UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA

CENTRO DE CIÊNCIAS AGRÁRIAS AMBIENTAIS E BIOLÓGICAS

EMBRAPA MANDIOCA E FRUTICULTURA

PROGRAMA DE PÓS-GRADUAÇÃO EM MICROBIOLOGIA AGRÍCOLA

CURSO DE MESTRADO

DIVERSIDADE E EVOLUÇÃO DO NONO E DÉCIMO PASSOS DA VIA BIOSSINTÉTICA DE PURINAS EM PROCARIOTOS

DENNIFIER COSTA BRANDÃO CRUZ

CRUZ DAS ALMAS – BAHIA

NOVEMBRO DE 2018

DIVERSIDADE E EVOLUÇÃO DO NONO E DÉCIMO PASSOS DA VIA BIOSSINTÉTICA DE PURINAS EM PROCARIOTOS

DENNIFIER COSTA BRANDÃO CRUZ

Bióloga

Universidade Federal do Recôncavo da Bahia, 2016

Dissertação submetida ao Colegiado do Programa de Pós-Graduação em Microbiologia Agrícola da Universidade Federal do Recôncavo da Bahia e Embrapa Mandioca e Fruticultura, como requisito parcial para obtenção do grau de Mestre em Microbiologia Agrícola.

Orientador: Phellippe Arthur Santos Marbach

CRUZ DAS ALMAS – BAHIA

NOVEMBRO DE 2018

FICHA CATALOGRÁFICA

C957d Cruz, Dennifier Costa Brandão.
Diversidade e evolução do nono e décimo passos da via biossintética de purinas em procariotos / Dennifier Costa Brandão Cruz._ Cruz das Almas, BA, 2018.
105f.; il.

Orientador: Phellippe Arthur Santos Marbach.

Dissertação (Mestrado) – Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias Ambientais e Biológicas.

1.Procariotos – Diversidade biológica. 2.Procariotos – Genética. I.Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas. II.Título.

CDD: 576.8

Ficha elaborada pela Biblioteca Universitária de Cruz das Almas – UFRB.
Responsável pela Elaboração – Antonio Marcos Sarmento das Chagas (Bibliotecário – CRB5 / 1615).
Os dados para catalogação foram enviados pela usuária via formulário eletrônico.

UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA

CENTRO DE CIÊNCIAS AGRÁRIAS AMBIENTAIS E BIOLÓGICAS

EMBRAPA MANDIOCA E FRUTICULTURA

PROGRAMA DE PÓS-GRADUAÇÃO EM MICROBIOLOGIA AGRÍCOLA

CURSO DE MESTRADO

COMISSÃO EXAMINADORA DA DEFESA DE DISSERTAÇÃO DE

DENNIFIER COSTA BRANDÃO CRUZ

Prof. Dr. Phellippe Arthur Santos Marbach

Universidade Federal do Recôncavo da Bahia - UFRB

(Orientador)

Prof. Dr. Jorge Teodoro de Souza

Universidade Federal de Lavras - UFLA

Prof. Dr. Aristóteles Góes Neto

Universidade Federal de Minas Gerais - UFMG

“Dissertação homologada pelo Colegiado do Programa de Pós–Graduação em Microbiologia Agrícola em _____ conferindo o grau de Mestre em Microbiologia Agrícola em _____.”

DEDICATÓRIA

“Dedico este trabalho a minha mãe Maria Helena, sem ela eu nada seria.”

AGRADECIMENTOS

Agradeço primeiramente a Universidade Federal do Recôncavo da Bahia, ao Programa de Pós-Graduação em Microbiologia Agrícola a sua administração e todo o corpo docente que contribuíram grandemente para que eu chegasse até aqui. Agradeço também a CAPES pelo financiamento da bolsa que possibilitou os meus estudos.

Agradeço a minha mãe Maria Helena e ao meu noivo Geovane, por sempre estarem ao meu lado me apoiando e incentivando em todas as etapas da minha vida e nas horas difíceis não me deixarem desistir.

Ao meu querido e caríssimo orientador, meu muito obrigado, por toda dedicação, incentivo, atenção, amizade e consideração. As suas palavras e ensinamentos ao longo desses 6 anos me acompanharão pela vida, você é parte fundamental da profissional e pessoa que eu sou hoje, obrigada!

