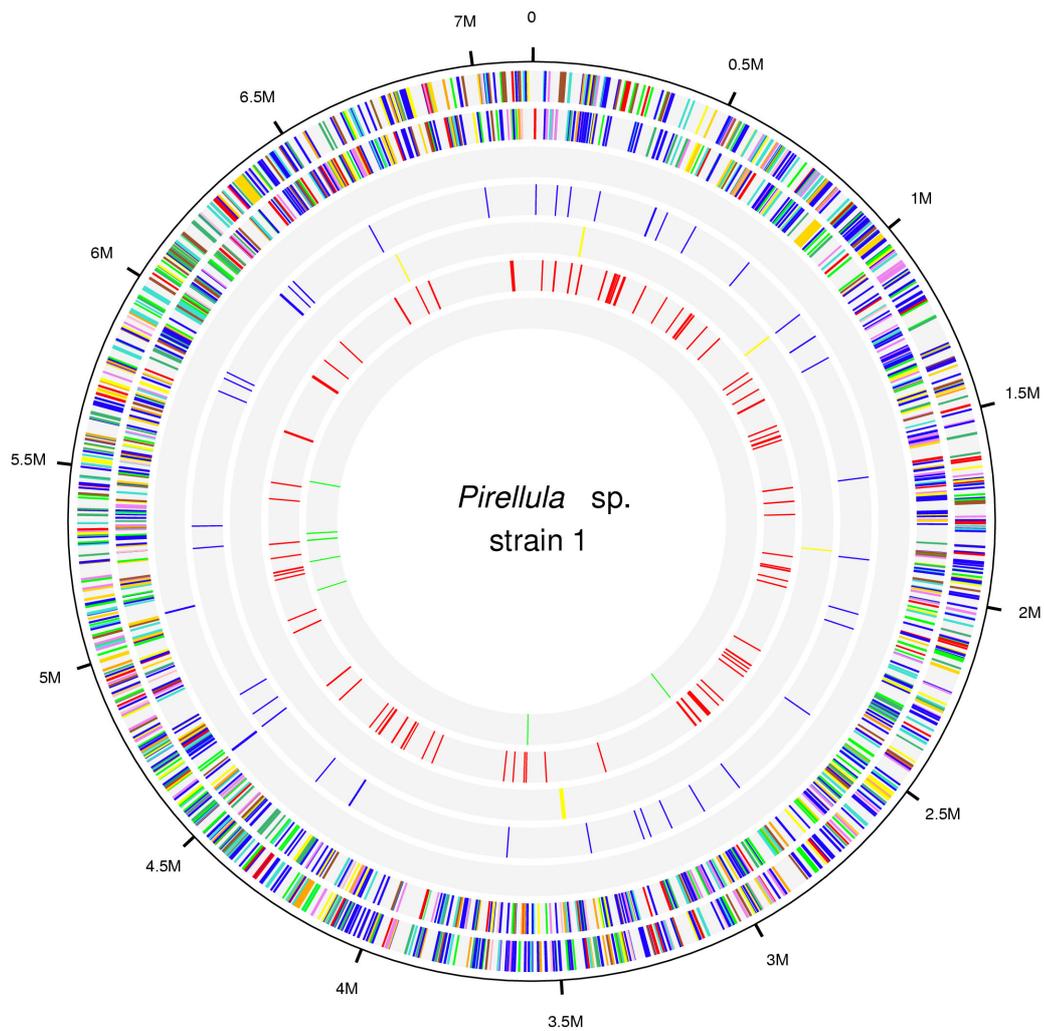


The genome of the free-living, marine *Planctomycete*  
*Pirellula* sp. strain 1 (“*Rhodopirellula baltica*”):  
Bioinformatics and Biology



Thesis for the attainment of the grade of PhD in natural sciences - Dr. rer. nat. -  
of the faculty of Biology/Chemistry of the University of Bremen  
submitted by Thierry Lombardot

Supervisors:

Prof. Frank Oliver Glöckner, Dr. Marga Bauer and Prof. Rudolf Amann

Max Planck Institute for Marine Microbiology - Department of Molecular Ecology, February 2004

### *Acknowledgments*

*I would like to thank Prof. Frank Oliver Glöckner and Dr. Marga Bauer for their supervision during my work and Prof. Rudolf Amann who gave me the opportunity to work at the Department of Molecular Ecology; Dr. Uta Bohnebeck and Prof. Dietmar Blohm who accepted to participate in the evaluation of this work; Michael Richter, Hauke Pfeffer, Thomas Otto, Tim Frana, Stella Koufou and Andreas Schmitz for their contribution to genome annotation and Hanno Teeling for further analysis and discussions; Michael Kube for the sequencing of *Pirellula* sp. strain 1; Folker Meyer, Alexander Goesmann and Burkhard Linke for their support with the GenDB system; Dörte Gade for her expertise in proteomics and Marc Mussmann for providing metagenomic data for test purposes; Furthermore, I would like to thank the members of the Department of Molecular Ecology for support and discussions: Anke Meyerdierks, Falk Warnecke, Sebastian Behrens, Chris Würdemann and all the others. Major funding of this project was provided by the BMBF (German Federal Ministry of Education and Research). Further supports came from the Max Planck Society. This work was done between January 2001 and February 2004 at the Department of Molecular Ecology of the Max Planck Institute for Marine Microbiology in Bremen.*

*Thierry Lombardot  
Bremen, February 2004*

## Abstract

The complete genome analysis of environmentally relevant microorganisms constitute a new emerging field of marine research. Within the frame of this work, the bioinformatic infrastructures adapted to microbial genomics data have been locally established at the Department of Molecular Ecology. This infrastructure includes powerful hardware systems consisting of a computing cluster and dedicated servers. The appropriate software for data storage, access and analysis has been completely integrated. Open source solutions were selected for these tasks, which allows an optimal control of the system at very low exploitation costs. Additionally, the adoption of the corresponding open standards facilitate data exchange with other research institutes and constitute a solid basis for national and international collaborations. These computer infrastructures were built to annotate and analyze the genome of *Pirellula* sp. strain 1, but will also be used for the analysis of upcoming complete genomes of environmentally relevant microorganisms at the MPI-Bremen. Currently, local metagenomics projects already take advantage of the established annotation pipeline, as the bioinformatic methods needed for genomics and metagenomics are very similar.

The genome of *Pirellula* sp. strain 1 was analyzed with bioinformatic methods and revealed the blueprint of this organism containing unexpected findings. The occurrence of an outstanding number of sulfatases gene copies and of an unexpected *Archaea*-like C<sub>1</sub> metabolism are some examples of annotation highlights. Based on these new data, hypothesis concerning the lifestyle and the evolution of this organism, and more broadly of *Planctomycetes*, could be formulated. This constitutes a proof of principle of the important role that whole genome analysis will play in the field of ecology in the future: the role of an “hypothesis generator”.

Gene expression level predictions were calculated for *Pirellula* sp. strain 1, linking *in silico* analysis and functional genomics experiments. Transcriptome and proteome expression data will be compared to these predicted expression levels, which represents an untouched field for environmental organisms as such comparisons were only made for classical model organisms.

The availability of the first complete genome of a *Planctomycete* allowed to reassess the phylogenetic position of this phylum by using “genome trees”. This new approach did not support the previously suggested deepest branching position of the *Planctomycetes* within the bacterial domain by particular 16S rDNA analysis and retained the thermophilic organisms as deepest branching. The “genome trees” can be considered as a new, exploratory phylogenetic method and the differences between alternative phylogenetic reconstruction methods (e.g. 16S rRNA vs. full genomes) illustrates again the complexity of phylogenetic studies on *Planctomycetes*.

The importance of the emerging field of environmental metagenomic studies as an indispensable companion of the whole genome sequencing of isolated strain becomes evident today. Just as the number of complete genomes, the number of available metagenomics sequences is expected to explode within the next years. A geographic information system (GIS) for metagenomics data was developed and represents a first prototype of specialized database and tools to handle this new kind of data.

## Table of content

### 1. Introduction

- 1.1. Prokaryotic genomes sequencing and bioinformatics
- 1.2. Genomes from the environment: The REGX project
- 1.3. The *Planctomyces*
  - 1.3.1. Environmental relevance
  - 1.3.2. Cellular biology
  - 1.3.3. Phylogeny
- 1.4. *Pirellula* sp. strain 1
  - 1.4.1. Pre-genomic era: physiological description
  - 1.4.2. Whole genome sequencing (MPI-Berlin)

### 2. Material and methods

- 2.1. Locally maintained bioinformatic tools and databases
- 2.2. Genome analysis: annotation
  - 2.2.1. Gene prediction
  - 2.2.2. Software package Pedant Pro
  - 2.2.3. Software package GenDB
  - 2.2.4. Computation clustering
  - 2.2.5. Public BLAST server
- 2.3. Consistent genomes comparisons
  - 2.3.1. The Pfam database
  - 2.3.2. Profile hidden Markov models
- 2.4. Codon usage analysis
  - 2.4.1. Codon Adaptation Index (CAI)
  - 2.4.2. Karlin-Mrazek (PHX/PA)
- 2.5. Genome trees: new phylogenetic reconstruction strategies
- 2.6. Geographic information system

### 3. Results and discussion

- 3.1. Genome annotation pipeline
  - 3.1.1. Pedant Pro and GenDB database systems comparison
    - 3.1.1.1. Software design comparison
    - 3.1.1.2. Databases architecture comparison
    - 3.1.1.3. Consequences for future systems
  - 3.1.2. Gene prediction
  - 3.1.3. Automatic annotation / Manually refined functional assignment
- 3.2. *Pirellula* sp. strain 1 genome interpretation
  - 3.2.1. DNA compositional asymmetries
  - 3.2.2. General genetic potential: an overview

- 3.2.3. Annotation highlights: unexpected findings
  - 3.2.3.1. Sulfatases high copy number
  - 3.2.3.2. Special enzymes for C<sub>1</sub> metabolism
- 3.3. Consistent cross-genomes comparisons
  - 3.3.1. Systematic study of environmentally relevant gene groups
    - 3.3.1.1. Sulfatases
    - 3.3.1.2. Glycosyl hydrolases
    - 3.3.1.3. Transporters
    - 3.3.1.4. Transposases / integrases
    - 3.3.1.5. Signal peptides
  - 3.3.2. Transcriptional regulators pool
    - 3.3.2.1. Quantitative comparisons
    - 3.3.2.2. Qualitative comparisons: ECF sigma factors
    - 3.3.2.3. Phylogenetic study of ECF sigma factors
- 3.4. Gene expression prediction based on codon usage
  - 3.4.1. Analysis according to PHX genes clusters
  - 3.4.2. PA genes
  - 3.4.3. Analysis of selected PHX gene groups
- 3.5. Genome trees as a tool for phylogenetic reconstruction
- 3.6. Metagenomes mapserver (prototype)
  - 3.6.1. Database design
  - 3.6.2. Towards a geo/ecological analysis of genomic fragments

## 4. Conclusions

## 5. Annexes

## 6. References

## 1. Introduction

### 1.1. Prokaryotic genomes sequencing and bioinformatics

Since a few years, it is possible to access rapidly the complete genome information of any living organism by reading its entire DNA sequences. Today, at the very beginning of 2004, the count of organisms whose genome has been sequenced still lies under 200, but the availability of thousands of genomes is only a question of time. This breakthrough in the field of biology gives for the first time direct access to the genetic blueprints of life. The first hundreds of genomes only constitute the premises of the genomic revolution which will have profound impact on the society, the economy and the way we understand the biological diversity surrounding us. In this context, the present work proposes to study an environmentally relevant microorganism - *Pirellula* sp. strain 1 - through its complete genome.

The first landmark of this revolution was set in an unexpected way. In July 1995, a spectacular announcement spread all over the scientific community: the first bacterial genome (*Haemophilus influenzae*, 1.8 Mbp) was finished earlier than expected using a new approach<sup>1,2</sup>. What was previously thought to be a multiyear, multimillion-dollar project had been finally accomplished by a team of scientists within 13 months at low costs using the so-called "shotgun sequencing" methodology. Previous strategies were based on the laborious segmentation of a genome into ordered, overlapping segments. On the contrary, the innovative shotgun approach for *H. influenzae* was based on the sequencing of more than 24,000 random genome fragments following computer assisted reassembly. This alternative approach proved to be faster and cheaper than any other. Interestingly, this ambitious project had failed to be founded by the US National Institute of Health (NIH) in its early phase, because serious doubt had been raised about the feasibility of such a chaotic approach.

The team of 36 scientists who successfully sequenced *H. influenzae*, headed by Craig Venter at the TIGR center (The Institute for Genome Research) and the Nobel prize winner Hamilton Smith from the Johns Hopkins University defined the current framework for a successful genome project. The shotgun sequencing approach has since then become a *de facto* standard for almost all genome projects, increasing the pace and lowering the cost of the whole discipline.

The genome sequencing project of *Escherichia coli* K-12 was started at the same period as *H. influenzae*, but was finished in 1997 because of its larger size (4.6 Mbp)<sup>3</sup>. This constituted the second landmark of the genomic era, by providing a real guide-book for this intensively studied model organism. Since this period, the number of publicly available prokaryotic genomes increased exponentially (Fig. 1).

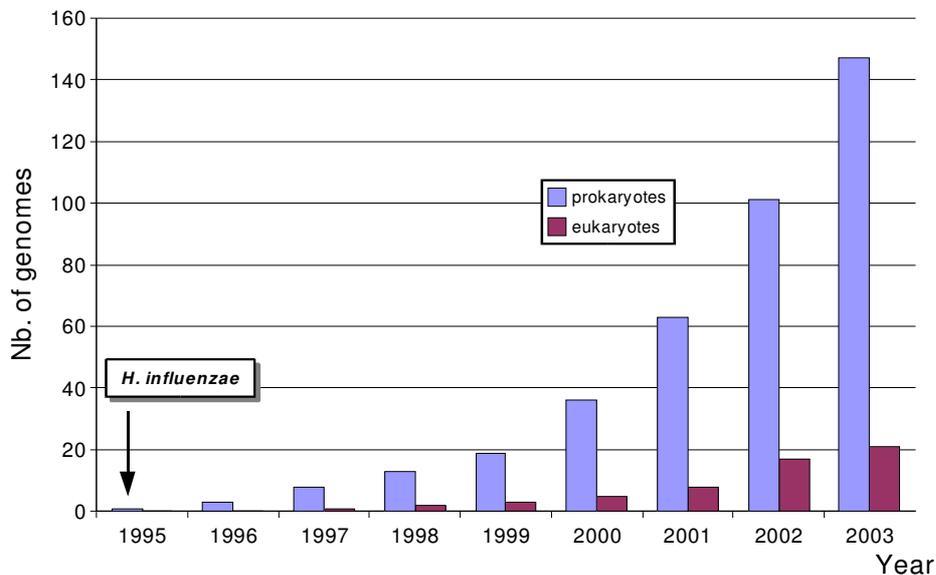


Fig. 1: Number of fully sequenced prokaryotic and eukaryotic organisms in the public databases since 1995. Source: EMBL<sup>4</sup> and GOLD<sup>5</sup> databases.

At the beginning of this work, in 2001, a total of 36 *Bacteria* and *Archaea* were fully sequenced. Today, with 135 *Bacteria* and 17 *Archaea* (January 2004), it becomes even difficult to keep trace of every new genome released. This high growth rate is not expected to drop within the next years, as a total of 428 more prokaryotic genome projects are currently running<sup>5</sup> (January 2004).

The costs of genome sequencing are highly dependent on the quality standard in terms of allowed sequencing errors or number of contigs that want to be achieved. Based on this fact, numerous genome sequencing projects have been started with the aim to reach only a low quality in order to significantly lower the costs. This so-called "draft-sequencing" leads to genome sequences usually consisting of 100 to 1000 contigs (unlinked fragments) containing a higher sequence error rate than finished, complete genomes. This approach is commonly used in industrial projects for the screening of new enzymes, but recently, the Joint Genome Institute (JGI) also focused on this approach. The number of draft genomes released by JGI already exceeded 40 at the end of 2003. Draft sequencing radically contrasts with the traditional whole genome approach initiated by the TIGR center. The sequence quality as a function of sequencing coverage follows an asymptotic behavior (Fig. 2). The cost saving factor between a draft or a complete genome is only 2 times according to TIGR<sup>6</sup>, but might reach 4 times according to the Joint Genome Institute<sup>7</sup>. According to these sources, the price of a draft genome ranged between 3 to 4 US cents/base and the cost of a complete, high quality genome ranged between 8 to 10 US cents/base in 2002 - 2003.

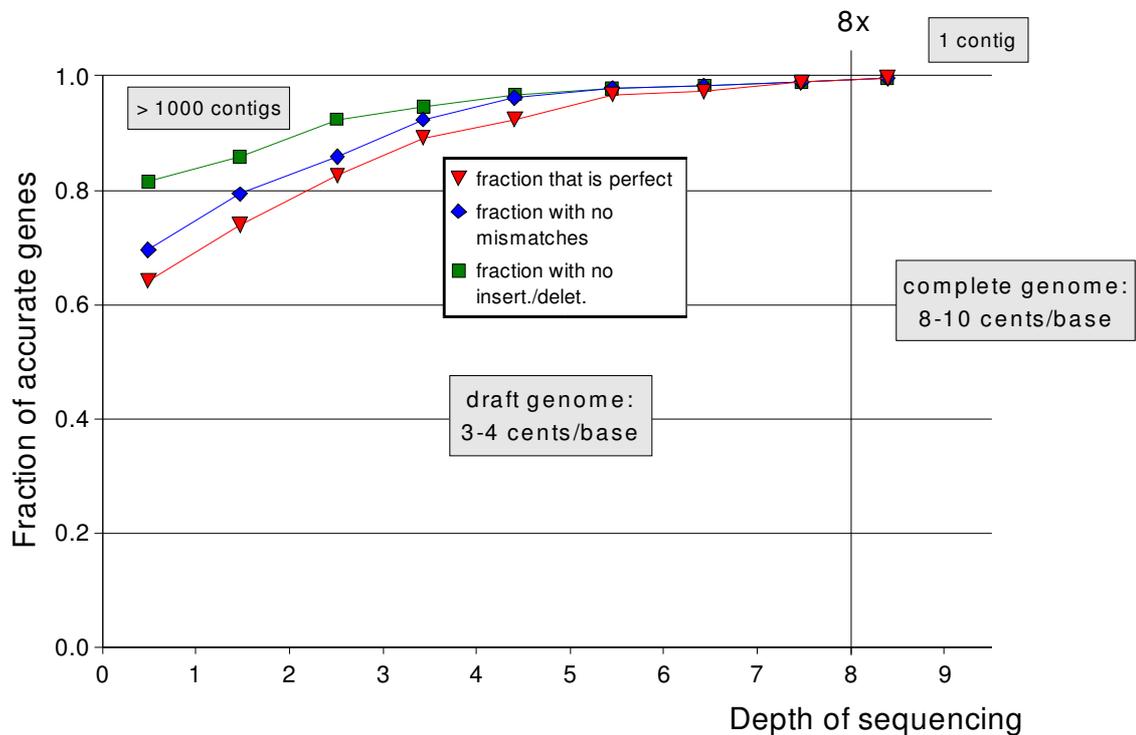


Fig. 2: Quality of gene sequences in a genome according to the coverage (depth of sequencing). The corresponding prices and contigs numbers are indicated. Prices are given in US cents. Sources: TIGR<sup>6</sup> and JGI<sup>7</sup>.

A draft sequence of a prokaryotic genome saves costs and is up to 10 times faster as compared to a closed genome. However, its scientific value is significantly lower for the following reasons: i) functional genomics (transcriptomics, proteomics) needs accurate datasets for probe design or protein identification; ii) genome organization can only be studied with a single, complete sequence and iii) comparative genomics needs complete gene repertoires. These axes of research constitute the mainframe of the post-genomic era and need complete genomes of high quality as solid basis. Moreover, a finished genome constitute a permanent, valuable resource for the scientific community.

The high number of complete and draft prokaryotic genomes released in the public DNA and protein databases contribute to their exponential growth. The high pace of genomic sequencing rises the problem of data integration and annotation quality. As shown in Fig. 3, the growth of the TrEMBL database, containing all non-redundant proteins derived from single gene or whole genome sequencing, is exponential. However, the SWISS-PROT database, containing manually curated and annotated proteins is more and more left behind by only growing linearly (Fig. 3).

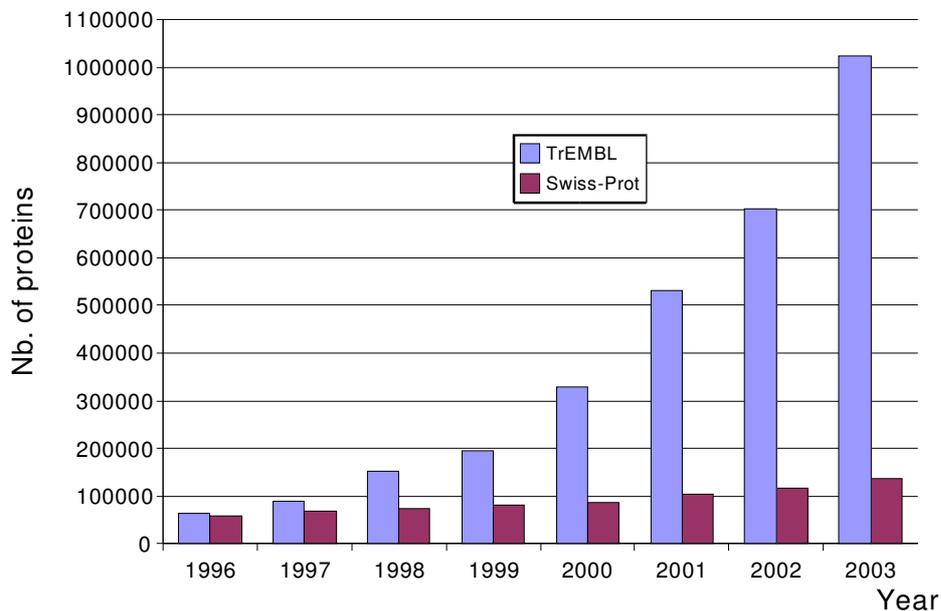


Fig. 3: Growth of the public protein databases since 1996. Completeness vs. accuracy: the TrEMBL database growth is exponential while the SWISS-PROT progression is almost linear<sup>8</sup>.

The discrepancy between the TrEMBL and the SWISS-PROT databases illustrates the gap between information and knowledge in modern molecular biology. While the quantity of information is growing exponentially with the new incoming sequences, the development of the knowledge that is extracted from this data source in the form of organized and classified function is rather linear. This observation highlights the fact that the real challenge of the post-genomic era will be to organize the flood of information coming from high-throughput sequencing projects in unified structures to improve our understanding of biological systems. This data integration step will constitute the next landmark of the genomic revolution.

In this context, the demand for adapted bioinformatics tools is growing in order to interpret the deluge of sequence data. Bioinformatics can be defined as a new emerging field focusing on the acquisition, storage, access, analysis, modeling and distribution of the many types of information embedded in DNA and protein sequences. Bioinformatics rely on mathematically intensive methodologies and exponentially growing databases. Therefore, a corresponding computational power is needed. Fortunately, the field of computing technology is also growing exponentially, as initially predicted by the "Moore's law". Gordon E. Moore, a physicist working in the field of electronic engineering, announced already in 1965 that in computer development, the number of transistors per integrated circuit followed an exponential growth over time and predicted that this trend would continue<sup>9</sup>. This estimation was shown to hold true over the last decades, as exemplified by the design of low-cost processing units (Fig. 4).

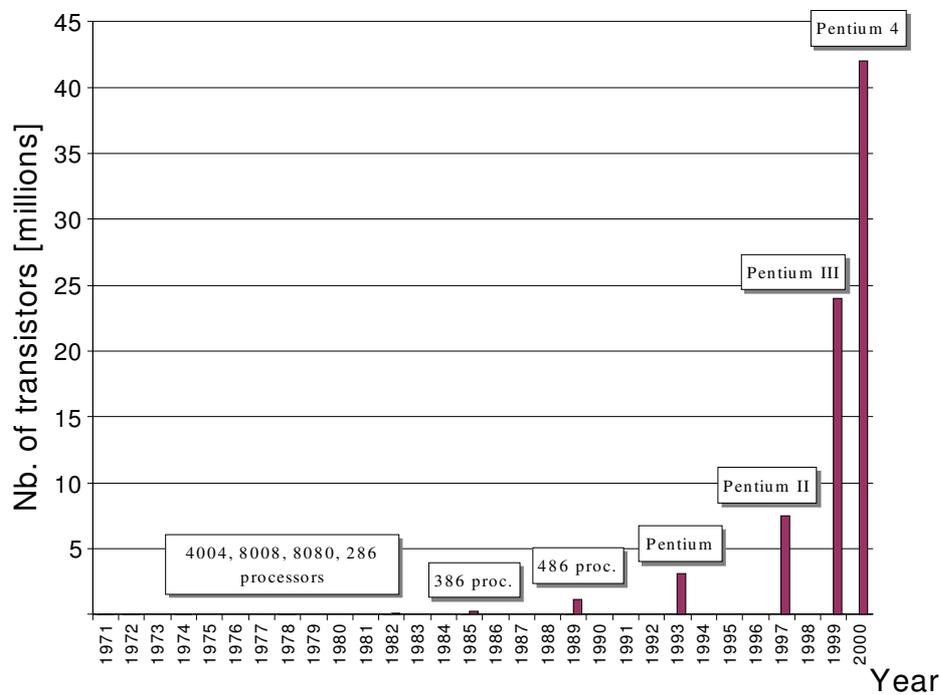


Fig. 4: Growth of computational processing power according to the number of transistors in low-cost CPUs (Intel processors) - an illustration of the Moore's law.

Whether the Moore's law is going to hold during the next decades is subject to intense controversy in the field of computational sciences. Critical overheating and a minimal size for silicon based transistors will definitely be reached in the future. This is likely to happen within the next decade<sup>10</sup>, but optimistic studies concluded that the technological improvements based on the conventional silicon nanoelectronics might continue longer and lead to chips containing more than 1 trillion transistors<sup>11</sup>.

The trends shows that the deluge of sequence data coming from genome projects is technically manageable from a computing capacity perspective. However, the data quality in terms of unified structures and classification is still sparse and will need better integration. Interestingly, a new project for centralized biological sequences management has been started very recently (December 2003). This project, called UniProt<sup>12</sup>, is going to unify all public protein databases into a central resource using consistent annotation and classification tools. Such efforts will constitute the urgently needed backbone to improve our understanding of biological systems based on genome data in the future.

## 1.2. Genomes from the environment: The REGX project

The first wave of prokaryotic genome sequencing selected exclusively microorganisms of medical and biotechnological interest. For the first time, the REGX project (ReaEnvironmental GenomiX), which was initiated at the Department of Molecular Ecology during the years 2000-2001, selected environmentally relevant microorganisms for an integrated genomic approach. This included the set up of a bioinformatics, transcriptomics and proteomics backbone for environmental genomics and the sequencing of three microorganisms: *Pirellula* sp. strain 1 (a *Planctomycete*), *Desulfobacterium autotrophicum* strain HRM2 and *Desulfotalea psychrophila* strain LSV54 (two sulfate reducers) (Fig. 5 and 6). This environmental genomics approach constitute a new emerging field in marine research.

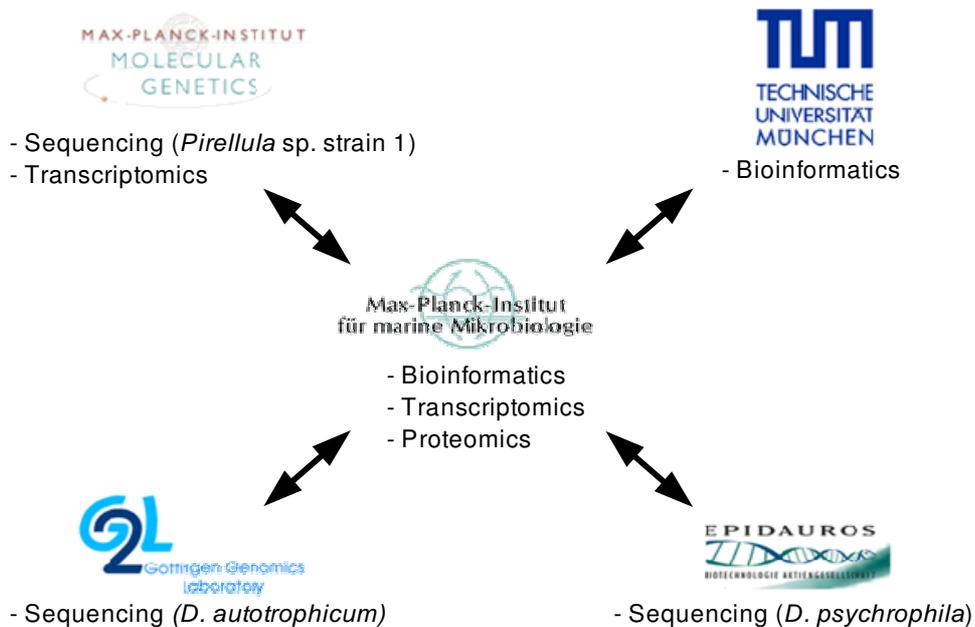
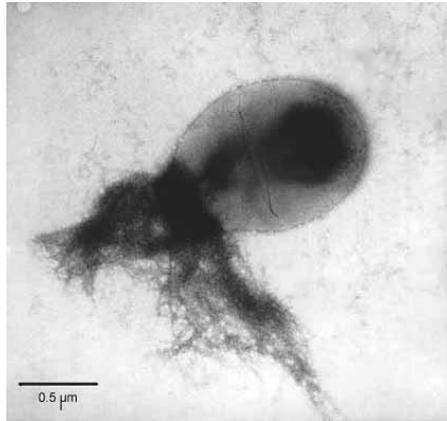
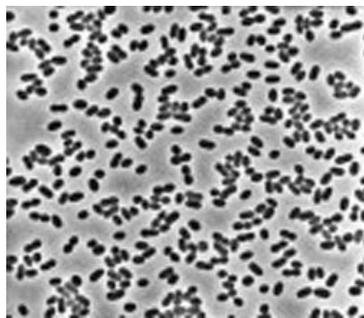


Fig. 5: The five scientific partners involved in the REGX project and tasks distribution.



a) *Pirellula* sp. strain 1



b) *Desulfobacterium autotrophicum*  
strain HRM2



c) *Desulfotalea psychrophila*  
strain LSv54

Fig. 6: The three environmentally relevant marine *Bacteria* selected for a whole genome approach by the REGX-project. a): electron microscopy picture of *Pirellula* sp. strain 1 (“*Rhodopirellula baltica*”), H. Schlesner; b) and c): phase contrast microscopy pictures of the two sulfate reducing *Bacteria* (SRB).

The selected organisms are representatives of important bacterial groups in the environment. They contribute to carbon fluxes in marine systems by mineralizing organic carbon molecules originating from the primary production in the upper layers of seas and oceans (Fig. 7). *Planctomycetes* are mainly located in the water column - freely floating or attached to phytodetrital aggregates (marine snow) - and degrade organic molecules (e.g. sugars) to CO<sub>2</sub> aerobically. Sulfate reducing *Bacteria* are located within the anaerobic layer of marine sediments and oxidize fermentation products (fatty acids) to CO<sub>2</sub> using inorganic sulfate as electron acceptor.

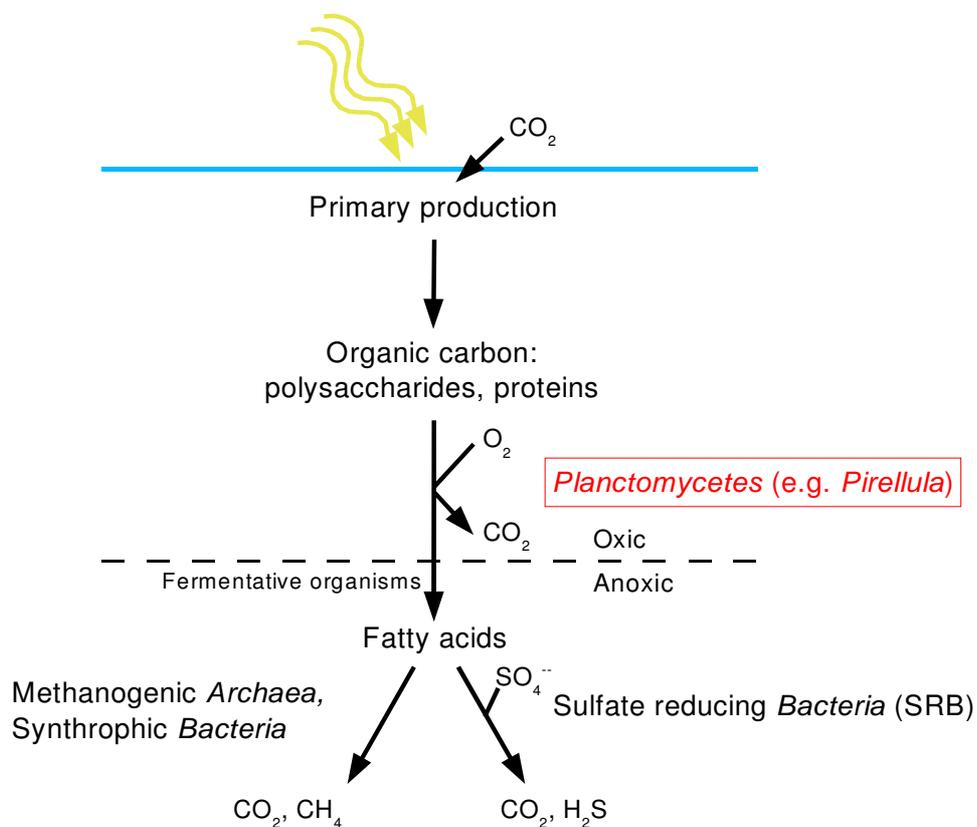


Fig. 7: Carbon flux in marine systems (simplified overview).

The aim of the REGX project - by combining bioinformatics, transcriptomics and proteomics approaches - is to establish an integrated platform for genome analysis of environmental bacteria within an ecological context. This constitutes a new emerging field in marine research.

The present work focuses on the bioinformatic analysis of the first organism of this project, *Pirellula* sp. strain 1 (“*Rhodopirellula baltica*”) and on the establishment of a bioinformatic pipeline for upcoming environmental genomics projects.

### 1.3. The *Planctomycetes*

*Pirellula* sp. strain 1 is a marine representative of the *Planctomycetes*, a globally distributed and environmentally relevant bacterial phylum (order *Planctomycetales*). This phylum includes four validated genera: *Pirellula*, *Gemmata*, *Isosphaera* and *Planctomyces*. Another group of organisms which was not yet isolated in pure cultures, the so-called "Anammox"-bacteria, also belongs to the *Planctomycetes*, but has no validly described genus. The main properties of the *Planctomycetes* are presented shortly in the

following sections.

### 1.3.1. Environmental relevance

Bacteria belonging to the phylum *Planctomycetes* have been isolated from terrestrial, freshwater and marine habitats<sup>13,14,15</sup>. This shows that representatives of this phylum successfully colonized a broad spectrum of ecological niches. However, as the majority of the bacterial diversity present in environmental samples cannot be cultivated and isolated under laboratory condition, the currently available isolated *Planctomycete* strains probably show only a partial picture of their physiology and ecological roles. Nevertheless, modern molecular methods allow the *in situ* detection and quantification of microorganisms in environmental samples without prior cultivation<sup>16</sup>. Studies based on these techniques repeatedly confirmed the occurrence of *Planctomycetes* in terrestrial and aquatic habitats<sup>17,18,19</sup> (Fig. 8). Further surveys reported the presence of *Planctomycetes* in marine-snow particles<sup>20</sup> (phytodetrital macroaggregates of the water column) and also in marine sediments<sup>22</sup>. In these studies, *Planctomycetes* were shown to represent 1 to 5% of the *in situ* bacterial community. These results show that their environmental relevance resides more in their ubiquitousness than in their local abundance.

*Planctomycetes* were originally thought to be specialized in the mineralization of organic carbon in natural habitats, but the “anammox” process - the anaerobic ammonia oxidation in freshwater and marine systems - has recently been attributed to new members of the *Planctomycetes*<sup>21,23,24</sup>. However, no isolate corresponding to this process is available so far.

As *Planctomycetes* are widespread in natural habitats and participate in the degradation of organic carbons or the anaerobic oxidation of ammonia, they play an important role in the carbon and nitrogen cycles. By their presence in marine systems, which covers around 70% of the earth surface, the *Planctomycetes* contribute to the fluxes of these elements between the hydrosphere and the atmosphere. A better understanding of these fluxes and their regulation is needed as a basis to assess the impacts of the human activities on natural cycles.

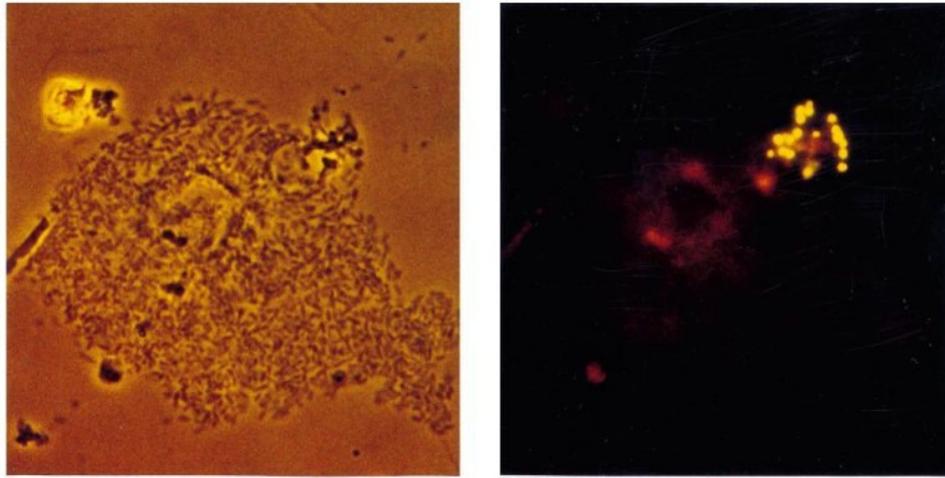


Fig. 8: Illustration of the natural occurrence of *Planctomyces*, as reported in the literature. **Left panel:** phase-contrast microscopy photomicrograph of an organic aggregate from the water of the Elbe river (Germany); **Right panel:** *in situ* labeling of single cells belonging to the *Planctomyces* with a 16S rRNA targeted oligonucleotides probe (same field, epifluorescence). (Original pictures: Bockelmann et al.<sup>19</sup>).

### 1.3.2. Cellular biology

*Planctomyces* show unique cellular biology features which are unexpected for prokaryotic organisms. A striking property of all *Planctomyces* is the occurrence of internal cellular compartmentalization<sup>25,26,27</sup>. In the genera *Pirellula*, *Isosphaera* and *Planctomyces*, a single intracytoplasmic membrane (ICM) separates two compartments. The innermost compartment, termed the riboplasm (R), with respect to the probable occurrence of most ribosomes, also contains the genetic material in the form of a condensed, fibrillar nucleoid. In the genus *Pirellula*, the riboplasm compartment was originally called the “pirellulosome”. In all *Planctomyces*, the outer compartment is termed the paryphoplasm (P) and seems to contain no or few ribosomes (Fig. 9). In members of the genus *Gemmata*, an additional double membrane surrounding the nucleoid has been observed, and anammox organisms contain an additional inner compartment, the anammoxosome (Fig. 9, NE and A). The newly discovered anammox process is thought to take place through the membrane of the anammoxosome<sup>23,24</sup>. However, the biological function of the cellular compartmentalization in the other *Planctomyces* remains unknown.

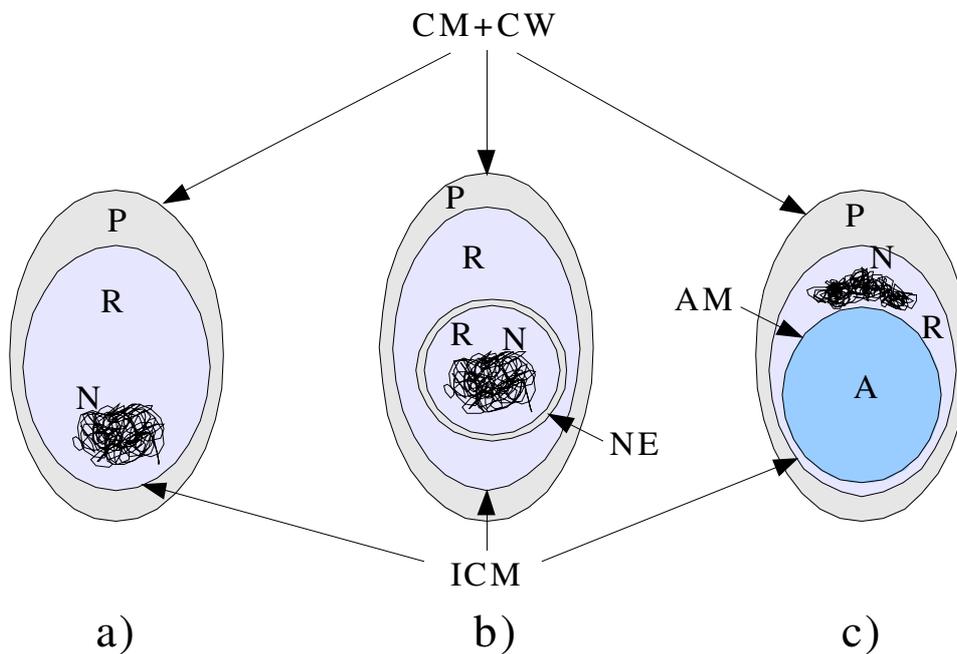


Fig. 9: Cellular compartmentalization in *Planctomycetes* (schematic representation). a) *Pirellula*, *Isosphaera* and *Planctomyces*; b) *Gemmata* and c) Anammox organisms. (CM+CW: cytoplasmic membrane and cell wall (proteinaceous); ICM: intracytoplasmic membrane; NE: nuclear body envelope; P: paryphloplasm; R: riboplasm; A: Anammoxosome; N: condensed fibrillar nucleoid)

Further cellular biology particularities of the *Planctomycetes* include a polar cell organization, a yeast-like cell division and intriguing crateriform structures on some parts of the cell surface, whose function is still unknown. The cell wall contains no peptidoglycan but is stabilized by a proteinaceous layer.

A common property of most *Planctomycetes* is the ability to attach to surfaces with cellular appendages or secreted material (Fig. 10). In some isolated strains, stalks are observed and lead to the formation of rosettes (spherical cells aggregates) or the attachment to natural surfaces. Other isolates have been shown to produce holdfast structures (secreted polymeric substances) also leading to surface attachment. The ability to attach to surfaces in natural habitats might provide an efficient way to access the nutrients of particular niches, such as the marine snow particles in the water column (sinking phytodetrital macroaggregates).

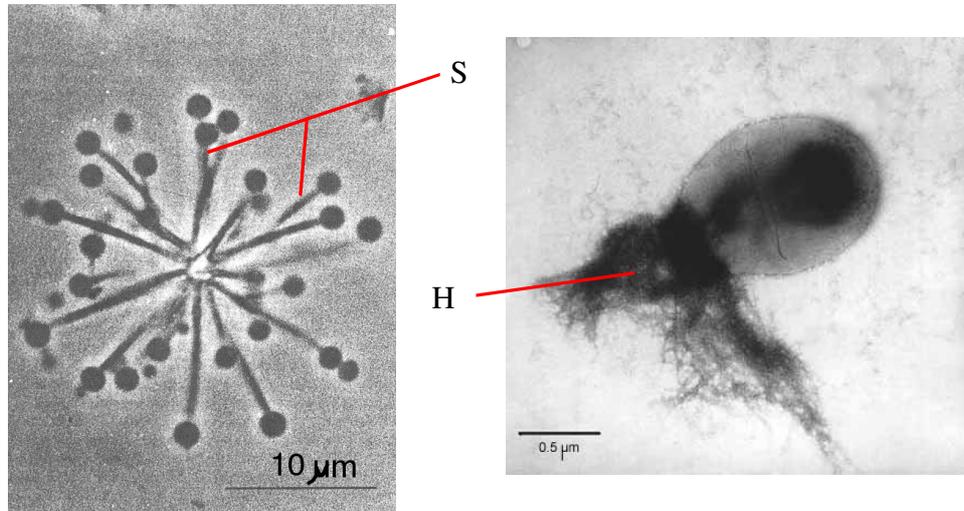


Fig. 10: The ability of the *Planctomycetes* to attach to surfaces in natural habitats is illustrated by the occurrence of cellular appendages like stalks (S) or the production of holdfast structure (H) composed of excreted biopolymers. **Left panel:** phase-contrast micrograph of a rosette of *Planctomyces bekefi*, showing stalks (S) (Original picture: Fuerst JA<sup>15</sup>); **Right panel:** electron micrograph of a single cell of *Pirellula* sp. strain 1, showing the secreted holdfast structure (H) (Original picture: Schlesner H<sup>13</sup>).

### 1.3.3. Phylogeny

According to 16S rDNA-based studies, *Planctomycetes* constitute an independent, monophyletic phylum of the bacterial domain<sup>28,29</sup>. The diversity within this phylum is particularly large. The four genera initially classified according to morphological characteristics (*Pirellula*, *Gemmata*, *Isosphaera* and *Planctomyces*) form distinct phylogenetic clusters (Fig. 11). The high diversity within every genus is shown by 16S rDNA sequence similarities as low as 85-88% between single strains. The phylogenetic position of *Pirellula* sp. strain 1 is shown in Fig. 11.

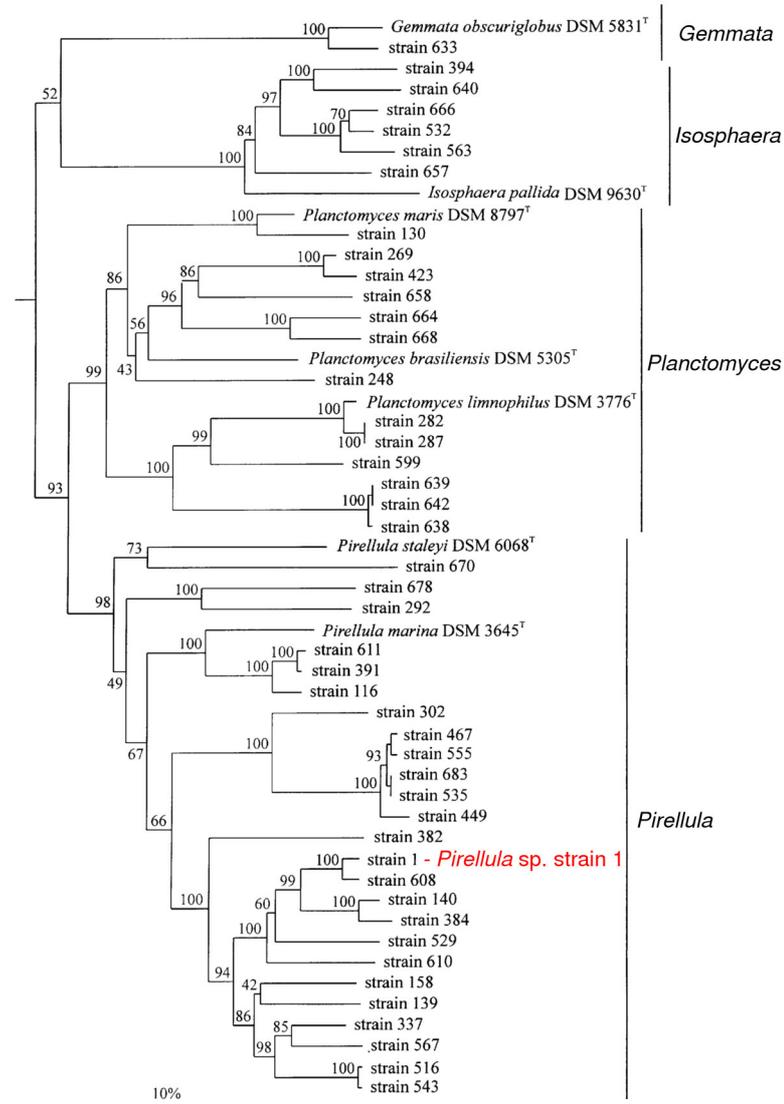


Fig. 11: Phylogenetic diversity within the cultivated *Planctomyces* (source: Gripenburg et al.<sup>28</sup>). The organism selected by this project, *Pirellula* sp. strain 1 (indicated in red), can be regarded as a representative of its genus. Analysis based on 16S rDNA gene sequences. The bar indicates 10% sequence divergence and bootstrap support is indicated for every node.

The four described genera corresponding to cultivated strains only represent a subset of the *Planctomyces* diversity present in the environment. Environmental 16S rDNA clones revealed that at least four new phylogenetic clusters belonging to uncultivated *Planctomyces* exist whose physiology is unknown<sup>23,30</sup>. Interestingly, the physiologically distinct “anammox” organisms constitute one or more additional clusters within the *Planctomyces* (Fig. 12).

