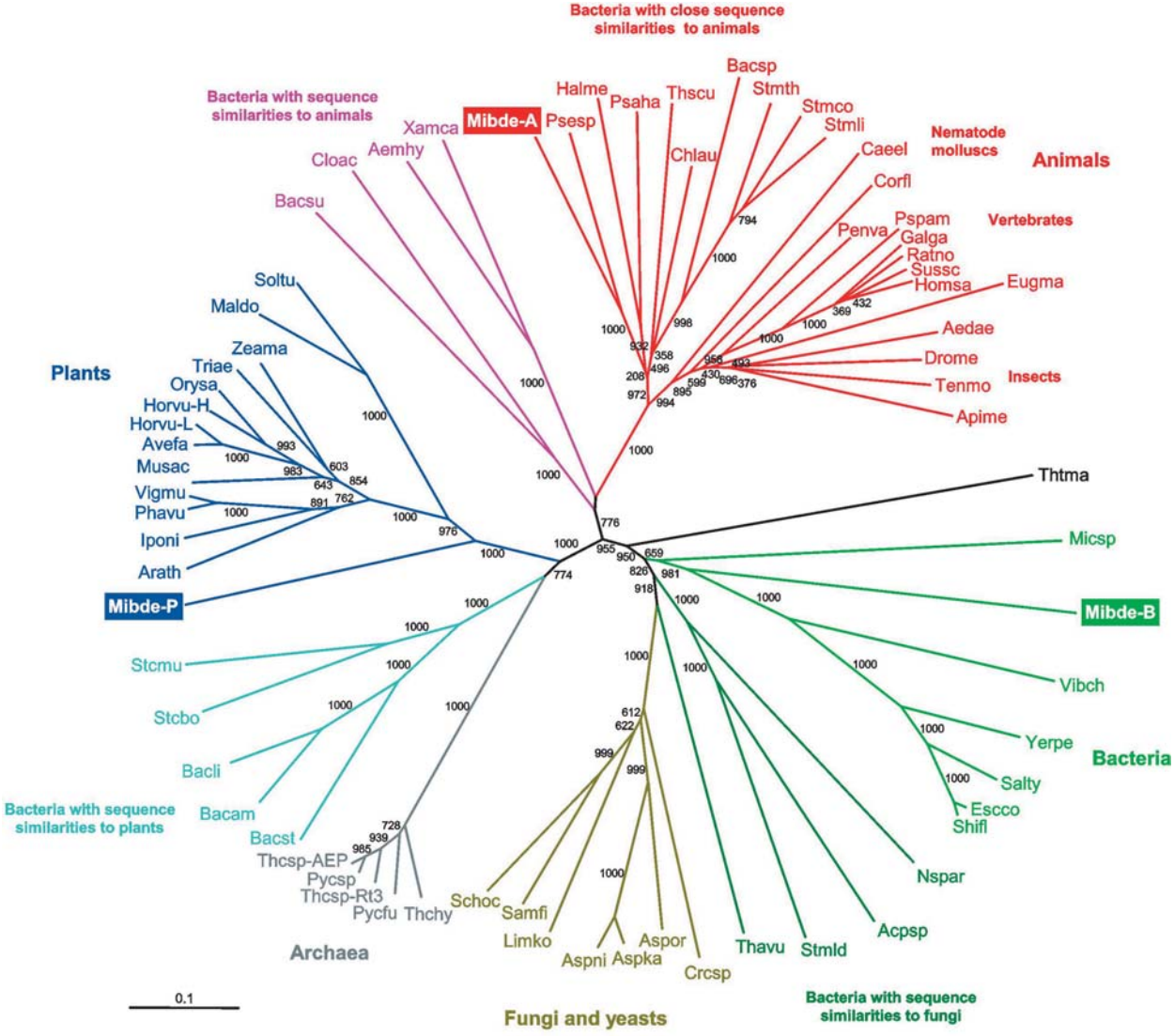

Figure 1. Unrooted (circle) bootstrap tree of α-amylases from the different living kingdoms. Species, abbreviations and accession numbers are indicated in table 1. The three α-amylases of *Microbulbifer degradans* are boxed at the tip of dashed branches. The alignment method used for drawing the tree is described in the text.