Ao professor Jorge Teodoro, meu muito obrigado, a sua colaboração, empenho e dedicação foi fundamental nesse processo, e contribuiu grandemente na produção desse trabalho.

Aos colegas e amigos que direta ou indiretamente participaram dessa trajetória, obrigada.

À todos o meu muito obrigada!

LISTA DE TABELAS

CAPÍTULO 1

CAPÍTULO 2

LISTA DE FIGURAS

CAPÍTULO 1

CAPÍTULO 2

LISTA DE SIGLAS

OTU - Operational Taxonomical Units

AICARFT - 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase

IMPCH – IMP cyclohydrolase

PRPP - Fosforribosil pirofosfato

IMP - Inosina monofosfato

PBP- Purine biosynthetic pathway

VBP- Via biossintética de purinas

NCBI - National Centre for Biotechnology Information

BLAST – Basic Local Alignment Search Tool

ML - Maximum likelihood

LPSN – List of Prokaryotic names with Standing in Nomenclature

AICAR- 5-Aminoimidazole-4-carboxamide ribonucleotide

FAICAR- 5-Formamidoimidazole-4-carboxamide ribotide

ÍNDICE

CAPÍTULO 2

RESUMO

Cruz, DCB. Diversidade e evolução do nono e décimo passos da via biossintética de purinas em procariotos

A via biossintética de purinas (VBP) inicia a partir do fosforibosil pirofosfato (PRPP) que é convertido à inosina monofosfato (IMP) em 10 etapas enzimáticas, em 4 dessas etapas as enzimas envolvidas podem variar de acordo com o grupo taxonômico. Nas bactérias e eucariotos o nono e décimo passo são realizados preferencialmente pela enzima PurH e nas archaeas pela PurP e PurO respectivamente. A PurH tem dois domínios, o AICARFT que realiza o nono passo e o IMPCH que realiza o decimo passo, esses domínios recentemente foram encontrados em espécies do Domínio *Archaea* como genes independentes não fusionados. A PurP e PurO são análogas aos domínios da PurH e até então consideradas assinaturas do Domínio *Archaea*. No primeiro capítulo desse trabalho foi realizada um estudo da distribuição dos genes relacionados com os últimos passos da VBP (*purH, purP, purO* e os genes que codificam AICARFT e IMPCH nomeados de *purV* e *purJ*, respectivamente) em 1405 genomas completamente sequenciados dos Domínios *Archaea* e *Bacteria*. As análises de genômica comparativa indicaram que a PurH foi preferencialmente a solução evolutiva do Domínio *Bacteria* para catalisar as reações enzimáticas das últimas etapas da VBP. Homólogos dos genes *purO, purV* e *purJ* foram encontrados em genomas bacterianos indicando que esse genes não são exclusivos do domínio *Archaea*. Também foi observado que existe padrão taxonômico para ocorrência desses genes e que em algumas espécies eles estão no mesmo contexto genômico que outros genes da VBP. Apesar de apresentarem o mesmo padrão de conservação da estrutura primária, os homólogos da PurO do Domínio *Archaea* e *Bacteria* não estão relacionados e formaram grupos distintos na árvore filogenética, devido a isso podemos inferir que a PurO já existia no ancestral

comum de todos os seres vivos. Por outro lado a PurP parece ter surgido após a divergência das archaeas, e suas isoformas foram originadas a partir de eventos de duplicação gênica. O objetivo do trabalho apresentado no segundo capítulo foi compreender a história evolutiva das PurVs e PurJs e suas relações evolutivas com os domínios AICARFT e IMPCH das PurHs. A genômica comparativa foi realizada com 2735 genomas, apesar do número de genomas analisados ser quase o dobro, os resultados observados para a distribuição de *purH, purV, purJ*, *purP* e *purO* nas linhagens procarióticas foram similares aos apresentados no primeiro capítulo. A análise dos aminoácidos do sítio ativo mostrou que apesar de algumas variações, existem aminoácidos conservados entre AICARFTs/PurVs e IMPCHs/PurJs, inclusive aminoácidos que foram descritos como essenciais para a atividade do AICARFT. De acordo com a literatura a maioria dessas variações são possíveis do ponto de vista físico-químico, além disso, elas são equivalentes entre AICARFTs/PurVs e IMPCHs/PurJs. A topologia da árvore filogenética das PurHs mostra uma clara separação entre Gram positivas e Gram Negativas sugerindo que as PurHs atuais originaram de um único evento de fusão gênica. As análises filogenéticas indicaram que as PurHs das archaeas são filogeneticamente relacionadas com as PurHs de Gram Negativas o que indica que elas foram adquiridas por transferência horizontal. Esse resultado é coerente com a hipótese de que o evento de fusão gênica que originou a PurH ocorreu no Domínio *Bacteria*. A partir da filogenia podemos inferir que a PurV provavelmente tenha sido originada a partir do domínio ancestral que originou a PurH e a PurJ a partir de quebras do gene da PurH ao longo da evolução e diversificação das linhagens procarióticas.