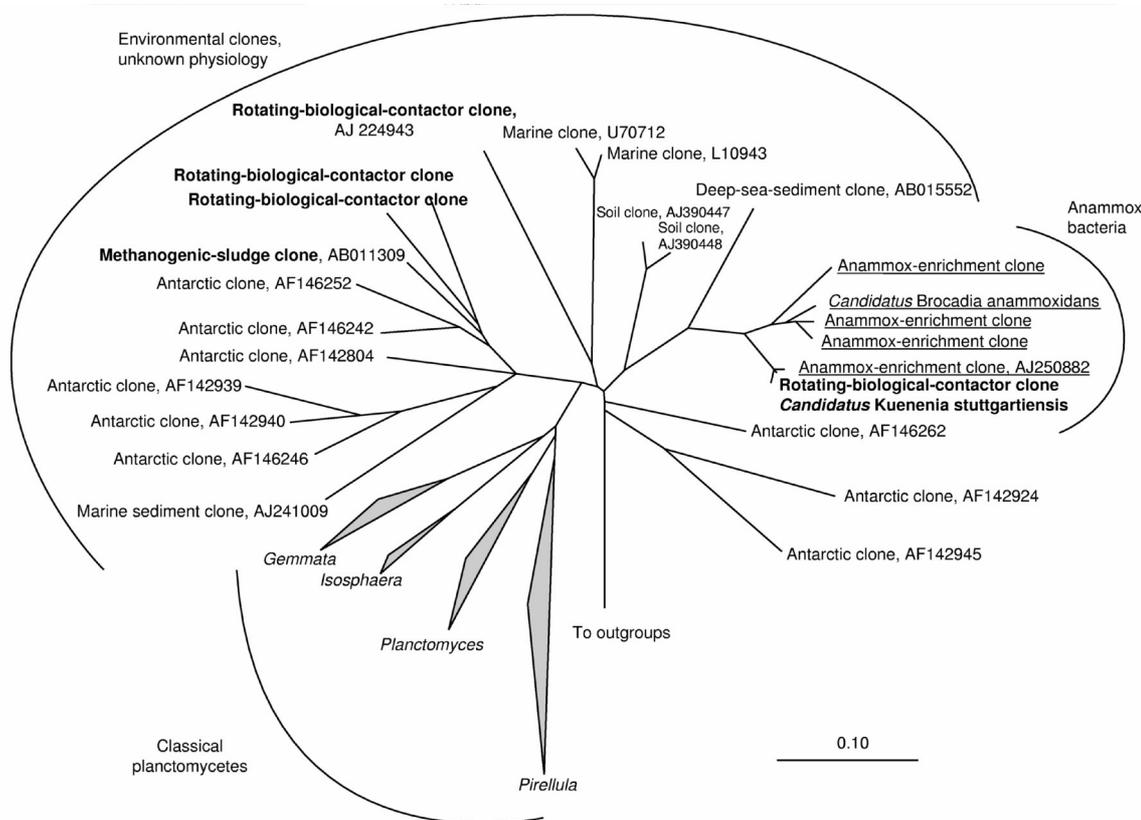


Fig. 12: Phylogenetic diversity within the *Planctomycetes* as revealed by 16S rRNA genes of cultivated and uncultivated organisms (source: Jetten et al.<sup>30</sup>). The four validated genera whose members have been successfully isolated (*Pirellula*, *Gemmata*, *Isosphaera* and *Planctomyces*) seems to represent only a limited part of the natural diversity.

While *Planctomycetes* are clearly monophyletic, their exact branching position within the bacterial domain varies according to the selected phylogenetic reconstruction method. Therefore, the exact branching position of the *Planctomycetes* is still a subject of intense discussions in the literature. Initial 16S rDNA analysis suggested a possible relationship to *Chlamydia*<sup>31,32</sup>, but extensive analysis on larger 16S and 23S datasets did not confirm these results<sup>33</sup>.

Analysis based on EF-Tu, an alternative phylogenetic marker, confirms the monophyly of the *Planctomycetes* as described by 16S rDNA, but also fails to clearly establish the branching position of this phylum within the domain Bacteria<sup>34</sup>.

Recently, the *Planctomycetes* have been assigned a deepest branching position within the *Bacteria* based on a particular selection of slowly evolving nucleotide positions in 16S rDNA genes<sup>36,37</sup>. However, a later analysis relying on alternative 16S positions rather supported a branching of the *Planctomycetes* after thermophilic members of the *Bacteria* (*Thermotogales* or *Aquificales*)<sup>38</sup>. The heterogeneity of the results of these studies shows

that the phylogeny of the *Planctomycetes* is still a challenging question. Therefore, the main interest on the *Planctomycetes* not only resides on their environmental relevance, but also on evolutionary aspects. The availability of the first complete genome of a representative of this phylum, *Pirellula* sp. strain 1, constitute a new data source to discuss the phylogenetic position of this interesting bacterial phylum.

#### 1.4. *Pirellula* sp. strain 1

*Pirellula* sp. strain 1 was isolated from the water column of the Kiel Fjord (Baltic Sea). It is in the process of being described as “*Rhodopirellula baltica*” (Rho.do.pi.rel'lu.la N.L fem. n. *rhodon* the rose; L.n. *pirum* pear; M.L. dim. ending *ella*; M.L. dim. ending *ula*. *Rhodopirellula* very small red pear; bal'ti.ca. L. fem. adj. pertaining to the Baltic Sea, the place of isolation)<sup>39</sup>.

##### 1.4.1. Pre-genomic era: physiological description

*Pirellula* sp. strain 1 is a marine, aerobic and heterotrophic member of the *Planctomycetes*<sup>39</sup>. The cell shape is ovoid, ellipsoidal or pearshaped and the size range is 1.0-2.5 x 1.2-2.3  $\mu$ m. Buds are formed at the broader cell pole. A flagellum is observed at the proximal pole. The optimal growth temperature lies between 28 and 30 °C and no growth is observed above 32 °C. Doubling time is around 10 hours with glucose as carbon and energy source<sup>40</sup>.

*Pirellula* sp. strain 1 seems to be an exclusive marine bacteria, as no growth could be observed in freshwater media. Growth was observed in media containing 12-175% of ASW (artificial sea water, 100% ASW = 34.5 ‰ salinity).

The list of substrates utilized as carbon and energy source by this organism is given in Table 1. A large spectrum of monosaccharides, as well as some di- and polysaccharides are utilized by *Pirellula* sp. strain 1, while the most naturally abundant polysaccharides cellulose and chitin are not hydrolyzed. Ammonia, nitrate and N-acetyl-glucosamine are utilized as nitrogen source. Glucose fermentation was not observed, and nitrate could not serve as electron acceptor.

In summary, these physiological tests describe *Pirellula* sp. strain 1 as a marine, aerobic, carbohydrate specialist.

Table 1: Substrates tested on *Pirellula* sp. strain 1 (data kindly provided by H. Schlesner).

Substrates used as carbon source	Substrates <u>not</u> used as carbon source
<u>monosaccharides:</u> C6: glucose, fructose, mannose, galactose, trehalose C5: lyxose, ribose, xylose <u>modified monosaccharides:</u> methylated: rhamnose N-acetylated: N-acetylglucosamine others: esculin, salicin <u>disaccharides:</u> cellobiose, lactose, maltose, sucrose, melibiose, amygdalin <u>trisaccharides:</u> melezitose, raffinose <u>polysaccharides:</u> chondroitine sulfate, gelatine, starch, dextrin <u>C1:</u> - <u>others:</u> glycerol, gluconate, glucuronate	<u>monosaccharides:</u> C6: sorbose C5: - <u>modified monosaccharides:</u> methylated: fucose N-acetylated: - others: - <u>disaccharides:</u> - <u>trisaccharides:</u> - <u>polysaccharides:</u> cellulose, chitin, alginate <u>C1:</u> methylamine, methylsulfonate, methanol <u>others:</u> ethanol, erythriol, adonitol, arabitol, dulcitol, inositol, mannitol, sorbitol, acetate, adipate, benzoate, caproate, citrate, formiate, fumarate, glutarate, lactate, malate, 2-oxoglutarate, phtalate, propionate, pyruvate, succinate, tartrate, norleucine, ornithine, urea, indole, inulin, pectin, casein, tween 80, all 20 amino acids

#### 1.4.2. Whole genome sequencing (MPI-Berlin)

The whole genome sequence of *Pirellula* sp. strain 1 was determined at the Max Planck Institute for Molecular Genetics in Berlin<sup>41</sup>. A shotgun sequencing strategy with extensive gap closure was applied and resulted in a single circular chromosome of 7.15 Mb. The final overall sequencing redundancy reached 8x coverage, which represents a high quality standard.

The availability of the genome of *Pirellula* sp. strain 1 offers, for the first time, the opportunity to study the complete genetic blueprint of a representative organism of the *Planctomycetes*. The genome of *Pirellula* sp. strain 1 is complete, reaches high quality

standards (8x coverage) and therefore constitute a valuable, permanent resource for the scientific community.

## 2. Material and methods

### 2.1. Locally maintained bioinformatic tools and databases

The establishment of the appropriate structures for bioinformatic analysis of whole genomes and metagenome fragments requires specialized software. Most algorithms and databases specialized in biological sequences emerged from academic projects and are freely accessible through the Internet. However, large scale analysis, optimal data access and performance can only be achieved with local installations. DNA or protein information is distributed in primary and secondary databases (Fig. 13). While most primary databases are a comprehensive source of original sequence information with uncurated descriptions, secondary databases (knowledge databases) build meta-information by grouping, classifying and modeling primary information e.g. according to protein families and domains. Annotation softwares usually rely on both types of databases to allow both comprehensiveness and accuracy. A list of the maintained bioinformatic tools and associated databases in the Microbial Genomics Group of the Department of Molecular Ecology is given in Table 2. The annotation of the genome of *Pirellula* sp. strain 1 provided a practical experience to select the most useful available tools. All local databases are updated on a regular basis in order to re-analyze genomes or metagenomic fragments of various ongoing projects.

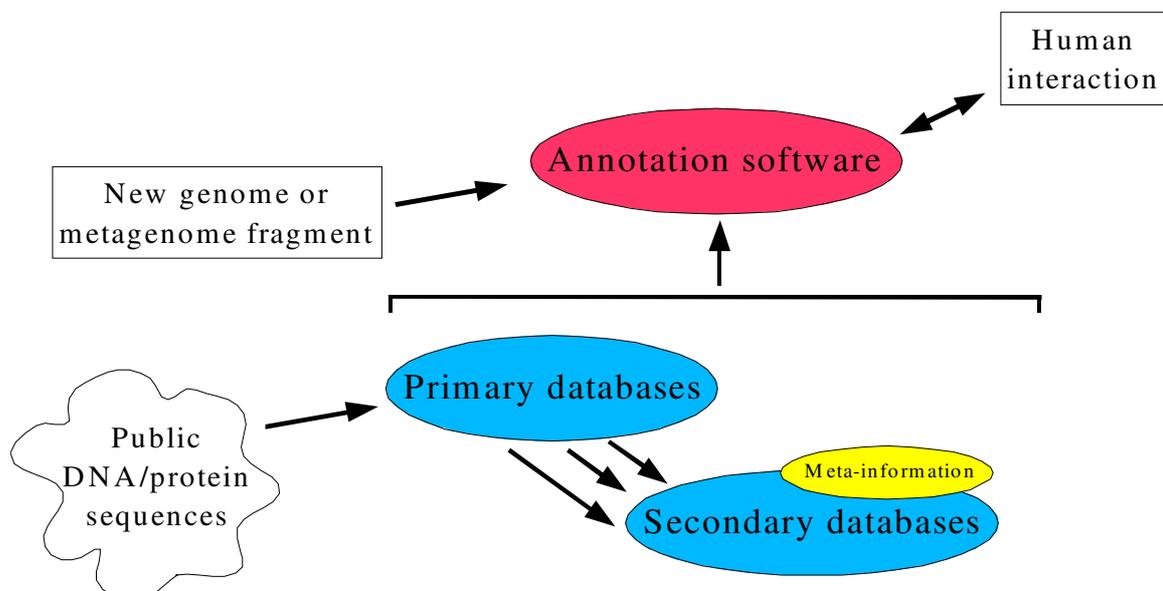


Fig. 13: Primary and secondary databases for genome annotation.

Table 2: Bioinformatic tools and databases for whole genome and metagenomic analysis. All components are installed and maintained locally on a computer cluster in the Department of Molecular Ecology (see section 2.2.4).

Tool / database	Description	Web page / FTP
Data integration:		
GenDB	Annotation software: open database standard, job distribution and graphical user interface	<a href="http://gendb.genetik.uni-bielefeld.de/">http://gendb.genetik.uni-bielefeld.de/</a>
Primary tools/databases:		
BLAST	Pairwise similarity search for nucleotides and protein sequences ( <u>B</u> asic <u>L</u> ocal <u>A</u> lignment <u>S</u> earch <u>T</u> ool)	<a href="http://www.ncbi.nlm.nih.gov/BLAST/">http://www.ncbi.nlm.nih.gov/BLAST/</a> <a href="ftp://ftp.ncbi.nlm.nih.gov/blast/executables/">ftp://ftp.ncbi.nlm.nih.gov/blast/executables/</a>
NCBI-nt	Non-redundant nucleotide database of the <u>N</u> ational <u>C</u> enter for <u>B</u> io <u>t</u> echnology <u>I</u> nformation	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=nucleotide">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=nucleotide</a> <a href="ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/">ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/</a>
NCBI-nr	Non-redundant protein database of the <u>N</u> ational <u>C</u> enter for <u>B</u> io <u>t</u> echnology <u>I</u> nformation	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=protein">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=protein</a> <a href="ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/">ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/</a>
Swiss-Prot	Manually curated protein database of the <u>E</u> xpasy center ( <u>E</u> xp <u>e</u> rt <u>P</u> rotein <u>A</u> nalysis <u>S</u> ystem)	<a href="http://www.expasy.ch/sprot/">http://www.expasy.ch/sprot/</a> <a href="ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb/fasta/">ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb/fasta/</a>
TrEMBL	Non-redundant protein database of the <u>E</u> uropean <u>M</u> olecular <u>B</u> iology <u>L</u> aboratory ( <u>T</u> ranslated <u>E</u> MBL nucleotide database)	<a href="http://www.ebi.ac.uk/trembl/">http://www.ebi.ac.uk/trembl/</a> <a href="ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb/fasta/">ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb/fasta/</a>
PIR	Manually/automatically curated protein database ( <u>P</u> rotein <u>I</u> nformation <u>R</u> essource)	<a href="http://pir.georgetown.edu/">http://pir.georgetown.edu/</a> <a href="ftp://nrfa.georgetown.edu/pir_databases/psd/fasta/">ftp://nrfa.georgetown.edu/pir_databases/psd/fasta/</a>
COG	<u>C</u> lusters of <u>O</u> rthologous <u>G</u> enes database, containing automatically classified gene products of full genomes	<a href="http://www.ncbi.nih.gov/COG">http://www.ncbi.nih.gov/COG</a> <a href="ftp://ftp.ncbi.nih.gov/pub/COG/COG/">ftp://ftp.ncbi.nih.gov/pub/COG/COG/</a>
Prokaryotic genomes	Comparative genomics: complete database of archaeal and bacterial genomes with original annotation information from EMBL, imported in GenDB databases.	<a href="http://www.ebi.ac.uk/genomes">http://www.ebi.ac.uk/genomes</a> <a href="ftp://ftp.ebi.ac.uk/pub/databases/genomes">ftp://ftp.ebi.ac.uk/pub/databases/genomes</a>
Secondary tools/databases:		
HMMER	Search tool based on profile HMM ( <u>H</u> idden <u>M</u> arkov <u>M</u> odels)	<a href="http://hmmer.wustl.edu/">http://hmmer.wustl.edu/</a> <a href="ftp://ftp.genetics.wustl.edu/pub/eddy/hmmer/">ftp://ftp.genetics.wustl.edu/pub/eddy/hmmer/</a>
InterPro	Integrative resource including major protein signatures databases (Prosite, Pfam, Prints, ProDOM, Smart, Tigrfams) and associated metatool (InterProScan)	<a href="http://www.ebi.ac.uk/interpro/">http://www.ebi.ac.uk/interpro/</a> <a href="ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/">ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/</a>
GO	Controlled vocabulary according to the <u>G</u> O consortium ( <u>G</u> ene <u>O</u> ntology) - linked to InterPro	<a href="http://www.geneontology.org/">http://www.geneontology.org/</a> <a href="ftp://ftp.geneontology.org/pub/go/">ftp://ftp.geneontology.org/pub/go/</a>
Pfam	Search against a curated collection of protein families and domains using Markov models (profile HMM)	<a href="http://www.sanger.ac.uk/Software/Pfam/">http://www.sanger.ac.uk/Software/Pfam/</a> <a href="ftp://ftp.sanger.ac.uk/pub/databases/Pfam/">ftp://ftp.sanger.ac.uk/pub/databases/Pfam/</a>
Prosite	Search against a curated collection of protein patterns using regular expressions.	<a href="http://www.expasy.ch/prosite/">http://www.expasy.ch/prosite/</a> <a href="ftp://www.expasy.ch/databases/prosite/">ftp://www.expasy.ch/databases/prosite/</a>
TMHMM	Transmembrane regions prediction based on Markov models	<a href="http://www.cbs.dtu.dk/services/TMHMM/">http://www.cbs.dtu.dk/services/TMHMM/</a> *
SignalP	Signal peptide prediction based on neural networks and Markov models	<a href="http://www.cbs.dtu.dk/services/SignalP/">http://www.cbs.dtu.dk/services/SignalP/</a> *
Phylogeny softwares:		
Phylip	Software package for nucleotides and proteins phylogenetic inference (distance methods, maximum parsimony and maximum likelihood)	<a href="http://evolution.genetics.washington.edu/phylip.html">http://evolution.genetics.washington.edu/phylip.html</a>
Tree-puzzle	Nucleotides and proteins phylogenetic inference using heuristic (maximum likelihood)	<a href="http://www.tree-puzzle.de/">http://www.tree-puzzle.de/</a>

\*Non-commercial license agreement necessary

## 2.2. Genome analysis: annotation

### 2.2.1. Gene prediction

The first step of genome annotation is to search for all possible coding regions. A combination of three software tools for gene prediction was used for *Pirellula* sp. strain 1: GLIMMER<sup>42,43,44</sup>, CRITICA<sup>45</sup> and ORPHEUS<sup>46,47</sup>. The properties of the corresponding algorithms differ significantly and are summarized in Table 3. In order to merge the results of the three programs, the following rules were applied: i) if predicted genes with the same stop but different start positions differed in less than 10% in length, only the longest version was kept; ii) if the difference was more than 10%, both version were kept for manual inspection.

Table 3: Main properties of the three gene prediction software used for *Pirellula* sp. strain 1. (-: not implemented; +: implemented; ++: advanced implementation).

Program	Use of external database	Ab initio* method	Use of RBS** region	Overlapping resolution
GLIMMER	-	++ (Markov models)	-	relaxed
CRITICA	+(DNA)	+	+	strict
ORPHEUS	+(Proteins)	+	++	strict

\*Database independent search method

\*\*Ribosomal binding site

### 2.2.2. Software package Pedant Pro

Pedant Pro is a commercial software<sup>48,49</sup> for semi-automatic and manual genome annotation (Biomax informatics AG). It is an integrative package based on external and heterogeneous bioinformatic tools which are presented in a single and user-friendly web interface (Fig. 14). A detailed description of this package is given in section 3.1.3 in a comparative study with the GenDB package. The main tools and databases used by Pedant Pro are summarized in Table 4. While the annotation process was going on, these databases were updated by direct downloads from the corresponding public servers. The visualization software Artemis<sup>50</sup> has been used to complement the Pedant Pro package for dynamic genomic context visualization (Fig. 15).

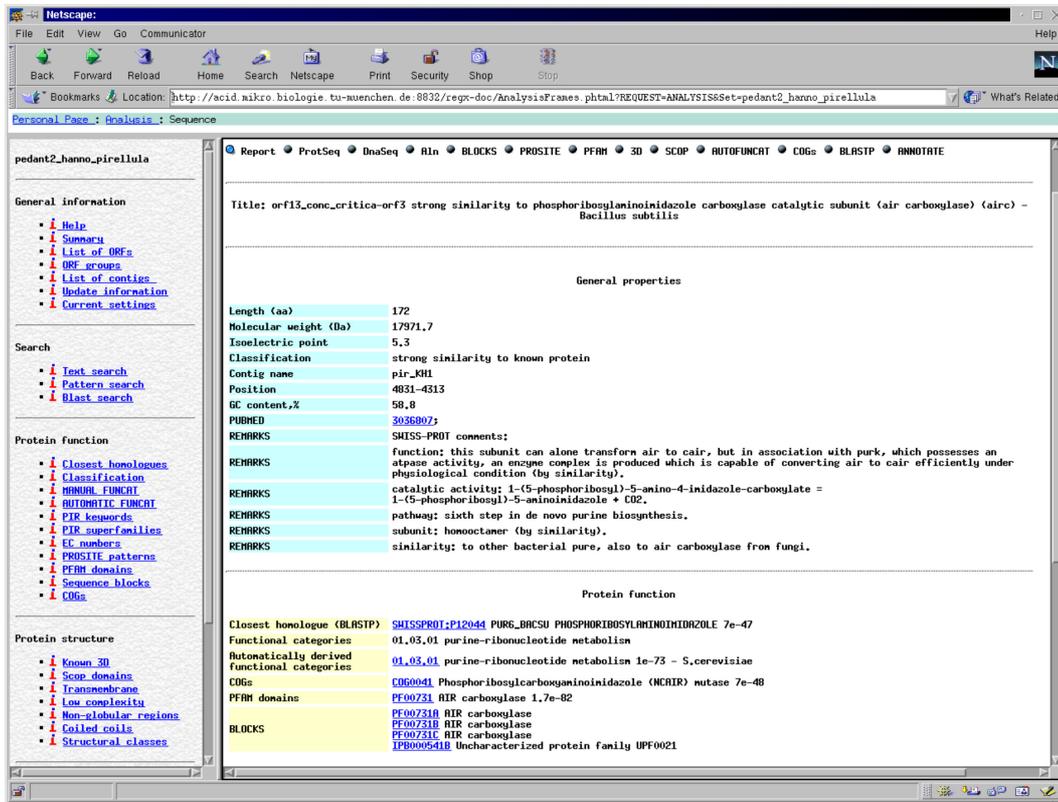


Fig. 14: The Pedant Pro package for genome annotation. This screenshot shows the best functional bioinformatic predictions (lower part) and the functional assignment (higher part) for a predicted gene of *Pirellula* sp. strain 1.

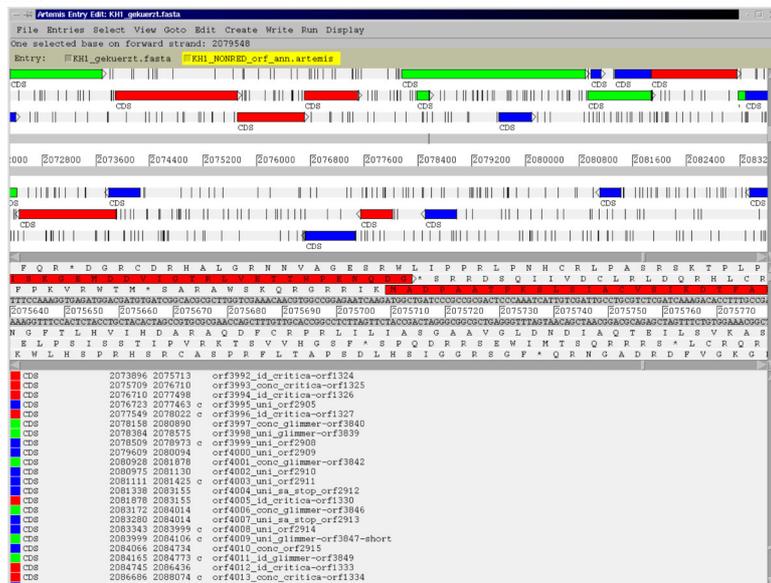


Fig. 15: The Artemis software for genomic context visualization. Higher section: overview of genes predicted by one or different programs; middle section: zoom at the sequence level; lower section: gene identifiers.

Table 4: Main external bioinformatic tools and databases integrated in the Pedant Pro package.

Tool / database	Description	Reference
BLASTP / protein databases	Pairwise similarity search against all known proteins (non-redundant list).	52
BLASTP / COG database	Pairwise similarity search against the <u>C</u> lusters of <u>O</u> rthologous <u>G</u> enes database, containing automatically classified gene products of full genomes.	174,175
Pfam	Search against a curated collection of protein families and domains using Markov models (profile HMM).	53
Prosite	Search against a curated collection of protein patterns using regular expressions.	176
Blocks	Search against an automatically generated collection of protein blocks (PSSM method).	177
ALOM2	Transmembrane regions prediction based on local hydrophobicity.	178
SignalP ver. 1	Signal peptide prediction based on neural networks.	144

### 2.2.3. Software package GenDB

GenDB is a genome annotation software package developed by the Department of Genetics of the University of Bielefeld<sup>51</sup>. It is an academic project and its use is free for research applications. A local GenDB version 1.1 has been installed on a specific computer cluster using jobs distribution. A detailed description of this package is presented in section 3.1.3 in a comparative study with the Pedant Pro package. The version 1.1 of GenDB offers a graphical user interface (GUI) which integrates the visualization of the genomic context, bioinformatic results and annotation (Fig. 16). The underlying databases are updated in a regular basis (Table 5).

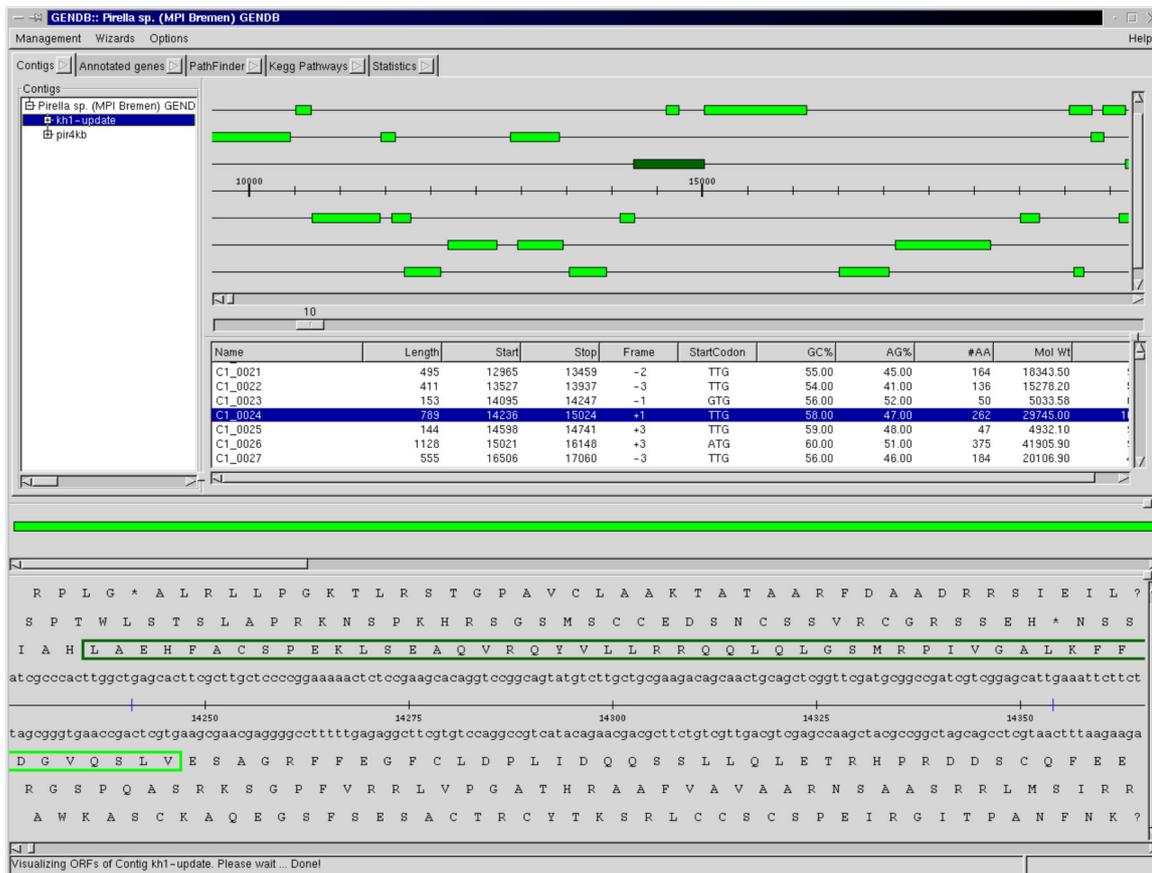


Fig. 16: The GenDB package for genome annotation. This screenshot shows the integrated genomic context visualization. Bioinformatics results and annotation windows are directly accessible for each predicted gene.

Table 5: Main external bioinformatic tools and databases integrated in the GenDB package (version 1.1).

Tool / database	Description	Reference
BLASTP / protein databases	Pairwise similarity search against all known proteins (non-redundant list).	52
Pfam	Search against a curated collection of protein families and domains using Markov models (profile HMM).	53
InterPro	Search against an integrative resource integrating major protein signature databases (Prosite, Pfam, Prints, ProDom, Smart and Tigrfams).	150
GO	Controlled vocabulary extraction from InterPro according to the GO consortium ( <u>Gene Ontology</u> ).	61,181
TMHMM	Transmembrane regions prediction based on Markov models.	179,180
SignalP ver. 2	Signal peptide prediction based on neural networks and Markov models.	144

#### 2.2.4. Computation clustering

The computational needs for whole genome analysis and genome comparisons are extensive and require specialized infrastructures. Moreover, metagenomic data generated by parallel projects in the laboratory of the Molecular Ecology Department constitute an additional computational load which has to be processed through the same pipeline. Therefore, a Linux-based computer cluster based on the job distribution system of GenDB has been set up (Fig. 17).

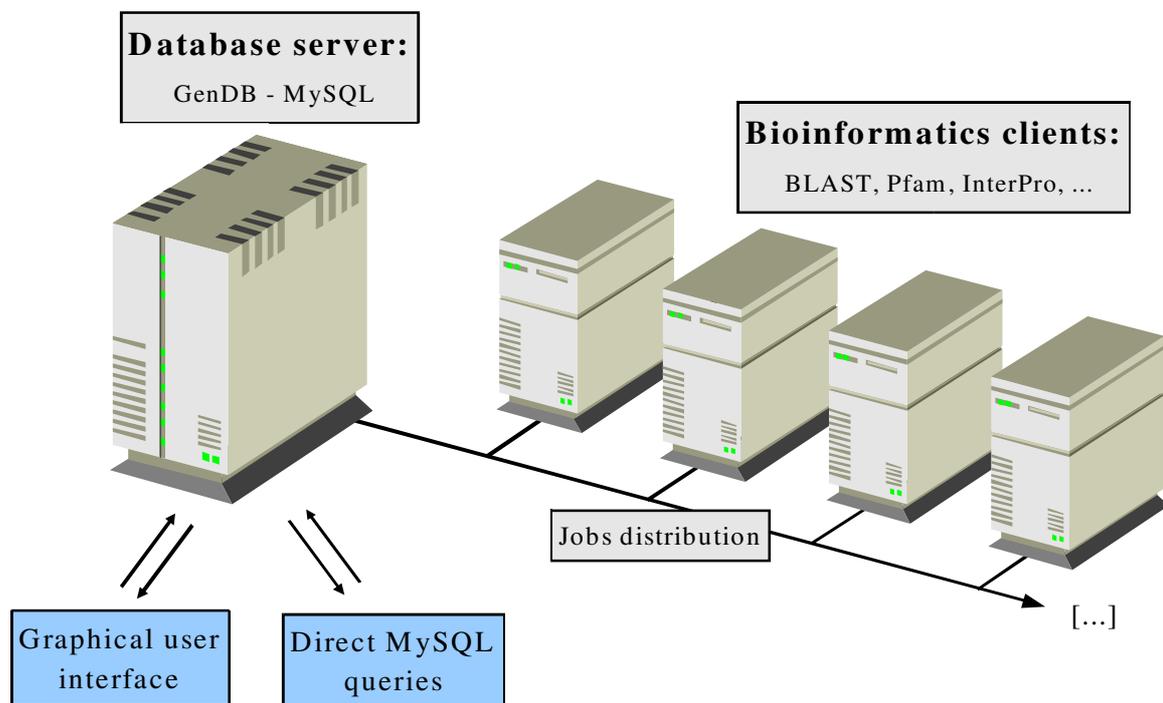


Fig. 17: Computer clustering solution for whole genome analysis. The server is specialized in database transaction for job distribution and interaction with the user. The jobs are distributed to the clients for each gene/tool combination and the results are parsed in the central MySQL database (server: Dual PIII - 700 Mhz, 1 GB RAM, SCSI RAID storage; clients: P4 - 2.4 GHz, 1.5 GB RAM or PIII - 700 Mhz, 512 MB RAM).

#### 2.2.5. Public BLAST server

In order to allow the scientific community to access and query the genomic data produced by the REGX project, a publicly available BLAST server has been set up. The wwwBLAST package, distributed by NCBI<sup>52</sup> has been installed on a low-cost web server (Intel P4 - 2.4 Ghz, 512 MB RAM). This configuration delivers the BLAST similarity search results within a few seconds (Fig. 18). Nucleotide or protein databases are available for the three genomes of the REGX project at the following address: [www.regx.de/blast](http://www.regx.de/blast).



Fig. 18: The public BLAST server of the REGX project ([www.regx.de/blast](http://www.regx.de/blast)). Any protein or nucleotide query sequence can be submitted against the genome of *Pirellula* sp. strain 1, *Desulfotalea psychrophila* or *Desulfobacterium autotrophicum*.

## 2.3. Consistent genome comparisons

### 2.3.1. The Pfam database

The Pfam database<sup>53</sup> is a curated collection of protein domains and protein families. Each Pfam entry is created in a two steps process. First, a multiple alignment of manually selected proteins or protein fragments of the same function is built with classical algorithms such as those implemented by the ClustalW<sup>54</sup> or the T-Coffee<sup>55</sup> software. The boundaries of a domain of known function is refined based on conserved residues and known 3D structures, if available. Second, a representative model for this multiple alignment is built with the hmmbuild program of the HMMER package<sup>56,60</sup>. Each Pfam entry thus consists of a multiple alignment of a protein domain and an associated model.

Any new sequence of unknown function can be aligned and scored against this model with the hmmpfam program and a predefined cut-off score is used to determine the significance of the match (Fig. 19).

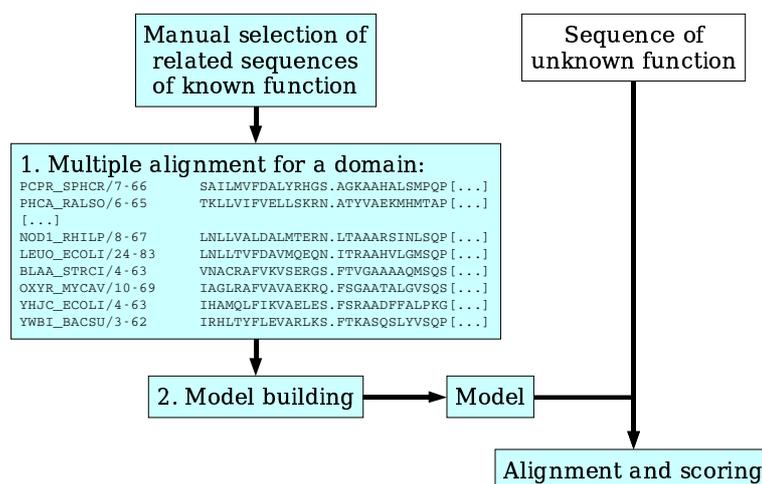


Fig. 19: Pfam database work flow. Instead of matching the sequence of unknown function against each sequence of known function, a model is build to summarize family or domain information.

### 2.3.2. Profile hidden Markov models

Profile hidden Markov models (HMM) constitute the core of the Pfam database. An HMM can be seen as a model that generates sequences based on a given model architecture and associated probabilities<sup>56,57,58,59</sup>. In a profile HMM for biological sequences, several possible states are defined for each sequence position. The three possible states are: match, insertion or deletion. Each position can have one of the three states which is associated with an “emission probability” for each character (one of the 20 amino acids). The “emission probability” is defined as the probability to see a given character in the given state. The transitions between the states are associated with “transition probabilities”. In the case of the HMMER software models used by Pfam, the number of possible transitions is seven. This property gave its name to the model architecture: Plan 7 (Fig. 20).

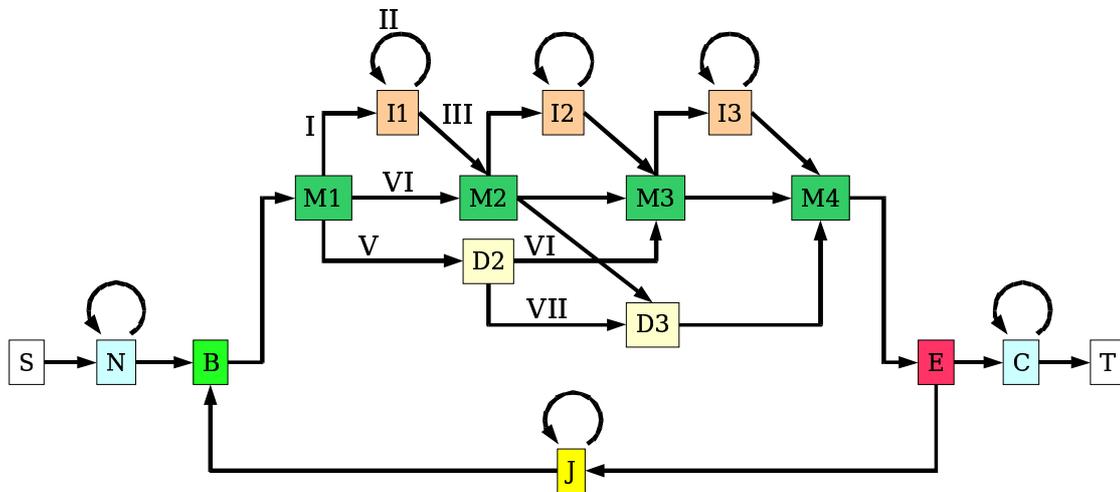


Fig. 20: The Plan 7 architecture - Markov model implementation of the HMMER 2.1 software package. Squares indicate possible states (S: start; N: N-terminus; B: begin, entering model; with  $i=1,2,3,4,\dots$  :  $M_i$ : match at position  $i$ ;  $I_i$ : insertion;  $D_i$ : deletion; E: end, exiting model; J: joining segment; C: C-terminus; T: terminated). The seven transition types of the main model are indicated by roman numbers (I:  $M \rightarrow I$ ; II:  $I \rightarrow I$ ; III:  $I \rightarrow M$ ; IV:  $M \rightarrow M$ ; V:  $M \rightarrow D$ ; VI:  $D \rightarrow M$ ; VII:  $D \rightarrow D$ ).

A Pfam model is thus defined by a set of position-specific “transition” and “emission probabilities” within the Plan 7 architecture. Qualitatively, the “transitions probabilities” define the possible paths across the model and the “emission probabilities” reflect the amino-acid expected at each position. To build such a model from a multiple sequence alignment, the hmmbuild program converts the observed counts of transitions and amino-acids emissions for each position into probabilities. The probabilities are then turned into log-odd scores to take background amino-acids frequencies into account. Background amino-acids frequencies are those observed in the public databases of proteins.

The general log-odd score expression for an amino-acid  $x$  is:

$$\log\text{-odd score}_x = \frac{p_x}{f_x}$$

where:

$p_x$ : emission probability of amino-acid  $x$  for the match/insertion state of this position (according to the multiple alignment);

$f_x$ : expected background frequency of amino-acid  $x$ , as observed in public sequence databases.

For practical programming reasons, the log-odd scores are calculated as follow in HMMER 2.1:

$$K_x = \text{integer}[\text{floor}(0.5 + (\text{INTSCALE} * \log_2(\frac{p_x}{f_x})))]$$

where:

$K_x$ : rounded log-odd score for amino-acid x;

integer, floor: rounding functions;

INTSCALE: scaling value arbitrary set to 1000;

$p_x$ : as above;

$f_x$ : expected background frequency of amino-acid x, as observed in the SWISS-PROT database (a Plan 7 null model is used for this value<sup>60</sup> and  $f_x$  is set to 1 for state transitions).

A model contains 47 log-odd scores (K values) for each position of the original multiple alignment (20 for the match state, 20 for the insertion state and 7 for each possible transition). A typical Pfam model in the HMMER ASCII file format is shown in Figure 21.

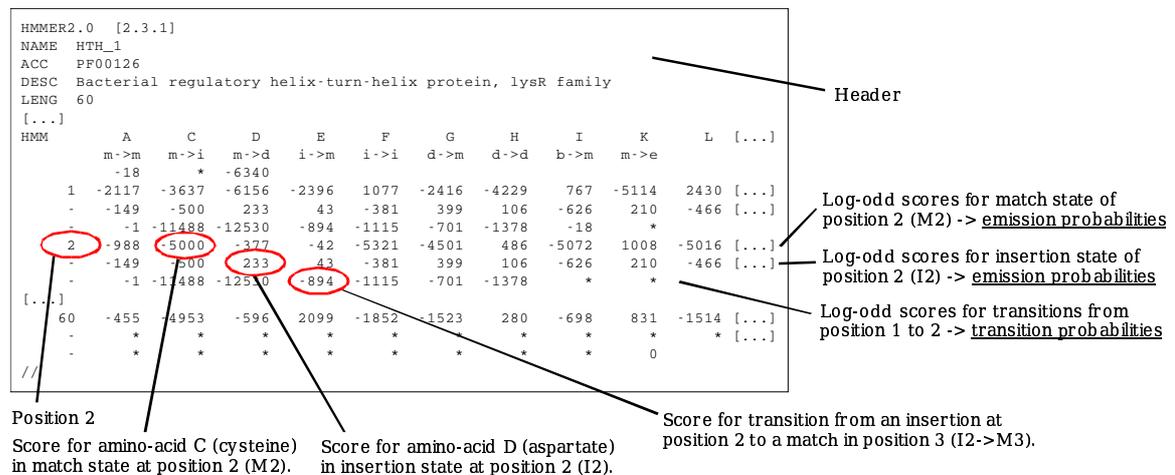


Fig. 21: A HMMER model (Pfam entry PF00126). Only part of the header and some positions are shown.

Once the model is built, alignment and scoring against any single sequence of unknown function is possible. For this purpose, the optimal path through the model which generates the sequence is searched.

The probability that the sequence is generated by the model using a given path is:

$$\text{Prob}(s, \text{path} \mid \text{model}) = t(B \mid M_1) t(M_N \mid E) \prod_{i=1}^N [t(Q_{i-1} \mid Q_i) p(x_{1(i)} \mid Q_i)]$$

where:

s: sequence containing amino-acids  $x_1..x_L$ ;

i: position in the model ( $i=1..N$ );

$t(B \mid M_1)$ : transition probability from the begin state to the first match state;

$t(M_N \mid E)$ : transition probability from the last match state (position N) to the end state;

$Q_i$ : a state of the model at position i;

$t(Q_{i-1} \mid Q_i)$ : transition probability between two states;

$p(x_{1(i)} \mid Q_i)$ : emission probability of amino-acid  $x_1$  in the state  $Q_i$ .

According to the Viterbi algorithm<sup>57</sup>, the following distance between the sequence and the model is searched for a minimum, trying each path through the model:

$$\text{dist}(s, \text{model}) = \min_{\text{paths}} [ -\log \text{Prob}(s, \text{path} \mid \text{model}) ]$$

where:

$\text{Prob}(s, \text{path} \mid \text{model})$ : probability that the sequence is generated by the model using a given path, as defined above.

When the optimal alignment between the sequence and the model is found, the overall score is finally reported in bits (log-odd score) and e-value.

In order to perform consistent cross genomes comparisons, relevant Pfam entries have been selected and grouped according to GO terms associations (Gene Ontology, controlled vocabulary<sup>61</sup>). Selected groups included Pfam entries related to sulfatases, polysaccharide degradation, transport and regulation. All protein encoding genes of fully sequenced *Bacteria* and *Archaea* were searched for significant hits to these groups (hmmpfam, e-value <  $10^{-3}$ ). Results were normalized when necessary to genome size or gene content of each organism for quantitative comparisons.

## 2.4. Codon usage analysis

### 2.4.1. Codon Adaptation Index (CAI)

The large amount of single gene sequences from model organisms already available in the pre-genomic era allowed statistical analysis of codon usage. Because the genetic code is based on 64 possible DNA triplets (codons) encoding only 20 amino acids, distinct but "synonymous" codons can be used for the same amino acid along a gene. Interestingly, the occurrence of synonymous codons is not random: a clear correlation between

experimentally verified gene expression levels and codon preference bias at the DNA sequence level has been observed. For example, among known highly expressed genes (e.g. ribosomal protein genes in *Escherichia coli*), a set of codons is preferentially used. This bias was assigned to a selection for translational efficiency corresponding to the tRNA pool present in the cell. To quantify this early observation, the Codon Adaptation Index<sup>62</sup> (CAI) was introduced in 1987.

In order to calculate the CAI of a single gene, a codon usage normalization is first carried out:

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}}$$

where:

$RSCU_{ij}$ : relative synonymous codon usage for the  $j$ th codon of the  $i$ th amino acid;  
 $X_{ij}$ : number of occurrence of the  $j$ th codon of the  $i$ th amino acid;  
 $n_i$ : number of alternative codons for the  $i$ th amino acid (one to six);

The CAI for a gene is then defined as the following ratio:

$$CAI = \frac{CAI_{obs}}{CAI_{max}}$$

and  $CAI_{obs}$  and  $CAI_{max}$  are calculated as the geometric mean of the RSCU values for the observed gene and the reference genes:

$$CAI_{obs} = \left( \prod_{k=1}^L RSCU_k \right)^{\frac{1}{L}}$$

$$CAI_{max} = \left( \prod_{k=1}^L RSCU_{kmax} \right)^{\frac{1}{L}}$$

where:

$RSCU_k$ : RSCU values for the  $k$ th codon in the observed gene;  
 $RSCU_{kmax}$ : the maximum RSCU value for the amino acid encoded by the  $k$ th codon in the observed gene, as found in the reference highly expressed genes (ribosomal proteins);  
 $L$ : number of codon in the gene.

For convenience, the relative adaptiveness of a codon is introduced:

$$w_{ij} = \frac{RSCU_{ij}}{RSCU_{imax}} = \frac{X_{ij}}{X_{imax}}$$

where:

$w_{ij}$ : relative adaptiveness of the  $j$ th codon for the  $i$ th amino acid

$X_{ij}$ : number of occurrence of the  $j$ th codon of the  $i$ th amino acid;

$X_{imax}$ : number of occurrence of the optimal codon for the  $i$ th amino acid, as defined by the reference genes.

Therefore, the value of  $w$  will range from zero to one. The optimal codon for each amino acid has  $w = 1$ . The above CAI definition for a gene is then equivalent to:

$$CAI = \left( \prod_{k=1}^L w_k \right)^{\frac{1}{L}}$$

where:

$w_k$ :  $w$  value for the  $k$ th codon in the gene;

$L$ : number of codon in the gene.

The CAI value ranges from zero to one. Genes with codon usage similar to highly expressed reference genes will show higher CAI values. For the genome of *Pirellula* sp. strain 1, reference ribosomal proteins were extracted from the manual annotation and the software package codonW was applied to derive CAI values for each gene<sup>63</sup>.

#### 2.4.2. Karlin-Mrazek (PHX/PA)

The Karlin-Mrazek calculation of gene expression prediction<sup>64</sup> ( $E(g)$ ) is an extension of the codon adaptation index (CAI), taking advantage of the whole genome sequence of an organism. For this calculation, a larger dataset of highly expressed reference genes is used as compared to CAI. This reference includes a weighted set of ribosomal proteins (RP), chaperones (CH) and translation factors (TF), instead of only ribosomal proteins. Beside this minor modification, the most important new parameter is the inclusion of a normalization based on all genes, allowing to classify genes with extreme codon usages as PHX (Predicted Highly Expressed) or PA (Putative Alien).

In order to calculate predicted expression level for each gene of a genome, the notion of codon usage difference is introduced:

$$B(F|G) = \sum_a p_a(F) \left[ \sum_{x,y,z=a} |f(x,y,z) - g(x,y,z)| \right]$$

where:

F, G: two different gene groups;

B(F|G): codon usage difference between gene groups F and G;

$p_a(F)$ : average frequency of amino acid a encoded by the genes of group F;

(x,y,z): codon triplet coding for amino acid a;

$f(x,y,z)$ : normalized average codon frequency for (x,y,z) in genes of group F;

$g(x,y,z)$ : normalized average codon frequency for (x,y,z) in genes of group G;

$f(x,y,z)$  and  $g(x,y,z)$  are normalized for each amino acid codon family according to:

$$\sum_{(x,y,z)=a} g(x,y,z) = 1$$

A gene which is highly expressed is expected to show low codon usage differences with the reference gene sets RP, CH and TF, but a high difference with the complete set of genes. Therefore, predicted expression levels with respect to reference gene sets are defined as:

$$E_{RP}(g) = \frac{B(g|C)}{B(g|RP)}, E_{CH}(g) = \frac{B(g|C)}{B(g|CH)}, E_{TF}(g) = \frac{B(g|C)}{B(g|TF)}$$

where:

g: one gene;

C: all genes;

RP: genes encoding ribosomal proteins;

CH: genes encoding chaperones;

TF: genes encoding translation factors.

A weighted combination of codon usage differences can be used to yield a general index:

$$E(g) = \frac{B(g|C)}{\frac{1}{2}B(g|RP) + \frac{1}{4}B(g|CH) + \frac{1}{4}B(g|TF)}$$

$E(g)$  is a general predicted expression level. A combination of the different predicted

expression levels and codon usage differences is used to classify genes with extreme codon usages as PHX or PA:

**PHX definition:** A gene qualifies as PHX (Predicted Highly Expressed) if the following conditions are fulfilled:

1. At least two of the three predicted expression levels  $E_{RP}(g)$ ,  $E_{CH}(g)$  and  $E_{TF}(g)$  exceed 1.05;
2. The general predicted expression level  $E(g)$  is greater or equal to 1.00.

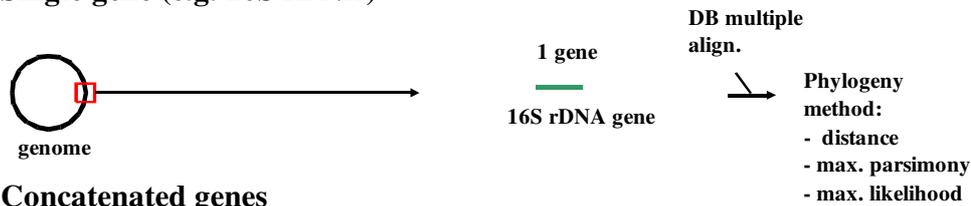
**PA definition:** A gene qualifies as PA (Putative Alien) if  $B(g|RP) > M + 0.15$ ,  $B(g|CH) > M + 0.15$ ,  $B(g|TF) > M + 0.15$  and  $B(g|C) > M + 0.12$ , where  $M$  is the median value of  $B(g|C)$  among all genes.