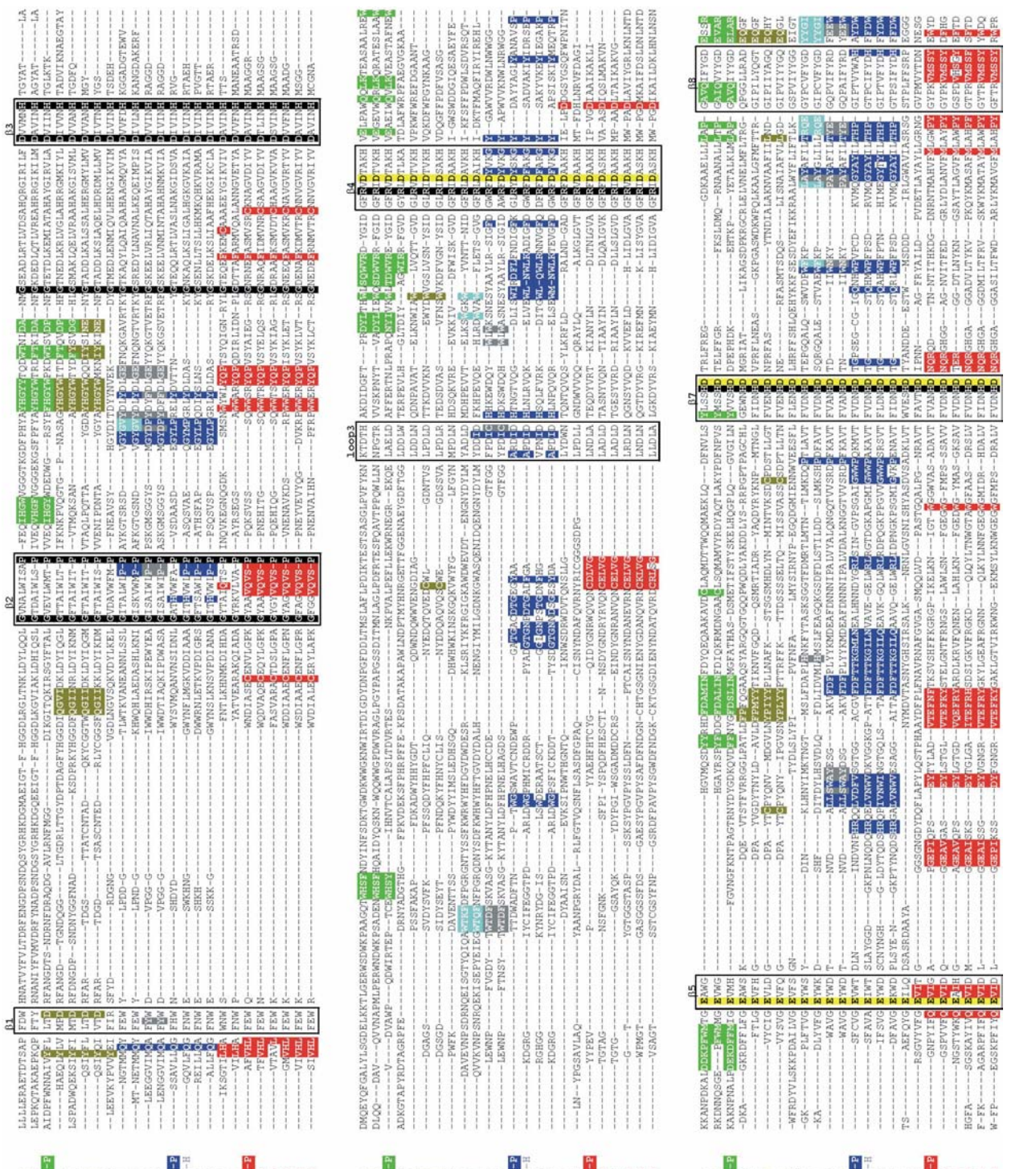

Figure 2. Alignment of α-amylase sequences of selected species representing the diverse kingdoms, along with bacterial sequences with similarities to these sequences. Highlighted residues: yellow, catalytic triad; black, invariant residues; light green, bacteria-like; gold, fungi-like; turquoise, liquefying- (intracellular)-like; gray, Archaea-like; blue, plant-like; red, animal-like. Abbreviations are as in table 1.

Taxon	Species	Animal type: Animal species' best match (E value)	Animal type: C-term. domain	Plant type: Plant species' best match (E value)	Plant type: C-term. domain	Bacterial type: Bacterial species' best match	Bacterial type: C-term. domain
Gamma Proteobacteria	*Microbulbifer degradans*	Drosophila U53478, e-103 ZP_00066069	CBM20	Hordeum J04202, e-83 ZP_00065690	AHA	Alkaliphilic Eubact. X53373, e-84 ZP_00065699	
	Pseudoalteromonas haloplanktis*	Sus scrofa AF064742, e-109 X58627	AHA	NA		NA	
	Halomonas meridiana*	Aedes U01209, e-107 AJ239061		NA		NA	
	Pseudomonas sp.* KFCC10818	Diabrotica AF208003, e-102 AF333075	**	NA		NA	
Actinobacteria	*Thermobifida fusca*	Apis AF259649, 2.e-97 ZP_00058434	CBM20	no		no	
	Streptomyces coelicolor	Drosophila AF136603, 2.e-99 AL352956	CBM20(partial?)	no		no	
	Streptomyces limosus*	Apis AF259649, 5.e-95 M18244	CBM20	NA		NA	
Green non sulfur	*Chloroflexus aurantiacus*	Drosophila U53699, e-124 ZP_00017646	CBM20	no		suspect of contamination	
Firmicutes	*Bacillus sp.** n°195	Apis AF259649, 5.e-88 AB006823	CBM25 (x2)	NA		NA	

Figure 3. Types of α-amylase genes found in Bacteria harboring non-bacterial α-amylases. For Bacteria whose genomes have already been sequenced, animal-type genes were searched with TBLASTN using *D. melanogaster* α-amylase (GenBank: X04569) as query; plant-type genes were searched with *Hordeum vulgare* AMYB (SwissProt: P04063); bacterial-type were searched with both *Bacillus subtilis* (SwissProt: P00691) and *Escherichia coli* (GenBank: AAC76595). White boxes, animal type; light-gray boxes, plant-type; dark-gray boxes, bacterial type. Under the boxes: accession numbers of the putative α-amylases found in bacteria; inside the boxes: best hits in reciprocal BLASTP search. *: no genome sequencing project available; AHA: C-terminal domain of *Pseudoalteromonas haloplanktis* or similar; **: C-terminal tail not identified, but also exists in *Microbulbifer* genome; NA: not available.