Palavras-chave: IMP, PRPP, nucleotídeos, filogenia, genômica comparativa.

ABSTRACT

Cruz, DCB. Diversity and evolution of the ninth and tenth steps of the purine biosynthetic pathway in prokaryotes

The purine biosynthetic pathway (PBP) starts from phosphoribosyl pyrophosphate (PRPP) which is converted to Inosine monophosphate (IMP) in 10 enzymatic steps, in 4 of these steps the enzymes involved may vary according to the taxonomic group. In *Bacteria* and eukaryotes the ninth and tenth steps are performed by the PurH enzyme and in the archaeas by the PurP and PurO respectively. PurH has two domains, the AICARFT that performs the ninth step and the IMPCH that performs the tenth step, these domains have recently been found in species of the *Archaea* Domain as independent unfused genes. The PurP and PurO are analogous to the PurH domains and until then considered signatures of the *Archaea* Domain. In the first chapter of this work a study of the distribution of the genes related to the last steps of the PBP (*purH, purP, purO* and the genes coding for AICARFT and IMPCH named *purV* and *purJ*, respectively) was carried out on 1405 completely sequenced genomes of the Domains *Archaea* and *Bacteria*. Analyzes of comparative genomics indicated that PurH was the evolutionary solution of the *Bacteria* Domain to catalyze the enzymatic reactions of the last stages of PBP. Homologues of the *purO*, *purV* and *purJ* genes were found in bacterial genomes indicating that these genes are not exclusive to the *Archaea* Domain. It was also observed that there is a taxonomic pattern for the occurrence of these genes and that in some species they are in the same genomic context as other genes of the PBP. In spite of presenting the same conservation pattern of the primary structure, the PurO homologs of the Domains *Archaea* and *Bacteria* are not related and formed distinct groups in the phylogenetic tree, due to this we can infer that the PurO already existed in the common ancestor of all living beings. On the other hand the PurP seems to have arisen after the divergence of archaeas, and its isoforms were originated from events of gene duplication. The

objective of the work presented in the second chapter was to understand the evolutionary history of the PurVs and PurJs and their evolutionary relations with the AICARFT and IMPCH domains of PurHs. The comparative genomics were performed with 2735 genomes, although the number of analyzed genomes was almost double, the results observed for the distribution of *purH*, *purV*, *purJ*, *purP* and *purO* in the prokaryotic lines were similar to those presented in the first chapter. The amino acid analysis of the active site showed that despite some variations, there are conserved amino acids between AICARFTs/PurVs and IMPCHs/PurJs, including amino acids which have been described as essential for AICARFT activity. According to the literature most of these variations are physically-chemically possible, in addition, they are equivalent between AICARFTs/PurVs and IMPCHs/PurJs. The topology of the phylogenetic tree of the PurHs shows a clear separation between Gram positive and Gram negative suggesting that the current PurHs originated from a single event of gene fusion. Phylogenetic analyzes indicated that archaeal PurHs are phylogenetically related to Gram Negative PurHs, indicating that they were acquired by horizontal transfer. This result is consistent with the hypothesis that the event of gene fusion that originated the PurH occurred in the *Bacteria* Domain. From the phylogeny we can infer that the PurV probably originated from the ancestral domain that originated PurH and PurJ from breaks of the PurH gene along the evolution and diversification of prokaryotic lines.

Key words: IMP, PRPP, nucleotides, phylogeny, comparative genomics.