These Karlin-Mrazek parameters have been calculated for each gene of *Pirellula* sp. strain 1 with a self-written Perl program (Annex 1). Reference gene groups for RP, CH and TF were extracted according to manually refined annotation and Pfam database searches<sup>53</sup>. For data mining, the results were imported in a *Pirellula* sp. strain 1 GenDB database with a short program based on Perl-O2DBI (Annex 2). The distribution of PHX and PA genes in functional categories was estimated according to manually refined “MIPS Funcat” assignments<sup>48</sup>. PHX clusters (2 or more genes) were visually inspected with the GenDB graphical user interface<sup>51</sup>.

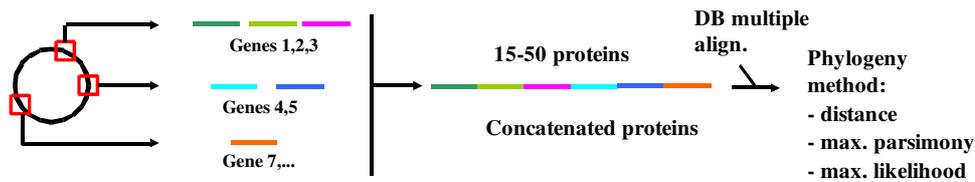
## 2.5. Genome trees: new phylogenetic reconstruction strategies

The genome sequences and genes predictions of 85 organisms (69 *Bacteria* and 16 *Archaea*) have been imported from the EMBL database<sup>65</sup> to a local GenDB system<sup>51</sup> for a total of 231'509 ORFs. Genome Trees have been calculated based on mean normalized BLASTP scores<sup>66</sup> using self-written Perl scripts: i) ORFs involved in fewer than 4 RBM (Reciprocal Best Match between pairs of genomes, BLASTP e-value  $< 10^{-10}$ ) have been excluded from the analysis. ii) matrices were built where distances between genome pairs were 1.0 minus the mean of normalized BLASTP scores of all ORFs of the query genome against the target genome (only e-values  $< 10^{-10}$ ). The normalization was calculated as follows: for each ORF of the query genome, the BLASTP bit-score against the best scoring ORF of the target genome was divided by the self matching bit-score. Tested parameters included: i) exchange of the BLASTP substitution matrix<sup>67</sup> (BLOSUM62, PAM70 and PAM250) and ii) different levels of RBM filtering (number of RBM  $> 4$  to 40). Trees were calculated on the obtained distance matrices by Fitch-Margoliash analysis with the PHYLIP software package<sup>160</sup> (global rearrangements, jumble 100). The particularities of the workflow for the genome tree approach as compared to more classical phylogenetic reconstruction methods is summarized in Figure 22.

◆ **Single gene (e.g. 16S rDNA)**



◆ **Concatenated genes**



◆ **Whole genome (Genome trees)**

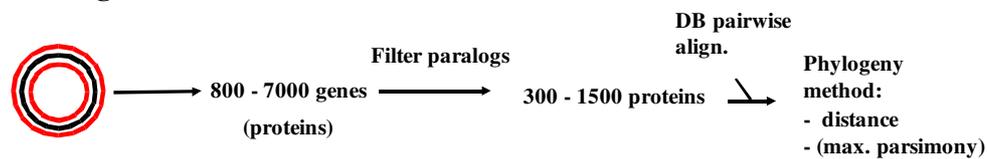


Fig. 22: Differences in the data-flow of the main available phylogenetic reconstruction methods.

## 2.6. Geographic information system

A prototype of a geographic information system (GIS) for marine metagenomics data has been created ("Metagenomes mapserver 0.3"). An Intel-based workstation (P4 - 2.4 Ghz, 1 GB RAM) was set up with the following software components: i) SuSE Linux 8.2 operating system; ii) HTTP server (Apache 1.3.27)<sup>68</sup>; iii) object-relational database server (PostgreSQL 7.3.4)<sup>69</sup>; iv) PostgreSQL extension for OpenGIS standard (PostGIS 0.7.5)<sup>70,71</sup> and v) University of Minnesota MapServer 4.0<sup>72</sup>.

A specific database structure was designed to integrate metagenomic data in a geographical context (see section 3.6). General GIS databases were imported from the GEO public portal of GRID-UNEP<sup>73</sup>. Metagenomic sequences and annotation were retrieved from EMBL<sup>65</sup> and a local project of the Departement of Molecular Ecology (data kindly provided by M. Mussmann). Additional specific information about sampling sites were compiled from the original metagenomics literature.

## 3. Results and discussion

### 3.1. Genome annotation pipeline

#### 3.1.1. Pedant Pro and GenDB database systems comparison

Two software packages for genome annotation have been mainly used during the REGX project. The first one is the proprietary product Pedant Pro<sup>48</sup>, which was licensed at the beginning of the project (Biomax Informatics AG). It seemed to be the best way to rapidly establish a productive system for the project without investing large initial efforts in software development. However, the flexibility of such a software package is limited: it is a static product protected by copyright statements and based on our contract, additional license costs are charged with any new genomic data processed through the system. Moreover, the need for additional external tools or customized applications increased significantly during the project to address biological questions. These requirements can only be matched by using a non-commercial product, which allows a better overall control of the data and which is freely expendable. The GenDB 1.1 software package from the Department of Genetics of the University of Bielefeld<sup>51</sup> was selected according to these criteria to successively replace the initial Pedant Pro installation. GenDB is a second generation software package, a total rewrite based on the experience and known problems associated with first generation solutions such as Pedant Pro. A comparative analysis of these two packages is detailed here to document what has been learned within the first years of genome analysis with respect to system design and architecture. Identifying clear problems and trends between first and second generation packages allows to gain insights into the next logical steps for genomes analysis software design.

##### 3.1.1.1. Software design comparison

The general trend in database systems applied in genome research is to increase complexity allowing better performances and flexibility. Nevertheless, the outdated flat-files based storage solutions consisting of simple ASCII text files (Fig 23, left) are still used for worldwide data exchange (e.g. EMBL genomes files<sup>65</sup>) or by genome visualization software such as Artemis<sup>74</sup>. Text files have the advantage of being easy to read, as long as the data complexity remains limited. They were also well suited to cope with small databases covering e.g. single genes. However, genome annotation requires complex structures to represent the interconnections between large amounts of data of different type produced by bioinformatics analysis. Relational or object-oriented databases constitute powerful tools meeting these requirements (Fig 23, center and right). They allow clear data structures by defining relationships between entities and very efficient query/update mechanisms by the use of indexes.

Pedant Pro and GenDB are closely related with respect to their storage strategies. The

core of both systems is a relational database system. However, Pedant Pro still contains some flat-files components stored in relational tables, while GenDB implements an object-oriented layer over the database. They represent the transition states around relational databases (Fig 23, white boxes). This trend puts fully object oriented databases as the central mechanism for future genome analysis systems.

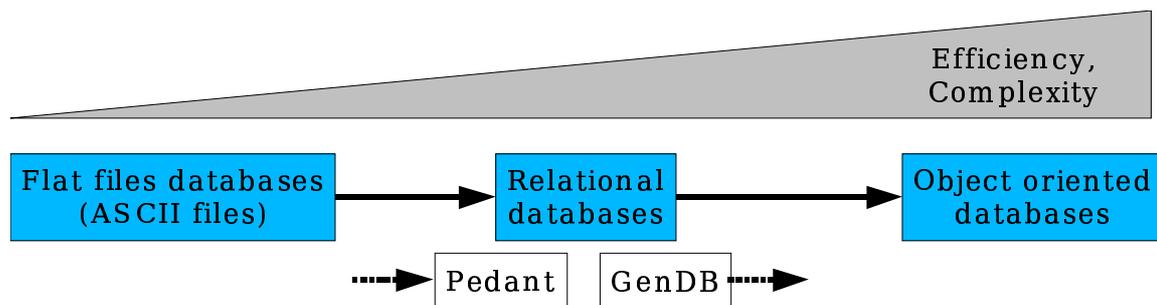


Fig. 23: Trends in bioinformatics database systems. The positions of Pedant Pro and GenDB packages are indicated with respect to their general architecture.

A closer look at the different software modules of Pedant Pro and GenDB is needed to point out the main design differences. As shown in Figure 24, the first difference resides in the core relational databases, implemented in MySQL<sup>75</sup>. In Pedant Pro, information from external bioinformatic tools such as gene-prediction or database searches are directly copied as plain text in the “raw-data” part of the database. When all analysis are computed, a parsing is done to create the “web” part of the database, suitable for access through the web-interface. In contrast, the core database of GenDB only contains a single version of parsed information extracted from the original results from the bioinformatic tools. The redundant information contained in original flat-files is selectively sorted out and only minimal information is stored. Such a strategy minimizes the database size by a factor of 5 to 10, decreases data complexity and increases performance. If the full original information is needed, on-the-fly computations of full results are possible and can be displayed on the user interface.

The second main design difference shown in Figure 24 resides in the Perl-layer surrounding the relational database in both systems. In Pedant Pro, each type of database transaction is operated by unlinked bundles of Perl-scripts, with direct dependencies within the database structure. Adding new features to the system is therefore complicated and error prone. To overcome this problem, GenDB implements a centralized, object oriented Perl-layer (Perl-O2DBI). Each database table is associated with a Perl-object via the classical DBI module (DataBase Interface). This constitutes an abstraction layer which allows to write new software modules without interacting directly with the core database.

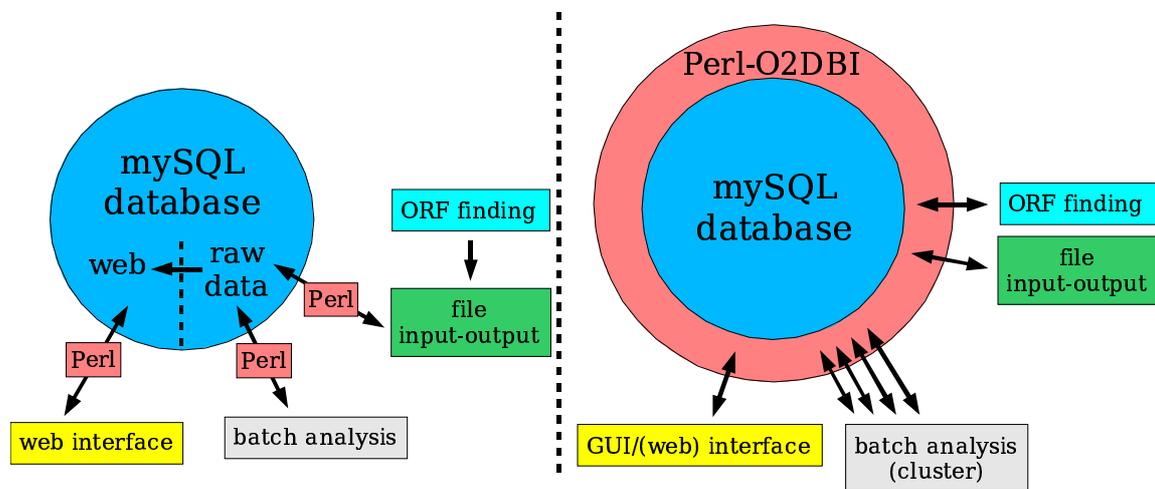


Fig. 24: Comparison of the general system architectures of Pedant Pro (left) and GenDB (right). The Perl-O2DBI layer of GenDB is shown, surrounding the relational database.

### 3.1.1.2. Databases architecture comparison

Structural differences between Pedant Pro and GenDB are also revealed by closely inspecting the relational (MySQL) database architectures. Entity-relationship diagrams (ER-diagrams) are an efficient way to visualize database redundancies or inconsistencies. The ER-diagram of Pedant Pro clearly shows the redundancy in data discussed in the previous section. All bioinformatic results are saved as plain text (Fig. 25, upper tables “\*\_data”) and parsed to summary tables for the web interface (Fig. 25, lower tables). This structure leads to the fact that a large part of the information is stored twice in different tables and fields. Further redundancy can be identified in the Pedant Pro database based on this representation. Many observed relationships between tables are not necessary, such as between contig and blast tables (Fig. 25, red fields and links). Original sequence data is also stored with redundancy: the sequence corresponding to a gene is for example present in 5 tables (contig, contig\_data, orf, orf\_data, prot\_data). Such database design can lead to severe problems while performing update operations as all duplicated information need to be modified at the same time. On the other hand, overall performances for database access are reduced by the larger size. In contrast to Pedant Pro, no redundant information or relationships could be identified in the GenDB ER-diagram (Fig. 26). For example, the sequence corresponding to a gene is only present in one copy (table contig) and DNA or amino-acids sequences are extracted according to the gene positions and the genetic code. Thus, changing e.g. an ORF-coordinate will only require the update of a single field.

Database design consistency can also be checked on the ER-diagrams of Pedant Pro and GenDB. Ideally, every type of information should be stored in a defined entity (table or

object). Both systems group each type of information in separate tables, with the exception of the annotation fields of Pedant Pro. In this case, annotation is distributed in three tables (rep, prot\_data, sel\_funcat), which is problematic for data management (Fig. 26, green fields). GenDB consistently groups all annotation information in a single table, even allowing user-specific multiple annotations for each ORF with a one-to-many relationship.

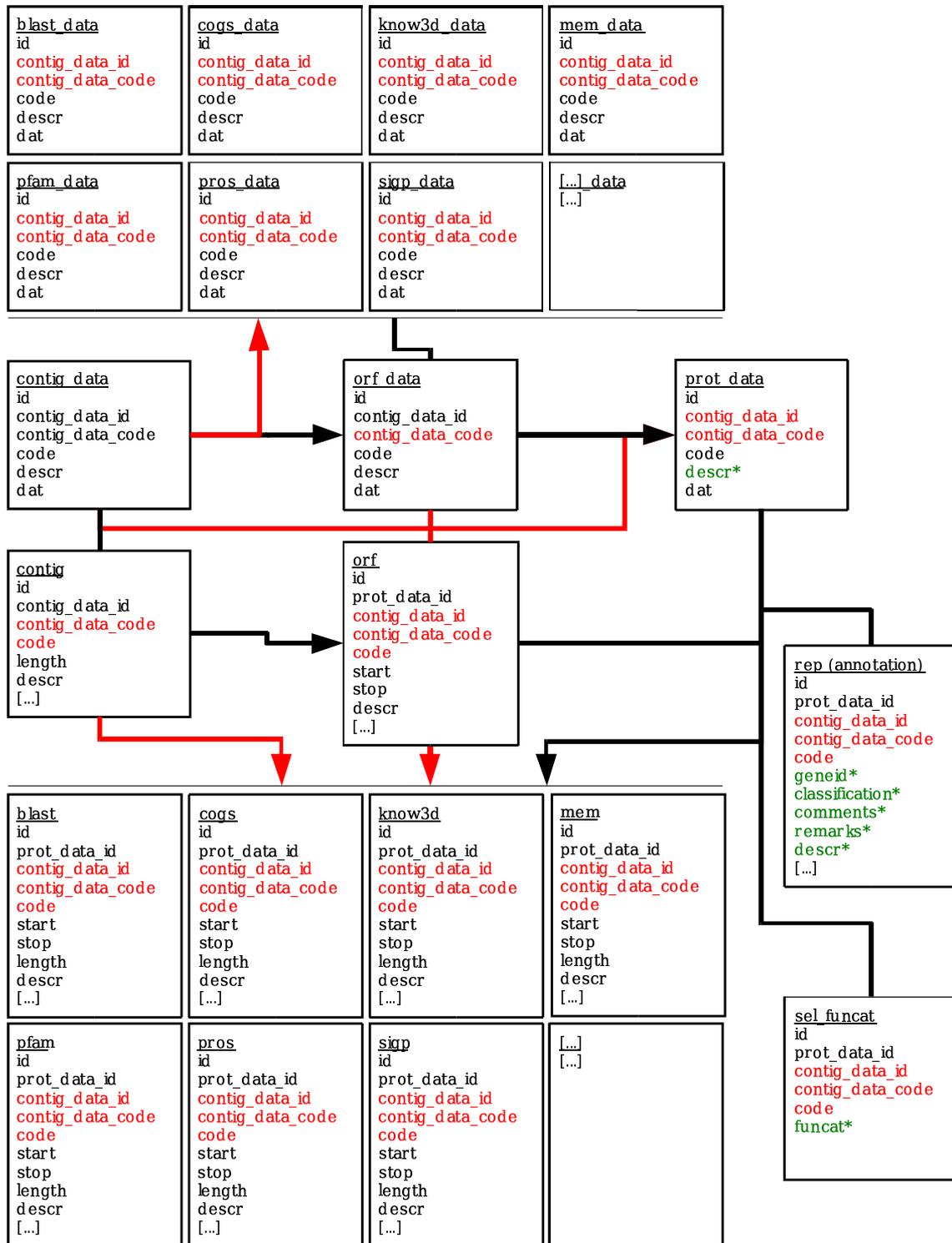


Fig. 25: Pedant Pro relational database architecture (main MySQL tables). Each box represents a database table. Links between tables are symbolized by arrows (one-to-many relationship) or simple line (one-to-one relationship). Redundant table fields or links are shown in red. Fields with annotation data are shown in green (asterisk).

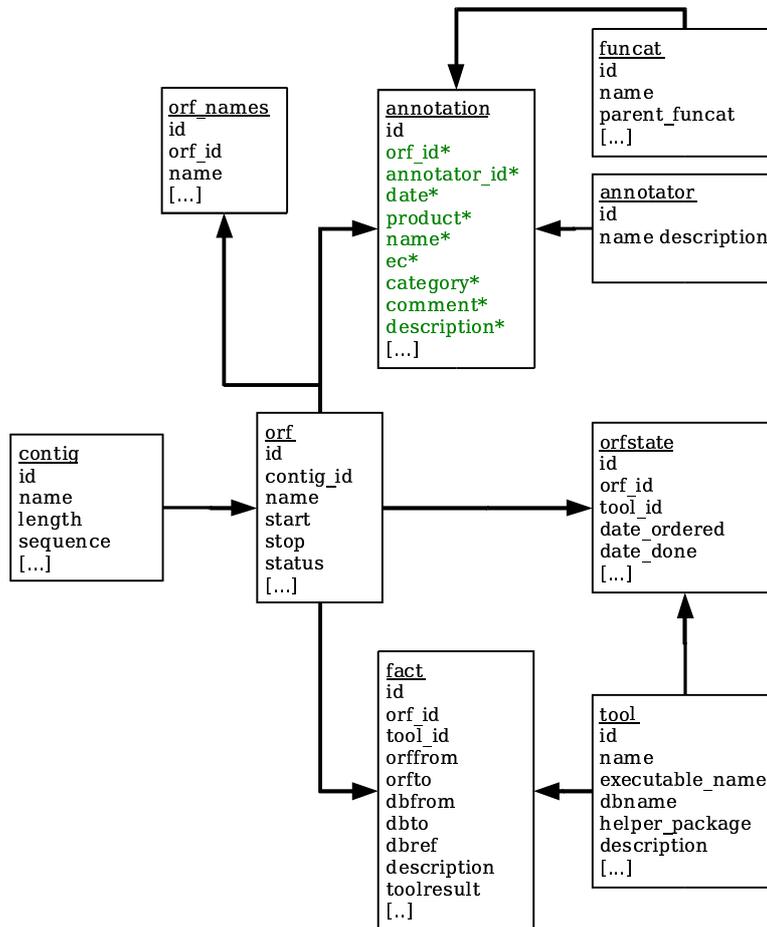


Fig. 26: GenDB relational database architecture (main MySQL tables). All relationships are of type one-to-many (arrows). Annotation fields (green and asterisk) are centralized in one table. No data redundancy can be observed.

### 3.1.1.3. Consequences for future systems

In summary, comparisons between the structures of first and second generation software packages for genome analysis reveal the path for next steps: a fully object oriented system. Key features of such systems will include an object oriented core database surrounded by a programming layer which is easily accessible. The core data structure will have to be strictly non-redundant and tightly interconnected in order to cope with the exponentially growing genome sequences data, and also to efficiently integrate transcriptome and proteome data. Finally, a continuously increasing calculation power will be needed to analyze new genomes and re-analyze older genomes in the context of exponentially growing datasets.

To date, the establishment of the GenDB system constitutes a valuable resource that will also enable to do whole genome analysis of further environmentally relevant

microorganisms. In parallel, the GenDB system is already used intensively for the analysis of genome fragments retrieved directly from environmental samples within the frame of diverse metagenomics projects in the Department of Molecular Ecology (e.g. Anaerobic methane oxidation and wadden sea sediments).

### 3.1.2. Gene prediction

The complete list of genes predicted in *Pirellula* sp. strain 1 by the three different programs was automatically reduced to a non-redundant list of genes and further refined to 7,325 ORFs (Open Reading Frames) by manual annotation. Two common problems could be identified in the gene finding process for *Pirellula* sp. strain 1: Overprediction and inaccuracy of gene start positions. These problems are generally recognized in prokaryotic genome research and recently, a new program with enhanced specificity/sensitivity was published: Z-curve<sup>76</sup>. Moreover, enhanced start position accuracy might now also be achieved with GS-Finder<sup>77</sup>. The use of these programs for further genome projects is expected to improve the quality of the gene finding procedure in combination with other advanced programs like GLIMMER<sup>44</sup>.

### 3.1.3. Automatic annotation / Manually refined functional assignment

The complete genome of a microorganism gives access to a new type of information which allows to describe its environmental potential. The comprehensiveness of such a description is limited by the current knowledge of gene functions contained in public databases and by the accuracy of information transfer to the genes of the newly sequenced organism. For *Pirellula* sp. strain 1, the lack of genetic characterization of *Planctomycetes* led to the fact that the annotation process was based almost exclusively on functional assignment transferred from sequences originating from other branches of the phylogenetic tree. The absence of very closely related sequences for the genes and proteins of *Pirellula* sp. strain 1 was problematic for annotation, but the use of a large set of tools and databases as implemented in the Pedant Pro or the GenDB packages, in combination with manual evaluation allowed to maximize the exploitation of the knowledge present in current databases.

The contribution of the different bioinformatic tools and databases in term of sequence coverage, reliability of the hits and consistency of the annotation vocabulary is summarized in Table 6. While primary protein or nucleotides databases offer the highest coverage, the high number of entries can only be search with methods including heuristics (BLAST<sup>52</sup>) which is only able to deliver a certain probability. Another problem is the lack of controlled vocabulary for protein names, which make the interpretation of the hits difficult. The COG database was shown to be useful for classification purposes, but since it is based on the same heuristic search method (BLAST) and the fact that the original COG classification system is done automatically leads to reliability problems. Secondary databases typically show a lower coverage, but the combination of advanced

modeling/search algorithm (e.g. Markov Models) and manually curated controlled vocabulary assures a very good reliability and consistency for the annotation.

Table 6: Relative contribution of the different bioinformatic tools/databases to the annotation process (++: excellent; +: good; -: low).

Tool / database	Coverage	Reliability	Consistency (e.g. controlled vocabulary)
Primary databases/tools:			
BLASTP / protein databases	++	+	-
BLASTP / COG database	+	+	+
Secondary databases/tools:			
Pfam (protein families)	+	++	++
Prosite (patterns)	-	+	++
InterPro (integrated sec. db)	+	+	++
TMHMM (transmembrane prediction)	+	+	+
SignalP (signal peptide prediction)	+	+	+

For the genes of *Pirellula* sp. strain 1, the automatic bioinformatic results have been confirmed manually and classified by at least two independent annotators to achieve a high quality standard. An overview of the annotation results is given in Table 7. The low proportion of genes with functional assignments reflects the distinct phylogenetic position of *Pirellula* sp. strain 1. The annotation results developed are discussed in section 3.2.

In the future, the availability of controlled vocabulary in primary databases (e.g. GO terms in Uniprot<sup>12</sup>) will increase the compatibility of primary and secondary databases results and facilitate automation. Ideally, 30% to 50% of the genes with common function will be automatically annotated according to GO terms without human intervention, allowing manually refined annotation to concentrate on the remaining data and annotation highlights.

Table 7: General annotation results of *Pirellula* sp. strain 1

Genome property	
Genome size	7,145,576 bp
GC content	55.4%
Number of predicted genes	7,325
Coding density	95%
Average gene length	939
rRNAs	single unlinked 23S-5S and 16S
tRNAs	70
Genes with similarities in databases (BLASTP e-value < $10^{-3}$ )	3,380 (46%)
Genes with functional assignments (BLASTP e-value < $10^{-3}$ )	2,582 (35%)

### 3.2. *Pirellula* sp. strain 1 genome interpretation

#### 3.2.1. DNA compositional asymmetries

Simple compositional indexes can be used to get an overview of prokaryotic genome structures. Such indexes can be calculated over a whole genome, irrespectively of gene-prediction. Therefore, they constitute tools that can be applied very early in a genome project, even before the raw sequencing stage has been finished.

Overall DNA compositional indexes which were shown to be informative in prokaryotic genome research are summarized in Table 8. In term of possible biological interpretation, the best index was shown to be the “cumulative GC skew”<sup>78,79</sup>. It is an integrative function of the original GC skew calculation applied on smaller DNA fragments. Qualitatively, GC skews ignore the biases associated with A and T positions along the sequence and therefore rely on partial information.

Table 8: Definitions and applications of the commonly used DNA compositional indexes (independent from gene prediction). Window sizes for calculation are usually 10 Kb.

Index name	Definition	Applications
GC content	$(G+C) / (A+T+G+C)$	localization of irregularities (HGT*, insertions)
GC skew	$(G-C) / (G+C)$	origin and terminus confirmation (locally)
cumulative GC skew	sum $[(G-C) / (G+C)]$	origin and terminus localization informations on replication mechanisms irregularities localization (HGT*, insertions)
keto excess	sum (GT) – sum (AC)	“
purine excess	sum (AG) – sum (TC)	“

\*HGT: Horizontal gene transfer.

The cumulative GC skew of *Pirellula* sp. strain 1 has been calculated and compared to the results obtained for diverse *Bacteria* and *Archaea*. The results for *Pirellula* sp. strain 1 are presented in the context of representative organisms (Fig. 27). The cumulative GC skew plot of *Escherichia coli* K-12 (Fig. 27, upper left) is very regular with sharp single maxima and minima, as described previously<sup>78</sup>. The maximum and minimum values of this plot has been reported to match the experimentally verified positions of the terminus and the origin of replication of this organism with an accuracy of 10 Kb, respectively<sup>79</sup>. This correlation was interpreted as a consequence of asymmetric DNA compositions between the leading and the lagging strand of replication<sup>80,81</sup>. Furthermore, small inversions in cumulative GC skew were pointed out in the literature<sup>78</sup> between different *Escherichia coli* strains. They could be mapped to small DNA inversions or regions acquired by horizontal gene transfer (cryptic prophages). Such observations were also confirmed in other genomes<sup>82</sup> and constitute a basis for the biological interpretation of local index irregularities in other prokaryotic genomes. Like *Escherichia coli* K-12, most known bacterial genomes display clear cumulative GC skew extrema that can be mapped to the origin and terminus of replication. However, some organisms reveal no clear trend for this parameter. As reported before<sup>83</sup>, the cumulative GC skew of *Synechocystis* sp. shows a very weak and blurred signal (Fig. 27, upper right). Two possible explanations for this observation have been proposed. As frequent genome rearrangements have been suggested by a relatively high number of transposases encoding genes within the *Synechocystis* sp. genome, a high genome plasticity might bias DNA compositional indexes<sup>84</sup>. Alternatively, a possible unknown replication mechanism involving more than one origin might lead to the observed irregularities<sup>78</sup>. However, multiple origins of replication, typical for eukaryotic organisms, could never be demonstrated experimentally in *Bacteria* or *Archaea*. Recent analysis of archaeal genomes lead to the hypothesis that a single representative of this phylum, *Halobacterium* sp. NRC-1, might have two origins of replication<sup>85</sup>, but these results are still awaiting experimental confirmation.

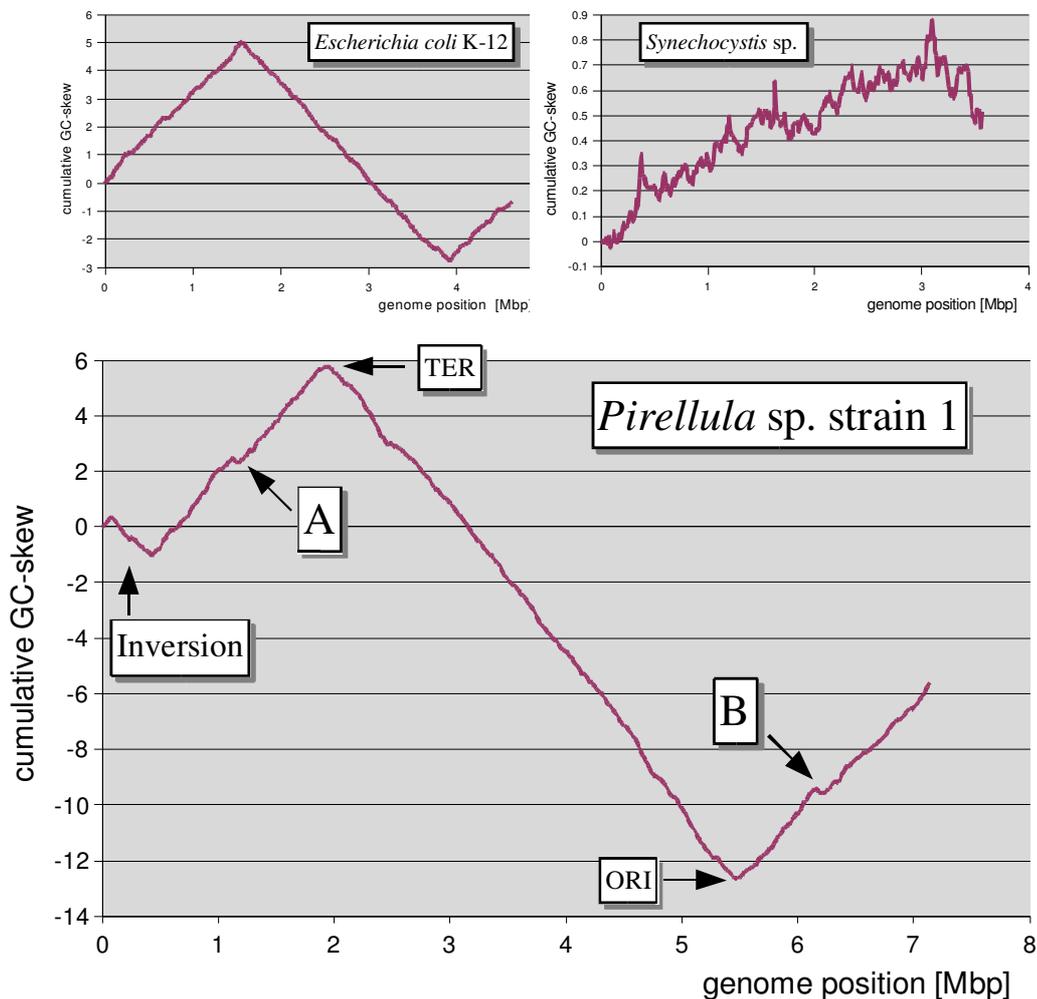


Fig. 27: GC skew of *Pirellula* sp. strain 1 and reference genomes (*Escherichia coli* K-12, *Synechocystis* sp.). Window size: 10 Kb. Scale has been adapted for *Synechocystis* sp. to show fine fluctuations. In *Pirellula* sp. strain 1, a large inversion is observed from position 100'000 to 440'000. A and B indicate notable small irregularities. ORI (pos. 5'460'000) and TER (pos. 1'950'000) indicates proposed origin and terminus of replication.

The cumulative GC skew of *Pirellula* sp. strain 1 shows interesting features (Fig. 27, bottom). Global maximum and minimum values clearly appear and can be proposed as terminus and origin of replication, respectively. Moreover, a large irregularity of a size of 340 Kb is observed within the middle of a replicore (positions 100'000 to 440'000). Possible explanations for the origin of this irregularity in compositional index include: i) a large insertion of foreign DNA; ii) a second origin of replication or iii) an internal chromosomal inversion. The first hypothesis would imply that a sequence of the size of a megaplasmid entered the genome in a single event. Such an event is highly unlikely, because annotation of this section of the *Pirellula* sp. strain 1 genome revealed the presence of necessary housekeeping genes. The second hypothesis for this irregularity – a

second origin of replication – can also be evaluated based on sequence information. The location of the origin of replication in *Bacteria* coincide with the occurrence of one or more genes related to the replication machinery, such as *dnaA*, *dnaN*, *dnaJ* *recF* or *gyrB*<sup>86</sup>. In *Pirellula* sp. strain 1, the global minimum value of cumulative GC skew occurs in the neighborhood of a *dnaN* gene (RB10108), which supports the assignment of the replication origin to this region (Fig. 27, ORI for *Pirellula* sp. strain 1). However, no such typical gene could be found in the neighborhood of the local minimum corresponding to the large irregularity, providing no support for this hypothetical second origin. Furthermore, a second origin at this location would result in a total of two pairs of replicores of different sizes, which would be a disadvantage for replication efficiency. The last and most probable scenario to explain this large irregularity is therefore an internal chromosomal inversion. This possibility is supported by the fact that such events seem to be common among the domain *Bacteria*. This was shown by independent reports describing the so-called “X-plots” or “X-alignments” of closely related *Bacteria* at the DNA or protein level<sup>87,88</sup>. In these plots, large chromosomal inversions between pairs of genomes were observed, often located around the origin or the terminus of replication. This rose the hypothesis that these inversions are related to the replication apparatus by yet unknown mechanisms<sup>87</sup>. The availability of more genomes related to *Pirellula* sp. strain 1 will give insights into such evolutionary mechanisms by using comparative genomics. However, the ongoing sequencing project for *Gemmata obscuriglobus* UQM 2246<sup>89</sup> and *Gemmata* sp. Wa1-1<sup>90</sup> might not provide a sufficient basis for comparison, because the large evolutionary distance which separates *Pirellula* and *Gemmata* within the *Planctomycetes* will most probably be reflected by a low overall sequence structure conservation.

The cumulative GC skew of *Pirellula* sp. strain 1 reveals two other irregularities of smaller size, which is a quite common observation in bacterial genomes (Fig. 27, A and B). The first irregularity (A) could not be correlated to any particular gene content or codon usage (CAI analysis), but the second (B) corresponds to a segment containing a high number of tRNA genes. Interestingly, genome rearrangements or insertion of foreign DNA has been reported to happen adjacent or into tRNA genes in some genomes<sup>91,92,93</sup>. The occurrence of 24 (34 %) tRNA genes in this region out of a total of 70 for the whole genome points this location as a possible hot spot for such events. However, codon usage analysis (CAI) and functional content revealed no particularity in this region, indicating that internal chromosomal rearrangements are more likely in this region than horizontal gene transfer.

As discussed above, GC skew measurements have been shown to be the most informative of all simple DNA composition indexes, but are relying on partial information. On the contrary, alternative indexes such as purine or keto excess make use of the full sequence. However, both usually display the same trends as GC skews, which is also true for *Pirellula* sp. strain 1 (data not shown). In general, the results obtained for these indexes confirm the predicted positions of origin and terminus of replication, the inversion and the two irregularities discussed above for the cumulative GC skew. Other indexes such as

GC content did not show interesting additional features.

The relative importance of the biological processes shaping the described DNA compositional asymmetries, also called the “chromosome polarization” of prokaryotic organisms, are still under discussion<sup>83,80</sup>. However, possible explanations have been formulated within the early years of genome research and still hold true on the large genome datasets available today. Two main parameters are generally considered: unequal mutation rates and selection biases between the leading and the lagging strand of replication (Table 9). In *Pirellula* sp. strain 1, the effect of differences in mutational rates due to the asymmetry of the replication fork might be more important than the effect of selection pressure on coding sequences. This view is supported by the fact that the distribution of genes on both strands is only weakly biased in this organism (~50% of genes on the leading strand). A potential selection pressure at the protein level reflected on the coding regions would have little effect on the observed DNA asymmetries. The same has been reported for several other *Bacteria*<sup>83,81</sup> (*E. coli* K-12: 54%, *Haemophilus influenzae* 54%, *Synechocystis* sp. 50%), but extreme cases of strand preference has been observed in some organisms (*Borrelia burgdorferi* 65%, *Mycoplasma genitalium* 78%). Again, the sequencing of more members of the *Planctomycetes* will allow to estimate if this observation holds true within this phylum.

Table 9: Mechanisms affecting DNA compositional asymmetries in prokaryotic genomes.<sup>80,81,82</sup>

Type	Parameter	Details	Consequences
selection pressure	replication / transcription co-orientation	more genes are located on the leading strand to avoid DNA and RNA polymerase collisions	codon usage can create DNA asymmetries between leading and lagging strand
	oligomer skew	some short sequences (e.g octamers) are preferentially located on the leading or the lagging strand	few, skewed oligomers usually represent a negligible part of the genome
mutational biases	T/G mispairing	T/G mismatch might occur preferentially during leading strand synthesis	more G than C on the leading strand
	+1G frameshifts	G insertion occurs preferentially on the leading strand during G series synthesis	More G on the leading strand
	cytosine deamination	C deamination to T happens more frequently on single stranded DNA	more G and T on the leading strand, less C and more A on the lagging strand (different time single stranded)
	purine/purine mispairing	purine/purine mismatch might occur preferentially during leading strand synthesis	more G and T on the leading strand
selective / mutational combination	Transcription coupled differential repair	more pyrimidine dimers repairs on antisense strand	complex
	Transcription bubble asymmetry	sense strand has a longer single stranded time, more C deamination can occur	complex

### 3.2.2. General genetic potential: an overview

The availability of a nearly complete list of the genes of *Pirellula* sp. strain 1 as revealed by whole genome annotation allows to describe the potential of this organism in a scale that no other approach can reach<sup>A1</sup>. An overview of the complete annotation according to functional categories and selected highlights is presented on a color-coded genome map (Fig. 28).

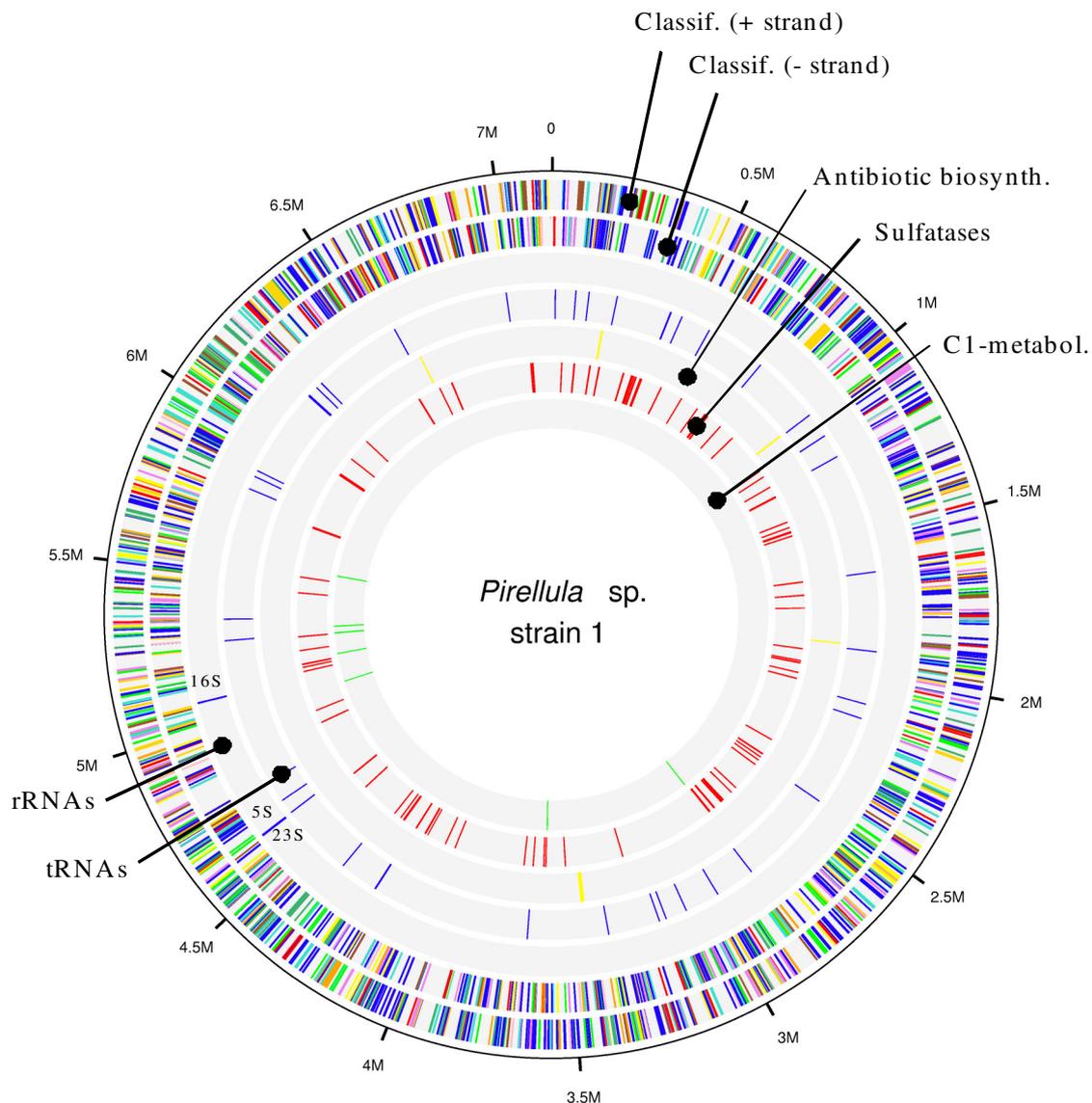


Fig. 28: Overview of the genome of *Pirellula* sp. strain 1, as revealed by the annotation process. Two outer circles: functional classification and distribution of genes on both strands. Inner circles indicate rRNAs (unlinked 5S/23S - 16S), tRNAs and genes involved in particular metabolisms (antibiotic biosynthesis, sulfatases, C1-metabolism). Sulfatases and the C1 metabolism enzymes are discussed in the next sections. Functional classification (two outer circles): blue: METABOLISM; yellow: ENERGY; red: CELL GROWTH, DIVISION AND DNA SYNTHESIS; green: TRANSCRIPTION; orange: PROTEIN SYNTHESIS; violet: PROTEIN DESTINATION; turquoise: TRANSPORT FACILITATION; pink: CELLULAR BIOGENESIS; sienna: CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION; seagreen: CELL RESCUE, DEFENSE, CELL DEATH AND AGEING; gold: CELLULAR ORGANIZATION; -: UNCLASSIFIED.

As *Pirellula* sp. strain 1 is described as an aerobic, heterotrophic bacteria, the corresponding standard pathways are expected to be encoded by its genome. These

pathways have been mapped on the basis of EC numbers assignments (Enzyme Commission). The advantage of this classification system is the fact that a corresponding pathway database is maintained by the KEGG project (Kyoto Encyclopedia of Genes and Genomes<sup>94,95</sup>).

EC numbers were manually confirmed during the annotation process and automatically mapped to the KEGG reference pathway maps within the GenDB system. This visualization gives an overview of metabolic networks. Furthermore, obvious missing links on pathway maps have been searched back into the genome, allowing an amelioration of the annotation quality.

This metabolic reconstruction process has first been done for the following standard pathways: glycolysis, pentose phosphate pathway and citrate cycle (Annex 3). All expected steps were found for these pathway maps, showing that *Pirellula* sp. strain 1 likely uses known enzymes for this central metabolism. Furthermore, the same metabolic reconstruction strategy showed that this organism has the potential to perform the synthesis of all amino acids (Annex 4).

According to laboratory tests, *Pirellula* sp. strain 1 is able to utilize a broad spectrum of monosaccharides, and some di- or polysaccharides as sources of carbon and energy. Again, the corresponding pathways were found in the genome and represented on pathway maps (Annex 5). Interestingly, *Pirellula* sp. strain 1 has the genetic potential to degrade cellulose although this could not be shown experimentally.

Some genes for antibiotic production were found in the genome (Fig. 28, 5th circle). They encode potential enzymes for polyketides or polypeptides antibiotic production. The exact nature of the potential products can not be estimated from sequence information alone.

Gene clusters encoding the constituents of a typical bacterial flagellum has been found, but no complete chemotaxis systems could be identified. The flagellum seems to contain the three components of a motor switch known from model organisms (FliN, FliM, FliG; RB9275, RB7360, RB5642). Therefore, a yet unknown chemotaxis system which triggers this classical motor switch might be present in *Pirellula* sp. strain 1.

A gene encoding a close homolog of fermentative lactate dehydrogenases (RB8859) has been identified, which suggests that *Pirellula* sp. strain 1 might survive in anoxic environments by fermentation. The sinking of marine snow particles along the water column to the sea floor might bury the associated microorganisms in anoxic layers of the sediments. Interestingly, *Planctomycetes* have been detected in such habitats by molecular techniques<sup>22</sup>. However, *Pirellula* sp. strain 1 could not grow in the absence of oxygen under laboratory conditions (H. Schlesner, personal communication). Expression profiling and *in situ* functional studies will be needed to quantify the possible contribution of *Planctomycetes* to fermentative processes in the environment.

*Pirellula* sp. strain 1 encodes typical bacterial transporters, such as for example a pool of ABC-type transporters and a PTS (phosphotransferase system) for sugar import. The specificity of ABC transporters includes oligopeptides, amino acids, manganese, nitrate, sulfate, phosphate and ribose. A graphical overview of the metabolic and transport

potential of *Pirellula* sp. strain 1 is shown in Figure 29.

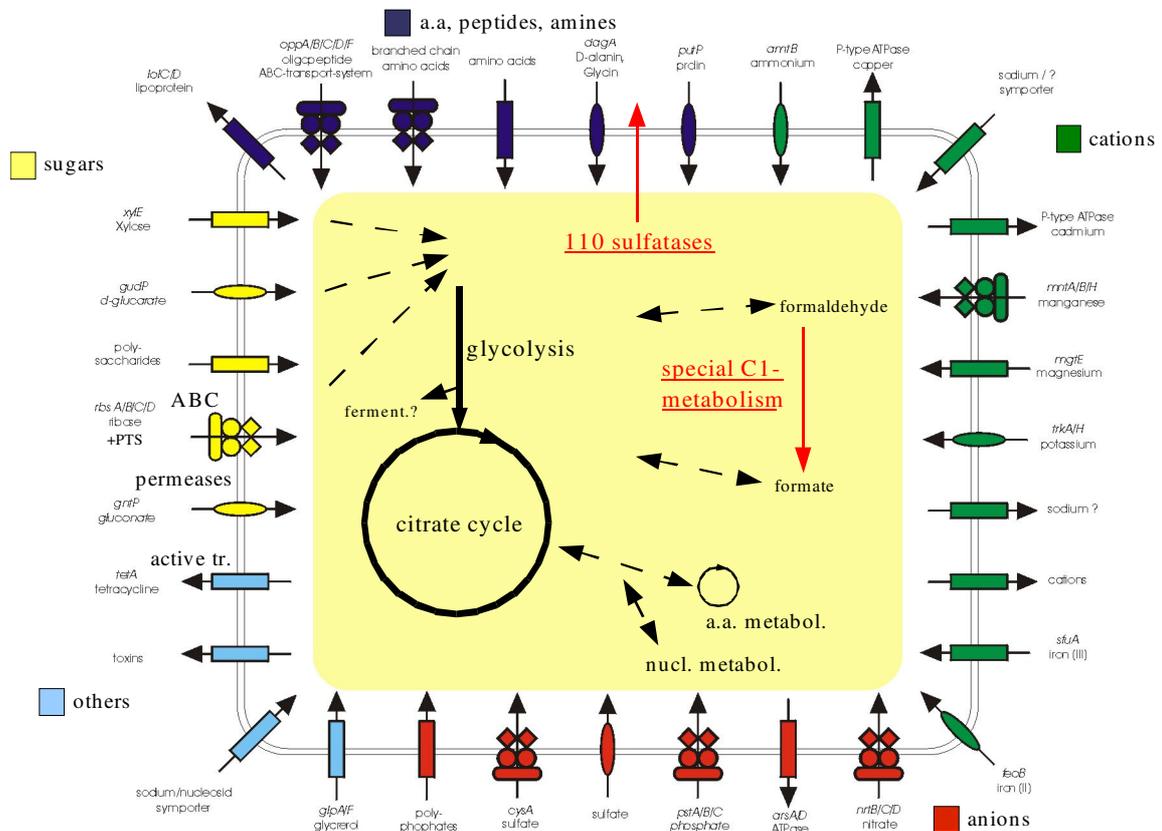


Fig. 29: Simplified overview of the metabolic potential of *Pirellula* sp. strain 1, as revealed by the annotation process. The pool of transporters of known specificity and the main metabolic components are schematically drawn. Specific metabolic pathways are described in more detailed in Annexes 3-5. Sulfatases and the C1 metabolism enzymes are discussed in the next sections.

This list of pathways reconstructed from the genome sequence is reliably matching the available experimental observations and constitutes a proof of principle for *in silico* metabolic reconstruction. An interesting consequence of this reliability is the fact that unexpected pathways predicted by genome analysis might really occur in the environment, even if no corresponding evidence could be identified under laboratory conditions. Such possibilities are further discussed in the next sections.

### 3.2.3. Annotation highlights: unexpected findings

#### 3.2.3.1. Sulfatases high copy number

The most striking fact observed during the annotation process was the regular occurrence of genes showing database hits to sulfatases along the genome. Finally, a total of 110

genes of *Pirellula* sp. strain 1 have been predicted to encode sulfatases based on combined bioinformatic evidences (see overview on Fig. 28). Such a high number was never reported for a prokaryotic genome before.

Sulfatases functional assignments were supported by both protein pairwise comparisons (BLASTP) and results from Markov models. Statistical support was high, with 106 of the 110 annotated sulfatases sequences showing e-value below  $10^{-10}$  against the Pfam database entry PF00884 (sulfatases), which is specific for this function.