mophilic green non-sulfur bacterium *Chloroflexus aurantiacus*, and the γ-proteobacterium *Microbulbifer degradans*. The expect values are very low (i.e., high exponent). The best eukaryote hits in GenBank using these putative proteins as BLASTP queries are always insects, which is probably due in part to the high representation of insect α-amylase sequences in the database. All these putative α-amylases found in Bacteria lack a stretch of nine amino acids in a loop typical of Vertebrates and some non-insect α-amylases [4]. In the three Bacteria, the animal-like protein ends with a starch-binding domain (classified as the CBM-20), which is known to be present in some other bacterial α-amylases, and also in β-amylases and glucoamylases [23, 24]. As mentioned above, the well-studied AHA harbors a very different C-terminal domain, which had no counterpart until now in bacterial α-amylases, but which was found by our screening in two animals: the nematode *Caenorhabditis elegans* (acc. No. CAB02856) and the freshwater bivalve *Corbicula fluminea* (new data from this study, AF468016). We screened the bacterial genomes with this domain (sequence from *P. haloplanktis*). Surprisingly, a similar domain was found only once, in the C-terminal position of a putative protein of *M. degradans*. This protein appears to be an α-amylase, too, but of a plant-type, most similar to that of barley (figs 1, 3). We thus checked for other possible occurrences of plant-type α-amylase in Bacteria, using the barley α-amylase as a query. No other convincing plant-type α-amylase was found in the bacterial genomes (see also below). The AHA C-terminal domains were aligned with the MAP program (fig. 4). The alignment shows that similarities are stronger between the two animals on the one hand, and between the two Bacteria on the other (shared indels), which suggests a common history within each group.

As shown in figure 1, a single 'bacterial type' of α-amylase cannot be simply defined. However, we had also to assess whether bacterial-type α-amylase genes coexist in those Bacteria with non-bacterial α-amylase. We screened the genomes of *C. aurantiacus*, *M. degradans*, *T. fusca*, and *S. coelicolor* with two very different α-amylase sequences from *Escherichia coli* K12 (GenBank: AAC76595) and *Bacillus subtilis* (SwissProt: P00691). The only significant hit was for *M. degradans*, with *E. coli* as a query (*E. coli* and *M. degradans* are both γ-proteobacteria). Unfortunately, there are no large genome data for Bacteria such as *P. haloplanktis* and other previously detected cases of animal-type α-amylase, so that checking whether these species do have a bacterial-type α-amylase was not possible.

Thus, *M. degradans* exhibits a conspicuous composite in the α-amylase family made of three α-amylases, one of

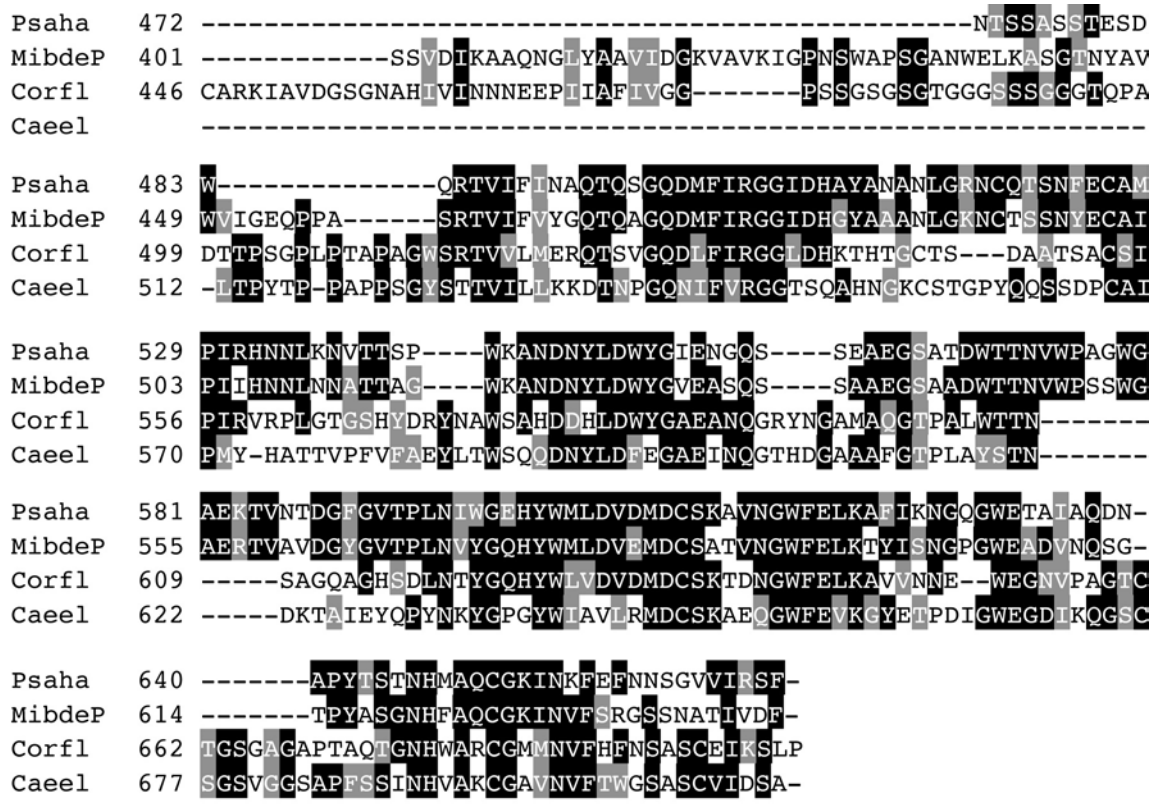

Figure 4. Alignment of the C-terminal domains of the AHA type, made with the MAP program, and adjusted by hand. Black background, identical residues (50% consensus); gray background, similar residues. Abbreviations are as in table 1.

bacterial-type, one of animal-type (with a bacterial C-terminal starch-binding domain), and one of plant-type (with a bacterial/animal C-terminal extra domain).

Given this unexpected composition of the α-amylase family, we checked whether other non-bacterial α-amylases (i.e., fungal) were present in the bacterial genomes investigated. We used the *Aspergillus niger* sequence (GenBank: A35282). The result was negative with our cutoff value.