INTRODUÇÃO

Nas ultimas décadas o estudo de vias biossintéticas, seja voltado para a biotecnologia quanto para a evolução, têm assumido papel relevante na biologia. As vias biossintéticas são ricas fontes informacionais e devido a isso, são atualmente bem exploradas em diversos aspectos, fazendo com que essa seja uma área promissora e de muitas possibilidades. A evolução e diversificação das vias biossintéticas estão intimamente relacionadas com a evolução e diversificação dos seres vivos. Segundo Caetano-Anollés et al. 2007, a maioria das vias, principalmente aquelas que produzem moléculas fundamentais aos organismos, como nucleotídeos, carboidratos, aminoácidos, são de origem antiga e provavelmente surgiram nas primeiras formas de vida. Sendo assim estudar a evolução e diversidade das vias biossintéticas, é também estudar a composição metabólica e gênica dos primeiros seres vivos.

O estudo das vias biossintéticas tem auxiliado no desenvolvimento e produção em escala industrial de compostos de importância médica, alimentar e agrícola, assim como na prospecção de novos compostos bioativos. Hoje a manipulação genética de uma via biossintética possibilita, por exemplo, que uma bactéria produza quantidades comerciais de substancias como o succinato que tem alto valor comercial (Zhu et al. 2016). Nesse cenário a via biossintética de purinas (VBP) se destaca por ser uma das mais antigas na história evolutiva dos seres vivos, por estar presente na grande maioria dos organismos celulares produzindo compostos precursores dos ácidos nucléicos e de outras moléculas essenciais para o funcionamento celular (Xu et al. 2007), e também pela relevância das suas enzimas e intermediários.

As enzimas e intermediários da VBP são alvos potenciais para o desenvolvimento de fármacos anticâncer e antimicrobianos e já apresentam resultados satisfatórios nesse sentido (Xu et al. 2007; Firestine et al. 2009; Nours et al. 2011; Tranchimand et al. 2011; Baggott & Tamura 2015). As enzimas da VBP também são comumente relacionadas com o crescimento e virulência de bactérias fitopatogências, bem como desempenham papel fundamental nas

interações simbióticas entre microrganismos e plantas (Xie et al. 2005; Park et al. 2007; Yan & Wang 2012; Yuan et al. 2013).

A VBP tem como precursor o fosforribosil-pirofosfato (PRPP), que através de 10 passos enzimáticos é convertido a inosina-monofosfato (IMP). Originalmente existem 15 enzimas distintas que podem estar envolvidas nesse processo (PurF, PurD, PurN, PurT, PurL, PurLQS, PurM, PurEII, PurK, PurEI, PurC, PurB, PurH, PurP e PurO) (Zhang et al. 2008-a) O terceiro passo da VBP pode ser desempenhado pela PurN ou PurT, o quarto pela PurL ou pelo complexo PurLQS, o sexto pela PurEII ou pela PurK e PurEI e o nono e décimo passo pela PurH, ou pela PurP e PurO respectivamente. Essa variação enzimática citada ocorre em grupos taxonômicos distintos, eucariotos, bactérias e archaeas geralmente utilizam enzimas diferentes para realizar os mesmos passos da via (Zhang et al. 2008-a, Zhang et al. 2008-b, Brown et al. 2011).

Em 2011 Brown e Colaboradores fizeram uma abordagem da distribuição dos genes da VBP no Domínio *Archaea*, algo que ainda não foi realizado para o Domínio *Bacteria*. Nesse trabalho eles descreveram a existência de mais duas variações na VBP. A PurH enzima responsável pelos dois últimos passos da via, é uma enzima bifuncional que apresenta dois domínios o IMPCH e o AICARFT, Brown e Colaboradores relataram a existência dos domínios da PurH de forma não fusionada no Domínio *Archaea*, cada um como um gene independente e que possivelmente também teriam atividade na via, entretanto eles não estudaram qual seria a origem desses novos genes e se eles também ocorriam no Domínio *Bacteria*.

Devido a isso o principal objetivo desse trabalho foi estudar a diversidade e evolução das enzimas nono e décimo passos da VBP nos genomas procarióticos dos Domínios *Archaea* e *Bacteria*. Além disso, entender a origem desses novos genes descritos por Brown e Colaboradores 2011, e saber se eles realmente ocorrem apenas nas archaeas.

REFERÊNCIAS

Baggott JE, Tamura T. 2015. Folate-Dependent Purine Nucleotide Biosynthesis in Humans. Adv. Nutr. v.6, n.5, p.564-571.