More specific information about this protein family can be used to estimate whether these unexpected genes encode potentially functional proteins. The biochemical function of known sulfatases is the cleavage of sulfate from organic O-sulfate esters. The active site of these enzymes is known and a catalytic mechanism has been proposed based on 3D structure determinations<sup>96</sup>. A post-translational modification occurring on a single cysteine or serine residue has been shown to be essential for catalytic activity in prokaryotes, while in eukaryotes, modification is only allowed on a cysteine residue<sup>97</sup>. In both cases, a formylglycine (FGly) is generated by oxidation at this position. Almost all sulfatases can therefore be classified into Cys- or Ser-type according to the amino acid which is converted to FGly in the active enzyme. The only exception are alkylsulfatases, which might not need this post-translational modification by using a distinct catalytic mechanism<sup>98</sup>. Interestingly, the amino acid sequence downstream of the modified residue in Cys- and Ser-type sulfatases is conserved among prokaryotic and eukaryotic sequences, following the motif CXPXR or SXPXR<sup>99</sup>. This pattern is thought to direct the postranslational modification to formylglycine.

In order to search for this pattern within the predicted sulfatases of *Pirellula* sp. strain 1, a multiple sequences alignment was built using these sequences and sulfatases with experimentally verified function from other organisms (Annex 6). The CXPXR pattern could be found in 84 (76%) of the predicted sulfatases of *Pirellula* sp. strain 1 at the position corresponding to the active site of known proteins. Hence, the majority of the sulfatases genes of *Pirellula* sp. strain 1 are of the Cys-type and are potentially functional if a postranslational machinery is active. No Ser-type patterns could be observed in the *Pirellula* sp. strain 1 sequences. The 36 predicted sulfatases lacking the consensus patterns are likely pseudogenes, but might also be active using distinct catalytic mechanisms like the alkylsulfatases, which remains to be clarified<sup>98</sup>.

The specific postranslational machinery for the activation of sulfatases is partially known from model organisms and can be searched in the genome of *Pirellula* sp. strain 1. According to the literature, one type of known modification systems consists of a single enzyme (AtsB) capable of oxidizing the serine residue of the SXPXR pattern of the sulfatases to a formyl-glycine<sup>100</sup>. AtsB was discovered in *Klebsiella pneumoniae*, a  $\gamma$ -*Proteobacterium*, and is an oxidoreductase containing iron-sulfur centers. This protein is not encoded by eukaryotic organisms, but homologs have been found in diverse *Proteobacteria* and *Bacteroides thetaiotaomicron*<sup>101</sup>. However, no significant hits to AtsB could be found in the genome of *Pirellula* sp. strain 1 (BLASTP<sup>52</sup>, e-value cutoff:  $10^{-3}$ ). The absence of AtsB homolog is consistent with the fact that the sulfatases of *Pirellula*

sp. strain 1 are of the Cys-type, as AtsB has been shown to be specialized on Ser-type sulfatases<sup>100</sup>.

Recently, a human enzyme involved in the post-translational modification of Cys-type sulfatases has been identified<sup>102,103,104</sup>. This enzyme was named FGE (Formylglycine Generation Enzyme) and is encoded by the human gene SUMF1. As mutations in this human gene cause a lysosomal storage disease called MSD (Multiple Sulfatase Deficiency), this discovery in eukaryotes will provide new perspectives for gene therapy. The consequences for the field of microbiology are also interesting, because FGE homologs have been identified in prokaryotic organisms, defining a new protein family<sup>102</sup>. As sequences of unrelated functions share similarities to the FGE family, it has been suggested to use iterative methods and a conservative cut-off to search for FGE homologs<sup>102</sup> (PSI-BLAST<sup>52</sup>, e-value < 10<sup>-40</sup>). The genome of *Pirellula* sp. strain 1 revealed one homolog to the human FGE using this threshold (RB11498). Furthermore, a third mechanism capable of performing the same sulfatases post-translational modification in the absence of AtsB or FGE has been shown to exist in *E. coli*<sup>101,102</sup>, but the corresponding enzyme(s) and gene(s) remain to be identified. In summary, these results show that *Pirellula* sp. strain 1 encodes the genetic potential to produce 84 active Cys-type sulfatases (Table 10).

Table 10: Occurrence of sulfatases and corresponding posttranslational mechanisms in model organisms according to experimental evidences<sup>98,101,102</sup> as compared to the genetic potential of *Pirellula* sp. strain 1 (+: present; ++: present in high copy number; -: absent; ?: unknown). FGE and AtsB: Formylglycine generating enzymes specific for Cysteine and Serine, respectively. A bioinformatic comparison of sulfatase genes in all genomes is presented later in section 3.3.1.

Organism	Cys-type sulfatase(s)	FGE	Ser-type sulfatase(s)
<i>Pirellula</i> sp. strain 1	++	+	-
<i>Homo sapiens</i>	+	+	-
<i>Escherichia coli</i> K-12	+	-	+
<i>Pseudomonas aeruginosa/putida</i>	+	+	-
<i>Klebsiella pneumoniae</i>	-	?	+

The cellular localization of bacterial sulfatases has been suggested to be correlated with their classification in Cys- and Ser-type. The Cys-type sulfatases has been reported to be cytoplasmic (e.g. in *Pseudomonas aeruginosa*), whereas the Ser-type versions carry signal peptides and seem to be exported to the periplasm or secreted (e.g. in *Klebsiella pneumoniae*)<sup>173</sup>. However, among the 84 potentially functional Cys-type sulfatases of *Pirellula* sp. strain 1, 59 (70%) of them contain signal peptides prediction with a good statistical support (SignalP<sup>144</sup>, p > 0.75), indicating that most of these proteins are probably secreted.

The substrate specificity of sulfatases can hardly be determined by sequence analysis alone. Sequence comparisons revealed the following distribution: 97 (88%) of the 110

predicted sulfatases of *Pirellula* sp. strain 1 have their closest homolog in prokaryotic and 13 (12%) in eukaryotic organisms. (BLASTP<sup>52</sup> against NCBI-nr database, status of December 2003). This observation reflects the overall sequence conservation reported for prokaryotic and eukaryotic sulfatases in the literature<sup>98</sup>.

Different physiological roles for sulfatases in eukaryotic and prokaryotic organisms have been reported so far. Eukaryotic sulfatases are typically lysosomal enzymes, working in concert with glycosyl hydrolases for the degradation of sulfated macromolecules in this cellular compartment<sup>104</sup>. Recently, secreted eukaryotic sulfatases showing glycosaminoglycan degradation activity have also been observed<sup>105</sup>. In prokaryotes, some members of the *Bacteria* revealed that sulfatases expression and activity can be enhanced by low sulfate concentrations<sup>106,107</sup> (SSI sulfatases, Sulfate Starvation Induced). This suggests that *Bacteria* exposed to sulfate limited environments activate sulfatases to hydrolyse sulfate esters on various organic compounds to retrieve free sulfate needed by biosynthetic pathways. However, some bacterial sulfatases have been reported to be involved in the degradation of sulfated polysaccharides for growth and energy purposes<sup>108,109,110,111</sup>. In the marine bacterium *Altermonas carrageenovora*, it was observed that sulfatase expression was independent of the sulfate concentration in the medium, and the hypothesis that sulfatases are involved in the degradation of sulfated polysaccharides occurring in marine environments was formulated<sup>112</sup>.

Sulfate limitation is not expected to occur in the habitat of *Pirellula* sp. strain 1, as it was isolated from a sea water sample (Fjord of Kiel, Baltic Sea). Thus, a role of the sulfatases to retrieve free sulfate under limiting conditions is not expected. A scenario involving the use of an outstanding variety of sulfatases to provide an efficient access to the carbon skeleton of diverse sulfated substrate is more likely.

The genomic context of the genes encoding putative sulfatases in *Pirellula* sp. strain 1 provides insights into the possible specificity of these enzymes. The fact that 25 (29%) of these potential functional genes are located in the proximity of genes involved in carbohydrate metabolism, such as glycosyl hydrolases, supports an implication in sulfated polysaccharides metabolism (Fig. 30). Recently, chondroitin sulfate (CS) has been tested on *Pirellula* sp. strain 1 cultures and was shown to be an excellent growth substrate (H. Schlesner, personal communication). CS consists of linear chains of glucuronic acid and 4- or 6-sulfated N-acetylgalactosamine extracted from higher eukaryotes and is a representative sulfated polysaccharide that occurs in nature. The test of more related substrates of environmental relevance such as carrageen or heparin is problematic, because they are not commercially available in laboratory quality.

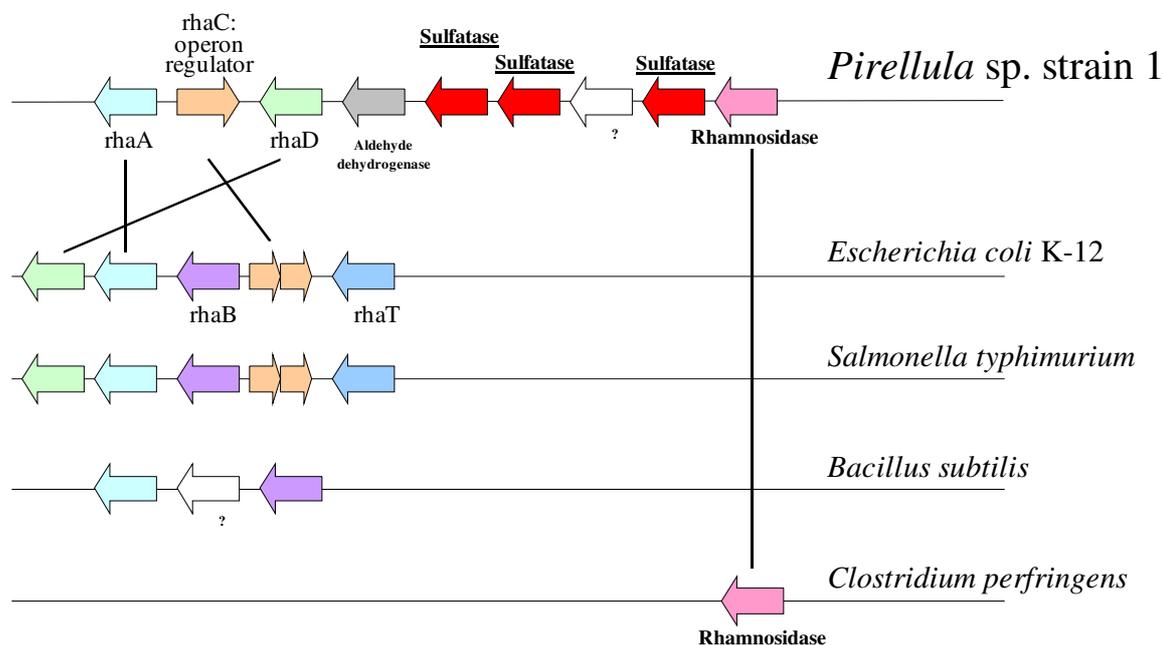


Fig. 30: An example of genomic context for genes encoding sulfatases in *Pirellula* sp. strain 1. These three sulfatases (in red) are located between an enzyme for polysaccharides degradation (rhamnosidase) and a monosaccharide degradation operon (rha operon). Colors indicate genes encoding orthologous proteins. (rhaA: rhamnose isomerase; rhaB: rhamnulose kinase; rhaC: rhamnose operon regulator; rhaD: rhamnulose-1-phosphate aldolase; rhaT: rhamnose transport system; sulfatases numbers in *Pirellula* sp. strain 1: RB684, RB686 and RB695).

An environmental interpretation for the high number of predicted sulfatase genes in *Pirellula* sp. strain 1 can be drawn from the previous considerations. An extended pool of sulfatases of different substrate specificities might play a key role in the degradation of a broad range of sulfated biopolymers present in e.g. macroscopic organic aggregates (the so-called marine snow particles<sup>113,114</sup>). Indeed, sulfated polysaccharides, constituting a large proportion of the cell wall of red algae for example, are found in high amount in such aggregates<sup>115,116,117</sup>. Organisms that belong to the *Planctomycetes* have been shown to be part of the microbial consortium associated with marine snow<sup>20</sup>. Furthermore, genomic fragments belonging to the genus *Pirellula* have been isolated by metagenomic approach e.g. from particles in sea water, which provide more evidences that members of this genus can efficiently colonize this habitat<sup>118</sup>. The lifestyle of *Pirellula* sp. strain 1 might therefore include the attachment to marine snow particles to take advantage of these particular substrates as sources of energy and growth. The sulfatases genes of *Pirellula* sp. strain 1 would then represent the fingerprint of an efficient adaptation of this environmental organism for the degradation of the corresponding complex substrates.

### 3.2.3.2. Special enzymes for C<sub>1</sub> metabolism

A set of very unexpected genes corresponding to a special type of one-carbon metabolism (C<sub>1</sub>) was revealed by the annotation of *Pirellula* sp. strain 1<sup>A2</sup> (see inner circle in Fig. 28). During the past, the genes encoding this particular set of enzymes were believed to be specific to methanogenesis (production of methane as a catabolic end product which is exclusively present in specialized anaerobic *Archaea*<sup>119,120</sup>). However, homologs of these genes were recently identified in methylotrophic *Proteobacteria*<sup>121,122</sup>. Further evidences demonstrated that the encoded enzymes, which are similar to the corresponding methanogenesis enzymes, are in fact involved in a new type of aerobic C<sub>1</sub> metabolism (methylotrophy) in these particular *Proteobacteria*<sup>123,124</sup>. Before the complete genome sequence of *Pirellula* sp. strain 1 was available, this common set of genes of methanogenic *Archaea* and some methylotrophic *Proteobacteria* was never found in other groups in the tree of life. Therefore, a scenario involving a single horizontal gene transfer from the *Archaea* to the *Proteobacteria* was proposed as the best evolutionary hypothesis in the literature<sup>121</sup>. The discovery of those genes in a member of the *Planctomycetes* (*Pirellula* sp. strain 1) shows that this hypothesis needs to be reformulated and provide new insights into the evolution of C<sub>1</sub> metabolic pathways.

In order to assess the potential of *Pirellula* sp. strain 1 to perform methanogenesis or methylotrophy according to these unexpected genes, metabolic reconstruction of the corresponding pathways was done. The methanogenesis pathway map for known organisms and *Pirellula* sp. strain 1 is presented in Figure 31. The results clearly show that the crucial last steps of this pathway could not be found in methylotrophic *Bacteria* or *Pirellula* sp. strain 1, as compared to the complete pathway present in methanogenic *Archaea* (Fig. 31, color boxes). Especially, the genes encoding the key methanogenesis enzymes Mcr (methyl coenzyme M reductase, EC 2.8.4.1) and Frh (coenzyme F420 hydrogenase, EC 1.12.99.1) are not present in *Pirellula* sp. strain 1 or any other member of the *Bacteria*. Therefore, these results indicate that *Pirellula* sp. strain 1 does not possess the genetic potential to perform methanogenesis, even if genes that can be involved in this metabolism are present (e.g. gene encoding Mch, EC 3.5.4.27).

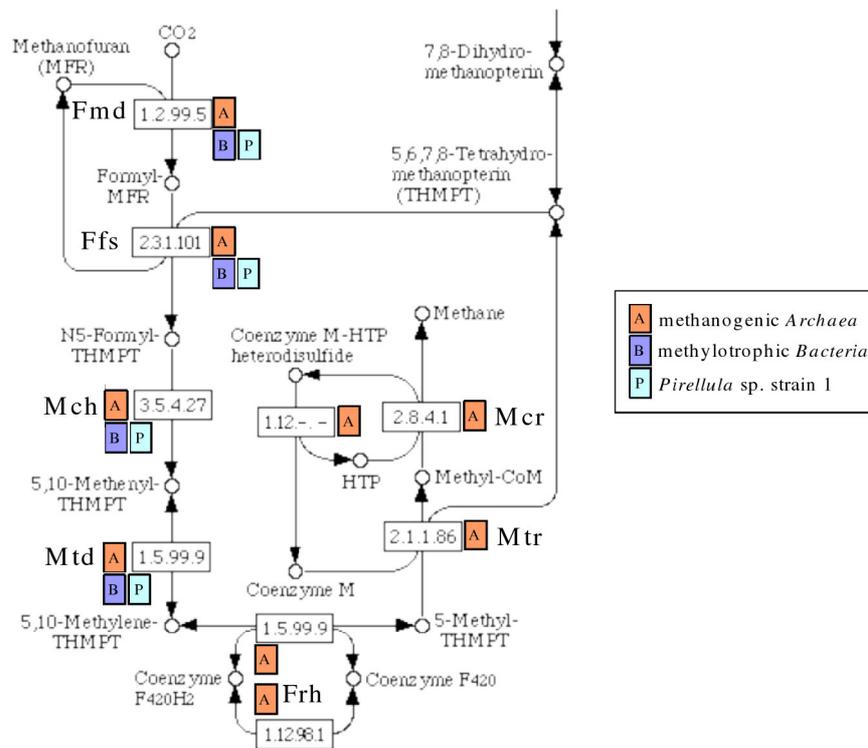


Fig. 31: Methanogenesis generic pathway map. The pathway is incomplete in *Pirellula* sp. strain 1 and methylotrophic *Proteobacteria* (according to the draft genome of *Methylobacterium extorquens* AM1). Color boxes indicate the presence of the corresponding gene and white boxes indicate enzymes designed by EC numbers (EC 1.2.99.5: formylmethanofuran dehydrogenase (**Fmd**); EC 2.3.1.101: formylmethanofuran-tetrahydromethanopterin N-formyltransferase (**Ffs**); EC 3.5.4.27: methenyltetrahydromethanopterin cyclohydrolase (**Mch**); EC 1.5.99.9: methylenetetrahydromethanopterin dehydrogenase (**Mtd**); EC 1.12.98.1: coenzyme F420 hydrogenase (**Frh**); EC 2.1.1.86: tetrahydromethanopterin S-methyltransferase (**Mtr**); EC 2.8.4.1: methyl coenzyme M reductase (**Mcr**); EC 1.12.-.-: oxidoreductase). The original map was taken from KEGG<sup>95</sup> and genome data from the EMBL database<sup>65</sup>.

In methylotrophic *Proteobacteria*, the newly discovered pathways involving those *Archaea*-like enzymes are currently subject of intensive studies in the model organism *Methylobacterium extorquens* AM1 (a methylotrophic  $\alpha$ -*Proteobacteria*)<sup>124,125,126,127,128</sup>. The current knowledge about this metabolism is shown in Figure 32, and the corresponding *Pirellula* sp. strain 1 genes are displayed. Genes corresponding to the first step, involving the oxidation of a C<sub>1</sub> substrate to formaldehyde, could not be identified in *Pirellula* sp. strain 1. Significant similarities to known methanol dehydrogenases are found for the gene product of RB10805 (BLASTP e-value < 10<sup>-42</sup>), but the specificity for

this substrate can not be affirmed by sequence comparison alone. However, the results clearly show that *Pirellula* sp. strain 1 possess the genetic potential to encode this particular C<sub>1</sub> pathway from formaldehyde to formate. Interestingly, the same set of genes was identified in the unfinished genome of *Gemmata obscuriglobus* UQM 2246, a fresh water isolate which belongs to a distinct subdivision of the *Planctomycetes* (preliminary draft sequences kindly provided by the TIGR center<sup>129</sup>). Despite the preliminary status of these sequences (211 contigs), this indicates that those particular "methylo trophy" genes might be widespread among the *Planctomycetes*.

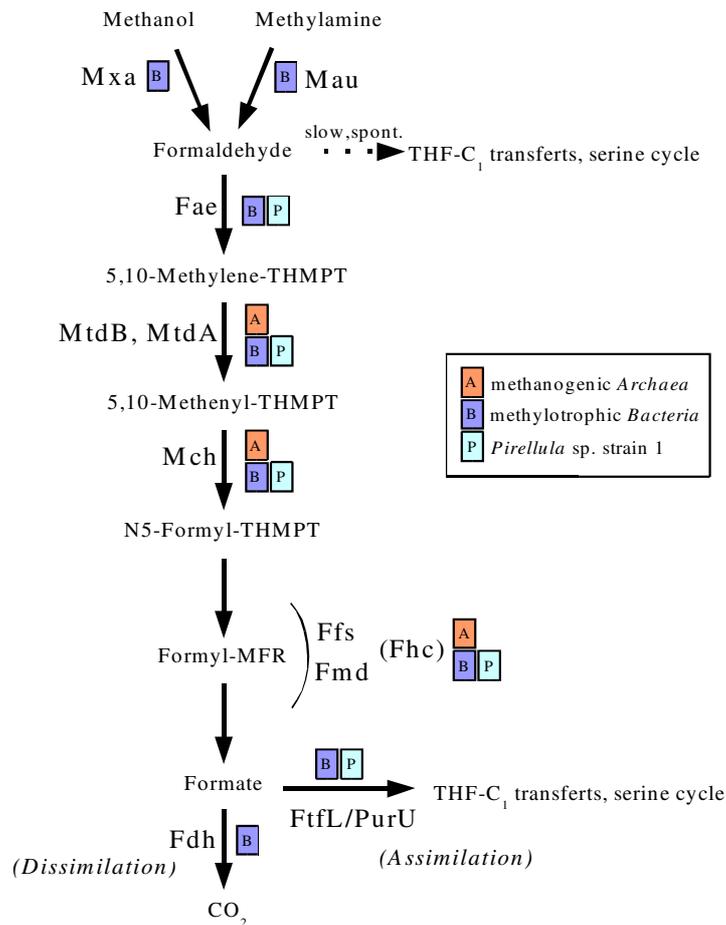


Fig. 32: Methylotrophy pathway involving archaeal-like enzymes and the cofactor MPT (methanopterin) in methylotrophic *Proteobacteria*<sup>127,128</sup> and presence of the corresponding genes in *Pirellula* sp. strain 1. This pathway is potentially functional in *Pirellula* sp. strain 1 starting from formaldehyde. Color boxes indicate the presence of the corresponding genes. In the case of the model methylotroph *Methylobacterium extorquens* AM1, possible C<sub>1</sub> substrate have been shown to include methanol or methylamine. (**Mxa**: methanol dehydrogenase; **Mau**: Methylamine dehydrogenase; **Fae**: formaldehyde-activating enzyme; **MtdB,(A)**: methylenetetrahydromethanopterin dehydrogenase; **Mch**: methenyltetrahydromethanopterin cyclohydrolase; **Fmd**: formylmethanofuran dehydrogenase; **Ffs**: formylmethanofuran-tetrahydromethanopterin N-formyltransferase; **Fhc**: formyltransferase/hydrolase complex; **Fdh**: Formate dehydrogenase; **Ftl/PurU**: Formate-tetrahydrofolate ligase / formyltetrahydrofolate deformylase).

The discovery of this unexpected methylotrophy-related C<sub>1</sub> pathway in the genome of *Planctomycetes* rises the question of its physiological role in this distinct phylum. No *Planctomycetes* were reported to degrade C<sub>1</sub> under laboratory conditions so far. Different C<sub>1</sub> sources were tested as carbon- and energy-source for *Pirellula* sp. strain 1 (methanol, methylamin, methylsulfonate), but no growth could be observed (H. Schlesner, D. Gade, personal communication). An alternative role for this pathway in *Planctomycetes* might be the detoxification of intracellular formaldehyde. This compound is toxic for all organisms due to its reactivity with biological macromolecules, and was shown to accumulate and inhibit growth in *Methylobacterium extorquens* AM1 mutants defective in Fae (Fig. 32, second step)<sup>125,130</sup>. However, formaldehyde accumulation is not expected to occur in *Pirellula* sp. strain or other *Planctomycetes*, because this compound is primarily formed by the degradation of C<sub>1</sub> carbon sources.

Beside a potential alternative biological role in *Planctomycetes*, this discovery provides hints about the evolution of this particular metabolism at the genetic level. These evolutionary aspects are under heavy discussion in the literature<sup>121,126</sup>. The key gene product Mch is a perfect candidate for phylogenetic analysis, because of its highly conserved sequence among methanogenic *Archaea* and methylotrophic *Proteobacteria*<sup>126</sup>. Moreover, Mch is a key enzyme of both methanogenesis or the related particular methylotrophy, acting respectively in the reductive or oxidative way. The phylogenetic analysis of all known Mch complete sequences was carried out, including the new sequence of *Pirellula* sp. strain 1 (Fig. 33). Interestingly, the Mch sequence of *Pirellula* sp. strain 1 does not group with the other bacterial sequences and branches between the *Archaea* and the *Proteobacteria*.

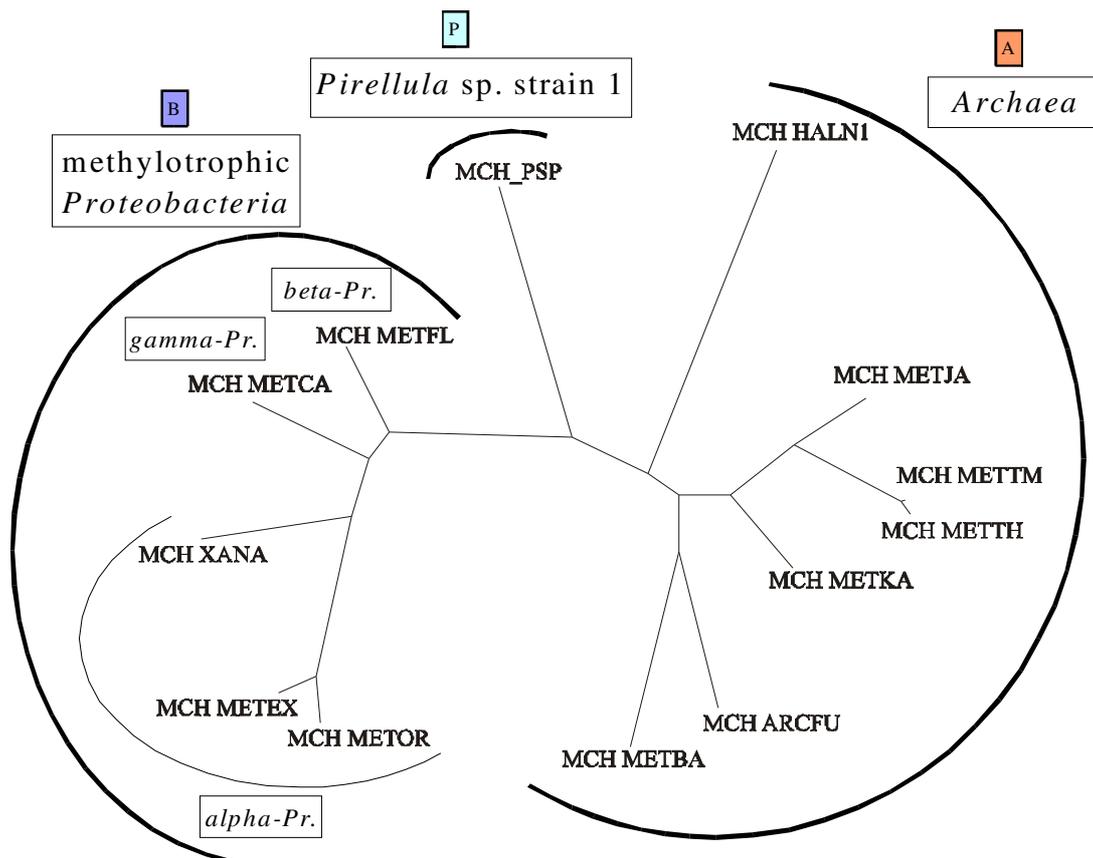


Fig. 33: Phylogenetic tree of Mch sequences (methenyltetrahydromethanopterin cyclohydrolase), the key enzyme of methanogenesis and bacterial methylotrophy involving archaeal-like enzymes. A clear separation between methylotrophic *Proteobacteria* and *Archaea* is observed, but *Pirellula* sp. strain 1 branches between them. Full length sequences were aligned using ClustalW 1.81 and phylogeny was inferred based on the maximum likelihood method (Proml, Phylip 3.6a<sup>160</sup>). The same tree topology was obtained using maximum parsimony (data not shown). The *Archaea* presented here are typical methanogens, except two species: *Archaeoglobus fulgidus* is a sulfate reducing *Archaea*, producing only traces of methane and *Halobacterium* sp. strain NRC-1. PSP: *Pirellula* sp. strain 1; ARCFU: *Archaeoglobus fulgidus*; HALN1: *Halobacterium* sp. strain NRC-1; METTH: *Methanobacterium thermoautotrophicum* str. Delta H; METTM: *Methanobacterium thermoautotrophicum* strain Marburg; METJA: *Methanococcus jannaschii*; METKA: *Methanopyrus kandleri*; METBA: *Methanosarcina barkeri*; METFL: *Methylobacillus flagellatum*; METEX: *Methylobacterium extorquens* AM-1; METOR: *Methylobacterium organophilum*; METCA: *Methylococcus capsulatus*; XAN: *Xanthobacter autotrophicus*.

The genomic context of these particular genes is a source of more evidences to further study their possible evolution. Figure 34 compares the genomic arrangement of the corresponding bacterial and archaeal genes in relevant complete or draft genomes. A conserved cluster can be observed between *Pirellula* sp. strain 1 and *Methylobacterium extorquens* AM1. This conservation is not observed in the known archaeal genomes.

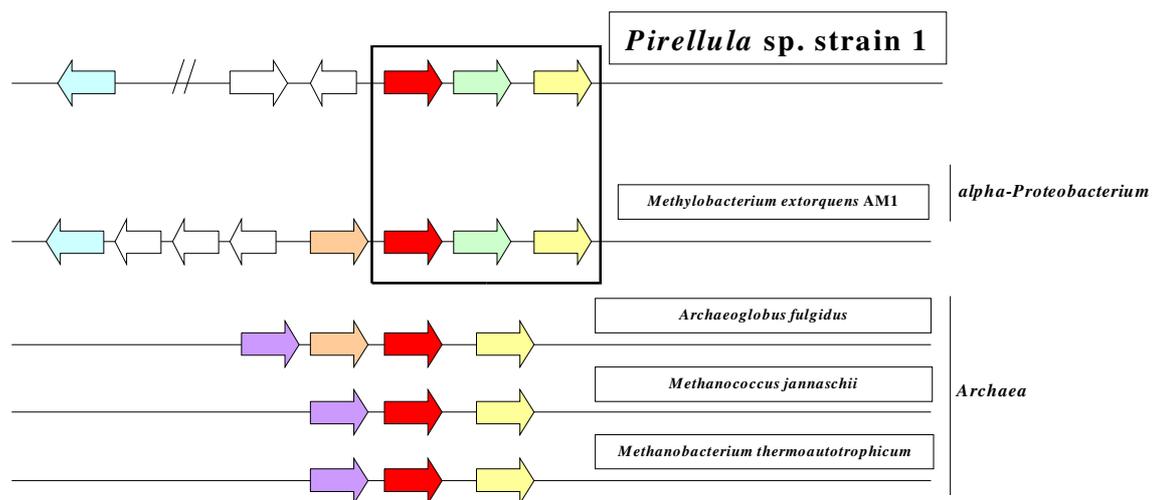


Fig. 34: Genomic context of the main methylotrophy/methanogenesis homologs in *Pirellula* sp. strain 1 and relevant members of the *Archaea* and another *Bacteria*. The conserved gene cluster among *Bacteria* is indicated. (Blue: mch; violet: fmd sub. D; orange: fmd sub. B; red: fmd sub. A; green: ffs; yellow: fmd sub. D). Isolated genes are not shown. Genome data were extracted from the EMBL database<sup>65</sup>.

These genomic context and phylogenetic results show that the previously proposed evolutionary scenario for these particular methylotrophy/methanogenesis genes, involving a single horizontal genes transfer event from *Archaea* to *Proteobacteria*<sup>121,126</sup>, is rather unlikely. Such a single event would place the Mch of *Pirellula* sp. strain 1 within the other known bacterial sequences. One might suggest a possible second gene transfer between a *Proteobacterium* and a *Planctomycete*, which would be supported by the observed gene order but is also rather incompatible with the Mch phylogeny. Therefore, these results suggest a new evolutionary scenario involving an horizontal genes transfer event between *Archaea* and the common ancestor of *Proteobacteria* and *Pirellula* sp. strain 1. As the position of the *Planctomycete* phylum is still a subject of intense discussions, the nature of this common ancestor remains uncertain. If the *Planctomyces* are indeed the deepest branching *Bacteria* as it has been recently suggested<sup>36</sup>, the ancestor of all *Bacteria*, or even LUCA (the last unique common ancestor of the tree of life) would have contained this complete set of genes already. However, new analysis of the phylogenetic position of *Pirellula* sp. strain 1 rather suggest that *Planctomyces* are branching later in the bacterial domain (see section 3.5).

In summary, the whole genome sequence of *Pirellula* sp. strain 1 allowed to extend the

current knowledge concerning the evolution of archaeal-like  $C_1$  genes. The very large number of genomes expected to be available in the near future might reveal a broader distribution of these genes. An interesting strategy might also include the specific screening for these genes in genomic fragments retrieved directly from the environment (metagenomic approach), in order to access the metabolic diversity present in uncultivated organisms.

### 3.3. Consistent cross-genomes comparisons

The availability of the genome of *Pirellula* sp. strain 1 allows, for the first time, to investigate the genetic potential of a member of the *Planctomycetes*. This new information has to be interpreted in the context of the growing number of other organisms whose genome sequences have been contributed by the scientific community. Particularly, comparisons of the abundance of specific gene families among the genomes can reflect the physiologic specialization of some organisms and overall life strategies. Unfortunately, the original annotation of prokaryotic organisms is provided by different annotators and is not directly comparable for three main reasons: i) the annotation was not performed at the same time, providing functional assignments which reflect different contents of public protein databases; ii) contrasting efforts are invested in the annotation of different genomes, according to funding or time pressures, which is reflected in the final quality; iii) there is a dramatic lack of common vocabulary for gene designation and classification. The strategy presented here for selected gene groups is based on Markov models (using the Pfam database<sup>53</sup>) to allow consistent cross genome comparison. As the calculation power needed by such an approach is high, the use of a computer cluster is needed (see material and method). Depending on the size of the targeted gene groups, calculation times were between 2 to 4 days on our hardware for each gene group.

#### 3.3.1. Systematic study of environmentally relevant gene groups

##### 3.3.1.1. Sulfatases

The first question which was addressed by cross-genome comparisons concerns the unexpected finding of 110 genes potentially encoding sulfatases in the genome of *Pirellula* sp. strain 1. Even if such a high number was never reported in the literature for any organism, the specific search for such genes in all available bacterial genomes has to be accomplished for confirmation. The screening for genes encoding predicted sulfatases in bacterial genomes is presented in Figure 35. Within this dataset, the highest number of sulfatases behind *Pirellula* sp. strain 1 is found in *Pseudomonas aeruginosa* PA01 (6 copies). These results clearly show that the number of sulfatases encoded by *Pirellula* sp. strain 1 is exceptional.

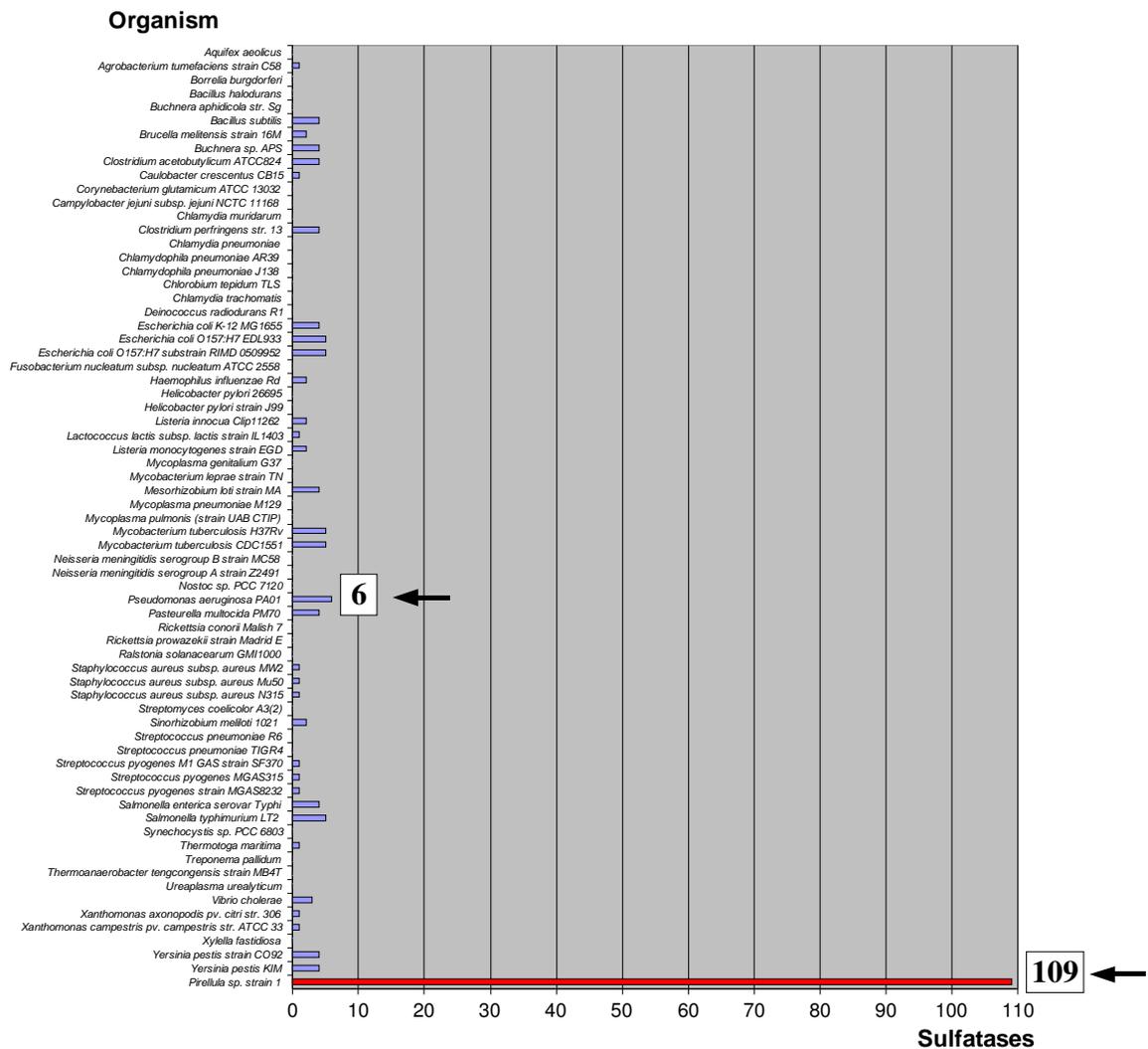


Fig. 35: Number of genes encoding predicted sulfatases in *Pirellula* sp. strain 1 (red) and other members of the *Bacteria*, as revealed by Markov model searches (Pfam: PF00884, e-value < 10<sup>-3</sup>).

Very recently, the whole genome of *Bacteroides thetaiotaomicron*, an organism that is able to degrade a broad range of polysaccharides and resides in the human intestine was

publicly released. The occurrence of genes encoding sulfatases in this organism was not discussed in the original genome publication<sup>131</sup> which focused principally on the large repertoire of genes encoding glycosyl hydrolases, the first step of polysaccharides degradation. This new genome was also included in the present survey and revealed an unusual total of 30 genes encoding sulfatases, which is less than *Pirellula* sp. strain 1, but remarkably more than any other bacterial genome sequenced so far. The total number of genes encoding sulfatases and glycosyl hydrolases in *Bacteroides thetaiotaomicron* and *Pirellula* sp. strain 1 based on Markov models is compared in Figure 36.

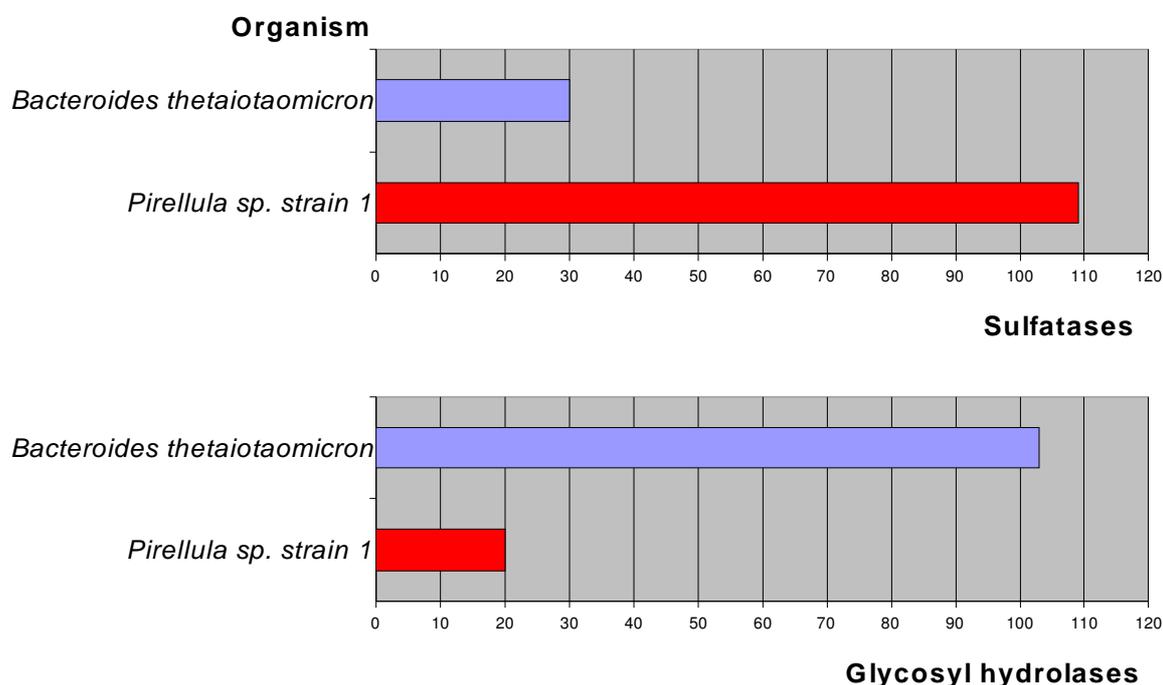


Fig. 36: Number of genes encoding predicted sulfatases and glycosyl hydrolases in *Bacteroides thetaiotaomicron* and *Pirellula* sp. strain 1 (red), as revealed by Markov model searches.

The exact function of the all 30 predicted sulfatases of *Bacteroides thetaiotaomicron* is still not known, but at least two of the corresponding proteins has been shown to participate in the degradation of a sulfated polysaccharide (chondroitine sulfate). Glycosyl hydrolases first degrade this polysaccharide to oligomers and monomers. In a second step, sulfatases remove the sulfate esters on specific positions on sugar units<sup>108,109</sup>. The sugar monomers can then enter central metabolism (glycolysis). Interestingly, this particular substrate (chondroitine sulfate) has been tested with *Pirellula* sp. strain 1 and was shown to provide high growth rates (H. Schlesner, personal communication). The extended pool of sulfatases in *Pirellula* sp. strain 1 and *Bacteroides thetaiotaomicron* might therefore constitute specialized enzymes for the optimal degradation of diverse sulfated polysaccharides.

More generally, *Bacteroides thetaiotaomicron* contains an exceptional number of genes encoding glycosyl hydrolases for diverse polysaccharides degradation (103), but in *Pirellula* sp. strain 1, these genes are less represented (20) (Fig. 36). A complete survey of glycosyl hydrolases in bacterial genomes is presented in the next section.

#### 3.3.1.2. Glycosyl hydrolases

*Pirellula* sp. strain 1 is able to degrade a broad spectrum of monosaccharides under laboratory conditions and is therefore physiologically described as a specialist in carbohydrate utilization (H. Schlesner, personal communication). Whether this organism is also specialized in polysaccharides degradation is not evident according to laboratory experiments. Only gelatin, starch and chondroitine sulfate are being hydrolyzed by *Pirellula* sp. strain 1, but not naturally abundant substrate such as alginate, cellulose or chitin. Within the last years, whole genome sequencing of *Bacteria* which are specialized in biopolymer degradation revealed that the ability to degrade a broad spectrum of polysaccharides is mirrored at the genomic level by a high number of genes encoding glycosyl hydrolases of different specificities<sup>131,132</sup>. The screening for genes encoding predicted glycosyl hydrolases in bacterial genomes reveals that *Pirellula* sp. strain 1 does not contain an exceptional number of such genes (Fig. 37).

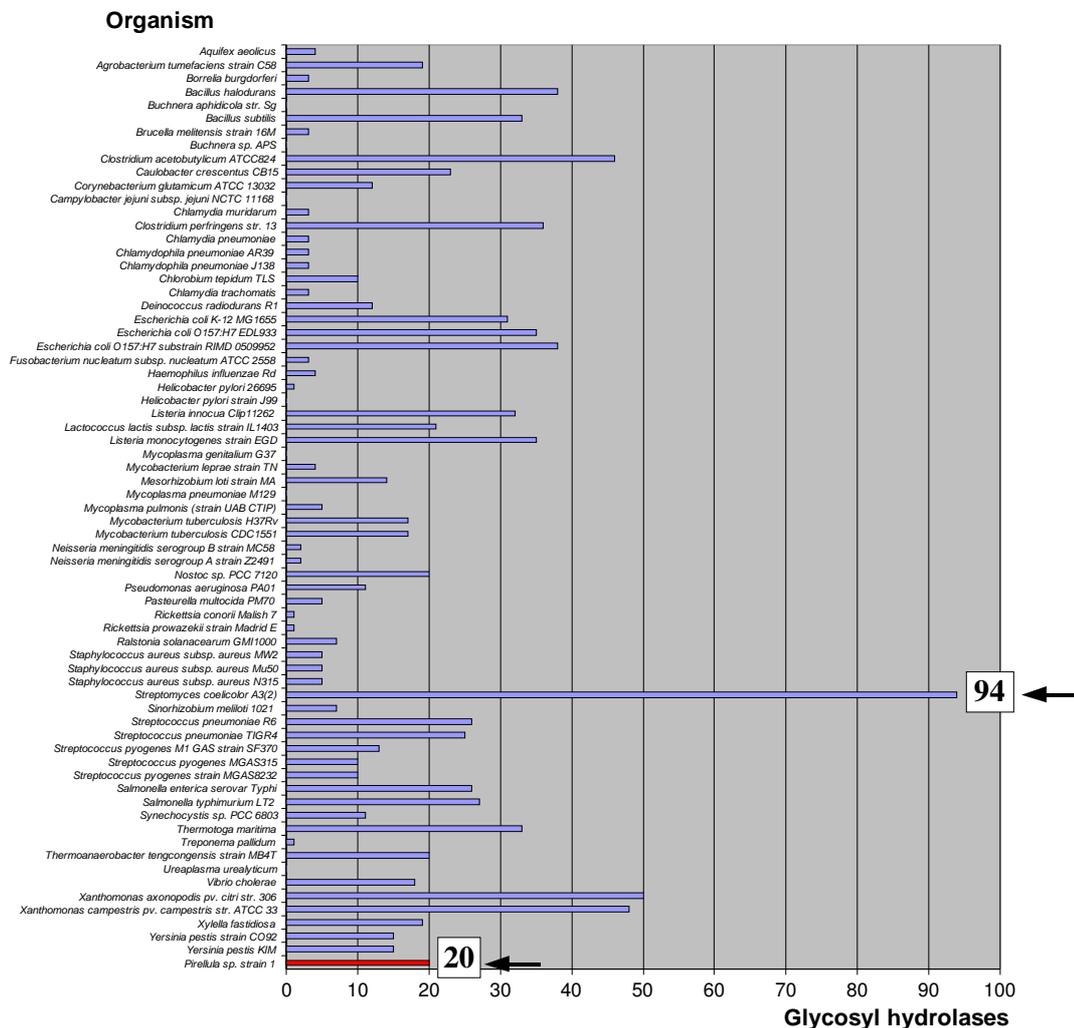


Fig. 37: Number of genes encoding predicted glycosyl hydrolases in *Pirellula* sp. strain 1 (red) and other members of the *Bacteria*, as revealed by Markov model searches (79 Pfam models,  $e$ -value  $< 10^{-3}$ ).

This comparison shows that members of the *Bacteria* that are specialized in polysaccharide degradation can clearly be identified by the high number of corresponding encoded enzymes (e.g. *Streptomyces coelicolor*: 94). A model organism, such as *Escherichia coli* K-12 reveals 31 glycosyl hydrolases using this analysis. With 20 genes encoding glycosyl hydrolases, *Pirellula* sp. strain 1 might be able to degrade diverse polysaccharides, but not such a broad spectrum as particular specialized organism.

Therefore, the results of cross genome comparisons for sulfatases and glycosyl hydrolases suggest that *Pirellula* sp. strain 1 is not specialized in polysaccharides degradation in general, but might have evolved optimized strategies to degrade sulfated polysaccharides.

### 3.3.1.3. Transporters

To further estimate how *Pirellula* sp. strain 1 copes with possible substrates from the environment, a comparison of the number of transporters in all *Bacteria* was performed. For this purpose, the ABC-type transporters were chosen as a representative target gene set. These transporters constitute a family that displays various substrate specificities and are widespread in *Bacteria*, but also in *Archaea* and *Eukarya*<sup>133</sup>. Moreover, ABC transporters have been experimentally characterized with specificities for small to large molecules and highly charged to highly hydrophobic substrates (e.g. inorganic ions, sugars, amino acids, proteins or complex polysaccharides)<sup>134</sup>.

The results reveal that with 55 copies, *Pirellula* sp. strain 1 does not encode an exceptionally large number of such transporters (Fig. 38). The highest number of ABC transporters in this dataset was found in *Mesorhizobium loti* (203) and *Agrobacterium tumefaciens* (180), two organisms that are interestingly sharing the same life-style: terrestrial habitats and plant interaction.

A larger genome is expected to contain the genetic potential to degrade a broader spectrum of substrates and to encode a larger number of corresponding transporters. The ratio of ABC transporter per Mbp in each genome has been calculated to study this correlation (Fig. 39). This normalization to genome size shows that this ratio is not constant among members of the *Bacteria*, with genomes encoding a minimum of 6.7 (*Mycobacterium leprae*) to a maximum of 36-38 ABC transporter per Mbp (*Agrobacterium tumefaciens*: 36.6, *Streptococcus pneumoniae* R6: 37.8). With 7.7 such genes per Mbp, *Pirellula* sp. strain 1 is located in the lower range. These results suggest that the pool of transporter of *Pirellula* sp. strain 1 has not been particularly extended during the evolution.