Other clusterings of bacterial α-amylases with eukaryotes

In addition to the clear clustering of the species analyzed above with animal α-amylases, figure 1 suggests some relationships of various α-amylases of Bacteria or Archaea with either plants or animals or fungi. This prompted us to check these branchings more carefully. As shown above, the only clear case of plant-type α-amylase in Bacteria is in *M. degradans*. However, some Bacteria of the bacilli group of firmicutes and a few Archaea closely related to each other cluster with plants. But there are few conserved motifs (highlighted in blue in fig. 2), compared to *M. degradans*. Moreover, the expect values in BLASTP are low. The best plant hit with *B. stearothermophilus* as a query has an expect value of $2.e^{-18}$. With the Archaeon *Pyrococcus furiosus,* the value is $2.e^{-28}$. Note that the other archaeal α-amylases known to date belong to another glycoside hydrolase family (GH57) and are therefore not considered here.

Connected to the well-established branch of animal and animal-like α-amylases, some other Bacteria from several taxa (firmicutes and γ-proteobacteria) form a loose branch. Here again, the conserved regions are scarce and the BLAST scores with animals are low: $3.e^{-22}$ with *B. subtilis* as a query; e^{-28} with *Xanthomonas campestris*. In addition, some typical animal motifs, such as VMSSY and CEHRW are absent.

Another loose branch of Bacteria is connected to Fungi. Using *Thermoactinomyces vulgaris* as a query, the best fungal hit (*Aspergillus*) has an expect value of e^{-71}. In this case, a real relationship is probable. On the other hand, the lower branches in this cluster are more questionable: the best fungal hit with *S. lividans* as a query is *Schizosaccharomyces pombe* (e^{-35}).

For a more comprehensive overview, we searched for similarities in protists. The protist world is very large highly diverse. Data are availabe for a few sequenced organisms, mainly parasitic. *Giardia lamblia* seems to have no α-amylase at all, as also may be the case for *Plasmodium falciparum*. Interestingly, in the free-living ciliate *Paramecium aurelia* (data available at the *P. aurelia* Genome Project at Genoscope: http://www.genoscope.cns.fr and http://paramecium. cgm.cnrs-gif.fr/ptblast), a putative α-amylase has been found, with good sequence similarity to Fungi (best hit with *A. nidulans*, e^{-61}). Another possible α-amylase could have some similarity with animals (best hit with *Drosophila*, $5.e^{-41}$), but the regions typical of animals are not well conserved, which casts doubt on a real relationship.

Discussion

Unrelated Bacteria have animal α-amylases

We found in the bacterial genome databases several new occurrences of animal-type α-amylase in bacterial species. Several points are of interest. The first is that the bacterial species, which harbor animal-type α-amylase, belong to more or less related phyla. Indeed, in some cases, they are not related at all, from both taxonomical and ecological points of view. Related taxa are, for example, a number of species of the single genus *Streptomyces* with an animal α-amylase. At a broader taxonomical level, several γ-proteobacteria from different families also have this type of enzyme (*P. haloplanktis*, *M. degradans*, *H. meridiana*, *Pseudomonas* sp. KFCC10818). *P. haloplanktis* and *M. degradans* are classified in the same family Alteromonadaceae. On the other hand, for example, *C. aurantiacus* is a green non-sulfur bacterium. *Bacillus* sp. No. 195 belongs to the firmicutes. The tree in figure 1 shows the relationship between the *Streptomyces* species α-amylases. Unexpectedly, the sequence from *Bacillus* sp. No. 195 (phylum firmicutes) is branched with them. Although the bootstrap value is low in this tree, a tree reconstruction made from an alignment of the complete sequences gives very high bootstrap values and a robust clustering with the other actinomycetes *T. fusca* and *T. curvata* (not shown). On the other hand, the relationships among γ-proteobacterial α-amylases are not clear. In addition, true animal α-amylases remain clustered together.

The bacterial species harboring animal-type α-amylases live in very different ecological conditions. *Streptomyces* species are soil bacteria; *T. fusca* and *T. curvata* are moderate thermophilic actinomycetes from composting plant material. *Pseudomonas* sp. KFCC1818 is an alkalophile. *P. haloplanktis* is a marine, psychrophilic species; *H. meridiana* is a salt-tolerant, mesophilic species; *M. degradans* is a marine mesophilic species; *C. aurantiacus* is a thermophile from hot springs. Of interest is that, with the exception of *Streptomyces*, all these bacteria display a specific adaptive character to the environment as far as temperature, pH, salinity and degradation capacity are concerned. In addition, some of them are true extremophiles.

Evidence for horizontal gene transfer in the animal-like α-amylases

There are several lines of evidence that the animal-type α-amylase genes in Bacteria result from horizontal gene

Table 2. Percent identity of α-amylase proteins between *D. melanogaster* and the animal-type α-amylases of Bacteria (some animals are also shown).

Species	% identity
Microbulbifer degradans	44
Chloroflexus aurantiacus	50
Thermobifida fusca	41
Streptomyces coelicolor	45
Pseudoalteromonas haloplanktis	46
Bacillus sp. No. 195	38
Pseudomonas sp. KFCC10818	43
Homo sapiens (human)	53
Euroglyphus maynei (mite)	49
Caenorhabditis elegans (nematode)	43

Values were computed with the BLAST2 server (http://www.ncbi.nlm.nih.gov/gorf/bl2.html) with no filter and the BLOSUM62 matrix, excluding the C-terminal domains where present.

transfer (HGT) from Eukarya (i.e., mainly animals, and in one case, a plant).