Brown AM, Hoopes SL, White RH, Sarisky CA. 2011. Purine biosynthesis in archaea: variations on a theme. Biology Direct. v.6, n.63. p.1-21.

Caetano-Anollés G, Kim SK, Mittenthal J E. 2007. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. PNAS. v. 104, p. 9358-9363.

Firestine SM, Paritala H, Mcdonnell JE, Thoden JB, Holden HM. 2009. Identification of inhibitors of N5-carboxyaminoimidazole ribonucleotide synthetase by high-throughput screening. Bio org Med Chem. v.17, n.9, p.3317-3323.

Nours JL, et al. 2011. Structural analyses of a purine biosynthetic enzyme from Mycobacterium tuberculosis reveal a novel bound nucleotide. The Journal of Biological Chemistry. v.286, n.47, p.40706–40716.

Park Y, et al. 2007. Analysis of virulence and growth of a purine auxotrophic mutant of Xanthomonas oryzae pathovaroryzae. FEMS Microbiol Lett. 276, p. 55–59.

Tranchimand S, Starks CM, Mathews II, Hockings SC, Kappock TJ. 2011. Treponema denticola PurE Is a bacterial AIR carboxylase. Biochemistry. v.50, p.4623–4637.

Xie B, et al. 2005. Symbiotic abilities of Sinorhizobium fredii with modified expression of purL. Appl Microbiol Biotechnol. 71, p. 505–514.

Xu L, et al. 2007. Structure-based design, synthesis, evaluation, and crystal structures of transition state analogue inhibitors of inosine monophosphate cyclohydrolase. The Journal of Biological Chemistry. v. 282, n.17, p.13033–13046.

Zhang Y, Morar M, Ealick SE. 2008-a. Structural biology of the purine biosynthetic pathway. Cell Mol Life Sci. 65(23):3699-724.

Zhang Y, White RH, Ealick SE. 2008-b. Crystal structure and function of 5-formaminoimidazole-4-carboxamide-1-β-d-ribofuranosyl 5′-monophosphate synthetase from Methanocaldococcus jannaschii. Biochemistry 47(1):205-217.

Zhu L-W, et al. 2016. Enhancing succinic acid biosynthesis in Escherichia coli by engineering its global transcription factor, catabolite repressor/activator (Cra). Scientific Reports. v.6.

CAPÍTULO 1

Different ways of doing the same: variations in the two last steps of the purine biosynthetic pathway in prokaryotes.

Artigo submetido à revista Genome Biology and Evolution (GBE)

Abstract

Cruz, DCB. Different ways of doing the same: variations in the two last steps of the purine biosynthetic pathway in prokaryotes

The last two steps of the purine biosynthetic pathway may be catalysed by different enzymes in prokaryotes. The genes that encode these enzymes include homologs of *purH*, *purP*, *purO* and those encoding the AICARFT and IMPCH domains of PurH, here named *purV* and *purJ*, respectively. In *Bacteria* these reactions are mainly catalysed by the domains AICARFT and IMPCH of PurH. In *Archaea* these reactions may be carried out by PurH and also by PurP and PurO, both considered signatures of this domain and analogous to the AICARFT and IMPCH domains of PurH, respectively. These genes were searched for in 1,405 completely sequenced prokaryotic genomes publicly available. Our analyses revealed taxonomic patterns for the distribution of these genes and anti-correlations in their occurrence. The analyses of bacterial genomes revealed the existence of genes coding for PurV, PurJ and PurO, which may no longer be considered signatures of the domain *Archaea*. Although genetically unrelated, the PurOs of *Archaea* and *Bacteria* show a high level of conservation in the amino acids of the active sites of the protein, allowing us to conclude that these enzymes are analogs. The gene *purO* was present in the common ancestor of all living beings, whereas the gene encoding PurP emerged after the divergence of *Archaea* and *Bacteria* and their isoforms originated in duplication events in the common ancestor of phyla *Crenarchaeota* and *Euryarchaeota*. The results reported here expand our understanding of the diversity and evolution of the last two steps of the purine biosynthetic pathway in prokaryotes.

Key words: *Archaea*, *Bacteria*, bioinformatics, comparative genomics, evolution, phylogeny.