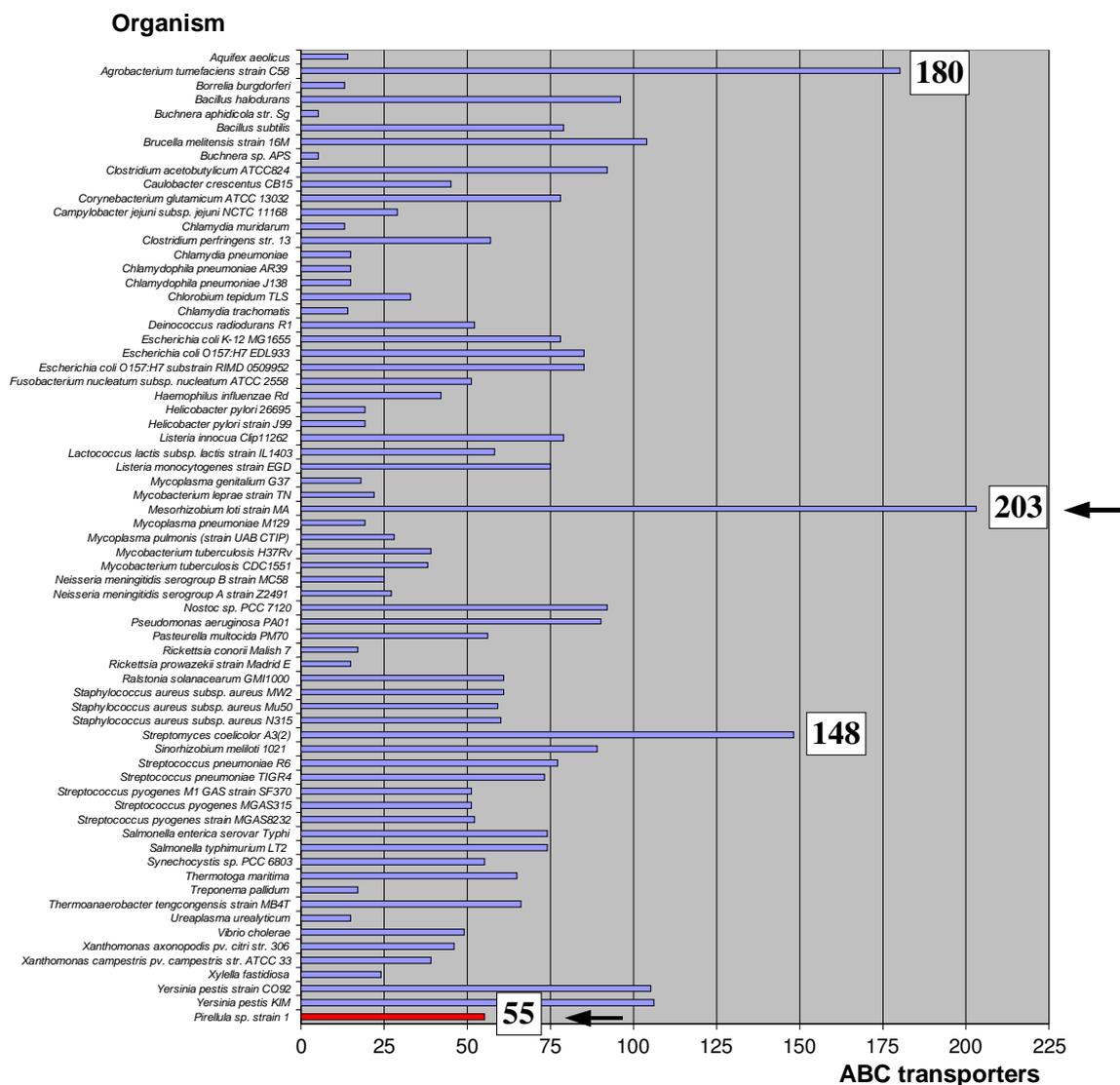


Fig. 38: Number of predicted ABC transporters in *Pirellula* sp. strain 1 (red) and other members of the *Bacteria*, as revealed by Markov model searches (ABC transporter, ATP-binding component - Pfam: PF00005, e-value < 10<sup>-3</sup>).

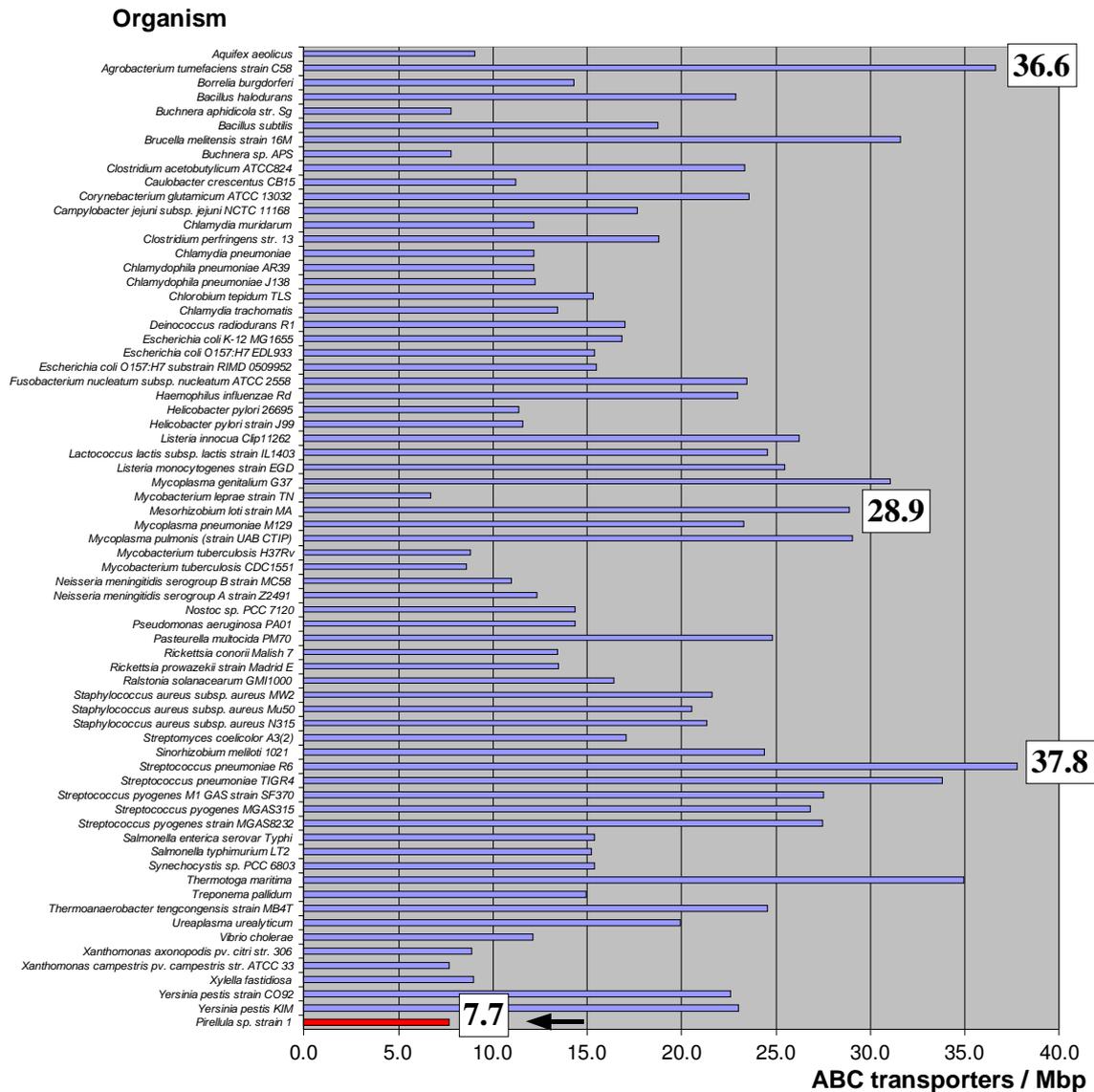


Fig. 39: Number of predicted ABC transporters (normalized to genome size) in *Pirellula* sp. strain 1 (red) and other members of the *Bacteria*, as revealed by Markov model searches. (Pfam: PF00005, e-value < 10<sup>-3</sup>).

### 3.3.1.4. Transposases / integrases

Despite the growing number of available genomes in public databases, the question of the extent and the quantification of horizontal gene transfer remains an open question in prokaryotic organisms<sup>135</sup>. Some authors claim that such events constitute the major force

in the evolution of prokaryotes, mainly based on the observation that many gene groups display a patchy distribution within the phylogenetic tree<sup>136</sup>. On the contrary, new reports based on intensive phylogenetic studies rather suggest that many genes are resistant to horizontal transfer<sup>137,138</sup>. Even if the quantification of such events in prokaryotic genomes is subject to controversy, the localization of transposases or integrases close to putatively horizontally transferred genes has been often reported in prokaryotic genomes<sup>139,140,141</sup>. Therefore, the screening of transposases and integrases in prokaryotic genomes might give insights into the extent of horizontal gene transfer. The results show very large differences between members of the *Bacteria*, ranging from 0 to 141 transposases and integrases (Fig. 40). Genomes which display high numbers of such genes in this analysis (*Yersinia pestis* strain CO92: 141, *Yersinia pestis* KIM: 121) have been previously reported to harbor numerous PAI (pathogenicity islands), genomic fragments putatively coming from horizontal gene transfer and containing transposases or integrases<sup>139</sup>. These fragments seem to allow the adaptation of some pathogenic organisms to new hosts.

*Pirellula* sp. strain 1, with a total of 43 genes predicted to encode transposases or integrases, is in the middle range as compared to other members of the *Bacteria*. However, with a large size of 7.15 Mbp, the normalization to genome size show that *Pirellula* sp. strain 1 contain a relatively low density of such genes (6 copies/Mbp). The highest transposase and integrase density is 5 times higher and is found again in some pathogenic *Bacteria* (*Yersinia pestis* strain CO92: 30.3 copies/Mbp, *Yersinia pestis* KIM: 26.3 copies/Mbp). The non-pathogenic *Bacillus halodurans* (22.1 copies/Mbp) and *Lactococcus lactis* (24.5 copies/Mbp) also show a high density of such genes (Fig. 41). This high transposases/integrases density has been reported before in both organisms, leading to the hypothesis that horizontal gene transfer events and internal rearrangement in those organisms played an important role in their molecular evolution<sup>142,143</sup>.

Therefore, the results obtained for *Pirellula* sp. strain 1 do not support an evolutionary scenario where a large part of the genome would have been acquired by multiple horizontal gene transfer, leading to the present large size of 7.15 Mbp. The quite low density of those genes also indicates that the genome plasticity of *Pirellula* sp. strain 1 might not be particularly high as compared to other members of the *Bacteria*.

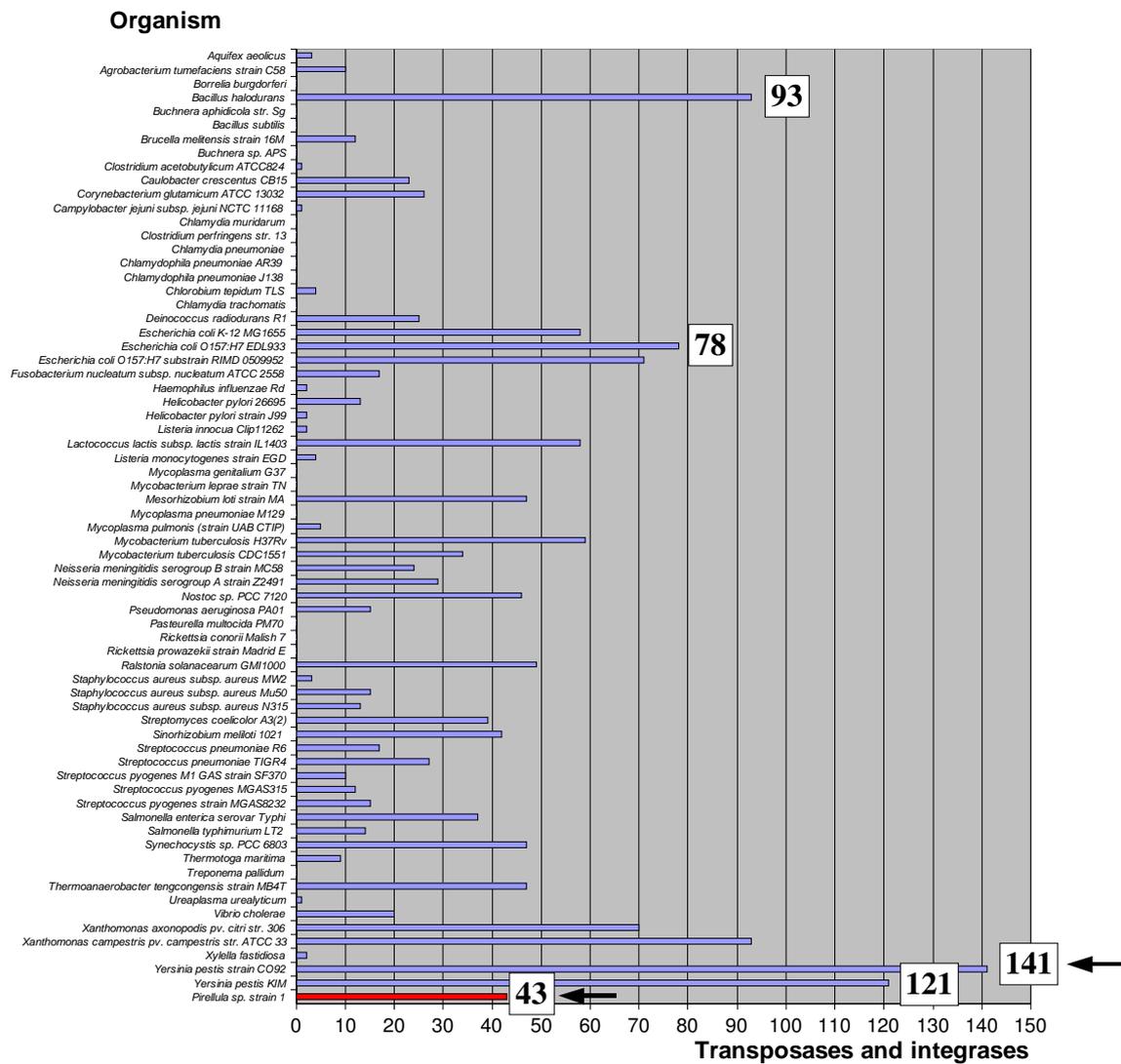


Fig. 40: Number of predicted transposases and integrases in *Pirellula* sp. strain 1 (red) and other members of the *Bacteria*, as revealed by Markov model searches (24 Pfam models, e-value < 10<sup>-3</sup>).

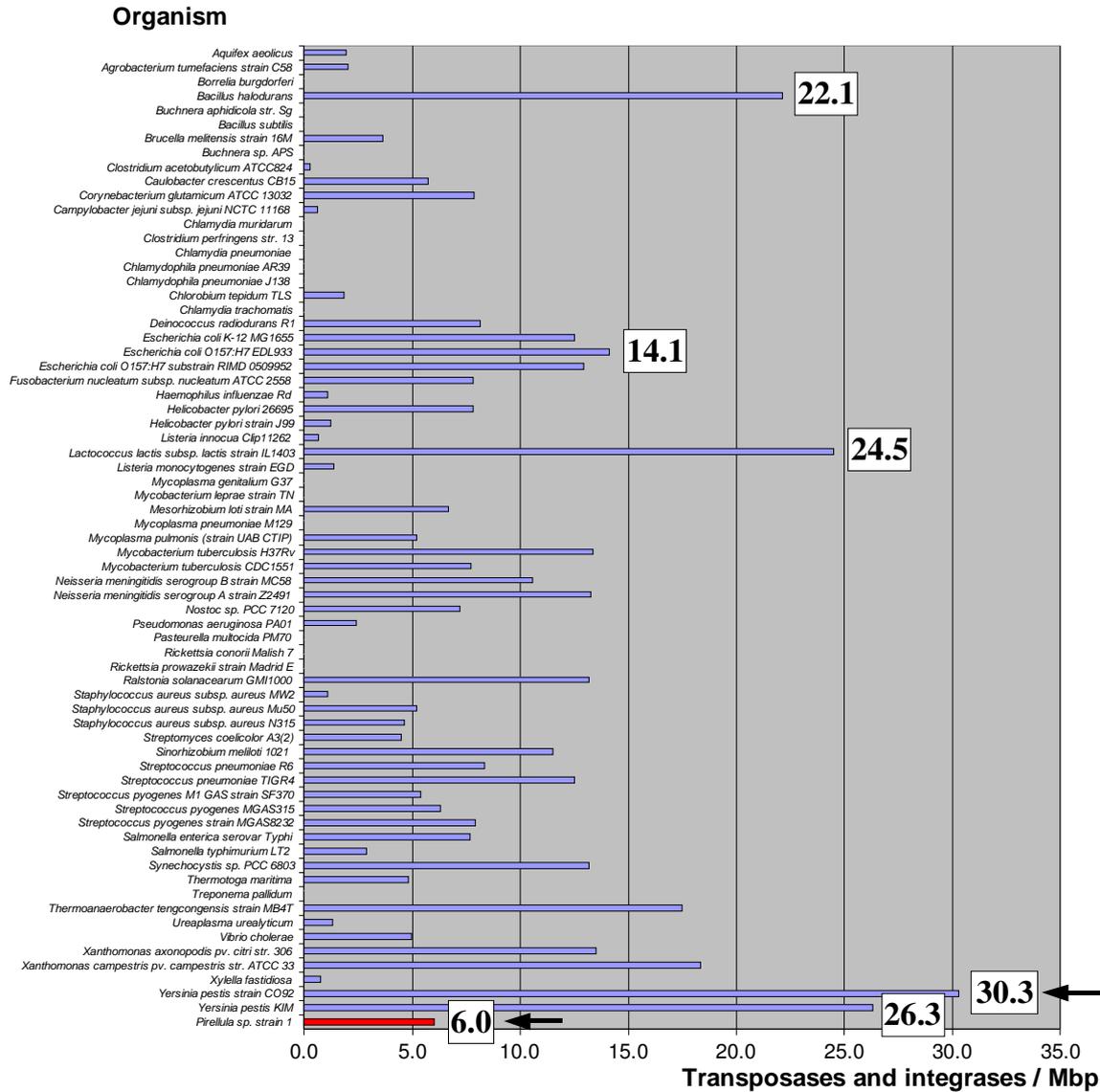


Fig. 41: Number of predicted transposases and integrases (normalized to genome size) in *Pirellula* sp. strain 1 (red) and other members of the *Bacteria*, as revealed by Markov model searches (24 Pfam models, e-value < 10<sup>-3</sup>).

### 3.3.1.5. Signal peptides

The use of specific signal sequences in proteins for efficient targeting is widespread in eukaryotic and prokaryotic organisms (signal peptides). In *Bacteria*, most membrane,

periplasmic targeted or secreted proteins contain such signals and can be searched with a Markov models based methodology<sup>144</sup>. The number of genes encoding proteins with predicted signal peptide in *Bacteria* was computed and is presented in Figure 42. The highest absolute number of signal peptides is found in *P. aeruginosa* (1277). Interestingly, *Pirellula* sp. strain 1 reveals the second highest number of signal peptides (1271). The normalization to genomic size shows that with 177.9 signal peptides per Mbp, *Pirellula* sp. strain 1 also has one of the highest values found in *Bacteria* (Fig. 43). The highest value is again found in *P. aeruginosa* (203.9 signal pep. / Mbp). As in *P. aeruginosa*, these results might reflect the potential adaption of *Pirellula* sp. strain 1 to a wide range of ecological niches. Moreover, the high proportion of signal peptide could also reflect protein targeting to cross intracellular compartmentalization, such as the intriguing pirellulosome. However, more experimental studies are needed to reveal the molecular mechanisms associated with the intracellular compartmentalization in *Planctomycetes*<sup>25,26</sup>.

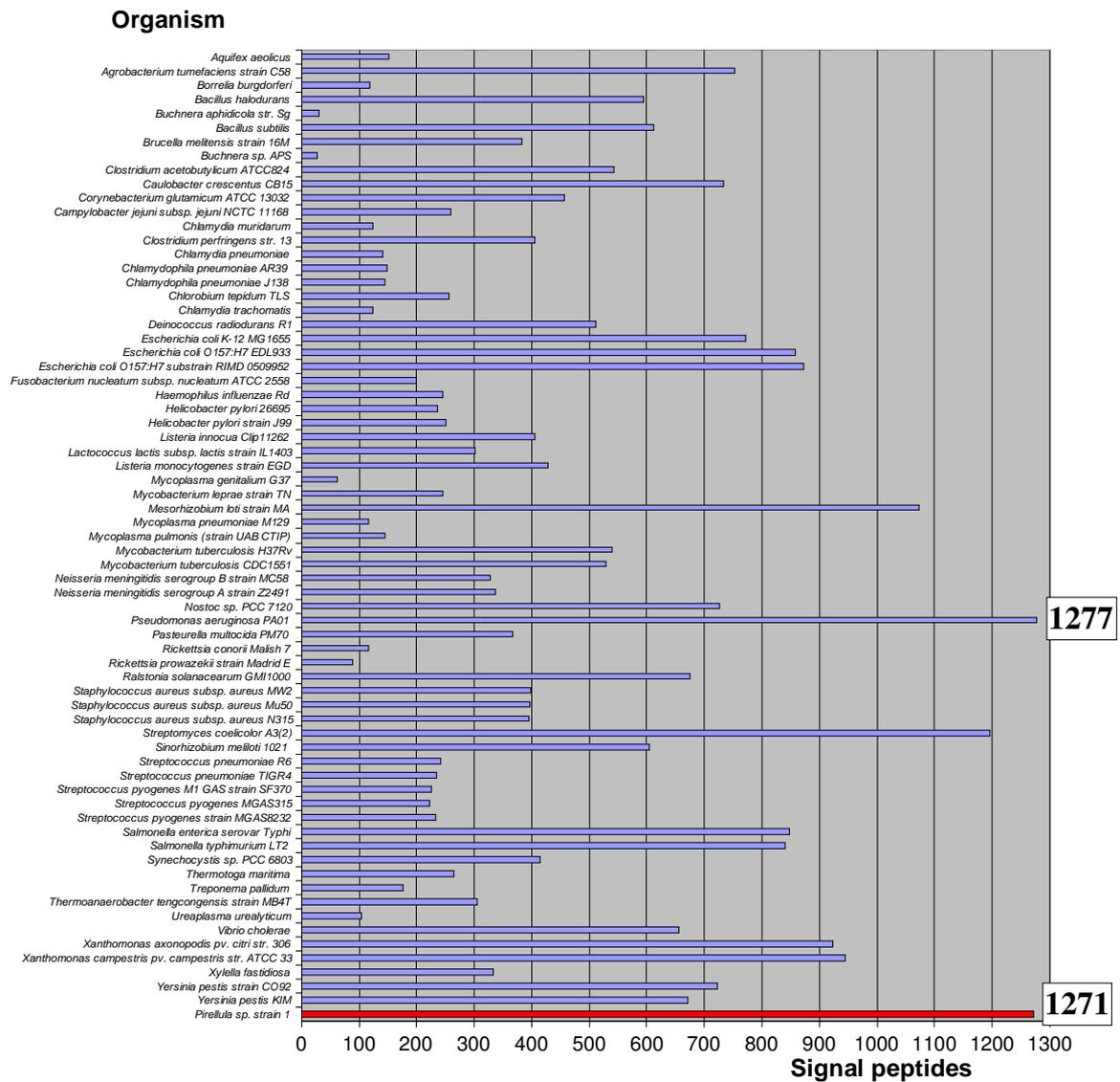


Fig. 42: Number of predicted signal peptides in the potential proteins of *Pirellula* sp. strain 1 (red) and other members of the *Bacteria* (SignalP 2.0<sup>144</sup>,  $p > 0.75$ ).

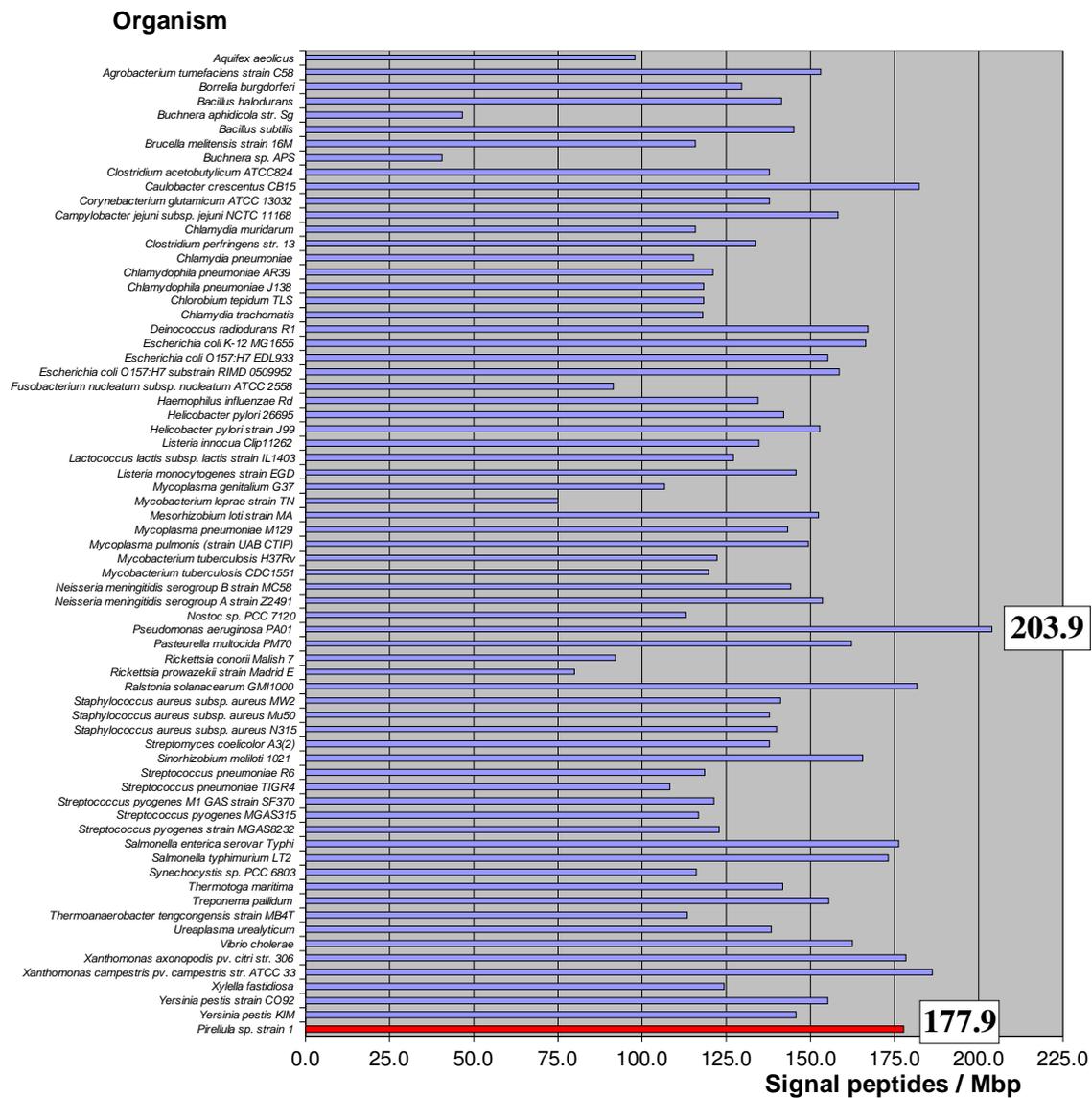


Fig. 43: Number of predicted signal peptides (normalized to genome size) in the potential proteins of *Pirellula* sp. strain 1 (red) and other members of the *Bacteria* (SignalP 2.0<sup>144</sup>,  $p > 0.75$ ).

### 3.3.2. Transcriptional regulators pool

With the completion of genomes from free-living organisms during the last years, it became evident that bacteria exposed to changing environmental conditions have larger genomes than organisms living in stable environments<sup>145,146,147</sup>. This rose the question of the regulation strategies in extended and complex gene pools when the organisms have to respond to external stimuli. From the currently available set of bacterial genomes it is now obvious that the proportion of genes encoding transcriptional regulators increases with genome size<sup>145,148</sup>. An intuitive explanation is that free-living bacteria exposed to rapidly changing environmental conditions need an extended sensing system for effective adaptation.

The completion of the large sized *Pirellula* sp. strain 1 genome represents an interesting new data source to address the validity of this quantitative correlation between genome size and the proportion of genes devoted to transcriptional regulation<sup>A3</sup>.

The increasing number of completely sequenced prokaryotic genomes delivers a valuable data source to address specific biological questions by genome comparison. To date (January 2004), 10 bacterial genomes with a genome size exceeding 6 Mb have been sequenced. This offers for the first time a reasonable dataset for the comparative investigation of large sized bacterial genomes and their associated transcriptional regulator pool.

#### 3.3.2.1. Quantitative comparisons

All available bacterial genomes including *Pirellula* sp. strain 1 have been investigated for the presence of transcriptional regulator domains. Large-sized genomes clearly show high numbers of genes encoding transcriptional regulators (e.g. *Streptomyces coelicolor* (615), *Pseudomonas aeruginosa* (446), *Bradyrhizobium japonicum* (482) and *Mesorhizobium loti* (433)). In *Pirellula* sp. strain 1, only 175 of such genes could be detected, which is unexpectedly low according to its large genome size. Typically, parasitic organisms encode less transcriptional regulators (e.g. *Borrelia burgdorferi* (15) and *Buchnera aphidicola* strain APS (10)), which may reflect the stable conditions surrounding them.

In order to systematically investigate the amount of regulators needed to cope with gene pools of different sizes, the absolute numbers of transcriptional regulators were normalized according to the gene content of each organism. As shown in Figure 44, a clear trend could be observed: The proportion of genes encoding transcriptional regulators clearly increases with genome size, ranging from less than 1% for small genomes (e.g. *Ureaplasma urealyticum*) to up to 8% for larger genomes like *P. aeruginosa* or *S. coelicolor*. This confirms earlier observations made on smaller datasets<sup>145</sup> or based on alternative comparison methods<sup>148</sup>. However, with only 2.4% of genes encoding transcriptional regulators for a genome of 7.15 Mb, *Pirellula* sp. strain 1

has the lowest proportion of transcriptional regulators within the currently sequenced bacteria with large sized genomes (Fig. 44, Psp). As a member of an evolutionary distinct lineage, *Pirellula* sp. strain 1 might contain yet unknown families of transcriptional regulators. Current knowledge about transcriptional regulators is still mainly based on model organisms such as *Escherichia coli*, *Bacillus subtilis*, *P. aeruginosa* or *S. coelicolor*. Therefore, potential regulatory protein families specific to *Planctomycetes* will be missed by the present analysis.

Recently, the genome of another marine organism, the cyanobacterium *Prochlorococcus marinus* strain CCMP13 (SS120) has been completely sequenced. Its 1.75 Mb genome contains around 1% of genes encoding transcriptional regulators (Fig. 44, Pm). The small genome size of this organism and its reduced pool of transcriptional regulators has been interpreted as an extreme adaptation to the stable conditions of the euphotic zone in the ocean<sup>149</sup>. Regarding their genome size, *P. marinus* and *Pirellula* sp. strain 1 seem to represent two opposing adaptation strategies within marine ecosystems: *P. marinus* successfully reduced its gene pool to a minimal phototrophic metabolism fitting in a given stable layer of the open ocean and *Pirellula* sp. strain 1 continuously has to cope with changing coastal conditions or sedimentation due to attachment to marine snow particles, taking advantage of an extended gene pool including metabolic diversity. The two times higher proportion of transcriptional regulators found in *Pirellula* sp. strain 1 than in *P. marinus* might also reflect these different strategies.

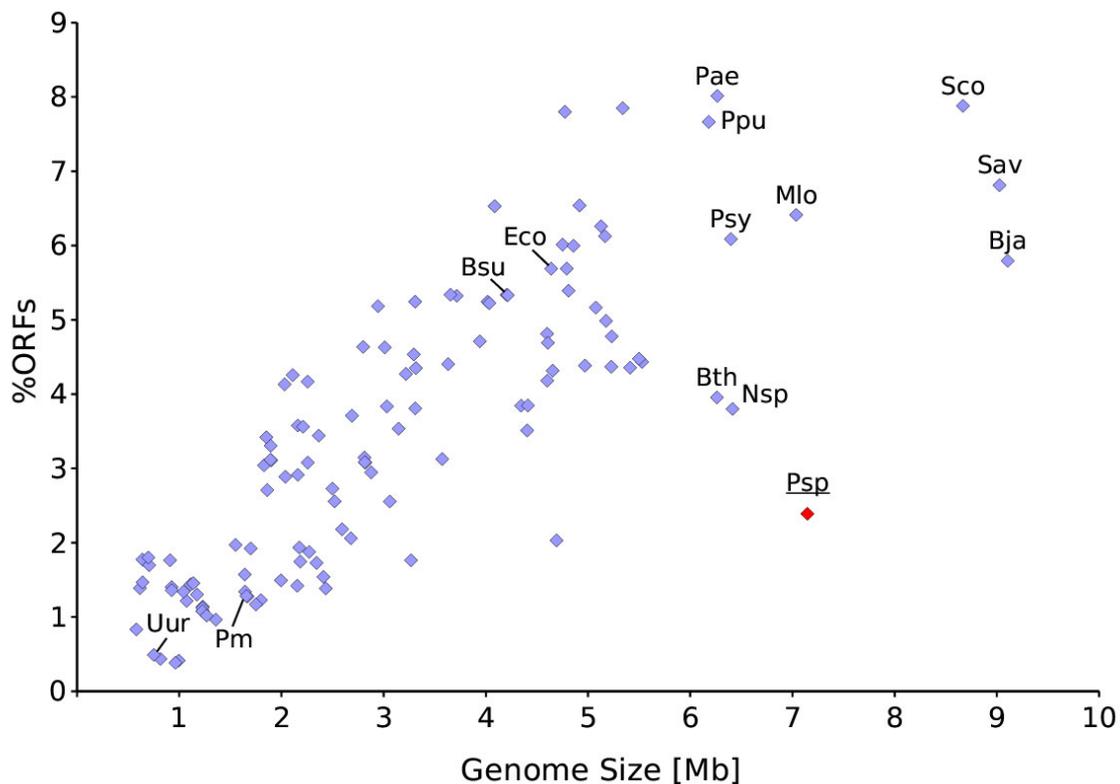


Fig. 44: Relative amount of encoded transcriptional regulators as compared to genome size, according to consistent comparisons of all available bacterial genomes. A total of 119 PFam models were selected for this analysis, based on Gene-Ontology (GO) associations<sup>150</sup> (GO0006355 : regulation of transcription, DNA-dependent). **Psp**: *Pirellula* sp. strain 1, **Bja**: *Bradyrhizobium japonicum*; **Bsu**: *Bacillus subtilis*; **Bth**: *Bacteroides thetaiotaomicron*; **Eco**: *Escherichia coli* K-12; **Mlo**: *Mesorhizobium loti*; **Nsp**: *Nostoc* sp.; **Pae**: *Pseudomonas aeruginosa*; **Pm**: *Prochlorococcus marinus* str. CCMP13; **Ppu**: *Pseudomonas putida*; **Psy**: *Pseudomonas syringae*; **Sav**: *Streptomyces avermitilis*; **Sco**: *Streptomyces coelicolor*; **Uur**: *Ureaplasma urealyticum*.

### 3.3.2.2. Qualitative comparisons: ECF sigma factors

The specificity of the Markov models selected for the analysis shown in Figure 44 allows to distinguish different classes of transcriptional regulators. The first question that can be addressed in the *Pirellula* sp. strain 1 genome is the origin of its regulation systems. According to the cellular compartmentalization observed in this bacterium, eukaryotic cellular processes might be expected. However, no significant hits were obtained in *Pirellula* sp. strain 1 with transcriptional regulator models built exclusively on eukaryotic or/and virus sequences. This observation is consistent with a prokaryotic ancestor for this organism, which is also supported by overall gene content and phylogenetic studies (see chapter 3.5).

In *Pirellula* sp. strain 1, sigma factors are the second largest group of transcriptional regulator domains, behind the two-components systems. The comparison of the total

number of predicted sigma factors in the 10 largest bacterial genomes and selected reference genomes (Fig. 45) shows an extended set of genes encoding sigma factors for *S. coelicolor* (68), *S. avermitilis* (64), *B. thetaiotaomicron* (49) and *Pirellula* sp. strain 1 (49). In contrast, large genomes belonging to the *Proteobacteria* or *Cyanobacteria* phylum show only moderate extensions of these genes: *P. aeruginosa* (26), *M. loti* (25), *P. putida* (24), *B. japonicum* (24), *P. syringae* (15) and *Nostoc* sp. (10). Genomes of less than 6 Mb showed a maximum of 22 predicted sigma factors (*Bacillus halodurans*). Based on this available data, a pronounced extension of sigma factors seems to have occurred only in selected phyla, such as *Actinobacteria*, *Bacteroides* or *Planctomycetes*. Interestingly, many members of these phyla share the ability to degrade a broad range of macromolecules such as polysaccharides<sup>147,131</sup> or, in the case of *Pirellula* sp. strain 1, most likely sulfated polysaccharides with an extended set of sulfatases (see chapter 3.2.3.1). These high number of sigma factors show that an extended regulatory system based on promoter specificity alteration of the RNA polymerase core enzyme for targeted genes sets seems to be a common strategy for many large bacterial genomes. Regulating gene expression by using many classes of modified promoters might be a more effective strategy than influencing gene expression with transcriptional regulators that bind to operators.

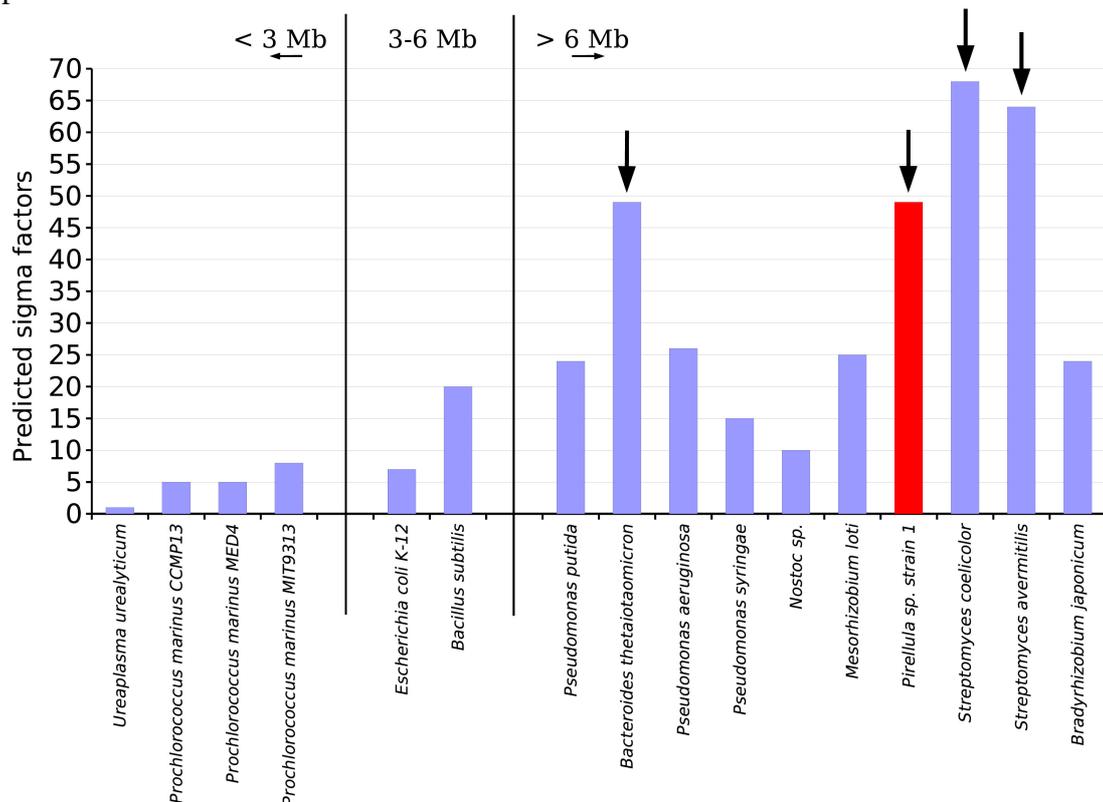


Fig. 45: Total number of predicted sigma factors compared to genome size in *Bacteria* using Markov models. All available genomes larger than 6 Mb, selected reference organisms of moderate (3-6 Mb) and small size (< 3 Mb) genomes are shown. *Pirellula* sp. strain 1 is shown in red. Arrows indicate genomes where a notable extension of sigma factors occurred.

The importance of sigma factors for the adaptation to changing environmental conditions in free-living bacteria is supported by a high proportion of ECF sigma factors (Extra Cytoplasmic Function, subfamily of sigma-70)<sup>147,151,131</sup>. With 37 out of 49 sigma factors predicted to belong to the ECF subfamily, *Pirellula* sp. strain 1 confirms this trend. The roles and mechanisms of regulation for many ECF sigma factors are largely unknown, but significant progress has been made in selected systems. Recent studies show that they are involved in processes such as cell-wall homeostasis, disulfide stress or aerial mycelium development in *S. coelicolor* and alginate biosynthesis in *P. aeruginosa*<sup>151,152,153,154</sup>. Furthermore, the detailed mechanism for the activation of the iron responsive FecI-type ECF sigma factor by transmembrane signaling has been revealed in *E. coli*. ECF sigma factors seem to be part of mechanisms for extra-cellular sensing that are even more complex than the typical two-component systems, involving three proteins: an outer membrane protein (sensor), a membrane associated anti-sigma and the sigma factor itself<sup>155,156</sup>.

### 3.3.2.3. Phylogenetic study of ECF sigma factors

To explore the diversity and evolution of the new ECF sigma factors found in *Pirellula* sp. strain 1, multiple alignments and phylogenetic trees have been calculated for the computationally best supported ECF sigma factors together with a selected dataset of bacterial sequences for which experimental evidences exist (Fig. 46). Clear phylogenetic clusters of known function can be observed: i) iron transport response in *Proteobacteria* (Fig. 46, cluster A); ii) stress response or virulence related in *Proteobacteria* (cluster B) and iii) disulfide stress response in *Actinobacteria* (cluster C). None of the ECF sigma factors of *Pirellula* sp. strain 1 assigns to either of those clusters already reported in previous studies<sup>151,156</sup>. On the contrary, two *Pirellula* sp. strain 1 specific clusters can be observed (clusters D and E). These results support the previous hypothesis that the ECF subfamily of sigma factors might have emerged by gene duplication in large prokaryotic genomes of selected phyla<sup>151</sup>.

An unexpected grouping is observed for SigZ of *B. subtilis* and RB6078 of *Pirellula* sp. strain 1 (Fig. 46, Cluster F). Sequence alignment reveals a highly conserved region four, which suggests a closely related -35 binding site (data not shown). Unfortunately, SigZ is one of the ECF sigma factor of *B. subtilis* whose function is still unknown<sup>157</sup>, and its exact DNA binding site remains to be determined. Targets of SigZ have been reported to include genes encoding for general stress response or unknown functions<sup>158</sup>. Furthermore, it is interesting that SigZ seems to be the only ECF sigma factor of *B. subtilis* which lacks autoregulation<sup>158</sup>. Further experimental evidence for this atypical sigma factor in *B. subtilis* might provide information about the function of its homolog in *Pirellula* sp. strain 1.

Considering the properties of ECF sigma factors, the strategy of using an extended set of

such regulators might provide significant advantages for free-living bacteria. First of all, the environmental changes are directly sensed on the surface of the organism, providing a fast response to activate the synthesis of transporters for available substrates or the appropriate extra-cytoplasmic enzymes for degradation of complex compounds. Second, a single factor of this subfamily is able to regulate up to 50 genes, as shown for SigX in *B. subtilis*<sup>158</sup>. Thus, a large set of such regulators might be an effective way to control the expression of large gene pools. Third, the relaxed specificity toward the -10 promoter regions clearly enables cross regulation of selected operons<sup>159</sup>. Such a strategy might provide a fine tuning for the regulation of genes, operons and regulons involved in environmental adaptation. In *Pirellula* sp. strain 1, the unexpected genetic potential revealed by complete genome analysis such as fermentative pathways, optimized degradation of sulfated polysaccharides or even possible C1-pathways require efficient metabolic switches. The number of ECF sigma factors in this organism and their general properties suggest them as perfect candidates for this task.



### 3.4. Gene expression prediction based on codon usage

Codon usage analysis has been shown to be a usefully tool for gene expression prediction in prokaryotic organisms<sup>64</sup>. During evolution, highly expressed genes undergo a more pronounced codon usage amelioration than other genes. This adaptation is mainly fitting the cellular tRNA pool of each organism in order to optimize protein synthesis, but other evolutionary pressures such as optimization of transcription rates might also play important roles in codon usage shaping<sup>64</sup>. The method of Karlin and Mrazek was applied to the predicted genes of *Pirellula* sp. strain 1, allowing to search for PHX and PA genes. PHX (Predicted Highly eXpressed) are genes showing highly optimized codon usage and PA (Putative Alien) are genes having unadapted or foreign codon usage.

A total of 767 PHX (14.2%) and 344 PA genes (6.4%) were found in the genome of *Pirellula* sp. strain 1 using this methodology (only genes larger than 300 nucleotides). These proportions are in the range of those observed in other prokaryotic organisms (PHX: 4 - 17% and PA: 6 - 7%)<sup>161,162,163,164,165</sup>.

The distribution of PHX and PA genes around the genome of *Pirellula* sp. strain 1 is shown in Figure 47. Particular clusters or gene categories are discussed in the next sections.

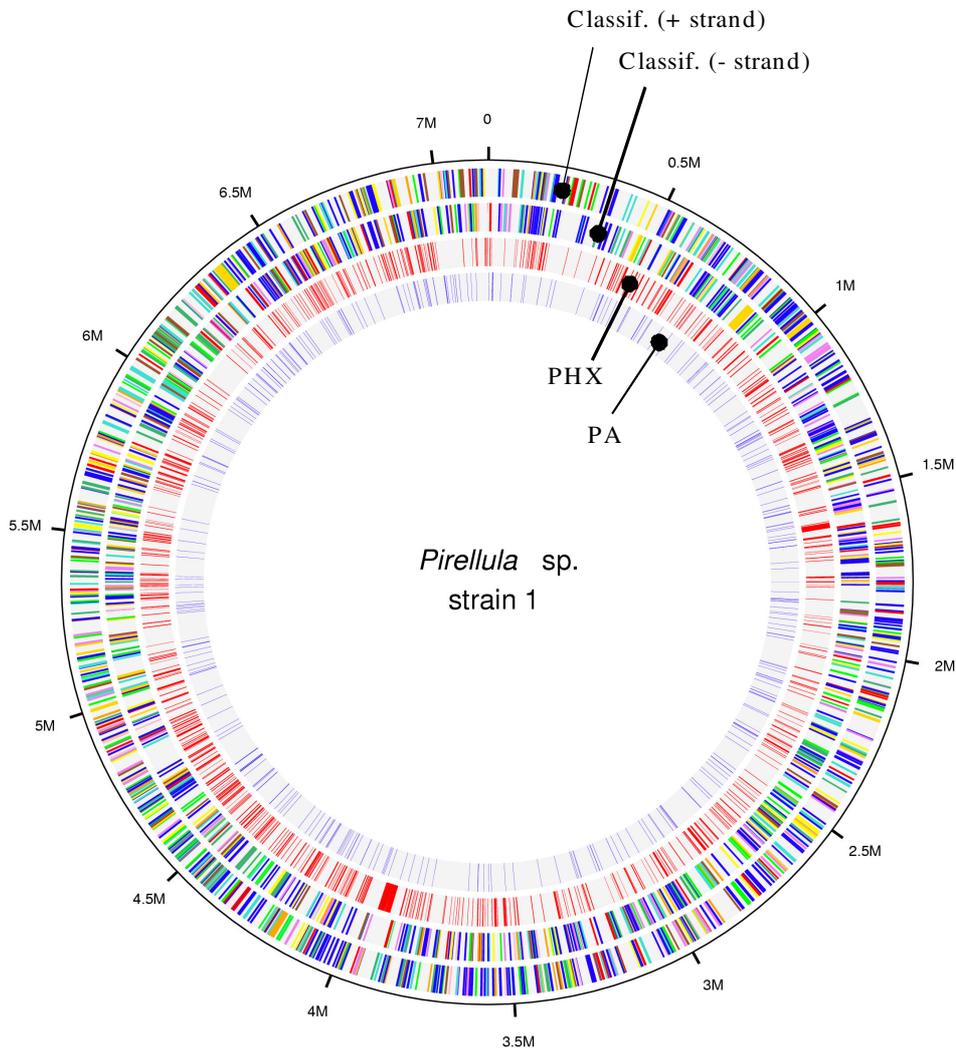


Fig. 47: Localization of predicted expressed (PHX) and putative alien genes (PA) in the genome of *Pirellula* sp. strain 1, as revealed by Karlin-Mrazek analysis (inner circles). Two outer circles: functional classification and distribution of genes on both strands (color codes, see Fig. 28).

### 3.4.1. Analysis according to PHX genes clusters

PHX genes (Predicted Highly expressed) were often reported to correspond to gene clusters, which might reflect functional coupling in the form of highly expressed operons<sup>64</sup>. A total of 99 clusters of at least two PHX genes could be found in the genome of *Pirellula* sp. strain 1. About the half of these clusters (47) encode ribosomal proteins (9 large clusters), information processing (RNA, DNA) (10) and central metabolism (28). This is in agreement with results reported for other organisms<sup>64,163</sup>. Gene clusters of particular environmental interest in *Pirellula* sp. strain 1 are summarized in Table 11.

Table 11: Selected gene clusters with high expression level predictions (PHX: Predicted Highly eXpressed).