A gene candidate for the status of horizontally transferred gene should meet several requirements regarding sequence similarity and phylogenetic distribution. The animal-type status of the bacterial genes has been established by the conservation of global sequence similarity, and especially by the presence of the typical animal stretches (fig. 2). As far as we know at present, all animal α-amylases share many amino acid motifs, which are absent from α-amylases of plants, fungi, and most Bacteria [3, 4]. We have to mention here that in our search, we found some α-amylase sequences which exhibited only one or few of the typical animal motifs. Therefore, these occurrences were not retained, all the more since the expect values were well below the threshold. Once the first clues, i.e., the similarities with animal sequences, have been identified, a second good indicator of putative HGT is the anomalous position of these α-amylases on the tree (fig. 1). The α-amylases studied here are without a doubt clustered with true animal α-amylases, and in one case, with the plant α-amylases. However, we have to take into account the phylogenetic distribution and the rarity of the species of interest. In the past few years, before a large number of bacterial genomes had been sequenced, whether animal-like α-amylases were widely spread in Bacteria was not clear. We now have a large enough sample to show that animal-type α-amylases are scarce, and scattered in a few, and often unrelated bacterial species. Therefore, the distribution of animal-type α-amylases is explained by HGT better than by massive gene loss.

According to Doolittle [37], this is still not rigorously sufficient to firmly establish the HGT event. Doolittle points out that the sequence identity between the putative recipient and donor should be high enough (at the protein level, since genome constraints may change the base composition quickly and significantly), above 60%. This criterion could be weighted by the identification of signature stretches of amino acids. Also, the range of variation within the donor taxon should be considered. In the case of animal α-amylases, sequence identity is often less than 60%. Table 2 shows the values of identity between *D. melanogaster* α-amylase and the α-amylases from Bacteria studied, but also with those from some animals. The identity value is only 43% with *C. elegans*, and as low as 39% between *C. elegans* and the mite Dermatophagoides pteronyssinus (not shown), which is due to diverged, 'derived' *Amy* sequences for these two species. The percent identity values between *D. melanogaster* and the candidate bacteria are within this range, or better. In addition, although sequence identity is obviously maximal just after HGT, the event may be ancient and the exogenous gene may have diverged quickly to fit genomic constraints or adaptive requirements of the recipient species. In this respect, the α-amylases studied here seem not to be eliminated by this criterion.

The case of bacterial α-amylases with lower sequence similarity to plants, fungi, or animals

The analysis of loose branches, containing α-amylases with similarity to either plants, animals or fungi, shows that, except in the case of *T. vulgaris*, one cannot reasonably conclude that HGT occurred. We can only suggest that these cases may be a remnant of old HGT, obscured by subsequent rearrangements with endogenous bacterial α-amylases or other glycoside hydrolases (different glycoside hydrolases may share significant sequence similarity). In addition, all the loose branches contain species from unrelated bacterial phyla – firmicules and γ-proteobacteria in the loose animal-like branch, firmicutes and Archaea in the group with similarity to plants, and actinomycetes, firmicutes and α-proteobacteria in the group with similarity to fungi. Thus, looking at the tree, a question arises: is there a true bacterial α-amylase type? The genuine bacterial type could be represented by the group with no connexion to eukaryotes. This group contains only γ-proteobacteria in our sample. However, the question deserves further attention. Given this uncertainty about the origin(s) of α-amylases in Eukaryotes and the need for additional data, explanations of the situation observed, in terms of HGT or gene loss, are still speculative.

Origins of the transferred genes

For the cases of bacterial α-amylases clearly related to animals, the question is not whether animal α-amylases were transferred in Bacteria, but how and how many times, and from which donors. The distribution of animal-type genes in several unrelated bacterial phyla (actinomycetes, γ-proteobacteria, firmicutes, green non-sulfur bacteria) suggests that it happened several times. However, HGT is also

frequent between Bacteria, and indeed, this might have happened provided that two species could meet each other, i.e., if their respective environments were similar. For example, *P. haloplanktis* and *M. degradans* are both marine bacteria, which belong to the same family. But since their ecological conditions are quite different, gene transfer between them is not probable. In contrast, both vertical origin and independent acquisition are theoretically possible. However, the tree topology suggests that the independent gain of an animal gene is more likely. We will see below that data from the C-terminal domains suggest alternative scenarios. Concerning actinomycetes (*Streptomyces*, *Thermomonospora*, *Thermobifida*), a single origin is likely, but a transfer clearly occurred toward the firmicute *Bacillus* sp. No. 195. The case of *C. aurantiacus* is also convincing of an independent gain of an animal α-amylase, since its phylogenetic position, but also its ecological conditions are quite different from the above-mentioned species. In summary, the fact that several unrelated phyla with ecological conditions so different that they could not easily come into contact suggests several independent occurrences of HGT. In turn, this suggests that acquisition of eukaryotic α-amylase may be of adaptive interest.

The donor taxa cannot yet be identified. Interestingly, we observed in animal-type α-amylases of Bacteria some evolutionary tendencies already observed among animal α-amylases, i.e., the presence/absence of an amino-acid stretch forming a glycine-rich loop [4]. This motif is present in *P. haloplanktis*, *H. meridiana* and *C. aurantiacus*, and absent in *M. degradans*, *Bacillus sp.*, *Streptomyces* and *T. fusca*. We cannot say whether this is an indication of the donor species, or a common evolutionary response to similar adaptive constraints. The only clue in the search for donors is the presence of the AHA C-terminal domain in two animals, which could be indicative of the origin of the *P. haloplanktis* gene, but exchanges of the C-terminal domain are possible and are discussed below. Indeed, the mechanism of HGT is unclear. Before being active in a bacterium, the transferred gene has to be cleared of its introns, sometimes numerous in animal α-amylases. This should probably occur before entering the bacterium. Despite this particular problem, a number of cases of HGT from animals toward bacteria have been reported [reviewed in ref. 38], showing the relatively high frequency of this phenomenon. For example, Bacteria have been recently proposed to commonly incorporate foreign DNA through electric shock from lightning [39]. However, this does not solve the problem of getting rid of introns if genomic DNA is absorbed.