Introduction

Purines and their derivatives are biomolecules essential to all living organisms as they play important roles in signalling pathways, carbohydrate metabolism and as precursors of nucleic acids (Smith and Atkins 2002). Additionally, the enzymes that catalyse their synthesis are important targets for antimicrobial and anticancer compounds (Kirsch and Whitney 1991). The enzymatic reactions involved in the *de novo* biosynthesis of purines were elucidated in the 1950's (Buchanan and Hartman 1959) and all genes that encode these enzymes were identified and their products biochemically characterized by the year 2000 (Zalkin 1983; Parker 1984; Schrimsher et al. 1986; Watanabe et al. 1989; Aiba and Mizobuchi 1989; Cheng et al. 1990; Inglese et al. 1990; He et al. 1992; Gu et al. 1993; Marolewski et al. 1994; Graupner et al. 2002; Hoskins et al. 2004; Ownby et al. 2005).

There are generally ten *pur* genes in the following order: *purF, purD, purN, purL, purM, purE, purK, purC, purB* and *purH*, each encoding an enzyme responsible for a step in the purine biosynthetic pathway (PBP). The majority of the *pur* genes were initially described in *Escherichia coli* and later, orthologs were found in other prokaryotes and eukaryotes, suggesting that these are the canonical genes encoding enzymes of the PBP in different evolutionary lineages (Chopra et al. 1991; Ni et al. 1991; Gu et al. 1992; Rayl et al. 1996; Nilsson and Kilstrup 1998; Peltonen and Mäntsälä 1999; Sampei et al. 2010; Liu et al. 2014). This scenario changed when some authors showed that the enzymatic reactions of three of the 10-11 steps of the PBP may be catalysed by different non-homologous enzymes in distinct microorganisms. For example, PurT, a novel enzyme involved in the PBP was described in 1994 as an analog of the already known PurN that catalyses the third step of the PBP (Marolewski et al. 1994). Two new enzymes, PurP and PurO were later described in the PBP of *Archaea* as analogs of PurH, until then considered as the canonical enzyme catalysing the two last steps of the PBP (Ownby et al. 2005; Graupner et al. 2002). PurP and PurO are currently considered signatures of the *Archaea* domain (Graupner et al. 2002; Ownby et al. 2005; Zhang et al. 2008-I; Zhang et al. 2008-II; Armenta-Medina et al. 2014). These two separate enzymes are analogous to the domains AICARFT and IMPCH of PurH, which contains them fused (fig. 1). The domain AICARFT

catalyses the penultimate reaction of the PBP, converting AICAR in FAICAR and the domain IMPCH catalyses the last reaction of the PBP, converting FAICAR into IMP, the final product of the pathway (Zhang et al. 2008-I).

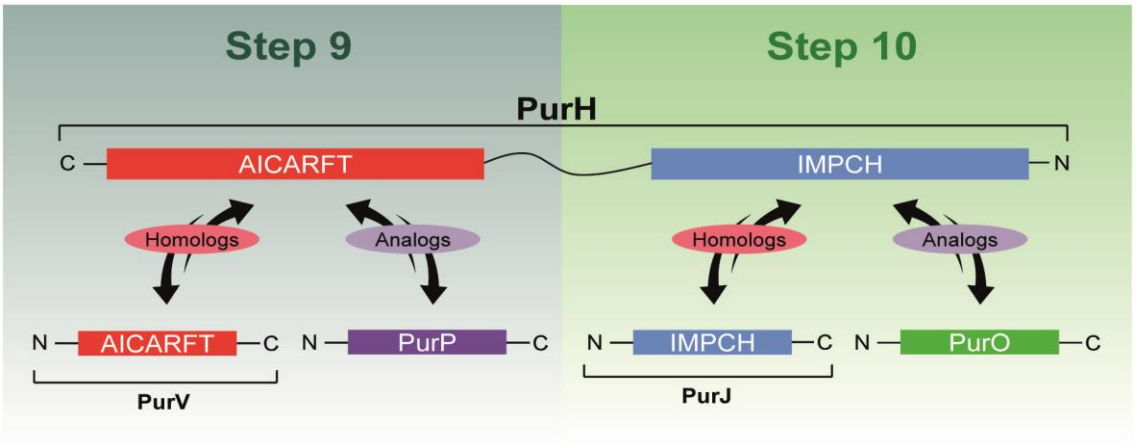


Fig. 1. Nomenclature and homology/analogy relationships among the proteins involved in the last two steps of the purine biosynthetic pathway. PurH is composed of two domains, IMPCH and AICARFT. The relationships of homology and analogy among the domains of PurH and other proteins are shown. PurH is shown in an inverted position to match the steps of the purine biosynthetic pathway.