PHX gene cluster(s)	Number of genes
nitrate transporter system + nitrate assimilation	4+2
PTS sugar transport system	3
fermentation	2
polysaccharide degradation	2
flagella	4 clusters
chaperonin	6
biopolymer transport	2

Two PHX gene clusters encoding a nitrate specific transporter and nitrate assimilation were found. The optimal expression of such a transporter might be an advantage in marine system, where low nitrate concentration is an important limiting growth factor. The only PTS transporter (phosphotransferase system) encoded by *Pirellula* sp. strain 1 also constitute a PHX cluster, which can reflect a central use of this system for sugar import in the cell in addition to other types of transporters (e.g. ABC-type).

The gene encoding the fermentative lactate dehydrogenase is predicted to be highly expressed, which supports the hypothesis that *Pirellula* sp. strain 1 might be active in anoxic habitat.

Additional interesting PHX genes can be found in the genome, such as 4 clusters encoding structural components of the flagella. This indicate that even if those genes are not located in a single operon, a concerted high expression might be achieved for the rapid assembly of this motility component.

#### 3.4.2. PA genes

The interpretation of PA genes (Putative Alien) in prokaryotic genomes has been shown to be difficult<sup>161</sup>. Initially, such PA genes were thought to be horizontally transferred, as their suboptimal codon usage might reflect original evolutionary constrains of foreign organisms where the genes come from. However, as most of these genes typically have no assigned functions and display no similarity to any other known genes, they might in fact mostly represent pseudogenes. Pseudogenes are genes that lost functionality because

of destructive mutations and are on the way to deletion. Due to the successive mutations within such disrupted genes, the apparent codon usage is randomized and appears unadapted to the translation machinery of the organism. In *Pirellula* sp. strain 1, 311 (90%) of the 344 PA genes have no similarity to any known sequence. This fraction most probably represents pseudogenes. The 33 remaining PA genes with homologies to known sequences represent potential horizontally transferred genes or genes with obsolete functions. In these 33 genes, 14 genes of unknown functions display homologies to genes from *Bacillus halodurans*, *Pseudomonas aeruginosa* or *Deinococcus radiodurans*. These genes are not clustered in the genome of *Pirellula* sp. strain 1, which provides no support for potential gene transfer. Within the few remaining genes, regulators of unknown functions, two transposases and an ATP-synthase subunit (epsilon subunit) are found. This subunit belongs to one of the two operons encoding ATP-synthases in *Pirellula* sp. strain 1. Both ATP-synthases are of the F<sub>0</sub>F<sub>1</sub> type. The first is related to typical bacterial sequences and displays the corresponding classical gene order. However, the second is closely related in term of sequence similarity and gene order to an ATP-synthase operon found in *Methanosarcina barkeri* (Archaea). The occurrence of this PA gene (RB4908) in the Archaea-like operon suggests that an horizontal gene transfer occurred and that this second ATP-synthase is most likely not expressed in *Pirellula* sp. strain 1 and is on the way to deletion.

### 3.4.3. Analysis of selected PHX gene groups

The analysis of the proportion of PHX in gene groups of particular interest can be used to deduce the environmental conditions the organism has commonly to cope with. For example, it was shown that the photosynthetic genes are mainly PHX in the genome of the cyanobacterium *Synechocystis* sp.<sup>161</sup>. In methanogenic Archaea, genes participating in methanogenesis are mainly PHX<sup>64</sup> and in *Deinococcus radiodurans*, which can survive to high doses of ionizing radiation, the largest number of PHX genes encoding detoxification or protease activities was observed<sup>162</sup>. Hence, PHX analysis can be correlated to the lifestyle of a microorganism.

The analysis of the PHX of particular groups of genes in *Pirellula* sp. strain 1 is shown in Table 12. The results show that the central metabolism enzymes in *Pirellula* sp. strain 1 (glycolysis, citrate cycle) are mainly PHX, which is in agreement with previous observations on aerobic, heterotrophic members of the *Bacteria*<sup>166</sup>. This confirms that these pathways are predominant in *Pirellula* sp. strain 1 and suggests that this organism mainly has an aerobic, heterotrophic lifestyle in the environment, as supposed from laboratory growth conditions (H. Schlesner, personal communication).

Table 12: Expression level predictions for selected gene groups or metabolic pathways. The proportion of PHX are high for central metabolism (highlighted in red). (PHX: Predicted Highly eXpressed)

Group of genes/metabolism	% PHX
all genes *	14.2
sulfatases	10.0
special C1 metabolism	only 1 gene
antibiotic biosynthesis	0
polysaccharides degradation	10.7
nucleotide metabolism	26.7
amino acids metabolism	27.2
glycolysis and gluconeogenesis	43.5
citrate cycle	64.7

Only one PHX gene involved in the special C<sub>1</sub> metabolism pathways was found in *Pirellula* sp. strain 1. From the genes corresponding to the C<sub>1</sub> pathway map presented in part 3.2.3.2 (Fig. 32), only the gene encoding the first enzymatic step (Fae) is PHX. This suggests that this particular pathway is not of central importance in this organism, but might be activated under special conditions.

The genes predicted to encode antibiotics biosynthesis (polyketides and polypeptides synthases) contain no PHX, which suggests that these particular pathways are not predominant.

The set of 110 predicted sulfatases in *Pirellula* sp. strain 1 contain 11 PHX. The corresponding expression prediction values E(g) show that 7 of these PHX copies are outstanding as compared to the other sulfatases (Fig. 48). Interestingly, 4 of the 11 PHX sulfatases do not contain the signature pattern CXPXR which is necessary for sulfatase activation in model organisms (see section 3.2.3.1), but a CXAXR pattern. This suggests that an unconventional activation signature might be active in the sulfatases of *Pirellula* sp. strain 1. Moreover, these results propose that this subset of particular sulfatases play a

predominant role under most environmental conditions. Their specificity is probably matching sulfated substrates that are abundant in some habitats such as the marine snow particles. Further investigations of the sulfatases of *Pirellula* sp. strain 1 are carried out experimentally by current projects (transcriptome and proteome analysis, C. Würdemann and D. Gade). At the moment, the *in silico* gene expression prediction (PHX) has been used as a criteria to select genes for partial transcriptome analysis. The presented gene expression prediction constitutes a bridge between *in silico* analysis and functional genomics experiments. The correlation of transcriptome and proteome data with the PHX predictions is an interesting perspective which remains to be tested on extended experimental datasets.

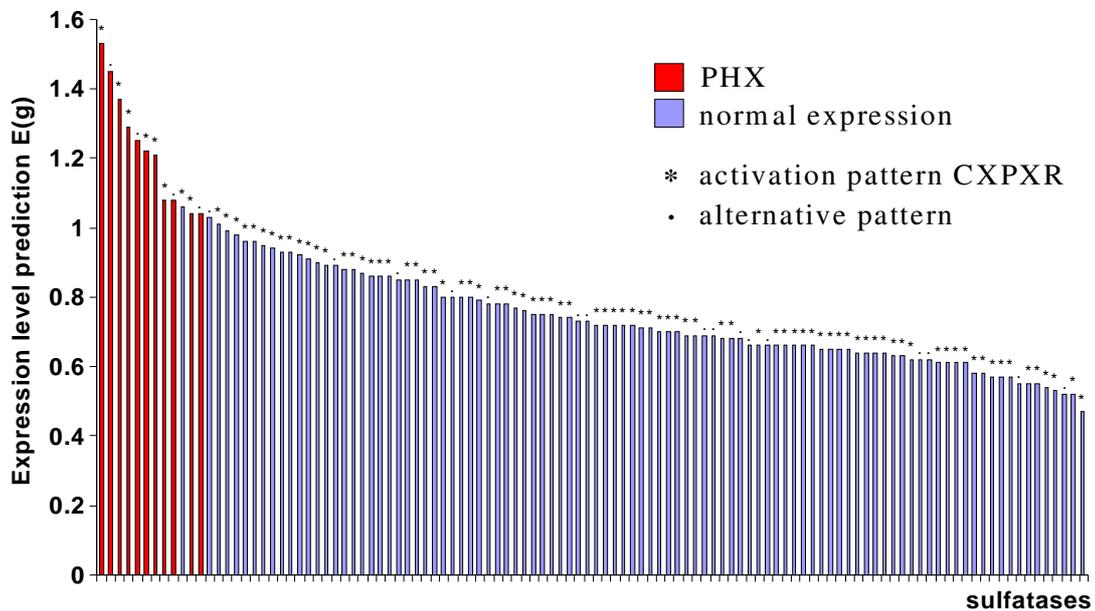


Fig. 48: Predicted expression levels among the 110 predicted sulfatases of *Pirellula* sp. strain 1. The E(g) values were calculated according to the method of Karlin and Mrazek. Genes are sorted by E(g) value and PHX copies (predicted highly expressed) are highlighted in red. Stars indicate the presence of a consensus pattern for activation (CXPXR). Dots indicate alternative patterns (e.g. CXAXR).

### 3.5. Genome trees as a tool for phylogenetic reconstruction

The availability of the complete genome of *Pirellula* sp. strain 1 offers a new data source for phylogenetic studies. The use of extended gene sets instead of single genes for phylogenetic reconstruction has been shown to be a promising approach.<sup>167,168,35</sup> For the first time, the position of a member of the *Planctomycetes* can be estimated based on whole genome information in "genome trees". This "genome tree" approach is computationally very intensive as it is based on normalized BLASTP scores of conserved homologs over full genomes. These calculations resulted in a stable tree topology with respect to parameter variations (Fig. 49).

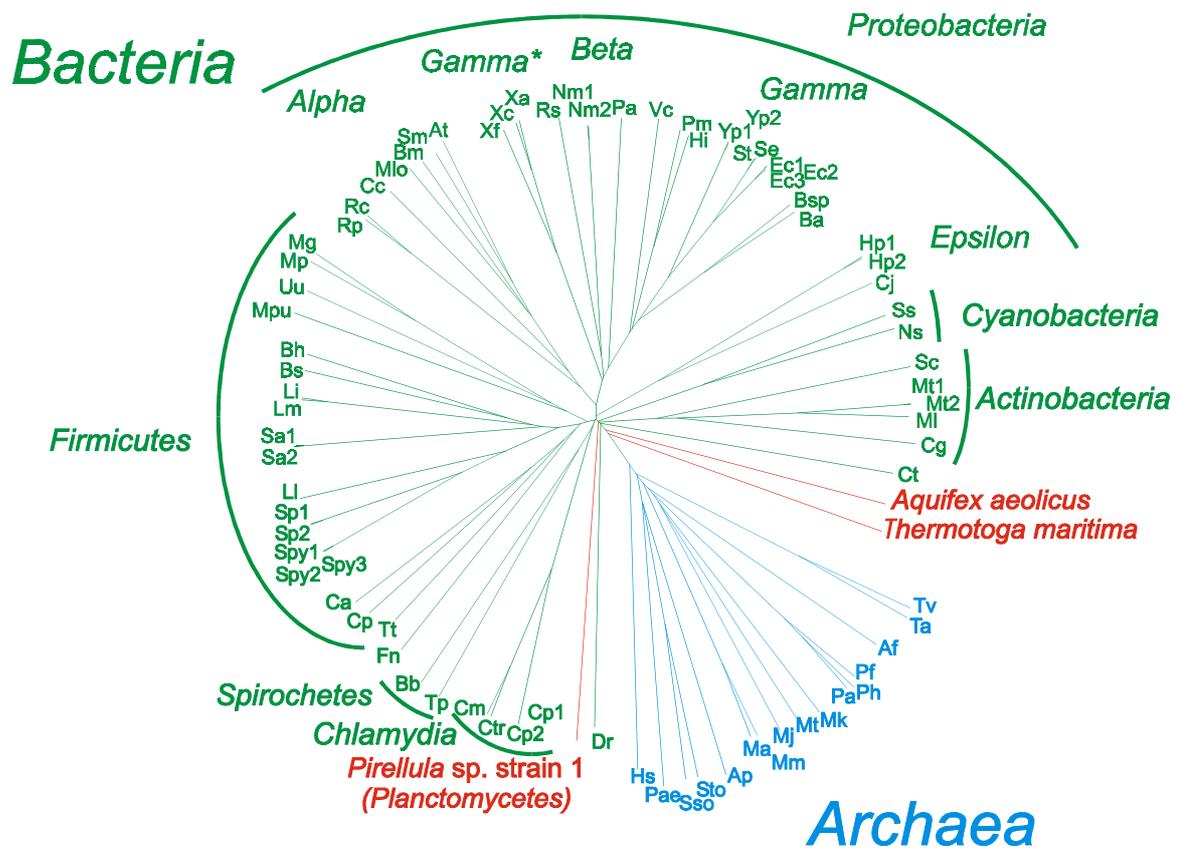


Fig. 49: Genome tree based on mean average best BLASTP hits. Red: *Pirellula* sp. strain 1 and the two thermophilic *Bacteria*; Green: other member of the *Bacteria*; Blue: *Archaea*. Parameters: BLASTP hits cutoff: e-value <  $10^{-10}$ , RBM cutoff: 4, BLOSUM62 matrix. Alternative settings (PAM250, PAM70 matrices, RBM cutoff: 10-40) confirmed the overall tree topology. Organisms: ***Bacteria***: At: *Agrobacterium tumefaciens* strain C58; Ba: *Buchnera aphidicola* str. Sg; Bb: *Borrelia burgdorferi*; Bh: *Bacillus halodurans*; Bm: *Brucella melitensis* strain 16M; Bs: *Bacillus subtilis*; Bsp: *Buchnera* sp. APS; Ca: *Clostridium acetobutylicum* ATCC824; Cc: *Caulobacter crescentus* CB15; Cg: *Corynebacterium glutamicum* ATCC 13032; Cj: *Campylobacter jejuni* subsp. jejuni NCTC 11168; Cm: *Chlamydia muridarum*; Cp: *Clostridium perfringens* str. 13; Cp1: *Chlamydophila pneumoniae* AR39; Cp2: *Chlamydophila pneumoniae* J138; Ct: *Chlorobium tepidum* TLS; Ctr: *Chlamydia trachomatis*; Dr: *Deinococcus radiodurans* R1; Ec1: *Escherichia coli* K-12 MG1655; Ec2: *Escherichia coli* O157:H7 EDL933; Ec3: *Escherichia coli* O157:H7 substrain RIMD 0509952; Fn: *Fusobacterium nucleatum* subsp. nucleatum ATCC 2558; Hi: *Haemophilus influenzae* Rd; Hp1: *Helicobacter pylori* 26695; Hp2: *Helicobacter pylori* strain J99; Li: *Listeria innocua* Clip11262; Lf: *Lactococcus lactis* subsp. lactis strain IL1403; Lm: *Listeria monocytogenes* strain EGD; Mg: *Mycoplasma genitalium* G37; Ml: *Mycobacterium leprae* strain TN; Mlo: *Mesorhizobium loti* strain MA; Mp: *Mycoplasma pneumoniae* M129; Mpu: *Mycoplasma pulmonis* (strain UAB CTIP); Mt1: *Mycobacterium tuberculosis* H37R; Mt2: *Mycobacterium tuberculosis* CDC1551; Nm1: *Neisseria meningitidis* serogroup B strain MC58; Nm2: *Neisseria meningitidis* serogroup A strain Z2491; Ns: *Nostoc* sp. PCC 7120; Pa: *Pseudomonas aeruginosa* PA01; Pm: *Pasteurella multocida* PM70; Rc: *Rickettsia conorii* Malish 7; Rp: *Rickettsia prowazekii* strain Madrid E; Rs: *Ralstonia solanacearum* GMI1000; Sa1: *Staphylococcus aureus* subsp. aureus MW2; Sa2: *Staphylococcus aureus* subsp. aureus Mu50; Sc: *Streptomyces coelicolor* A3(2); Se: *Salmonella enterica* serovar Typhi; Sm: *Sinorhizobium meliloti* 1021; Sp1: *Streptococcus pneumoniae* R6; Sp2: *Streptococcus pneumoniae* TIGR4; Spy1: *Streptococcus pyogenes* M1 GAS strain SF370; Spy2: *Streptococcus pyogenes* MGAS315; Spy3: *Streptococcus pyogenes* strain MGAS8232; Ss: *Synechocystis* sp. PCC 6803; St: *Salmonella typhimurium* LT2; Tp: *Treponema pallidum*; Tt: *Thermoanaerobacter tengcongensis* strain MB4T; Uu: *Ureaplasma urealyticum*; Vc: *Vibrio cholerae*; Xa: *Xanthomonas axonopodis* pv. citri str. 306; Xc: *Xanthomonas campestris* pv. campestris str. ATCC 33; Xf: *Xylella fastidiosa*; Yp1: *Yersinia pestis* strain CO92; Yp2: *Yersinia pestis* KIM; ***Archaea***: Af: *Archaeoglobus fulgidus*; Ap: *Aeropyrum pernix* K1; Hs: *Halobacterium* sp. NRC-1; Ma: *Methanosarcina acetivorans* str. C2A; Mj: *Methanococcus jannaschii*; Mk: *Methanopyrus kandleri* AV19; Mm: *Methanosarcina mazei* strain Goe1; Mt: *Methanobacterium thermoautotrophicum* delta H; Pab: *Pyrococcus abyssii*; Pae: *Pyrobaculum aerophilum* strain IM2; Pf: *Pyrococcus furiosus* DSM 3638; Ph: *Pyrococcus horikoshii* OT3 DNA; Sso: *Sulfolobus solfataricus*; Sto: *Sulfolobus tokodaii* strain 7; Ta: *Thermoplasma acidophilum*; Tv: *Thermoplasma volcanium*.

The presented genome tree approach placed the thermophilic *Bacteria* (*Thermotoga maritima* and *Aquifex aeolicus*) as the deepest branching members of this domain (Fig 49). The *Planctomycetes*, represented by *Pirellula* sp. strain 1, emerge as one of the first non thermophilic bacterial phylum, at the same level as the *Actinobacteria*, *Cyanobacteria* and *Deinococcus*. All known phyla described by 16S rRNA analysis are fully resolved by this method. The position of the different groups of *Proteobacteria* (*alpha*, *beta*, *gamma*, *gamma*\*, *delta* and *epsilon*) are for example precisely following classical 16S rRNA trees. Interestingly, the recently proposed *Spirochetes-Chlamydia* superclade<sup>167,168,35</sup> is supported by these genome tree results.

The phylogenetic position of the *Planctomycetes* is still a subject of discussions in the literature and the results of the genome trees approach presented here are in agreement with one of the previously proposed *Planctomycetes* position. While *Planctomycetes* are clearly monophyletic according to all types of analysis, their exact branching position within the bacterial domain varies according to the selected phylogenetic marker or method. In 1995, first analysis based on a limited number of 16S rRNA sequences suggested a possible relationship to *Chlamydia*<sup>31,32</sup>, but extensive analysis on larger 16S and 23S datasets did not support these earlier results<sup>33</sup>. Later, an EF-Tu-based analysis confirmed again the monophyly of the *Planctomycetes*, but could not propose a clear branching position<sup>34</sup>. In 2002, a new algorithm based on the selection of slowly evolving nucleotide positions in 16S rRNA sequences placed the *Planctomycetes* as the deepest branching phylum within the *Bacteria*<sup>37,36</sup>. This new tree topology was interesting, considering the eukaryotic-like cellular compartmentalization observed in *Planctomycetes*. However, a contradictory analysis based on the same 16S rDNA dataset, but using an alternative site selection methodology, rather suggested that the *Planctomycetes* are branching off after thermophilic members of the *Bacteria* (*Thermotogales* or *Aquificales*)<sup>38</sup>, which is in agreement with the presented genome tree approach. Interestingly, an approach based on ribosomal protein concatenation confirmed that the *Planctomycetes* are not deepest branching and suggested a relationship between *Planctomycetes* and *Chlamydia*<sup>A4</sup>.

The information used by the different phylogenetic reconstruction approaches and their associated bias are compared in Table 13. The genome trees constitute a new, experimental phylogeny whose potential is just being discovered. The proportion of single copy orthologs genes has been shown to be lower as previously believed, supporting the observation that genome trees are compatible with 16S rRNA based phylogeny<sup>137</sup>. However, no reliable high-throughput method for their identification has been described so far. In the future, the success of genome trees for phylogenetic inference will rely on effective filtering of horizontally transferred genes.

In summary, the new analysis based on extended gene sets (concatenation or genome trees) are in agreement with most of the studies performed earlier, but are in contradiction with the proposed deepest branching of the *Planctomycetes* within the bacterial domain.

Clearly, the debate concerning the phylogenetic position of *Planctomycetes* remains open and this question has to be constantly reevaluated on larger datasets or possible alternative methodologies.

Table 13: Comparison of the main available phylogenetic reconstruction methods (-: low; +:medium; ++: high).

Method	Number of available sequences	Information content	Horizontal gene transfer bias	Computation time	Interpretation
Single gene	++	-	-	-/+	Phylogeny
Concatenated proteins	-	+	+	+/**	Phylogeny
Genome trees	-	++	++	++	Phylogeny / Eco-physiology

### 3.6. Metagenomes mapserver (prototype)

The first wave of genome sequencing projects was oriented towards organisms of medical or biotechnological interest. *Pirellula* sp. strain 1 represents one of the first microorganisms that was selected for a whole genome approach because of its environmental relevance. Thanks to centralized efforts such as the Joint Genome Institute projects<sup>170</sup>, the number of complete and draft prokaryotic genomes of organisms playing crucial roles in the environment is now growing quickly. However, exploring the microbial diversity in the environment by whole genome sequencing is obviously limited by the fact that only a minor part of the microorganisms can be cultivated under laboratory conditions<sup>171</sup>. Therefore, the retrieval of prokaryotic genome fragments directly from the environment by metagenomic approaches is the indispensable companion of whole genome sequencing of isolates in the field of environmental genomics. The metagenomic approach allows to gain insights into the genetic potential of uncultivated microorganisms by cloning and sequencing 50-200 Kb genomic fragments. Genomic and metagenomic approaches are highly synergistic, because genome fragment of unknown identity can be compared to fully annotated reference genomes.

Interestingly, the metagenomic approach produces a new type of biological data. Currently, metagenomic fragments are integrated in public genes or proteins databases such as EMBL<sup>65</sup> or GenBank<sup>169,172</sup>, but the environmental context of the samples is lost and has to be searched in the literature for each independent entry. A genome oriented database structure would be more appropriate, but metagenomic data are not included in public genome databases because of their short length and the lack of clear assignment to any described microorganism. Therefore, there is a clear need for a new database structure to efficiently exploit the potential of this new data source. In this context, a

prototype of geographic information system (GIS) for environmental metagenomics data storage and data mining is presented here: the "Metagenomes mapserver". Such a system opens new opportunities such as: i) database query according to the sampling site from a geographic map; ii) sequence similarity searches with visual localization of the results (under development) and iii) search for particular functions and visual localization. The new possible data mining fluxes are shown in Figure 50.

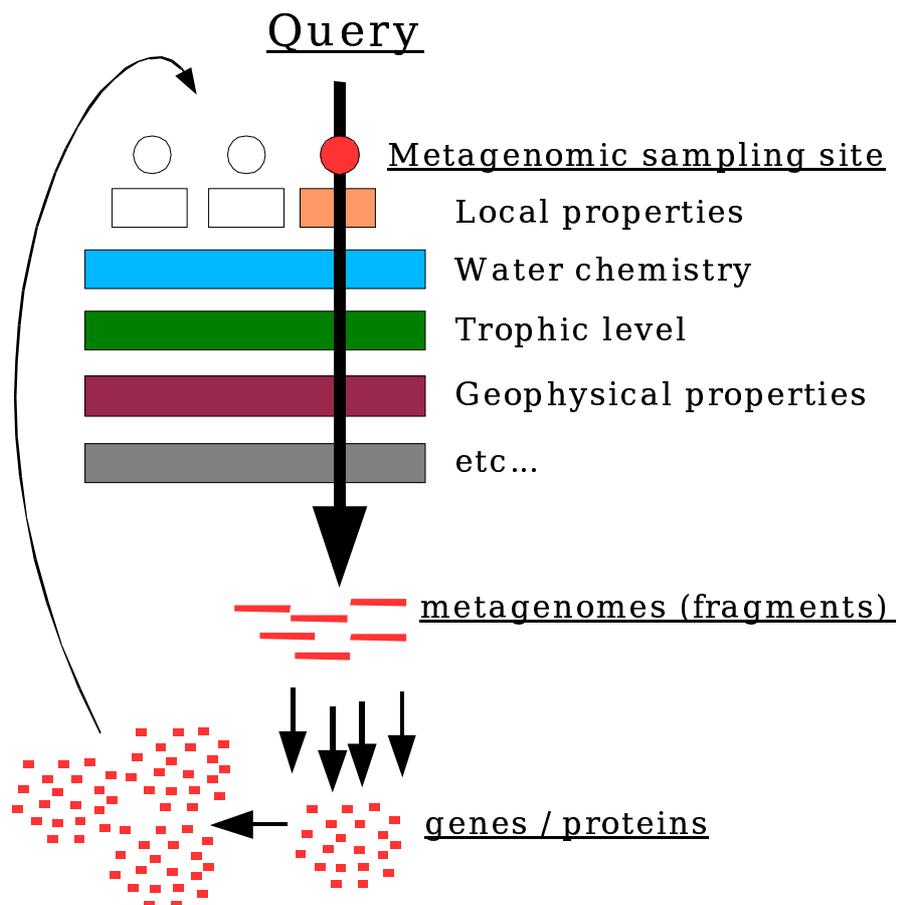


Fig. 50: Metagenomic data mining with geographical information. The "Metagenomes mapserver" prototype allows to access genes and genome fragments in the context of their original sampling site. In contrast, classical molecular biology databases only cover the lower part of this diagram.

### 3.6.1. Database design

The aim of the Metagenomes Mapserver prototype is to integrate geographic information with genomic fragments directly retrieved from the environment. Therefore, the underlying database structure has been designed in two main interconnected components (Fig. 51).

The first component is a GIS layer providing geographic data, following the OpenGIS

standard<sup>71</sup>. The use of this open standard is very advantageous, as any new geographical data conforming to these specifications can be directly integrated in the local system (as it is implemented in this prototype). Another possibility which will be more common in the near future is the retrieval of additional informations directly from remote standard GIS servers through the Internet.

The second component of the Metagenomes Mapservers consists of a typical genomic layer strictly containing classical sequence related information such as genes position (orf), annotation and additional computed bioinformatic facts for metagenomic fragments. Combined GIS / metagenomes data retrieval is possible through a new web interface which is presented in the next section.

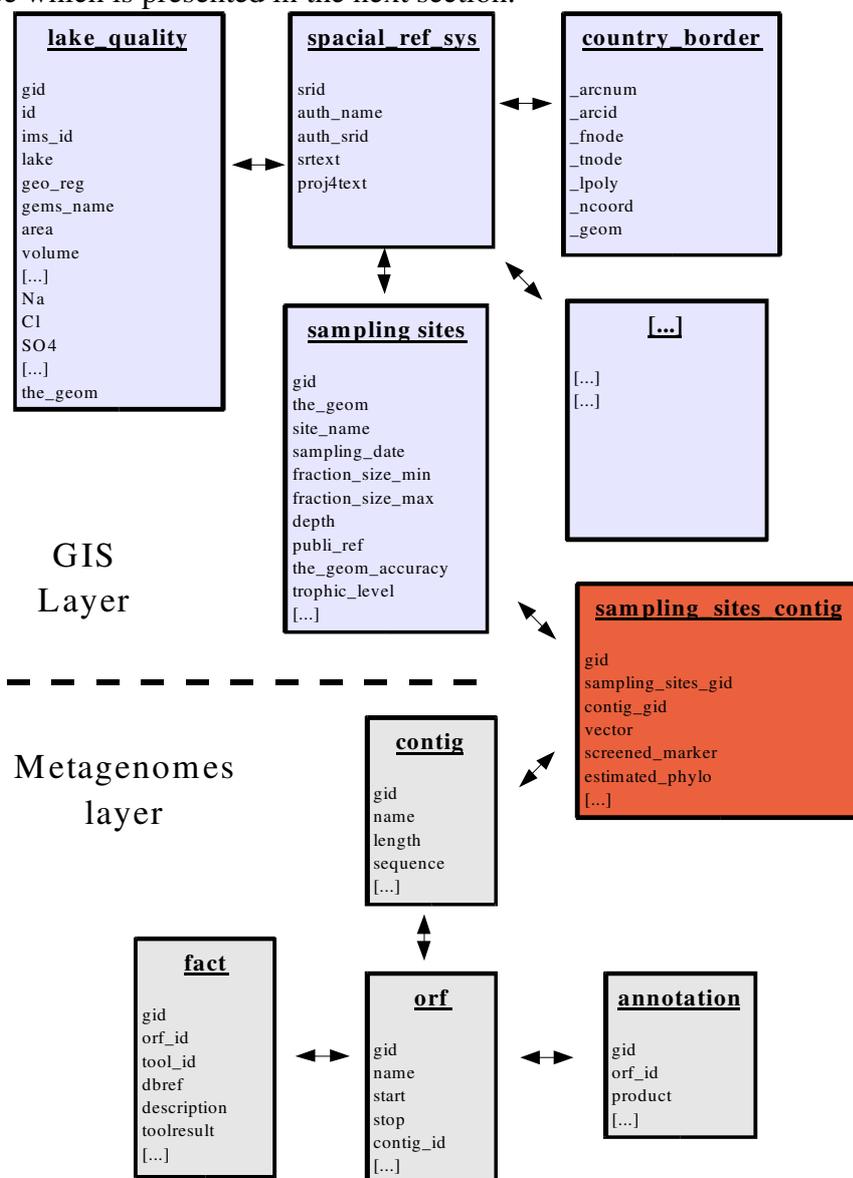


Fig. 51: Database structure of the Metagenomes Mapservers prototype. Blue tables: geographic information (e.g. sampling sites); grey tables: genomic or metagenomic data; red table: data cross-linking.

### 3.6.2. Towards a geo/ecological analysis of genomic fragments

The user interface of the Metagenomes Mapserver is implemented in the form of a classical web page which can be accessed from any web browser such as Internet Explorer or Netscape. The prototype version is still restricted to local access in the Department of Molecular Ecology for tests purposes, but free access from the Internet could be opened in the future, providing the scientific community with this new data-mining tool adapted to environmental metagenomics.

Figure 52 shows a screenshot of the Metagenome Mapserver interface. A world map is presented to the user as starting point to give an overview of the regions where natural samples were studied by metagenomic approach. The interface is dynamic: the user can zoom and move to selected regions of interest, add and remove information layers for optimal geographical data mining.

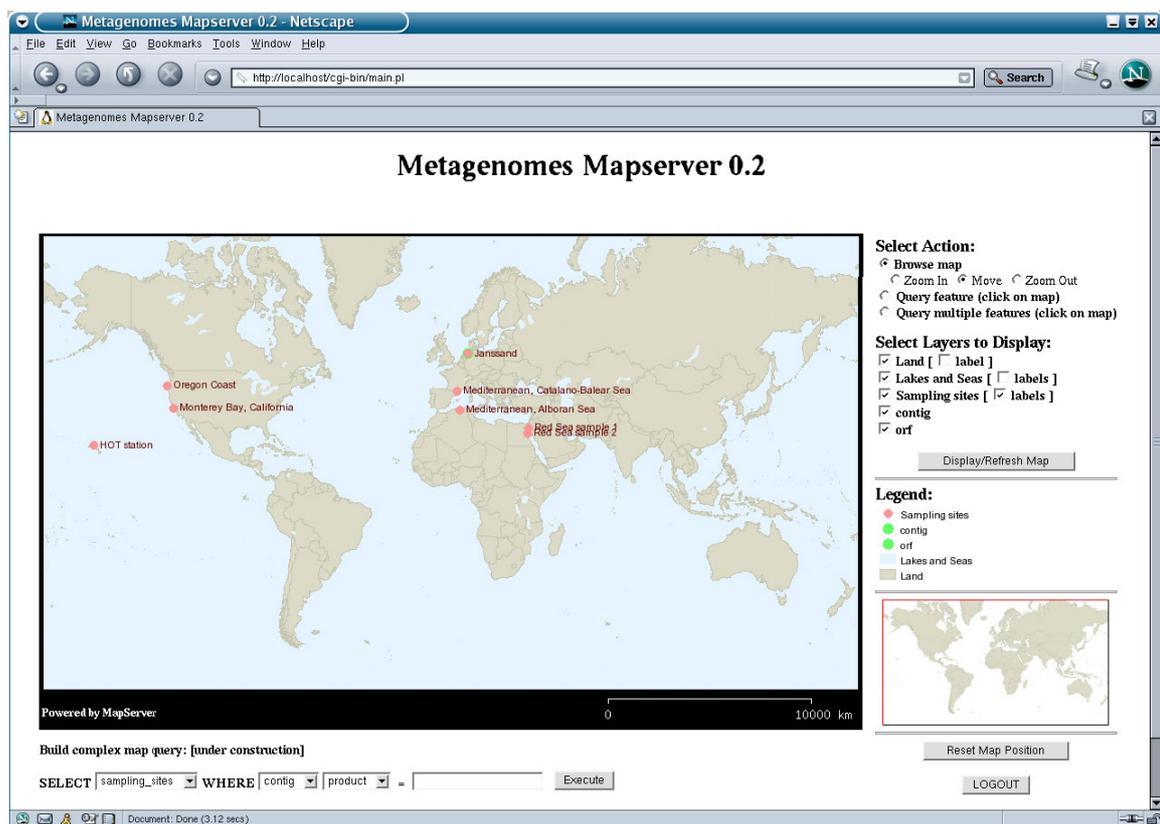
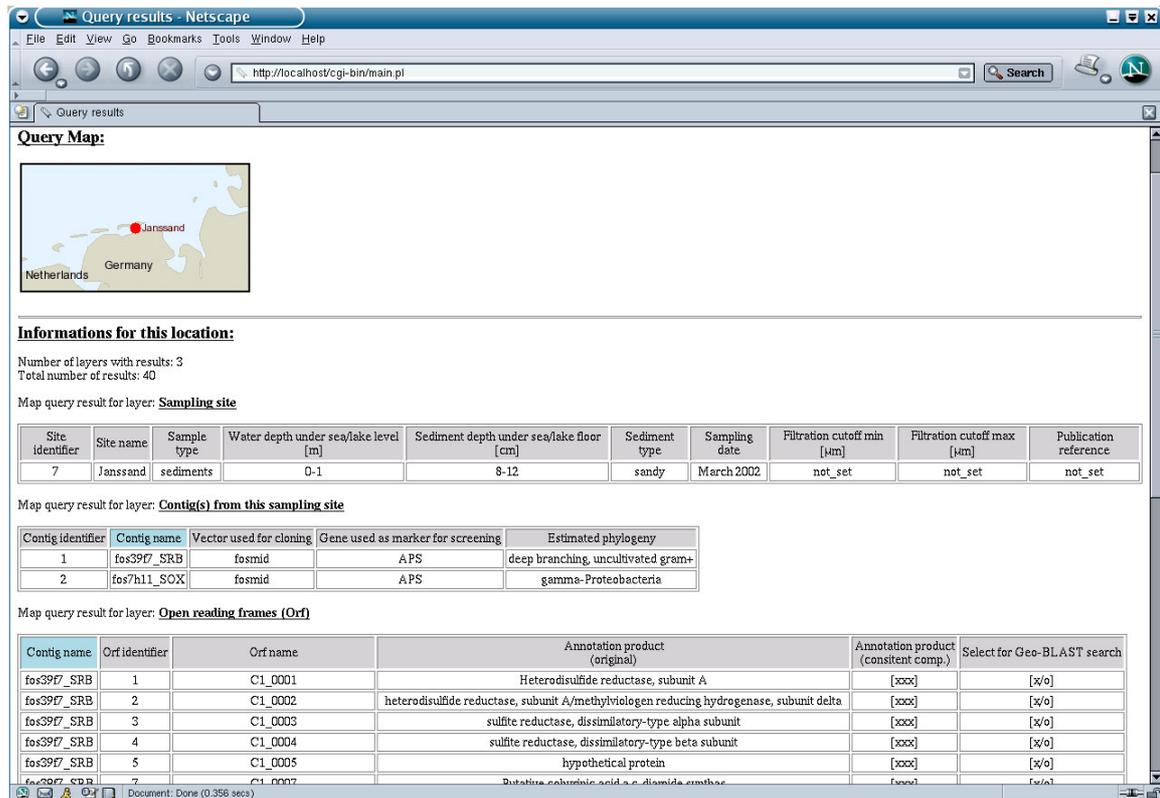


Fig. 52: Metagenomes mapserver screenshot (main map). A simple click on a sampling site (red dots) provides access to the associated data (see next Figure).

Sampling sites of interest can be selected and a query by simple mouse click gives access to the metagenomic information layer. Figure 53 shows an example of sampling site with associated metagenomics sequences. In this case, the samples were taken from a given

depth in marine sediments and two metagenomic fragments were sequenced. All the annotated genes of these fragments are listed, opening a window on the metabolic potential of the microorganisms in this particular ecological niche.



**Query Map:**

**Informations for this location:**  
 Number of layers with results: 3  
 Total number of results: 40  
 Map query result for layer: **Sampling site**

Site identifier	Site name	Sample type	Water depth under sea/lake level [m]	Sediment depth under sea/lake floor [cm]	Sediment type	Sampling date	Filtration cutoff min [µm]	Filtration cutoff max [µm]	Publication reference
7	Janssand	sediments	0-1	8-12	sandy	March 2002	not_set	not_set	not_set

Map query result for layer: **Contig(s) from this sampling site**

Contig identifier	Contig name	Vector used for cloning	Gene used as marker for screening	Estimated phylogeny
1	fos39f7_SRB	fosmid	APS	deep branching, uncultivated gram+
2	fos7hl1_SOX	fosmid	APS	gamma-Proteobacteria

Map query result for layer: **Open reading frames (Orf)**

Contig name	Orf identifier	Orf name	Annotation product (original)	Annotation product (consistent comp.)	Select for Geo-BLAST search
fos39f7_SRB	1	C1_0001	Heterodisulfide reductase, subunit A	[xxx]	[x/o]
fos39f7_SRB	2	C1_0002	heterodisulfide reductase, subunit A/methylviologen reducing hydrogenase, subunit delta	[xxx]	[x/o]
fos39f7_SRB	3	C1_0003	sulfite reductase, dissimilatory-type alpha subunit	[xxx]	[x/o]
fos39f7_SRB	4	C1_0004	sulfite reductase, dissimilatory-type beta subunit	[xxx]	[x/o]
fos39f7_SRB	5	C1_0005	hypothetical protein	[xxx]	[x/o]
fos39f7_SRB	7	C1_0007	Dutatinic coenzyme A: c. diamide synthase	[xxx]	[x/o]

Fig. 53: Metagenomes mapperserver screenshot (example of sampling site entry). The sampling location is highlighted and all related data are displayed: local properties, fragments which were retrieved from this location and annotated genes (sequence data kindly provided by M. Mussmann).

The presented Metagenomes Mapperserver prototype constitutes a new platform for genomic/metagenomic data mining in an ecological context. An interesting perspective for this prototype is the integration of specialized algorithms for automatic correlation of genomic and geographical/ecological data. At the moment, the available metagenomic data constitute a limiting factor. However, the number of publicly available metagenomic sequences retrieved directly from the environment is expected to grow exponentially within the next years, just as whole genomes data. Therefore, it is an urgent task to further develop the appropriate bioinformatic structures to ensure an optimal knowledge extraction from this fascinating new data source.

## 4. Conclusions

The bioinformatic analysis of the genome of *Pirellula* sp. strain 1 revealed the complete blueprint of life of this organism and some unexpected findings. This constitutes a proof of principle of the important role that whole genome analysis will play in the field of ecology as a “hypothesis generator” in the future. The establishment of *in silico* gene expression predictions represents a bridge between bioinformatic and transcriptome/proteomes studies to generate hypothesis which can then be confirmed or rejected by functional genomic studies. Furthermore, the availability of this first *Planctomycete* genome also allowed to contribute to the study of the evolution of this particular phylum.

Looking back, the choice of a member of an outstanding bacterial phylum like *Pirellula* sp. strain 1 for whole genome sequencing was risky, as only limited physiological information was available for *Planctomycetes*. However, this was also the opportunity for new and unexpected findings because this organism represents the first member of the *Planctomycetes* whose complete genome was sequenced. These results give insights into further judicious choices of further environmental bacterial strains for a whole genome approach. Of course, these organisms will have to be environmentally relevant, either because of their ubiquitousness or their local abundance in specific habitats. But an optimal amount of knowledge can be obtained if they belong to outstanding bacterial groups whose physiology is not well known and where no other representative has been sequenced, as exemplified by *Pirellula* sp. strain 1. However, the time window during which such choices will be possible is probably quite short from now on, as over 400 microbial genomes project are underway (status of January 2004). Most part of the phylogeny of cultivated microorganisms will soon be covered by one or more complete genomes. When this time point will be reached, the focus of environmental genomics might move to comparative studies of closely related microorganisms whose ecological niches are slightly different. Furthermore, the emerging field of environmental metagenomics represents an indispensable independent source of information for comparative studies. With the help of the appropriate computational structures, such genomic and metagenomic comparisons might lead to the definition of habitat-specific genes and will also deliver more background information about the correlations between the genetic potentials and the ecology of microorganisms.

## 5. Annexes

### **Annex 1:**

A Perl program to calculate Karlin-Mrazek parameters for every gene of a prokaryotic genome (documentation and listing of karlin\_mrazek.pl).

### **Annex 2:**

A typical import-export program to convert data from a text file to a GenDB database, using Perl-02DBI (listing: karlin\_mrazek\_gendb\_import.pl).

### **Annex 3:**

Central metabolism in *Pirellula* sp. strain 1: glycolysis, pentose phosphate pathway and citrate cycle.

### **Annex 4:**

Amino acids metabolism in *Pirellula* sp. strain 1.

### **Annex 5:**

Carbohydrate metabolism in *Pirellula* sp. strain 1.

### **Annex 6:**

Sulfatases: multiple sequence alignment.

## Annex 1:

A Perl program to calculate Karlin-Mrazek parameters for every gene of a prokaryotic genome (documentation and listing of karlin\_mrazek.pl).

```
#####  
# codon_mrazek.pl readme file #  
# Thierry Lombardot 10.07.2002 #  
#####
```

codon\_mrazek.pl is an implementation of the "codon usage difference" calculation introduced by Jan Mrazek in 2000 [see J. Bact. 182(18):5238-5250]. "Codon usage difference" is a valuable index for gene expression levels prediction.

This script runs an external program for codon tabulation (codonw, John Peden, <http://www.molbiol.ox.ac.uk/cu/codonW.html>) and then calculates Karlin-Mrazek parameters for each orf of the given dataset. The "codonw" executable has to be located within the same folder.

Input formats:

Open reading frames (ORFs) lists in FASTA format:

RP.fasta: a list of all ORFs encoding Ribosomal proteins.

TF.fasta: a list of all ORFs encoding Transcription and translation factors.

CH.fasta: a list of all ORFs encoding Chaperonins.

C.fasta: full list of ORFs

usage:

```
perl codon_mrazek.pl RP.fasta TF.fasta CH.fasta C.fasta outfile
```

output:

A single tab delimited file containing 6 columns:

column1: gene name

column2: B(g|C) value

column3: B(g|RP) value

column4: B(g|CH) value

column5: B(g|TF) value

column6: expression level prediction (PHX,N,PA). PHX= Predicted highly expressed, N= normally expressed (not PHX or PA) and PA= Putative alien.