As a matter of fact, since Bacteria such as the species studied here are in frequent contact with plants, that there is evidence of only one case of transfer of α-amylase from a plant, in *M. degradans*, is surprising. *M. degradans* may have acquired its plant-type α-amylase-like gene from its vegetal substrate. Interestingly, this bacterium has been isolated from the salt marsh grass *Spartina* (Poaceae) [40], and the best BLASTP hit with this α-amylase is a barley α-amylase, which belongs to the same family, Poaceae. Regarding *M. degradans* and its remarkable multikingdom α-amylase family, worth noting is that this bacterium is a polysaccharide-degrading species, for which a broad range of enzymatic activities is advantageous. Indeed, this species is known to degrade more complex carbohydrates than reported for any other Bacteria: agar, chitin, alginic acid, carragheenan, cellulose, β-glucan, laminarin, pectin, pullulan, starch, and xylan [40, 41].

Shuffling of C-terminal domains within genomes and between kingdoms

We have no evidence that the putative α-amylase genes mentioned in this study have no other specificities than starch-degrading activity. The C-terminal domains may be involved in substrate specificity. In this respect, the observed distribution of CBM-20 and AHA extra domains, suggesting possible interkingdom domain shuffling, deserves special attention. We have established firmly that the AHA C-terminal domain is present in two animal α-amylases: *C. fluminea* and *C. elegans*. To date, our search of the AHA C-terminal domain in other bivalves has been negative (not shown). On the other hand, searching in nematode data bases (http://www.nematode.net/BLAST) shows that this domain may be ancestral in Nematodes (not shown). In Bacteria, this domain was found only once (except in *P. haloplanktis*, where it was discovered): in *M. degradans*, a species that belongs to the same family Alteromonadaceae, but the domain was attached to an unrelated (plant-type) α-amylase gene. The origin of this domain is unknown, given the few occurrences in Bacteria as well as Eukarya. Both hypotheses (bacterial or eukaryote origin) may be considered. If the C-terminal domain is bacterial, it seems to have disappeared from most species, and may have been transferred to at least two, not closely related, animals. On the other hand, if the motif is of animal origin, it has been lost in most animals (but few have been fully sequenced, unlike bacteria). We need to point out that the genes coding for these domains in *C. elegans* and *C. fluminea* are interrupted by one and three introns, respectively, with a shared position (same position and phase). Unfortunately, this observation is not decisive, since introns may be gained, possibly at an identical position independently [see ref. e.g., 42]. Thus, the phylogenetic distribution of this domain deserves further investigation and raises an exciting question concerning its origin and its putative function in non-bacterial species. This extra domain might possibly be involved in membrane anchoring, because in *P. halopanktis* it is involved in outer membrane recognition and assists α-amylase secretion.

As mentioned above, most animal-like α-amylases of Bacteria possess a C-terminal starch-binding domain (the

CBM-20 type), and it is also the case for the new ones from our study. This domain serves as a raw-starch binding domain in a number of bacterial α-amylases, but is also found in other glycoside hydrolases not only from the α-amylase family, such as cyclodextrin glucanotransferase, β-amylase and glucoamylase [23–25]. To support an evolutionary scenario, we searched for other occurrences of a CBM-20 sequence in the *M. degradans* genome. Only one was found (ZP_00067465), in a putative protein of unknown function. This could have served as a donor through duplication and graft to the animal-like α-amylase of *M. degradans*, while the AHA domain was translocated to the plant-like gene. For the other species (*Streptomyces*, *Chloroflexus*), similar duplication/graft of CBM-20 may have occurred from other endogenous glycoside hydrolases. Interestingly, among various glycoside hydrolases that possess a CBM-20 C-terminal domain, the CBM-20 sequences have been shown to follow the species phylogeny, whereas the core (i.e., catalytic) sequences of the enzymes are clustered according to their function [24, 25]. This indicates an intragenomic origin and distribution of the C-terminal domains by duplication. The widespread distribution of CBM-20 in Bacteria strongly suggests its bacterial origin. However, two cases of CBM-20-like domains have been detected in mammalian proteins – laforin [43] and genethonin [44].

The history we propose for *P. haloplanktis* and *M. degradans* is illustrated in figure 5. We have hypothesized an animal origin for the AHA C-terminal domain, mainly because of its rarity in the large sample of bacterial genomes now sequenced. Most probably, the AHA domain was attached to an α-amylase gene in the donor, so that the animal-type gene should be of common origin in the two Bacteria, and not gained independently. The tree in figure 1 is not in favor of a common ancestry of the animal donor genes in these species, but quick adaptive divergence toward the particular environments may have lead to incorrect branching. Hopefully, data from other Alteromonadaceae will help reconstitute the true story.

Conclusion

Although a number of horizontal transfers from eukaryotes to bacteria have been shown or suspected [39], they are not very frequent. Considering the huge populations of bacteria and their promiscuity with living or dead eukaryote cells, absorption and integration of eukaryotic DNA may be pervasive. However, to become fixed, in-