## Annex 1 (continued):

```
#!/usr/bin/perl

#####
# This PERL script is an implementation of the modified codon usage index #
# introduced by Jan Mrazek for gene expression prediction. #
# See reference: J. Bact 182(18):5238-5250 #
# #
# See README file "codon_mrazek.pl.readme" #
# ----- #
# Thierry Lombardot 10.07.2002 #
#####

#####
#main program

#arguments check
if (not (defined($ARGV[4]))) {
    die("\n\nUSAGE:\nperl codon_mrazek.pl RP TF CH C OUT \n\n");
}
#for calculation time
$time_start = time;
print(STDERR "\n\nProcessing...\n\n");
runcodonw();
%RP= parse_codonw($ARGV[0]); #parsing codon usage of orfs coding Ribosomal proteins
%TF= parse_codonw($ARGV[1]); #parsing codon usage of orfs coding Transcription/Translation factors proteins
%CH= parse_codonw($ARGV[2]); #parsing codon usage of orfs coding Chaperonins proteins
%C= parse_codonw($ARGV[3]); #parsing codon usage of all orfs
$all_codon_file= "$ARGV[3]". "_all";
@ORF= parse_codonw_all($all_codon_file); #parsing codon usage of each orf
%RPn = normalize_codon(%RP); #normalization to 1 for each codon class
%TFn= normalize_codon(%TF);
%CHn= normalize_codon(%CH);
%Cn= normalize_codon(%C);
open(OUT, ">$ARGV[4]");
print_output_title();
for $orf_nb (0..($#ORF - 1)) { #for each orf
    %{$ORFn[$orf_nb]}= normalize_codon(%{$ORF[$orf_nb]});
    %{$ORFp[$orf_nb]}= aa_frequency(%{$ORF[$orf_nb]});
    $BgC[$orf_nb]= calculate_BgX(%Cn);
    $BgRP[$orf_nb]= calculate_BgX(%RPn);
    $BgTF[$orf_nb]= calculate_BgX(%TFn);
    $BgCH[$orf_nb]= calculate_BgX(%CHn);
    $ERPg[$orf_nb]= calculate_EXg($BgC[$orf_nb], $BgRP[$orf_nb]);
    $ETFg[$orf_nb]= calculate_EXg($BgC[$orf_nb], $BgTF[$orf_nb]);
    $ECHg[$orf_nb]= calculate_EXg($BgC[$orf_nb], $BgCH[$orf_nb]);
    $Eg[$orf_nb]= calculate_Eg($BgC[$orf_nb], $BgRP[$orf_nb], $BgTF[$orf_nb], $BgCH[$orf_nb]);
}
@test = (1,2,4,10,40,50);
%M = calculate_M(@BgC);
for $orf_nb (0..($#ORF - 1)) { #for each orf
    $exp_pred[$orf_nb]= define_PHX_PA($BgRP[$orf_nb], $BgTF[$orf_nb], $BgCH[$orf_nb], $BgC[$orf_nb], $ERPg[$orf_nb],
    $ETFg[$orf_nb], $ECHg[$orf_nb], $Eg[$orf_nb], %M);
    print_output_data();
}
close(OUT);
#for calculation time
$time_stop = time; #for calculation time
$elapsed_time_sec = $time_stop - $time_start;
$elapsed_time_min = ($elapsed_time_sec / 60);
printf(STDERR "\nDone...(in %d minutes)\n\n", $elapsed_time_min);

#main program end
#####

#####
#subroutines

sub runcodonw { #run codonw for codon tabulation (creating *.blk files)
    print(STDERR "\n*****running codonw...\n");
    system("./codonw $ARGV[0] -cutot -nomenu"); #codon usage tabulation for RP (rib. prot.)
    print(STDERR "\n*****running codonw...\n");
    system("./codonw $ARGV[1] -cutot -nomenu"); #codon usage tabulation for TF (transcr + transl. fact.)
    print(STDERR "\n*****running codonw...\n");
    system("./codonw $ARGV[2] -cutot -nomenu"); #codon usage tabulation for CH (chaperonins)
    print(STDERR "\n*****running codonw...\n");
    system("./codonw $ARGV[3] -cutot -nomenu"); #codon usage tabulation for C (all ORFs)
    print(STDERR "\n*****running codonw...\n");
    system("cp $ARGV[3] $ARGV[3]_all");
    system("./codonw $ARGV[3]_all -nomenu"); #codon usage tabulation for every single ORF.
    system("rm *.out $ARGV[3]_all"); #delete useless files generated by codonw
}
}
```

## Annex 1 (continued):

```
sub parse_codonw { #parsing of codonw output (parsing *.blk files to different hashes: e.g. RP{UUU} = 1000 means that
RP proteins have a total of 1000 UUU codons.
my ($file)= @_;
my %orfs;
open(FILE,"<$file.blk") || die ("can't open $file.blk");
while (<FILE>) {
# /([AUGC]{3})\s+(\d+).*([AUGC]{3})\s+(\d+).*([AUGC]{3})\s+(\d+).*([AUGC]{3})\s+(\d+)/;#regex for codon count
(codonw output format)
/([AUGC]{3})(.5).*([AUGC]{3})(.5).*([AUGC]{3})(.5).*([AUGC]{3})(.5)/; #corrected regex for high codon
counts
$orfs{$1}= $2;$orfs{$3}= $4;
$orfs{$5}= $6;$orfs{$7}= $8;
}
close(FILE);
return %orfs;
}

sub parse_codonw_all { #parsing of codonw output for an orf list (parsing *_all.blk file to an array of hashes: e.g.
orf1{UUU} = 1000 means that orf2 have a total of 1000 UUU codons.)
my ($file)= @_;
my @orfs;
my $orf_nb;
open(FILE,"<$file.blk") || die ("can't open $file.blk");
$orf_nb = 0;
while (<FILE>) {
if (/codons\s\sin\s+(.+)\(used/) {
$orfs[$orf_nb]{id}= $1;
$orf_nb++;
next;
}
/([AUGC]{3})\s+(\d+).*([AUGC]{3})\s+(\d+).*([AUGC]{3})\s+(\d+).*([AUGC]{3})\s+(\d+)/; #regex for codon count
(codonw output format)
$orfs[$orf_nb]{$1}= $2;$orfs[$orf_nb]{$3}= $4;
$orfs[$orf_nb]{$5}= $6;$orfs[$orf_nb]{$7}= $8;
}
close(FILE);
return @orfs;
}

sub normalize_codon { #Codon frequencies normalization: normalizing codon count to 1 for each codon group (codon for
one amino acid)
my (%orfs)= @_;
my %normalized_orfs, $aa, $i,$j;
$normalized_orfs{id}= $orfs{id};
%codons = ( #codons for each aa are written in hashes of arrays. e.g. $codons[Phe]= [UUU, UUC]
Phe=> [UUU, UUC],
Leu=> [UUA, UUG, CUU, CUC, CUA, CUG],
Ile=> [AUU, AUC, AUA,],
Met=> [AUG],
Val=> [GUU, GUC, GUA, GUG],
Ser=> [UCU, UCC, UCA, UCG, AGU, AGC],
Pro=> [CCU, CCC, CCA, CCG],
Thr=> [ACU, ACC, ACA, ACG],
Ala=> [GCU, GCC, GCA, GCG],
Tyr=> [UAU, UAC],
His=> [CAU, CAC],
Gln=> [CAA, CAG],
Asn=> [AAU, AAC],
Lys=> [AAA, AAG],
Asp=> [GAU, GAC],
Glu=> [GAA, GAG],
Cys=> [UGU, UGC],
Trp=> [UGG],
Arg=> [CGU, CGC, CGA, CGG, AGA, AGG],
Gly=> [GGU, GGC, GGA, GGG]
);
foreach $aa (keys %codons) { #for each amino acid
$sum_codons= 0;
for $i (0..${#{$codons{$aa}}}) { #for each codon coding for a single amino acid
$sum_codons= $sum_codons + $orfs{$codons{$aa}[$i]}; #sum all codons
}
for $j (0..${#{$codons{$aa}}}) { #for each codon coding for a single amino acid
if ($sum_codons == 0) { $normalized_orfs{$codons{$aa}[$j]}= 0; }
if ($sum_codons > 0) { $normalized_orfs{$codons{$aa}[$j]}= $orfs{$codons{$aa}[$j]} / $sum_codons; }
}
}
return (%normalized_orfs);
}

sub aa_frequency { #Calculation of amino acids (aa) frequency, based on codon frequency
my (%orfs)= @_;
my %orfs_aa;
my $sum_aa;

```

## Annex 1 (continued):

```
my $orf_length;
#amino acids count
$orfs_aa{Phe} = $orfs{UUU}+$orfs{UUC}; #Phe codons
$orfs_aa{Leu} = $orfs{UUA}+$orfs{UUG}+$orfs{CUU}+$orfs{CUC}+$orfs{CUA}+$orfs{CUG}; #Leu codons
$orfs_aa{Ile} = $orfs{AUU}+$orfs{AUC}+$orfs{AUA}; #Ile codons
$orfs_aa{Met} = $orfs{AUG}; # Met codon
$orfs_aa{Leu} = $orfs{GUU}+$orfs{GUC}+$orfs{GUA}+$orfs{GUG}; #Leu codons
$orfs_aa{Ser} = $orfs{UCU}+$orfs{UCC}+$orfs{UCA}+$orfs{UCG}+$orfs{AGU}+$orfs{AGC}; #Ser codons
$orfs_aa{Pro} = $orfs{CCU}+$orfs{CCC}+$orfs{CCA}+$orfs{CCG}; #Pro codons
$orfs_aa{Thr} = $orfs{ACU}+$orfs{ACC}+$orfs{ACA}+$orfs{ACG}; #Thr codons
$orfs_aa{Ala} = $orfs{GCU}+$orfs{GCC}+$orfs{GCA}+$orfs{GCG}; #Ala codons
$orfs_aa{Tyr} = $orfs{UAU}+$orfs{UAC}; #Tyr codons
$orfs_aa{His} = $orfs{CAU}+$orfs{CAC}; #His codons
$orfs_aa{Gln} = $orfs{CAA}+$orfs{CAG}; #Gln codons
$orfs_aa{Asn} = $orfs{AAU}+$orfs{AAC}; #Asn codons
$orfs_aa{Lys} = $orfs{AAA}+$orfs{AAG}; #Lys codons
$orfs_aa{Asp} = $orfs{GAU}+$orfs{GAC}; #Asp codons
$orfs_aa{Glu} = $orfs{GAA}+$orfs{GAG}; #Glu codons
$orfs_aa{Cys} = $orfs{UGU}+$orfs{UGC}; #Cys codons
$orfs_aa{Trp} = $orfs{UGG}; #Trp codon
$orfs_aa{Arg} = $orfs{CGU}+$orfs{CGC}+$orfs{CGA}+$orfs{CGG}+$orfs{AGA}+$orfs{AGG}; #Arg codon
$orfs_aa{Gly} = $orfs{GGU}+$orfs{GGC}+$orfs{GCA}+$orfs{GGG}; #Gly codon
#amino acids frequency calculation
$sum_aa= 0;
foreach $key (keys %orfs_aa) { $sum_aa= $sum_aa + $orfs_aa{$key} }; # amino acids counts normalization
foreach $key (keys %orfs_aa) { $orfs_aa{$key}= $orfs_aa{$key}/$sum_aa };
$orfs_aa{tot}= $sum_aa;
return(%orfs_aa);
}

sub calculate_BgX { #Codon usage difference calculation B(g|C) or B(g|RP) or B(g|TF) or B(g|CH)
my %X = @_;
my $BgC, $sum_codons_diff;
$BgC = 0;
foreach $aa (keys %codons) { #for each amino acid
    $sum_codons_diff = 0;
    for $i (0..#{%codons}{$aa}) { #for each codon coding for a single amino acid
        $sum_codons_diff = $sum_codons_diff + abs($ORFn{$orf_nb}{$codons{$aa}{$i}} - ($X{%codons}{$aa}{$i}}));
    }
    $BgC = $BgC + $ORFp{$orf_nb}{$aa}*$sum_codons_diff;
}
return ($BgC);
}

sub calculate_EXg { #expression prediction calculation EXg (relative to x). For ERPg, ETFg and ECHg
my ($BgC, $BgX) = @_;
my $EXg;
$EXg= $BgC/$BgX;
return ($EXg);
}

sub calculate_Eg { #general expression prediction calculation Eg
my ($BgC, $BgRP, $BgTF, $BgCH) = @_;
$Eg= $BgC / (0.5*$BgRP + 0.25*$BgTF + 0.25*$BgCH);
return ($Eg);
}

sub calculate_M { #calculates M: the median value of BgC for all genes
my(@values) = sort(@_);
my $median,$array_size;
print (STDERR "\n***WATCH BgC values: $values[0] $values[1] $values[2] $values[3] $values[4]...last: $values[ $#values]
\n");
$array_size= $#values + 1;
print (STDERR "\n***WATCH array size = $array_size\n");
if ($array_size%2 != 0) {
    $array_middle= $array_size/2 - 0.5;
    $median= $values[$array_middle];
}
else {
    $array_middle= $array_size/2;
    $median= ($values[$array_middle-1] + $values[$array_middle])/2;
}
print (STDERR "\n***WATCH median = $median\n");
return($median);
}

sub define_PHX_PA { #Evaluation of PHX and PA
my ($BgRP, $BgTF, $BgCH, $BgC,$ERPg, $ETFg, $ECHg, $Eg, $M) = @_;
if (((($ERPg > 1.05) && ($ETFg > 1.05)) || (($ERPg > 1.05) && ($ECHg > 1.05)) || (($ETFg > 1.05) && ($ECHg > 1.05)))
&& ($Eg >= 1.005)) {
    return("PHX"); #This orf is Predicted Highly eXpressed (PHX)
}
if ( ( ($BgRP > ($M + 0.15)) && ($BgTF > ($M + 0.15)) && ($BgCH > ($M + 0.15)) && ($BgC > ($M + 0.12)) ) ) {
    return("PA"); #This orf is predicted Putative Alien (PA)
}
}
```

## Annex 1 (continued):

```
return("-"); #This orf is not PHX and not PA - a "normal" orf.
}

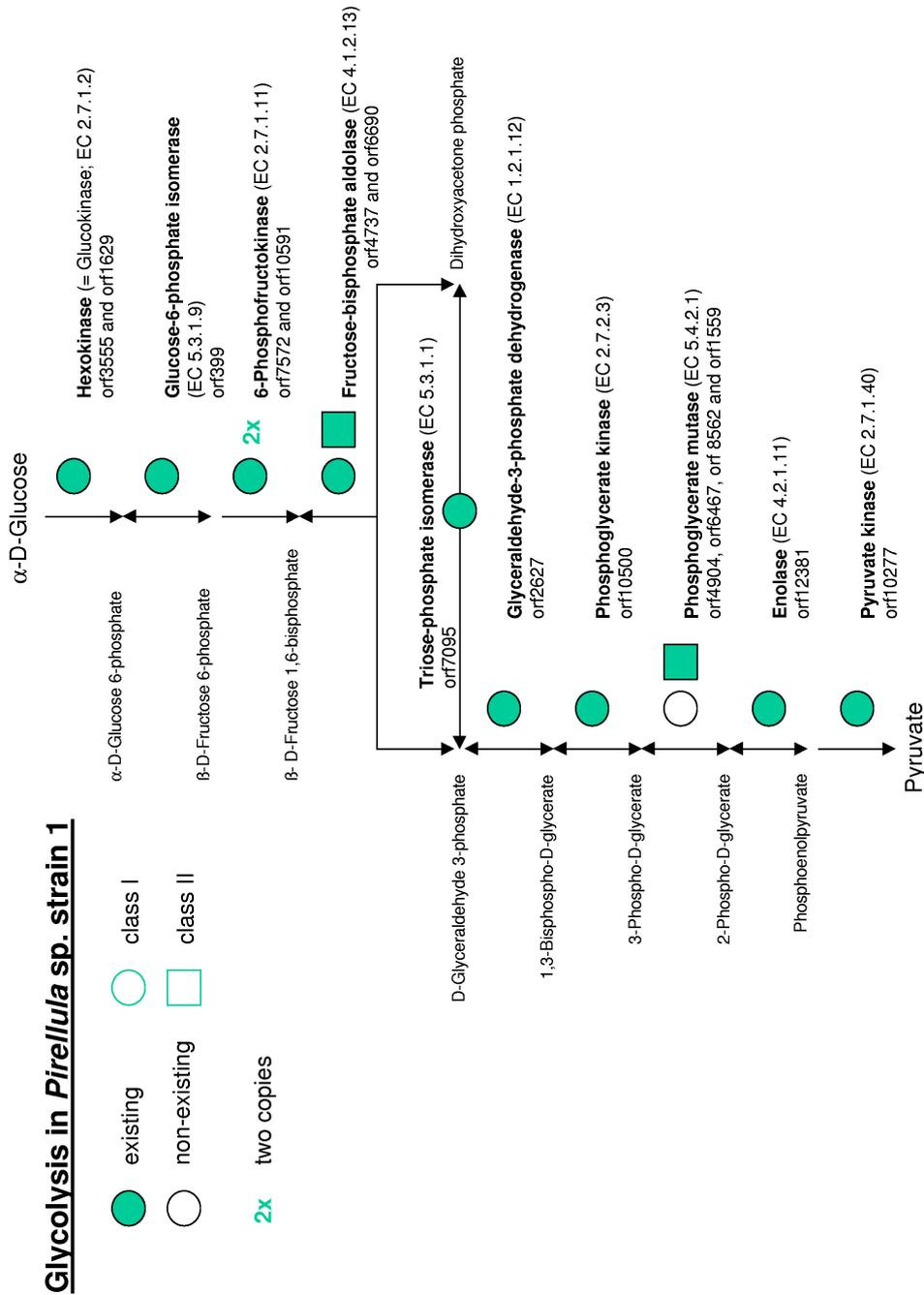
sub print_output_title { #writing output to STDOUT
# my $file = @_;
$title_column1= "orf name";
$title_column2= "leng.[aa]";
$title_column3= "B(g|C)";
$title_column4= "B(g|RP)";
$title_column5= "B(g|TF)";
$title_column6= "B(g|CH)";
$title_column7= "ERP(g)";
$title_column8= "ETP(g)";
$title_column9= "ECH(g)";
$title_column10= "E(g)";
$title_column11= "pred.";
$title_column12= "M";
printf(OUT "%-20s%10s%8s%8s%8s%8s%8s%8s%8s%8s%8s\n", $title_column1, $title_column2, $title_column3,
$title_column4, $title_column5, $title_column6, $title_column7, $title_column8, $title_column9, $title_column10,
$title_column11, $title_column12);
}

sub print_output_data { #writing output to STDOUT
# my $file = @_;
printf(OUT "%-20s%10d%8.2f%8.2f%8.2f%8.2f%8.2f%8.2f%8.2f%8s%8.2f\n", $ORF[$orf_nb]{id}, $ORFp[$orf_nb]{tot},
$BgC[$orf_nb], $BgRP[$orf_nb], $BgTF[$orf_nb], $BgCH[$orf_nb], $ERPg[$orf_nb], $ETPg[$orf_nb], $ECHg[$orf_nb], $Eg
[$orf_nb], $exp_pred[$orf_nb], $M);
}
```



### Annex 3:

Central metabolism in *Pirellula* sp. strain 1: glycolysis, pentose phosphate pathway and citrate cycle.



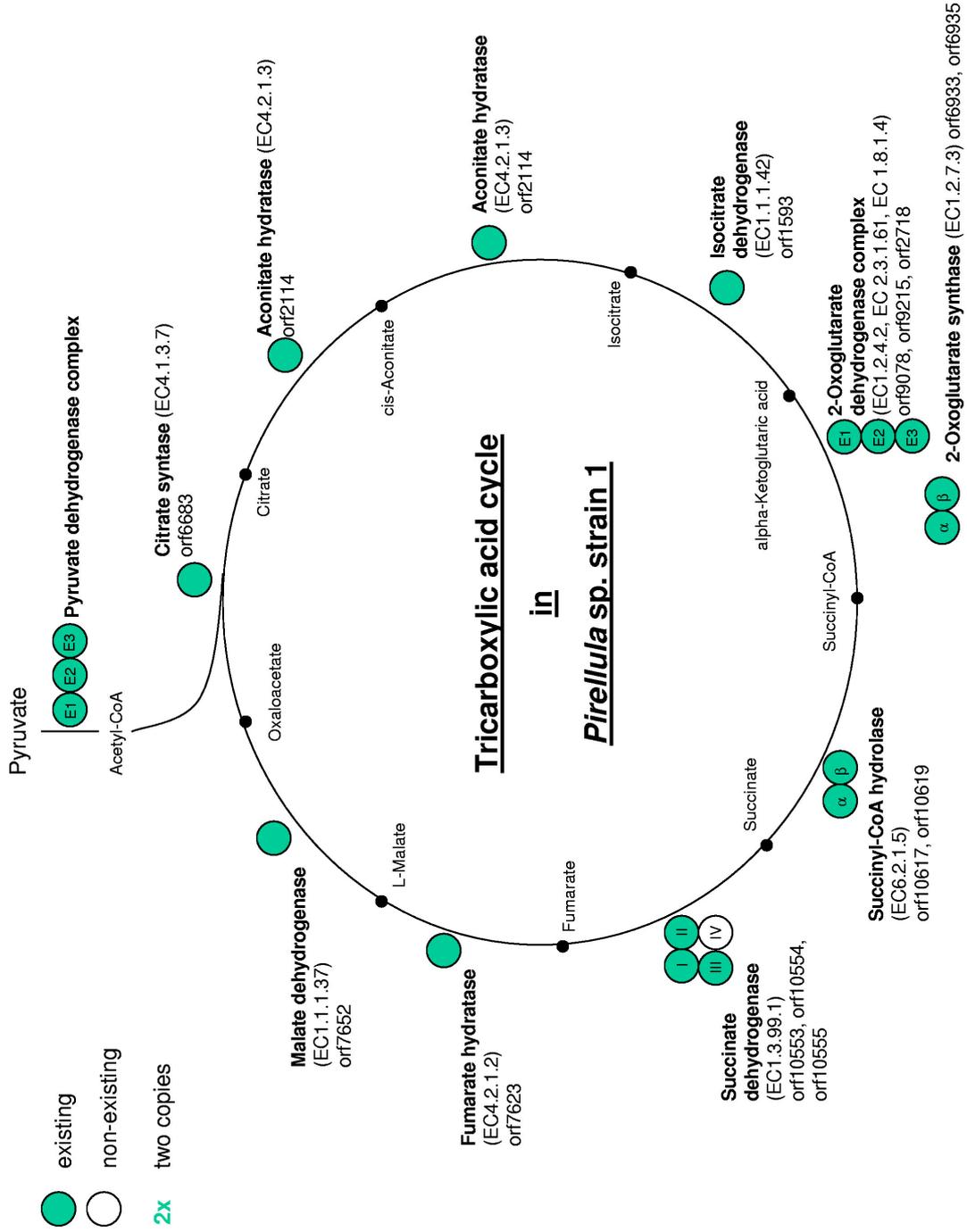
**Pentose phosphate pathway**  
**in *Pirellula* sp. strain 1**

- existing
- 2x two copies
- non-existing



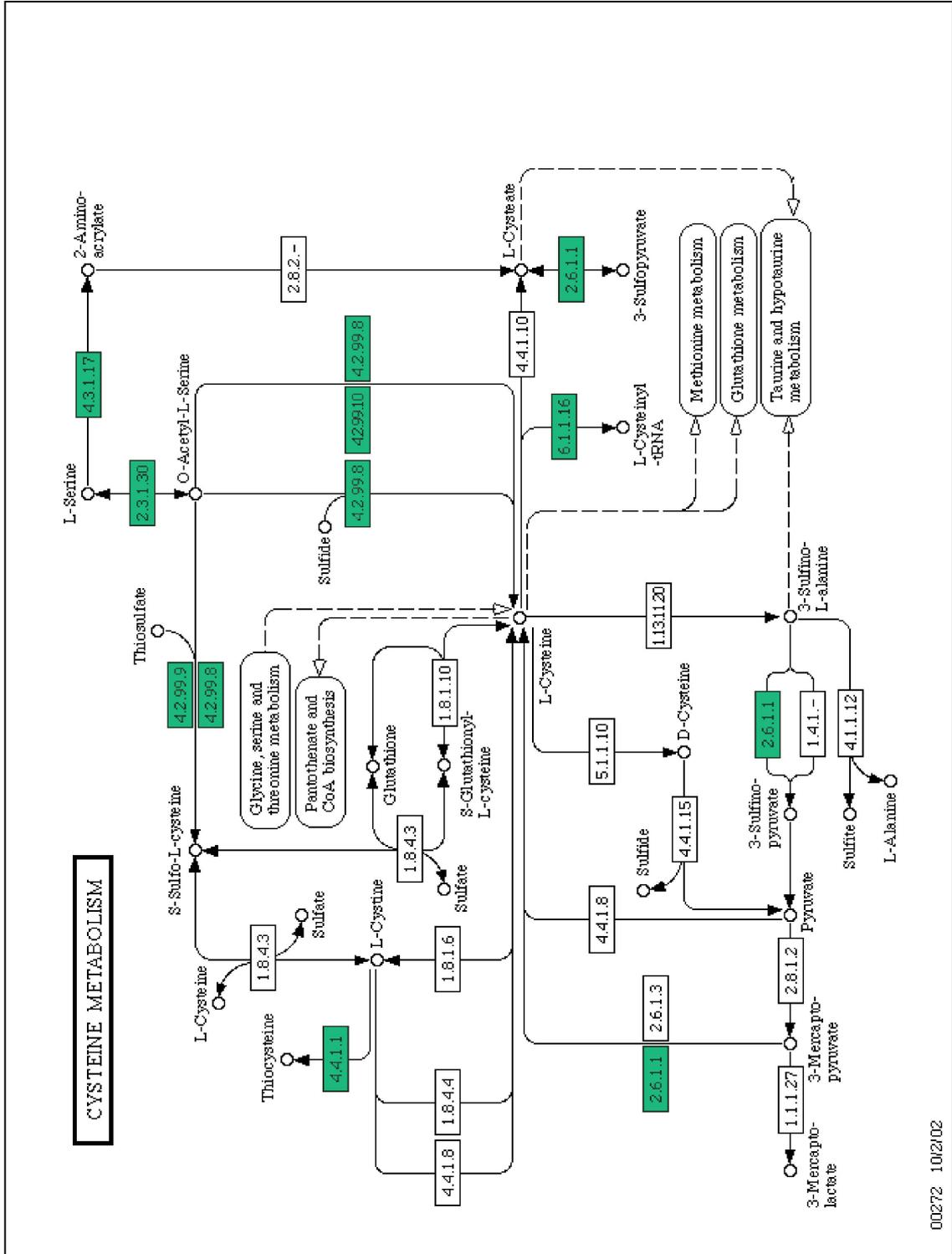
**Annex 3 (continued):**

**Annex 3 (continued):**



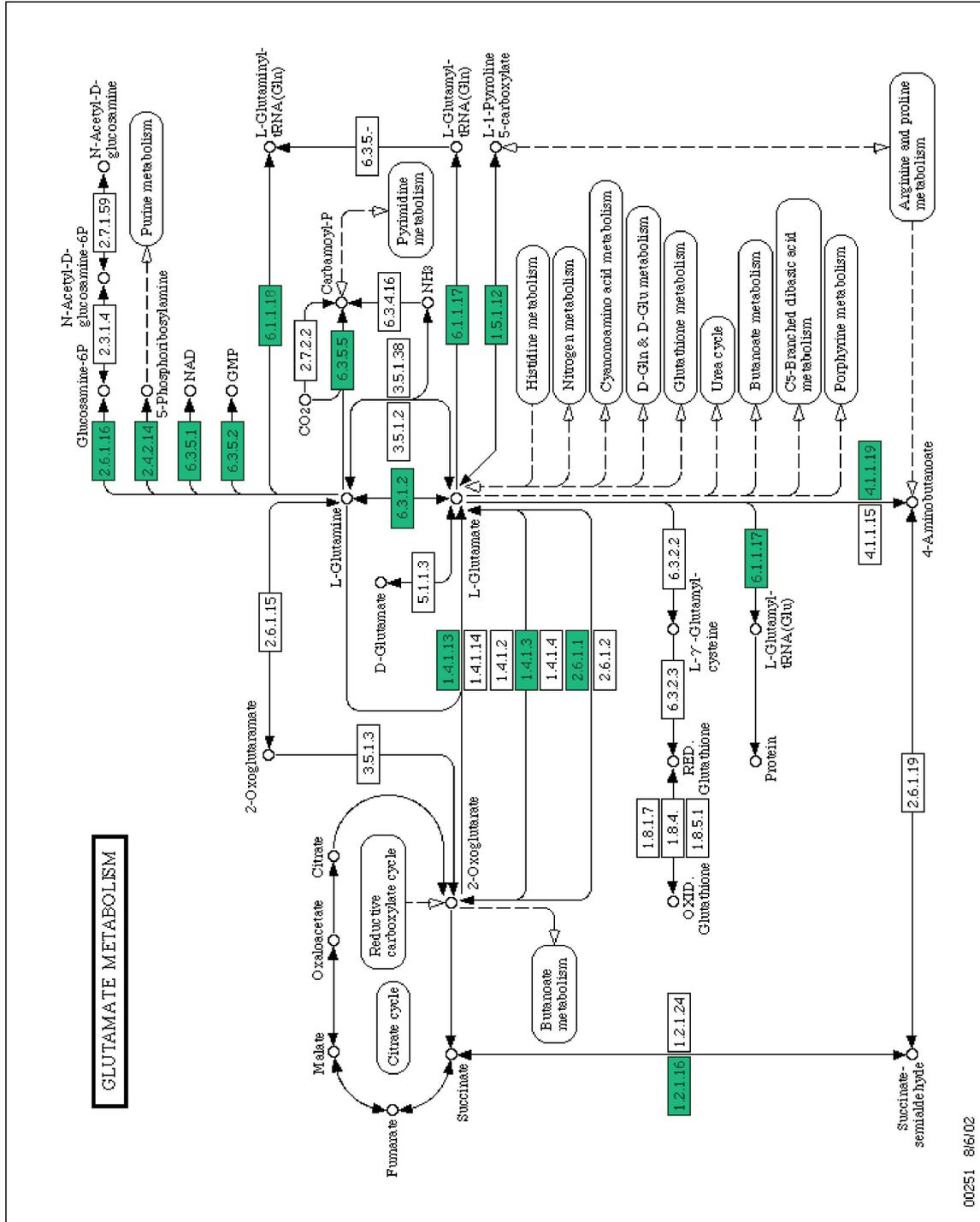


**Annex 4 (continued):**



00272 10/2/02

**Annex 4 (continued):**



00251 8/6/02



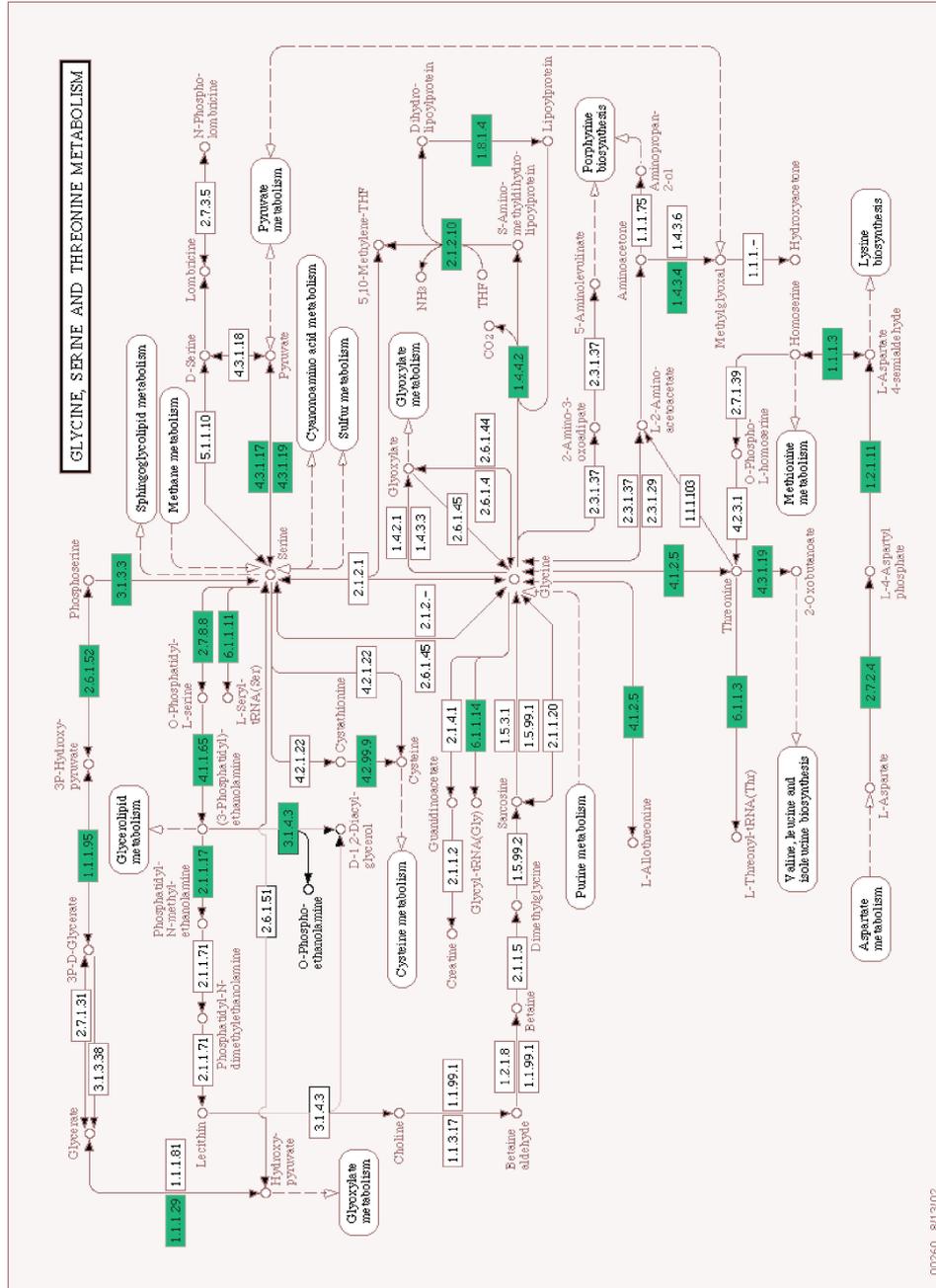








**Annex 4 (continued):**



00260 8/13/02



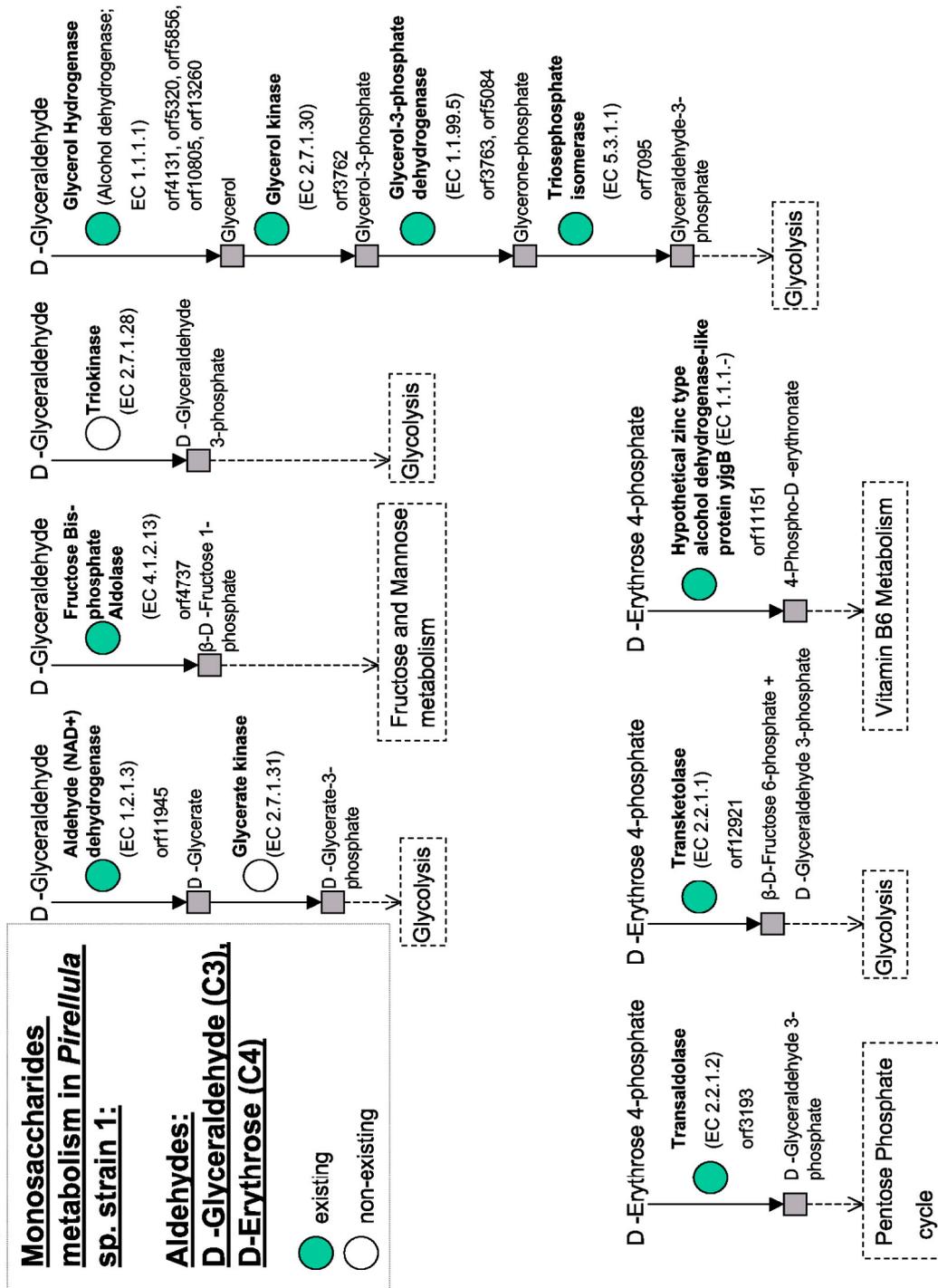
### Annex 5:

Carbohydrate metabolism in *Pirellula* sp. strain 1. Experimental evidence for substrate on which growth was observed were kindly provided by H. Schlesner and D. Gaude (+: present; -: not present; ?: unclear).

Monosaccharides:	experimental evidence	bioinformatic evidence	
<b>Aldehydes:</b>		Import via PTS	metabolism
D-Glyceraldehyde (C3)			+
D-Erythrose (C4)			+
D-Threose (C4)			?
D-Ribose (C5)	+		+
D-Arabinose (C5)			-
D-Xylose (C5)	+		+
D-Lyxose (C5)	+		+
D-Allose (C6)			?
D-Altrose (C6)			?
D-Glucose (C6)	+		+
D-Mannose (C6)	+		+
D-Gulose (C6)			?
D-Idose (C6)			?
D-Galactose (C6)	+		+
D-Talose (C6)			?
Monosaccharides:	experimental evidence	bioinformatic evidence	
<b>Ketones:</b>		Import via PTS	metabolism
Dihydroxyacetone (C3)			+
D-Erythrulose (C4)			?
D-Ribulose (C5)			+
D-Xylulose (C5)			+
D-Psicose (C6)			?
D-Fructose (C6)	+	+	+
D-Sorbose (C6)	-		?
D-Tagatose (C6)			-
Disaccharides:	experimental evidence	bioinformatic evidence	
		Import via PTS	metabolism
Sucrose	+		+
Lactose	+		+
Maltose	+		+
Isomaltose			+
Cellobiose	+		?
Polysaccharides:	experimental evidence	bioinformatic evidence	
		Import via PTS	metabolism
Cellulose	-		+
Starch	+		+
Chitin	-		?
Chondroitine sulfate	+		+
Others:	experimental evidence	bioinformatic evidence	
		Import via PTS	metabolism
D-Fucose	-		
Melibiose	+		
Dextrin			+
α-melezitose	+		
Raffinose	+		
L-Rhamnose	+		
Trehalose	+		?
N-acetylglucosamine	+		+

**Annex 5 (continued):**

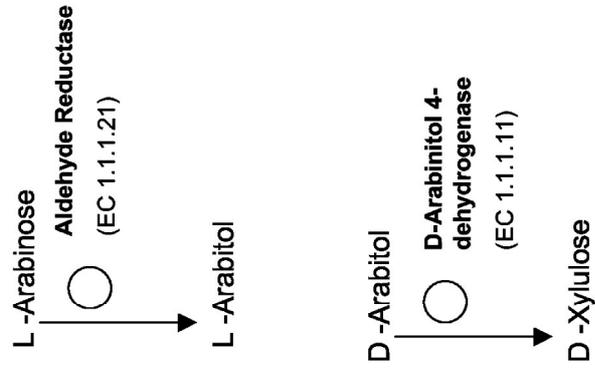
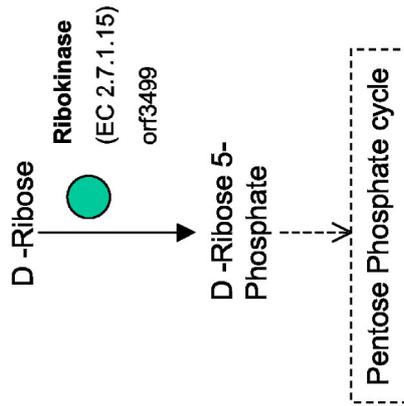
Carbohydrate metabolism in *Pirellula* sp. strain 1: Pathway maps.



**Monosaccharides**  
**metabolism in *Pirellula***  
**sp. strain 1:**

**Aldehydes:**  
**D -Ribose (C5),**  
**Arabinose (C5)**

-  existing
-  non-existing

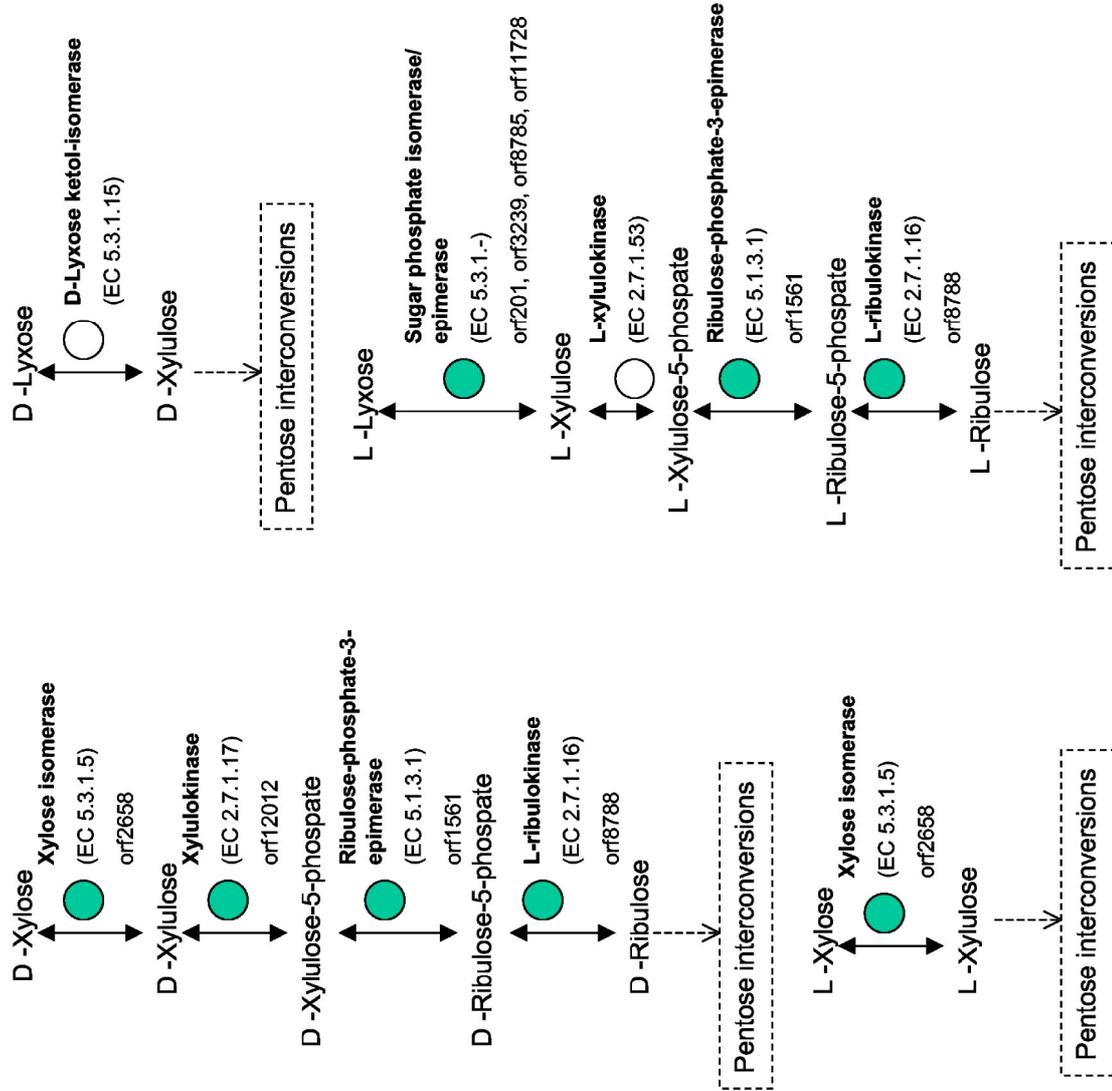


**Annex 5 (continued):**

**Monosaccharides**  
**metabolism in *Pirellula***  
**sp. strain 1:**

**Aldehydes:**  
**Xylose (C5),**  
**Lyxose (C5)**

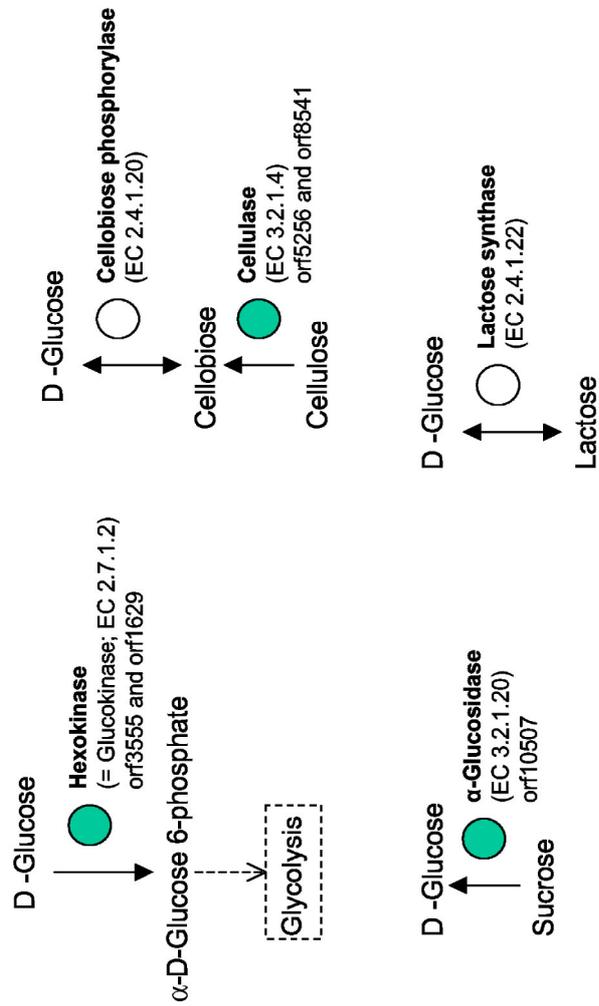
- existing
- non-existing



**Annex 5 (continued):**

**Monosaccharides**  
**metabolism in *Pirellula***  
**sp. strain 1:**  
**Aldehydes: Glucose (C6)**

-  existing
-  non-existing

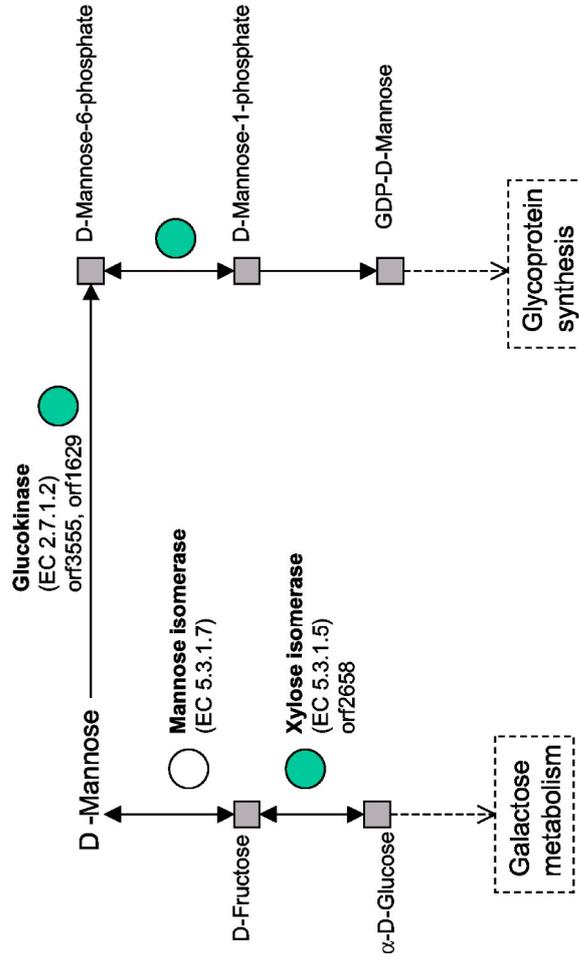


**Annex 5 (continued):**

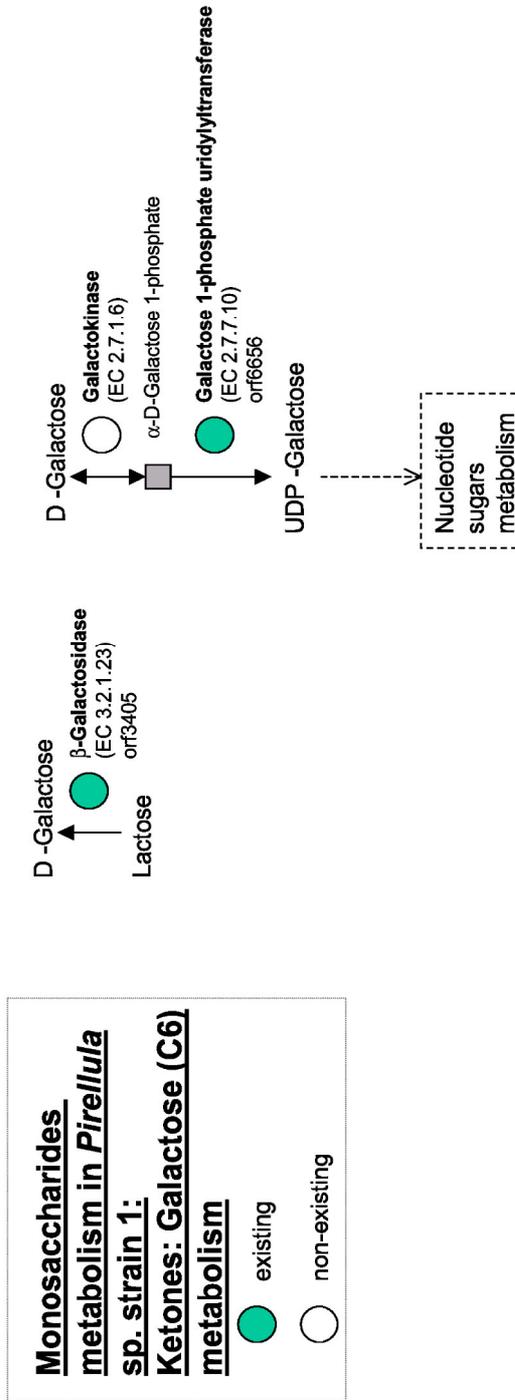
**Annex 5 (continued):**

**Monosaccharides**  
**metabolism in *Pirellula***  
**sp. strain 1:**  
**Aldehydes: Mannose (C6)**  
**metabolism**

● existing  
○ non-existing



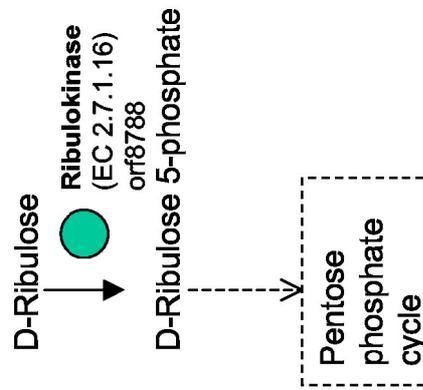
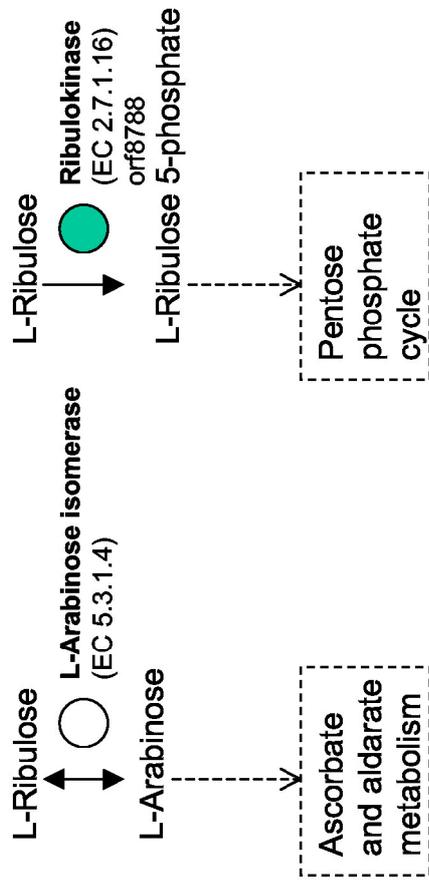
**Annex 5 (continued):**



**Monosaccharides**  
**metabolism in *Pirellula*:**  
**Ketones: Ribulose (C5)**

● existing

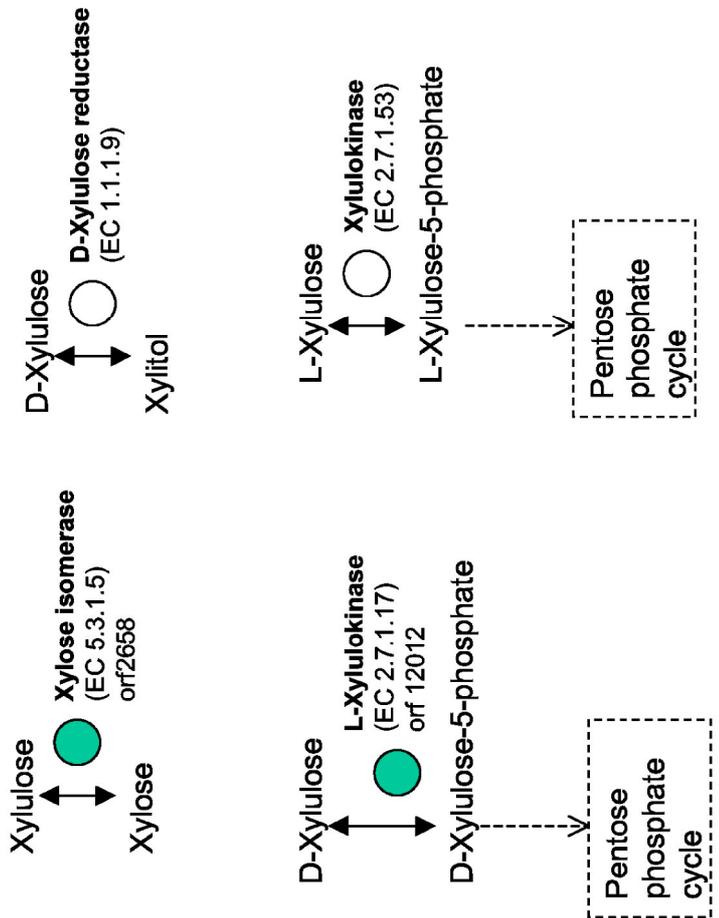
○ non-existing



**Annex 5 (continued):**

**Monosaccharides**  
**metabolism in *Pirellula*:**  
**Ketones: Xylulose (C5)**

● existing  
○ non-existing

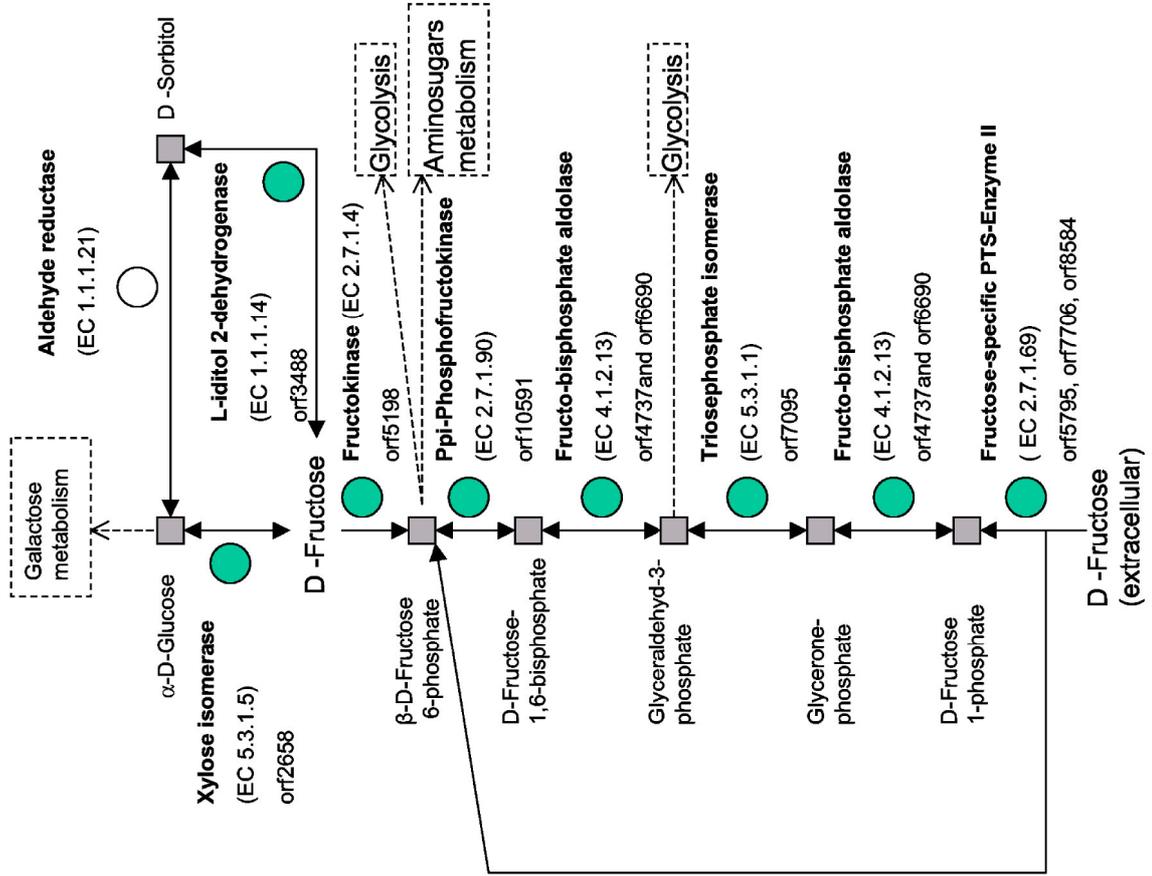


**Annex 5 (continued):**

**Monosaccharides metabolism in *Pirellula*: Ketones: Fructose (C6) metabolism**

● existing  
○ non-existing

**Annex 5 (continued):**



**Monosaccharides  
import in *Pirellula***

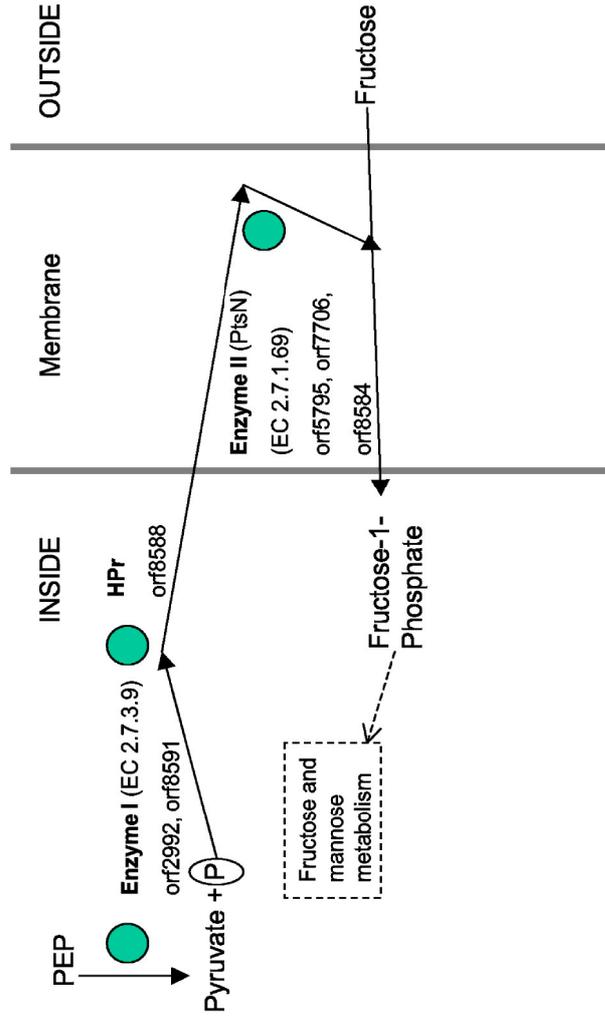
- existing
- non-existing

Note:

The KEGG Pathways show that Protein-N(pi)-phosphotransferase (= PTS-permease or Enzyme II; EC 2.7.1.69) is responsible for the import of extracellular substrates like D-Glucose, Arbutin, Salicin, N-Acetyl-D -Glucosamine, D-Glucosamine, L-Sorbose, D-Mannitol, D-Fructose, D-Mannose.

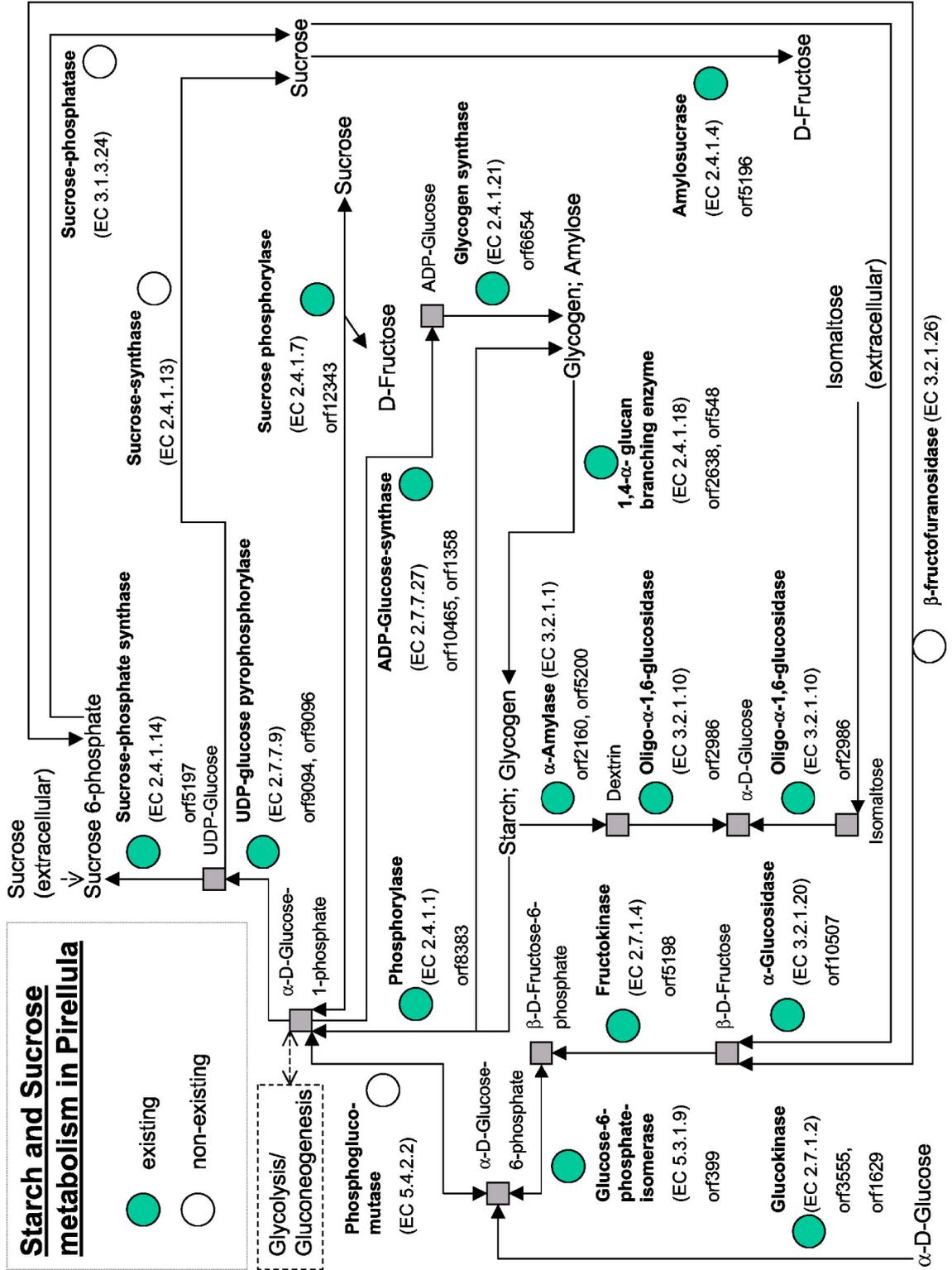
There are three ORFs (orf5795, orf7706, orf8584) that are identified as components of this enzyme, but they are only fructose-specific. Fructose is imported by this PTS and other sugars could be imported by perhaps a symporter.

*Pirellula* contains a fructose-specific phosphotransferase system:

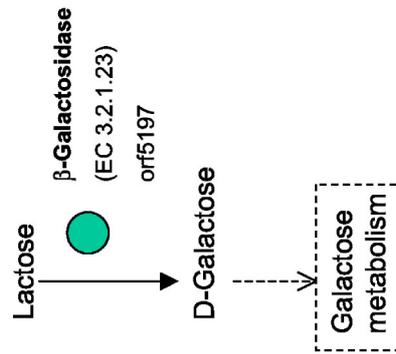
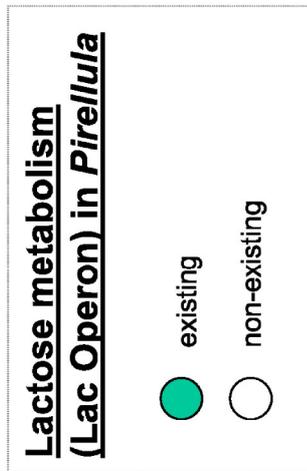


**Annex 5 (continued):**

**Annex 5 (continued):**



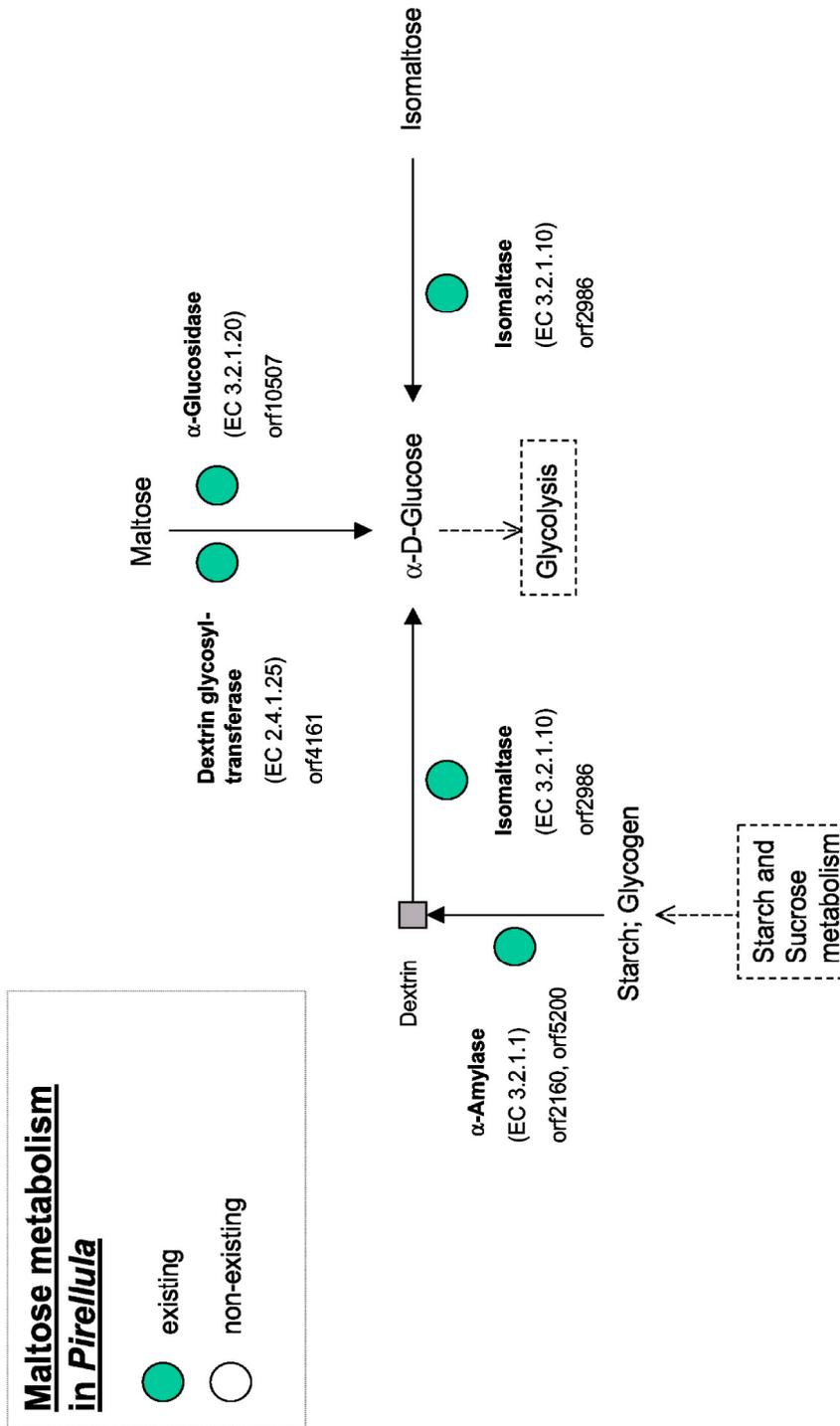
**Annex 5 (continued):**



**Note:** The Lactose Operon is normally consisting of three genes *lacZ* ( $\beta$ -Galactosidase ), *lacY* (lactose permease) and *lacA* (thiogalactosidase transacetylase).

*Pirellula* has some orfs that are corresponding to this enzymes:  $\beta$ -Galactosidase (EC 3.2.1.23; orf5197), lactose permease (possibly orf2209 and orf12406) and thiogalactosidase transacetylase (EC 2.3.1.18; orf2470 (, orf1235)).

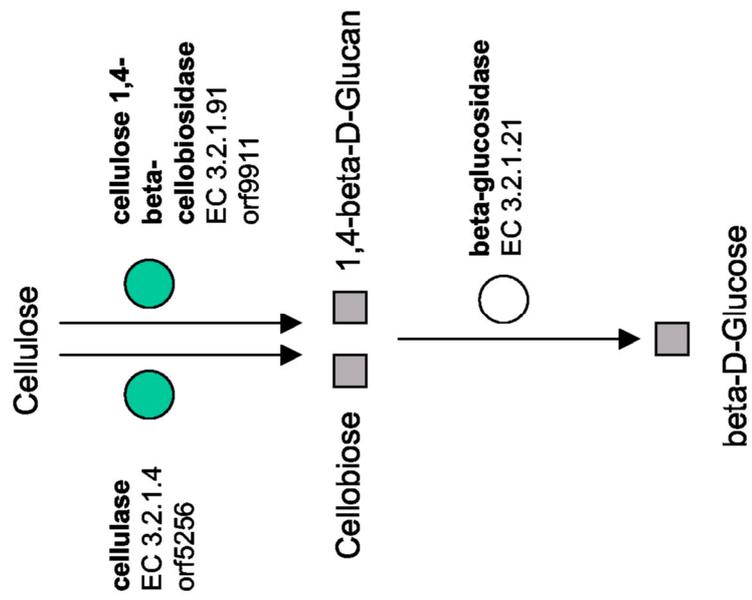
**Annex 5 (continued):**



**Cellulose metabolism**  
**in *Pirellula***

● existing

○ non-existing



**Annex 5 (continued):**



## 6. References

A1. Glöckner FO, Kube M, Bauer M, Teeling H, Lombardot T, Ludwig W, Gade D, Beck A, Borzým K, Heitmann K, Rabus R, Schlesner H, Amann R, Reinhardt R. Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1. *Proc Natl Acad Sci U S A*. 2003 Jul 8; 100(14): 8298-303.

A2. Bauer M, Lombardot T, Teeling H, Amann R, Ward NL, Glöckner FO. Phylogenetic implications of the unexpected presence of archaeal-like genes for C1 transfer enzymes in planctomycetes. *under revision*.

A3. Lombardot T, Bauer M, Teeling H, Amann R, Glöckner FO. Environmental interpretation of the transcriptional regulators pool of the marine bacterium *Pirellula* sp. strain 1 based on whole genome comparisons. *submitted*.

A4. Teeling H, Lombardot T, Bauer M, Ludwig W, Glöckner FO. Reevaluation of the phylogenetic position of the Planctomycetes by means of concatenated ribosomal proteins, mRNA polymerase subunits and whole genome trees. *accepted for publication*.

#####

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995 Jul 28; 269(5223): 496-512.

2. Nowak R. Bacterial genome sequence bagged. *Science*. 1995 Jul 28; 269(5223): 468-70.

3. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. The complete genome sequence of *Escherichia coli* K-12. *Science*. 1997 Sep 5; 277(5331): 1453-74.

4. Janssen P, Audit B, Cases I, Darzentas N, Goldovsky L, Kunin V, Lopez-Bigas N, Peregrin-Alvarez JM, Pereira-Leal JB, Tsoka S, Ouzounis CA. Beyond 100 genomes. *Genome Biol*. 2003; 4(5): 402.

5. Bernal A, Ear U, Kyrpides N. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res*. 2001 Jan 1; 29(1): 126-7.

6. Fraser CM, Eisen JA, Nelson KE, Paulsen IT, Salzberg SL. The value of complete microbial genome sequencing (you get what you pay for). *J Bacteriol.* 2002 Dec; 184(23): 6403-5.
7. Branscomb E, Predki P. On the high value of low standards. *J Bacteriol.* 2002 Dec; 184(23): 6406-9.
8. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 2003 Jan 1; 31(1): 365-70.
9. Moore GE. Cramming more components onto integrated circuits. *Electronics.* 1965 Apr; 38(8).
10. Lundstrom M. Applied physics. Moore's law forever? *Science.* 2003 Jan 10; 299(5604): 210-1.
11. Meindl JD, Chen Q, Davis JA. Limits on silicon nanoelectronics for terascale integration. *Science.* 2001 Sep 14; 293(5537): 2044-9.
12. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 2004 Jan 1;32:D115-9.
13. Schlesner H. The development of media suitable for the microorganisms morphologically resembling *Planctomyces* spp., *Pirellula* spp., and other *Planctomycetales* from various aquatic habitats using dilute media. *System Appl Microbiol.* 1994; 17:135-145.
14. Wang J, Jenkins C, Webb RI, Fuerst JA. Isolation of *Gemmata*-like and *Isosphaera*-like planctomycete bacteria from soil and freshwater. *Appl Environ Microbiol.* 2002 Jan;68(1):417-22.
15. Fuerst JA. The planctomycetes: emerging models for microbial ecology, evolution and cell biology. *Microbiology.* 1995 Jul;141:1493-1506.
16. Amann R, Ludwig W. Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. *FEMS Microbiol Rev.* 2000 Dec;24(5):555-65.

17. Chatzinotas A, Sandaa RA, Schonhuber W, Amann R, Daae FL, Torsvik V, Zeyer J, Hahn D. Analysis of broad-scale differences in microbial community composition of two pristine forest soils. *Syst Appl Microbiol*. 1998 Dec;21(4):579-87.
18. Neef A, Amann R, Schlesner H, Schleifer KH. Monitoring a widespread bacterial group: *in situ* detection of planctomycetes with 16S rRNA-targeted probes. *Microbiology*. 1998 Dec;144:3257-66.
19. Bockelmann U, Manz W, Neu TR, Szewzyk U. Characterization of the microbial community of lotic organic aggregates ('river snow') in the Elbe River of Germany by cultivation and molecular methods. *FEMS Microbiol Ecol*. 2000 Aug 1;33(2):157-170.
20. Rath J, Wu KY, Herndl GJ, DeLong EF. High phylogenetic diversity in a marine-snow-associated bacterial assemblage. *Aquat Microb Ecol*. 1998 Apr;14 (3):261-269.
21. Strous M, Fuerst JA, Kramer EH, Logemann S, Muyzer G, van de Pas-Schoonen KT, Webb R, Kuenen JG, Jetten MS. Missing lithotroph identified as new planctomycete. *Nature*. 1999 Jul 29;400(6743):446-9.
22. Llobet-Brossa E, Rossello-Mora R, Amann R. Microbial Community Composition of Wadden Sea Sediments as Revealed by Fluorescence In Situ Hybridization. *Appl Environ Microbiol*. 1998 Jul 1;64(7):2691-6.
23. Jetten MS, Sliemers O, Kuypers M, Dalsgaard T, van Niftrik L, Cirpus I, van de Pas-Schoonen K, Lavik G, Thamdrup B, Le Paslier D, Op den Camp HJ, Hulth S, Nielsen LP, Abma W, Third K, Engstrom P, Kuenen JG, Jorgensen BB, Canfield DE, Sinninghe Damste JS, Revsbech NP, Fuerst J, Weissenbach J, Wagner M, Schmidt I, Schmid M, Strous M. Anaerobic ammonium oxidation by marine and freshwater planctomycete-like bacteria. *Appl Microbiol Biotechnol*. 2003 Dec;63(2):107-14.
24. Kuypers MM, Sliemers AO, Lavik G, Schmid M, Jorgensen BB, Kuenen JG, Sinninghe Damste JS, Strous M, Jetten MS. Anaerobic ammonium oxidation by anammox bacteria in the Black Sea. *Nature*. 2003 Apr 10;422(6932):608-11.
25. Lindsay MR, Webb RI, Strous M, Jetten MS, Butler MK, Forde RJ, Fuerst JA. Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell. *Arch Microbiol*. 2001 Jun;175(6):413-29.
26. Lindsay MR, Webb RI, Fuerst JA. Pirellosomes: a new type of membrane-bounded cell compartment in planctomycete bacteria of the genus *Pirellula*. *Microbiol-UK*. 1997;143:739-748.

27. Butler MK, Wang J, Webb RI, Fuerst JA. Molecular and ultrastructural confirmation of classification of ATCC 35122 as a strain of *Pirellula staleyi*. *Int J Syst Evol Microbiol*. 2002 Sep;52:1663-7.
28. Gripenburg U, Ward-Rainey N, Mohamed S, Schlesner H, Marxsen H, Rainey FA, Stackebrandt E, Auling G. Phylogenetic diversity, polyamine pattern and DNA base composition of members of the order *Planctomycetales*. *Int J Syst Bacteriol*. 1999 Apr;49 Pt 2:689-96.
29. Ward N, Rainey FA, Stackebrandt E, Schlesner H. Unraveling the extent of diversity within the order *Planctomycetales*. *Appl Environ Microbiol*. 1995 Jun;61(6):2270-5.
30. Jetten MS, Wagner M, Fuerst J, van Loosdrecht M, Kuenen G, Strous M. Microbiology and application of the anaerobic ammonium oxidation ('anammox') process. *Curr Opin Biotechnol*. 2001 Jun;12(3):283-8.
31. Liesack W, Stackebrandt E. Occurrence of novel groups of the domain Bacteria as revealed by analysis of genetic material isolated from an Australian terrestrial environment. *J Bacteriol*. 1992 Aug; 174(15): 5072-8.
32. Ward-Rainey N, Rainey FA, Schlesner H, Stackebrandt E. Assignment of hitherto unidentified 16S rDNA species to a main line of descent within the domain *Bacteria*. *Microbiology* 1995 Dec;141:3247-3250.
33. Ward NL, Rainey FA, Hedlund BP, Staley JT, Ludwig W, Stackebrandt E. Comparative phylogenetic analyses of members of the order *Planctomycetales* and the division *Verrucomicrobia*: 23S rRNA gene sequence analysis supports the 16S rRNA gene sequence-derived phylogeny. *Int J Syst Evol Microbiol*. 2000 Nov;50:1965-72.
34. Jenkins C, Fuerst JA. Phylogenetic analysis of evolutionary relationships of the planctomycete division of the domain bacteria based on amino acid sequences of elongation factor Tu. *J Mol Evol*. 2001 May;52(5):405-18.
35. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. Universal trees based on large combined protein sequence data sets. *Nat Genet*. 2001 Jul;28(3):281-5.
36. Brochier C, Philippe H. Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature*. 2002 May 16;417(6886):244.
37. Gribaldo S, Philippe H. Ancient phylogenetic relationships. *Theor Popul Biol*. 2002 Jun;61(4):391-408.

38. Di Giulio M. The ancestor of the Bacteria domain was a hyperthermophile. *J Theor Biol.* 2003 Oct 7;224(3):277-83.
39. Schlesner H, Rathmann M, Bartels C, Tindall B, Gade D, Rabus R, Pfeiffer S, Hirsch P. Taxonomic heterogeneity within the *Planctomycetales* as derived by DNA/DNA-hybridization, description of *Rhodopirellula baltica* gen. nov., sp. nov. and transfer of *Pirellula marina* to the genus *Blastopirellula* gen. nov. as *Blastopirellula marina* com. nov. *submitted*.
40. Rabus R, Gade D, Helbig R, Bauer M, Glockner FO, Kube M, Schlesner H, Reinhardt R, Amann R. Analysis of N-acetylglucosamine metabolism in the marine bacterium *Pirellula* sp. strain 1 by a proteomic approach. *Proteomics.* 2002 Jun;2(6):649-55.
41. Kube M. Sequenzierung und Strukturen von *Pirellula* sp. Stamm 1, PhD thesis 2003 (german), MPI für molekulare Genetik, Berlin.
42. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 1998 Jan 15; 26(2): 544-8.
43. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 1999 Dec 1; 27(23): 4636-41.
44. Aggarwal G, Ramaswamy R. Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J Biosci.* 2002 Feb; 27: 7-14.
45. Badger JH, Olsen GJ. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol.* 1999 Apr; 16(4): 512-24.
46. Frishman D, Mironov A, Mewes HW, Gelfand M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.* 1998 Jun 15; 26(12): 2941-7.
47. Frishman D, Mironov A, Gelfand M. Starts of bacterial genes: estimating the reliability of computer predictions. *Gene.* 1999 Jul 8; 234(2): 257-65.
48. Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A, Mewes HW. Functional and structural genomics using PEDANT. *Bioinformatics.* 2001 Jan;17(1):44-57.
49. Frishman D, Mokrejs coM, Kosykh D, Kastenmuller G, Kolesov G, Zubrzycki I, Gruber C, Geier B, Kaps A, Albermann K, Volz A, Wagner C, Fellenberg M, Heumann K, Mewes HW. The PEDANT genome database. *Nucleic Acids Res.* 2003 Jan 1;31(1):

207-11.

50. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. Artemis: sequence visualization and annotation. *Bioinformatics*. 2000 Oct;16(10):944-5.

51. Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, Puhler A. GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res*. 2003 Apr 15;31(8):2187-95.

52. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997 Sep 1;25(17):3389-402.

53. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. *Nucleic Acids Res*. 2002 Jan 1;30(1):276-80.

54. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*. 2003 Jul 1;31(13):3497-500.

55. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. 2000 Sep 8;302(1):205-17.

56. Eddy SR. Profile hidden Markov models. *Bioinformatics*. 1998;14(9):755-63.

57. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*. 1994 Feb 4;235(5):1501-31.

58. Kanehisa M. Post-Genome Informatics. Oxford University Press, 1st edition.2000 Mar 15; pp. 95-115.

59. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press. 1999 Jul 1; pp. 46-60.

60. Eddy SR, Birney E. HMMER: sequence analysis using profile hidden Markov models. <http://hmmer.wustl.edu>

61. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A,

- Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000 May; 25(1): 25-9.
62. Sharp PM, Li WH. The Codon Adaptation Index -- a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 1987 Feb 11;15(3):1281-95.
63. Peden J. CodonW. <http://www.molbiol.ox.ac.uk/cu/codonW.html>
64. Karlin S, Mrazek J. Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol.* 2000 Sep;182(18):5238-50.
65. Stoesser G, Baker W, van den Broek A, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, Nardone F, Stoehr P, Tuli MA, Tzouvara K, Vaughan R. The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res.* 2003 Jan 1;31(1):17-22.
66. Clarke GD, Beiko RG, Ragan MA, Charlebois RL. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J Bacteriol.* 2002 Apr; 184(8): 2072-80.
67. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A.* 1992 Nov 15; 89(22): 10915-9.
68. The Apache Software Foundation. <http://www.apache.org>
69. PostgreSQL, open source database software. <http://www.postgresql.org>
70. PostGIS: support for geographic objects to the PostgreSQL object-relational database. <http://postgis.refrains.net>
71. The Open GIS Consortium (OGC): a member-driven, non-profit international trade association that is leading the development of geoprocessing interoperability computing standards. <http://www.opengis.org>
72. University of Minnesota MapServer CGI package. <http://mapserver.gis.umn.edu>
73. GEO Data Portal - Online Environmental Database of UNEP.Net (United Nations Environment Network). <http://geodata.grid.unep.ch>
74. Berriman M, Rutherford K. Viewing and annotating sequence data with Artemis.

*Brief Bioinform.* 2003 Jun;4(2):124-32.

75. MySQL: The World's Most Popular Open Source Database. <http://www.mysql.com>

76. Guo FB, Ou HY, Zhang CT. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.* 2003 Mar 15; 31(6): 1780-9.

77. Ou HY, Guo FB, Zhang CT. GS-Finder: a program to find bacterial gene start sites with a self-training method. *Int J Biochem Cell Biol.* 2004 Mar; 36(3): 535-44.

78. Grigoriev A. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* 1998 May 15;26(10):2286-90.

79. Grigoriev A. Genome arithmetic. *Science* 1998 Sep 25;281:1923-1923a.

80. Frank AC, Lobry JR. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene.* 1999 Sep 30;238(1):65-77.

81. Lobry JR, Louarn JM. Polarisation of prokaryotic chromosomes. *Curr Opin Microbiol.* 2003 Apr;6(2):101-8.

82. Grigoriev A. Graphical genome comparison: rearrangements and replication origin of *Helicobacter pylori*. *Trends Genet.* 2000 Sep;16(9):376-8.

83. McLean MJ, Wolfe KH, Devine KM. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol.* 1998 Dec;47(6): 691-6.

84. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 1996 Jun 30;3(3): 109-36.

85. Zhang R, Zhang CT. Multiple replication origins of the archaeon *Halobacterium* species NRC-1. *Biochem Biophys Res Commun.* 2003 Mar 21;302(4):728-34.

86. Bailey CC, Bott KF. An unusual gene containing a dnaJ N-terminal box flanks the putative origin of replication of *Mycoplasma genitalium*. *J Bacteriol.* 1994 Sep;176(18): 5814-9.

87. Eisen JA, Heidelberg JF, White O, Salzberg SL. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* 2000;1(6).
88. Suyama M, Bork P. Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet.* 2001 Jan;17(1):10-3.
89. The Institute for Genome Research: Microbial genomes and chromosomes in progress, status of December 2003. <http://www.tigr.org/tdb/mdb/mdbinprogress.html>
90. Integrated Genomics: GOLD, Genomes Online Database, status of December 2003. <http://wit.integratedgenomics.com/GOLD/>
91. Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, Sasamoto S, Watanabe A, Idesawa K, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M, Matsuno A, Mochizuki Y, Nakayama S, Nakazaki N, Shimpo S, Sugimoto M, Takeuchi C, Yamada M, Tabata S. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Res.* 2000 Dec 31; 7(6): 331-8.
92. Sullivan JT, Trzebiatowski JR, Cruickshank RW, Gouzy J, Brown SD, Elliot RM, Fleetwood DJ, McCallum NG, Rossbach U, Stuart GS, Weaver JE, Webby RJ, De Bruijn FJ, Ronson CW. Comparative sequence analysis of the symbiosis island of *Mesorhizobium loti* strain R7A. *J Bacteriol.* 2002 Jun; 184(11): 3086-95.
93. Sullivan JT, Ronson CW. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc Natl Acad Sci U S A.* 1998 Apr 28; 95(9): 5145-9.
94. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* 2002 Jan 1;30(1):42-6.
95. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 2004 Jan 1;32(1):D277-D280.
96. Boltes I, Czapinska H, Kahnert A, von Bulow R, Dierks T, Schmidt B, von Figura K, Kertesz MA, Uson I. 1.3 A structure of arylsulfatase from *Pseudomonas aeruginosa* establishes the catalytic mechanism of sulfate ester cleavage in the sulfatase family. *Structure.* 2001 Jun;9(6):483-91.
97. Dierks T, Miech C, Hummerjohann J, Schmidt B, Kertesz MA, von Figura K.

Posttranslational formation of formylglycine in prokaryotic sulfatases by modification of either cysteine or serine. *J Biol Chem*. 1998 Oct 2;273(40):25560-4.

98. Kertesz MA. Riding the sulfur cycle--metabolism of sulfonates and sulfate esters in gram-negative bacteria. *FEMS Microbiol Rev*. 2000 Apr;24(2):135-75.

99. Dierks T, Lecca MR, Schlotterhose P, Schmidt B, von Figura K. Sequence determinants directing conversion of cysteine to formylglycine in eukaryotic sulfatases. *EMBO J*. 1999 Apr 15;18(8):2084-91.

100. Marquardt C, Fang Q, Will E, Peng J, von Figura K, Dierks T. Posttranslational modification of serine to formylglycine in bacterial sulfatases. Recognition of the modification motif by the iron-sulfur protein AtsB. *J Biol Chem*. 2003 Jan 24;278(4):2212-8.

101. Szameit C, Miech C, Balleininger M, Schmidt B, von Figura K, Dierks T. The iron sulfur protein AtsB is required for posttranslational formation of formylglycine in the *Klebsiella* sulfatase. *J Biol Chem*. 1999 May 28;274(22):15375-81.

102. Landgrebe J, Dierks T, Schmidt B, von Figura K. The human SUMF1 gene, required for posttranslational sulfatase modification, defines a new gene family which is conserved from pro- to eukaryotes. *Gene*. 2003;316(Oct 16):47-56.

103. Cosma MP, Pepe S, Annunziata I, Newbold RF, Grompe M, Parenti G, Ballabio A. The multiple sulfatase deficiency gene encodes an essential and limiting factor for the activity of sulfatases. *Cell*. 2003 May 16;113(4):445-56.

104. Dierks T, Schmidt B, Borissenko LV, Peng J, Preusser A, Mariappan M, von Figura K. Multiple sulfatase deficiency is caused by mutations in the gene encoding the human C(alpha)-formylglycine generating enzyme. *Cell*. 2003 May 16;113(4):435-44.

105. Morimoto-Tomita M, Uchimura K, Werb Z, Hemmerich S, Rosen SD. Cloning and characterization of two extracellular heparin-degrading endosulfatases in mice and humans. *J Biol Chem*. 2002 Dec 20;277(51):49175-85.

106. Cerbelaud EC, Conway LJ, Galliher PM, Langer RS, Cooney CL. Sulfur regulation of heparinase and sulfatases in *Flavobacterium heparinum*. *Appl Environ Microbiol*. 1986 Mar;51(3):640-6.

107. Kertesz MA, Leisinger T, Cook AM. Proteins induced by sulfate limitation in *Escherichia coli*, *Pseudomonas putida*, or *Staphylococcus aureus*. *J Bacteriol*. 1993 Feb;175(4):1187-90.

108. Hwa V, Salyers AA. Analysis of two chondroitin sulfate utilization mutants of *Bacteroides thetaiotaomicron* that differ in their abilities to compete with the wild type in the gastrointestinal tracts of germfree mice. *Appl Environ Microbiol.* 1992 Mar;58(3): 869-76.
109. Hwa V, Salyers AA. Evidence for differential regulation of genes in the chondroitin sulfate utilization pathway of *Bacteroides thetaiotaomicron*. *J Bacteriol.* 1992 Jan;174 (1):342-4.
110. Raman R, Myette JR, Shriver Z, Pojasek K, Venkataraman G, Sasisekharan R. The heparin/heparan sulfate 2-O-sulfatase from *Flavobacterium heparinum*. A structural and biochemical study of the enzyme active site and saccharide substrate specificity. *J Biol Chem.* 2003 Apr 4;278(14):12167-74.
111. Myette JR, Shriver Z, Claycamp C, McLean MW, Venkataraman G, Sasisekharan R. The heparin/heparan sulfate 2-O-sulfatase from *Flavobacterium heparinum*. Molecular cloning, recombinant expression, and biochemical characterization. *J Biol Chem.* 2003 Apr 4;278(14):12157-66.
112. Barbeyron T, Potin P, Richard C, Collin O, Kloareg B. Arylsulphatase from *Alteromonas carrageenovora*. *Microbiology.* 1995 Nov;141 ( Pt 11):2897-904.
113. Grossart HP, Kiorboe T, Tang K, Ploug H. Bacterial colonization of particles: growth and interactions. *Appl Environ Microbiol.* 2003 Jun;69(6):3500-9.
114. Kiorboe T, Tang K, Grossart HP, Ploug H. Dynamics of microbial communities on marine snow aggregates: colonization, growth, detachment, and grazing mortality of attached bacteria. *Appl Environ Microbiol.* 2003 Jun;69(6):3036-47.
115. Bouarab K, Potin P, Correa J, Kloareg B. Sulfated Oligosaccharides Mediate the Interaction between a Marine Red Alga and Its Green Algal Pathogenic Endophyte. *Plant Cell.* 1999 Sep;11(9):1635-1650.
116. Michel G, Chantalat L, Fanchon E, Henrissat B, Kloareg B, Dideberg O. The iota-carrageenase of *Alteromonas fortis*. A beta-helix fold-containing enzyme for the degradation of a highly polyanionic polysaccharide. *J Biol Chem.* 2001 Oct 26;276(43): 40202-9.
117. Michel G, Helbert W, Kahn R, Dideberg O, Kloareg B. The structural bases of the processive degradation of iota-carrageenan, a main cell wall polysaccharide of red algae. *J Mol Biol.* 2003 Nov 28;334(3):421-33.

118. Vergin KL, Urbach E, Stein JL, DeLong EF, Lanoil BD, Giovannoni SJ. Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order *Planctomycetales*. *Appl Environ Microbiol*. 1998 Aug;64(8):3075-8.
119. Thauer RK. Biochemistry of methanogenesis: a tribute to Marjory Stephenson. *Microbiology*. 1998 Sep;144:2377-406.
120. Ferry JG. Enzymology of one-carbon metabolism in methanogenic pathways. *FEMS Microbiol Rev*. 1999 Jan;23(1):13-38.
121. Chistoserdova L, Vorholt JA, Thauer RK, Lidstrom ME. C1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic *Archaea*. *Science*. 1998 Jul 3;281(5373):99-102.
122. Graham DE, Overbeek R, Olsen GJ, Woese CR. An archaeal genomic signature. *Proc Natl Acad Sci U S A*. 2000 Mar 28;97(7):3304-8.
123. Vorholt JA. Cofactor-dependent pathways of formaldehyde oxidation in methylotrophic bacteria. *Arch Microbiol*. 2002 Oct;178(4):239-49.
124. Vorholt JA, Chistoserdova L, Lidstrom ME, Thauer RK. The NADP-dependent methylene tetrahydromethanopterin dehydrogenase in *Methylobacterium extorquens* AM1. *J Bacteriol*. 1998 Oct;180(20):5351-6.
125. Vorholt JA, Marx CJ, Lidstrom ME, Thauer RK. Novel formaldehyde-activating enzyme in *Methylobacterium extorquens* AM1 required for growth on methanol. *J Bacteriol*. 2000 Dec;182(23):6645-50.
126. Vorholt JA, Chistoserdova L, Stolyar SM, Thauer RK, Lidstrom ME. Distribution of tetrahydromethanopterin-dependent enzymes in methylotrophic bacteria and phylogeny of methenyl tetrahydromethanopterin cyclohydrolases. *J Bacteriol*. 1999 Sep;181(18):5750-7.
127. Pomper BK, Vorholt JA. Characterization of the formyltransferase from *Methylobacterium extorquens* AM1. *Eur J Biochem*. 2001 Sep;268(17):4769-75.
128. Pomper BK, Saurel O, Milon A, Vorholt JA. Generation of formate by the formyltransferase/hydrolase complex (Fhc) from *Methylobacterium extorquens* AM1. *FEBS Lett*. 2002 Jul 17;523(1-3):133-7.

129. Ward NL et al. The Institute of Genome Research. web site: <http://www.tigr.org/tdb/mdb/mdbinprogress.html>.
130. Marx CJ, Chistoserdova L, Lidstrom ME. Formaldehyde-detoxifying role of the tetrahydromethanopterin-linked pathway in *Methylobacterium extorquens* AM1. *J Bacteriol.* 2003 Dec;185(24):7160-8.
131. Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, Chiang HC, Hooper LV, Gordon JI. A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. *Science.* 2003 Mar 28;299(5615):2074-6.
132. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang CH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabinowitsch E, Rajandream MA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature.* 2002 May 9;417(6885):141-7.
133. Higgins CF. ABC transporters: physiology, structure and mechanism--an overview. *Res Microbiol.* 2001 Apr-May;152(3-4):205-10.
134. Schmitt L. The first view of an ABC transporter: the X-ray crystal structure of MsbA from *E. coli*. *ChemBiochem.* 2002 Mar 1;3(2-3):161-5.
135. Lawrence JG, Ochman H. Reconciling the many faces of lateral gene transfer. *Trends Microbiol.* 2002 Jan;10(1):1-4.
136. Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ, Doolittle WF. Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet.* 2003;37:283-328.
137. Lerat E, Daubin V, Moran NA. From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the *gamma-Proteobacteria*. *PLoS Biol.* 2003 Oct;1(1):E19.
138. Daubin V, Lerat E, Perriere G. The source of laterally transferred genes in bacterial genomes. *Genome Biol.* 2003;4(9):R57.
139. Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebahia M, James KD, Churcher C, Mungall KL, Baker S, Basham D, Bentley SD, Brooks K, Cerdeno-Tarraga AM, Chillingworth T, Cronin A, Davies RM, Davis P,

Dougan G, Feltwell T, Hamlin N, Holroyd S, Jagels K, Karlyshev AV, Leather S, Moule S, Oyston PC, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*. 2001 Oct 4;413(6855):523-7.

140. Radnedge L, Agron PG, Worsham PL, Andersen GL. Genome plasticity in *Yersinia pestis*. *Microbiology*. 2002 Jun;148(Pt 6):1687-98.

141. Deng W, Burland V, Plunkett G 3rd, Boutin A, Mayhew GF, Liss P, Perna NT, Rose DJ, Mau B, Zhou S, Schwartz DC, Fetherston JD, Lindler LE, Brubaker RR, Plano GV, Straley SC, McDonough KA, Nilles ML, Matson JS, Blattner FR, Perry RD. Genome sequence of *Yersinia pestis* KIM. *J Bacteriol*. 2002 Aug;184(16):4601-11.

142. Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, Fuji F, Hiramata C, Nakamura Y, Ogasawara N, Kuhara S, Horikoshi K. Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. *Nucleic Acids Res*. 2000 Nov 1;28(21):4317-31.

143. Bolotin A, Mauger S, Malarne K, Ehrlich SD, Sorokin A. Low-redundancy sequencing of the entire *Lactococcus lactis* IL1403 genome. *Antonie Van Leeuwenhoek*. 1999 Jul-Nov;76(1-4):27-76.

144. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*. 1997 Jan;10(1):1-6.

145. Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrenner P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E, Westbrook-Wadman S, Yuan Y, Brody LL, Coulter SN, Folger KR, Kas A, Larbig K, Lim R, Smith K, Spencer D, Wong GK, Wu Z, Paulsen IT, Reizer J, Saier MH, Hancock RE, Lory S, Olson MV. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*. 2000 Aug 31;406(6799):959-64.

146. Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res*. 2001 Oct 31;8(5):205-13; 227-53.

147. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang

CH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream MA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*. 2002 May 9;417(6885):141-7.

148. Cases I, de Lorenzo V, Ouzounis CA. Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol*. 2003 Jun;11(6):248-53.

149. Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V, Duprat S, Galperin MY, Koonin EV, Le Gall F, Makarova KS, Ostrowski M, Oztas S, Robert C, Rogozin IB, Scanlan DJ, Tandeau de Marsac N, Weissenbach J, Wincker P, Wolf YI, Hess WR. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A*. 2003 Aug 19;100(17):10020-5.

150. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*. 2003 Jan 1; 31(1): 315-8.

151. Helmann JD. The extracytoplasmic function (ECF) sigma factors. *Adv Microb Physiol*. 2002, 46: 47-110.

152. Manganelli R, Voskuil MI, Schoolnik GK, Dubnau E, Gomez M, Smith I. Role of the extracytoplasmic-function sigma factor sigma(H) in *Mycobacterium tuberculosis* global gene expression. *Mol Microbiol*. 2002 Jul;45(2):365-74.

153. Cao M, Helmann JD. Regulation of the *Bacillus subtilis* bcrC bacitracin resistance gene by two extracytoplasmic function sigma factors. *J Bacteriol*. 2002 Nov;184(22): 6123-9.

154. Li W, Stevenson CE, Burton N, Jakimowicz P, Paget MS, Buttner MJ, Lawson DM, Kleanthous C. Identification and structure of the anti-sigma factor-binding domain of the disulphide-stress regulated sigma factor sigma(R) from *Streptomyces coelicolor*. *J Mol Biol*. 2002 Oct 18;323(2):225-36.

155. Braun V, Mahren S, Ogierman M. Regulation of the FecI-type ECF sigma factor by transmembrane signalling. *Curr Opin Microbiol*. 2003 Apr;6(2):173-80.

156. Visca P, Leoni L, Wilson MJ, Lamont IL. Iron transport and regulation, cell signalling and genomics: lessons from *Escherichia coli* and *Pseudomonas*. *Mol Microbiol*. 2002 Sep;45(5):1177-90.
157. Cao M, Salzberg L, Tsai CS, Mascher T, Bonilla C, Wang T, Ye RW, Marquez-Magana L, Helmann JD. Regulation of the *Bacillus subtilis* extracytoplasmic function protein sigma(Y) and its target promoters. *J Bacteriol*. 2003 Aug;185(16):4883-90.
158. Asai K, Yamaguchi H, Kang CM, Yoshida K, Fujita Y, Sadaie Y. DNA microarray analysis of *Bacillus subtilis* sigma factors of extracytoplasmic function family. *FEMS Microbiol Lett*. 2003 Mar 14;220(1):155-60.
159. Huang X, Fredrick KL, Helmann JD. Promoter recognition by *Bacillus subtilis* sigmaW: autoregulation and partial overlap with the sigmaX regulon. *J Bacteriol*. 1998 Aug;180(15):3765-70.
160. Felsenstein J. PHYLIP -- Phylogeny Inference Package. *Cladistics*. 1989; 5:154-66.
161. Mrazek J, Bhaya D, Grossman AR, Karlin S. Highly expressed and alien genes of the *Synechocystis* genome. *Nucleic Acids Res*. 2001 Apr 1; 29(7): 1590-601.
162. Karlin S, Mrazek J. Predicted highly expressed and putative alien genes of *Deinococcus radiodurans* and implications for resistance to ionizing radiation damage. *Proc Natl Acad Sci U S A*. 2001 Apr 24; 98(9): 5240-5.
163. Karlin S, Mrazek J, Campbell A, Kaiser D. Characterizations of highly expressed genes of four fast-growing bacteria. *J Bacteriol*. 2001 Sep; 183(17): 5025-40.
164. Karlin S, Barnett MJ, Campbell AM, Fisher RF, Mrazek J. Predicting gene expression levels from codon biases in alpha-proteobacterial genomes. *Proc Natl Acad Sci U S A*. 2003 Jun 10; 100(12): 7313-8.
165. Mrazek J, Karlin S. Detecting alien genes in bacterial genomes. *Ann N Y Acad Sci*. 1999 May 18; 870: 314-29.
166. Karlin S, Mrazek J, Gentles AJ. Genome comparisons and analysis. *Curr Opin Struct Biol*. 2003 Jun; 13(3): 344-52.
167. Wolf YI, Rogozin IB, Grishin NV, Koonin EV. Genome trees and the tree of life. *Trends Genet*. 2002 Sep;18(9):472-9.

168. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol.* 2001 Oct 20;1(1):8.
169. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res.* 2003 Jan 1;31(1):23-7.
170. DOE Joint Genome Institute, Microbial genomics programs, status of December 2003. [http://www.jgi.doe.gov/JGI\\_microbial/html/index.html](http://www.jgi.doe.gov/JGI_microbial/html/index.html)
171. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev.* 1995 Mar; 59 (1): 143-69.
172. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: update. *Nucleic Acids Res.* 2004 Jan 1; 32(1): D23-6.
173. Schirmer A, Kolter R. Computational analysis of bacterial sulfatases and their modifying enzymes. *Chem Biol.* 1998 Aug;5(8):R181-6.
174. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 2000 Jan 1; 28(1): 33-6.
175. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics.* 2003 Sep 11; 4(1): 41.
176. Hulo N, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, Bairoch A. Recent improvements to the PROSITE database. *Nucleic Acids Res.* 2004 Jan 1; 32(1): D134-7.
177. Henikoff JG, Pietrokovski S, Henikoff S. Recent enhancements to the Blocks Database servers. *Nucleic Acids Res.* 1997 Jan 1; 25(1): 222-5.
178. Klein P, Kanehisa M, DeLisi C. Prediction of protein function from sequence properties. Discriminant analysis of a database. *Biochim Biophys Acta.* 1984 Jun 28; 787 (3): 221-6.
179. Moller S, Croning MD, Apweiler R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics.* 2001 Jul; 17(7): 646-53.

180. Kall L, Sonnhammer EL. Reliability of transmembrane predictions in whole-genome data. *FEBS Lett.* 2002 Dec 18; 532(3): 415-8.

181. Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, Mulder N, Oinn T, Maslen J, Cox A, Apweiler R. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* 2003 Apr; 13(4): 662-72.