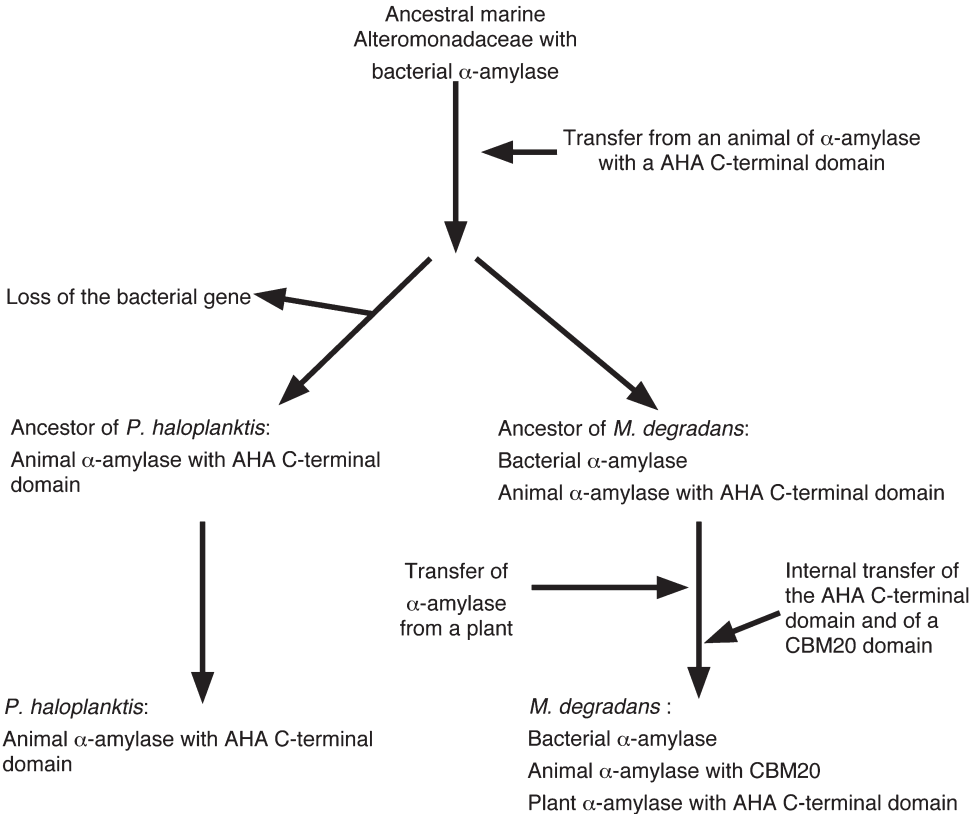

Figure 5. A scenario, among several, for the evolution of α-amylases in *Pseudoalteromonas haloplanktis* and its relative *Microbulbifer degradans* (γ-proteobacteria: Alteromonadaceae).

corporated coding DNA must be devoid of introns, and in order to be active, the transferred gene (intronless genomic DNA or cDNA) should be full-length or at least a complete domain. Second, it would have to bring a selective advantage. Thus, there are important obstacles to overcome. Once they have been surmounted, we may observe rapid adaptation of codon usage, signal peptide, and promoter. One of the results may be the eventual loss of the original bacterial gene.

Acknowledgements. We are grateful to two anonymous referees for their comments. We thank Linda Sperling and the Genoscope for sharing data on *Paramecium* before publication. S. J. thanks VEGA for grant No. 2/2057/23. J.-L. D. L. and G. F. thank the CNRS and the CGRI-FNRS for financial support.

1 Janecek S. (1997) α-Amylase family: molecular biology and evolution. Prog. Biophys. Mol. Biol. **67:** 67–97
2 Pujadas G. and Palau J. (2001) Evolution of α-amylases: architectural features and key residues in the stabilization of the (β/α)8 scaffold. Mol. Biol. Evol. **18:** 38–54
3 Janecek S. (1994) Sequence similarities and evolutionary relationships of microbial, plant and animal α-amylases. Eur. J. Biochem. **224:** 519–524
4 Da Lage J.-L., van Wormhoudt A. and Cariou M.-L. (2002) Diversity and evolution of the α-amylase genes in animals. Biol. Bratislava **57 (suppl. 11):** 181–189
5 Janecek S., Lévêque E., Belarbi A. and Haye B. (1999) Close evolutionary relatedness of α-amylases from Archaea and plants. J. Mol. Evol. **48:** 421–426
6 Long C. M., Virolle M.-J., Chang S.-Y., Chang S. and Bibb M. J. (1987) α-Amylase gene of *Streptomyces limosus*: nucleotide sequence, expression motifs, and amino acid sequence homology to mammalian and invertebrate α-amylases. J. Bacteriol. **169:** 5745–5754
7 Mazodier P. and Davies J. (1991) Gene transfer between distantly related bacteria. Annu. Rev. Genet. **25:** 147–171
8 Vigal T., Gil J. A., Daza A., Garcia-Gonzalez M. D. and Martin J. F. (1991) Cloning, characterization and expression of an α-amylase gene from *Streptomyces griseus* IMRU3570. Mol. Gen. Genet. **225:** 278–288
9 Petricek M., Tichy P. and Kuncova M. (1992) Characterization of the α-amylase-encoding gene from *Thermomonospora curvata.* Gene **112:** 77–83
10 Bahri S. M. and Ward J. M. (1993) Sequence of the *Streptomyces thermoviolaceus* CUB74 α-amylase-encoding gene and its transcription analysis in *Streptomyces lividans*. Gene **127:** 133–137
11 Sumitani J. I., Nagae H., Kawaguchi T. and Arai M. (1998) *Bacillus* animal type α-amylase: cloning and sequencing of the gene, and comparison of the deduced amino acid sequence with that of other amylases. J. Ferment. Bioeng. **85:** 428–432
12 Coronado M. J., Vargas C., Mellado E., Tegos G., Drainas C., Nieto J. J. et al. (2000) The α-amylase gene *amyH* of the moderate halophile *Halomonas meridiana*: cloning and molecular characterization. Microbiology **146:** 861–868
13 Feller G., Lonhienne T., Deroanne C., Libioulle C., Van Beeumen J. and Gerday C. (1992) Purification, characterization and nucleotide sequence of the thermolabile α-amylase from the antarctic psychrotroph *Alteromonas haloplanktis* A23. J. Biol. Chem. **267:** 5217–5221
14 Feller G., Le Bussy O., Houssier C. and Gerday C. (1996) Structural and functional aspects of chloride binding to *Alteromonas haloplanctis* α-amylase. J. Biol. Chem. **271:** 23836–23841
15 Feller G., Payan F., Theys F., Quian M., Haser R. and Gerday C. (1994) Stability and structural analysis of α-amylase from the antarctic psychrophile *Alteromonas haloplanctis* A23. Eur. J. Biochem. **222:** 441–447
16 Aghajari N., Feller G., Gerday C. and Haser R. (1998) Crystal structures of the psychrophilic α-amylase from *Alteromonas haloplanctis* in its native form and complexed with an inhibitor. Protein Sci **7:** 564–572
17 Aghajari N., Feller G., Gerday C. and Haser R. (1998) Strucures of the psychrophilic *Alteromonas haloplanctis* α-amylase give insight into cold adaptation at a molecular level. Structure **6:** 1503–1516
18 Svensson B. (1994) Protein engineering in the α-amylase family: catalytic mechanism, substrate specificity, and stability. Plant Mol. Biol. **25:** 141–157
19 D'Amico S., Gerday C. and Feller G. (2000) Structural similarities and evolutionary relationships in chloride-dependent α-amylases. Gene **253:** 95–105
20 Matsuura Y., Kusunoki M., Harada W. and Kakudo M. (1984) Structure and possible catalytic residues of Taka-amylase A. J. Biochem. (Tokyo) **95:** 697–702
21 Machius M., Wiegand G. and Huber R. (1995) Crystal structure of calcium-depleted *Bacillus licheniformis* α-amylase at 2.2 Å resolution. J. Mol. Biol. **246:** 545–559
22 Kadziola A., Sogaard M., Svensson B. and Haser R. (1998) Molecular structure of a barley α-amylase-inhibitor complex: implications for starch binding and catalysis. J. Mol. Biol. **278:** 205–217
23 Svensson B., Jespersen H., Sierks M. R. and McGregor E. A. (1989) Sequence homology between putative raw-starch binding domains from different starch-degrading enzymes. Biochem. J. **264:** 309–311
24 Janecek S. and Sevcik J. (1999) The evolution of starch-binding domain. FEBS Lett. **456:** 119–125
25 Janecek S., Svensson B. and MacGregor E. A. (2003) Relation between domain evolution, specificity, and taxonomy of the α-amylase family members containing a C-terminal starch-binding domain. Eur. J. Biochem. **270:** 635–645
26 Feller G., D'Amico S., Benotmane A. M., Joly F., Van Beeumen J. and Gerday C. (1998) Characterization of the C-terminal propeptide involved in bacterial wall spanning of α-amylase from the psychrophile *Alteromonas haloplanktis*. J. Biol. Chem. **273:** 12109–12115
27 Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J., Rapp B. A. and Wheeler D. L. (2002) GenBank. Nucleic Acids Res. **30:** 17–20
28 Altschul S. F., Gish W., Miller W., Myers E. W. and Lipman D. J. (1990) Basic local alignment search tool. J. Mol. Biol. **215:** 403–410
29 Boer P. H. and Hickey D. A. (1986) The α-amylase gene in *Drosophila melanogaster*: nucleotide sequence, gene structure and expression motifs. Nucleic Acids Res. **14:** 8399–8411
30 Tatusova T. A. and Madden T. L. (1999) Blast 2 sequences – a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. **174:** 247–250
31 Huang X. (1994) On global sequence alignment. CABIOS **10:** 227–235
32 Thompson J. D., Higgins D. G. and Gibson T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680
33 Bairoch A. and Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28:** 45–48
34 Janecek S. (2002) How many conserved sequence regions are there in the alpha-amylase family? Biol. Bratislava **57 (suppl. 11):** 29–41

35 Saitou N. and Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425
36 Page R. D. (1996) TreeView: an application to display phylogenetic trees on personal computers. CABIOS **12:** 357–358
37 Doolittle R. F. (2002) Gene transfers between distantly related organisms. In: Horizontal Gene Transfer, pp. 269–275, Syvanen, M. & Kado, C. I. (eds), Adademic Press, London
38 Koonin E. V., Makarova K. S., Wolf Y. I. and Aravind L. (2002) Horizontal gene transfer and its role in the evolution of prokaryotes. In: Horizontal Gene Transfer, pp. 277–304, Syvanen, M. & Kado, C. I. (eds), Adademic Press, London
39 Demanèche S., Bertolla F., Buret F., Nalin R., Sailland A., Auriol P. et al. (2001) Laboratory-scale evidence for lightning-mediated gene transfer in soil. Appl. Environ. Microbiol. **67:** 3440–3444
40 Gonzales J. M. and Weiner R. M. (2000) Phylogenetic characterization of marine bacterium strain 2-40, a degrader of complex polysaccharides. Int. J. Syst. Evol. Microbiol. **50:** 831–834
41 Howard M. B., Ekborg N. A., Taylor L. E., Weiner R. M. and Hutcheson S. W. (2003) Genomic analysis and initial characterization of the chitinolytic system of *Microbulbifer degradans* strain 2-40. J. Bacteriol. **185:** 3352–3360
42 Stoltzfus A., Logsdon J. M., Palmer J. D. and Doolittle W. F. (1997) Intron 'sliding' and the diversity of intron positions. Proc. Natl. Acad. Sci. USA **94:** 10739–10744
43 Minassian B. A., Ianzano L., Meloche M., Andermann E., Rouleau G. A., Delgado-Escueta A. V. et al. (2000) Mutation spectrum and predicted function of laforin in Lafora's progressive myoclonus epilepsy. Neurology **55:** 341–346
44 Janecek S. (2002) A motif of microbial starch-binding domain found in human genethonin. Bioinformatics **18:** 1534–1537