A recent comparative genomic analysis of the *Archaea* domain showed that in some free-living species of the phylum *Euryarchaeota* the domains AICARFT and IMPCH of PurH are encoded by distinct genes. These archaeal species do not contain genes encoding PurH nor its analogs PurP and PurO (Brown et al. 2011). In this study, the authors also showed that species of the phylum *Crenarchaeota* do not possess the genes coding for PurH, nor homologs of its domains found in *Euryarchaeota* or its analogs PurP and PurO (Brown et al. 2011). This study indicates that the diversity of enzymes involved in the PBP of *Archaea* is higher than previously thought. In their study, Brown et al. (2011) did not include the domain *Bacteria*, where most of the prokaryotic diversity resides.

Purine biosynthesis is among the most ancient metabolic pathways and probably evolved in the LUCA (Caetano-Anollés et al. 2007). According to the hypothesis of Horowitz (1945), enzymes in the last steps of biosynthetic pathways are the first to be recruited. Curiously, the last two steps of the PBP show the highest variation.

Nowadays, the availability of completely sequenced genomes in public databases representing most of the diversity of prokaryotic higher taxa potentially provide a comprehensive picture of the diversity and evolution of biological processes. In this study, we report on a genomic analysis concerning the diversity and evolution of the two last steps of the PBP in prokaryotic lineages. The results are presented in a taxonomical framework that includes the currently accepted phylogenetic classification of the prokaryotes.

Material and Methods

Searches for purine biosynthetic genes (pur) in prokaryotic genomes

A total of 1,405 completely sequenced prokaryotic genomes deposited in the NCBI (*National Centre for Biotechnology Information*) database were used in this study. Bacterial and archaeal strains identified at the genus or species levels were treated as distinct Operational Taxonomical Units (OTUs) in the analyses. The program TBLASTN were used to perform searches for *purH*, *purP* and *purO* and for genes that code for the domains AICARFT and IMPCH in the nucleotide collection (nr/nt), RefSeq Representative genomes (refseq_representative_genomes) and ReFseq Genome (refseq_genomes) databases. The genes that code for the domains AICARFT and IMPCH were named hereafter as *purV* and *purJ*, respectively (fig. 1). Additionally, the program BLASTP was used to perform searches for PurH, PurP, PurO, PurV and PurJ in the non-redundant protein sequences (nr) database.

BLAST searches were done individually in each completely sequenced genome and the presence of conserved domains typical of the searched proteins were used as criteria of homology to recover the sequences. Aminoacid sequences of PurH (GI:16131836) of *Escherichia coli* and PurP (GI: 15668306) and PurO (GI:34588137) of *Methanocaldococcus jannaschii* were used as queries. The aminoacid sequence of PurH was used as query to perform searches for PurH, PurV and PurJ and for the genes that encode these proteins.

Diversity of the last two steps of the purine biosynthetic pathway (PBP)

During the BLAST searches, the presence or absence and the number of copies of *purH*, *purV*, *purJ*, *purP* and *purO* in all the genomes analysed and the genomic context in relation to the other genes of the PBP were registered. The search for taxonomic patterns and the occurrence of these genes in prokaryotic genomes were done in all taxonomic categories, from domain to species.

Multiple alignments and phylogeny of PurP and PurO

The aminoacid sequences of the genes *purH*, *purV*, *purJ*, *purP* and *purO* recovered in the BLAST searches were code-aligned in the guidance server with the MAFFT algorithm (Penn et al. 2010). The program MEGA 6 (Tamura et al. 2013) was used to edit the multiple alignments of the proteins PurP and PurO, choose the substitution matrix, and to perform the phylogenetic analyses with the maximum likelihood (ML) method. The phylogenetic trees were visualized and edited in the program Archaeopterix (Han and Zmasek 2009).

Aminoacid sequence analyses

Sliding window plot analyses were done with the program SWAAP 1.0.2 (Pride 2000) using the multiple alignment of proteins. The average identity of the sequences was calculated with the model K2P in a sampling window of 10 aminoacids with only one aminoacid displaced along the multiple alignments. The logos were produced in the program Web Logo (Crooks et al. 2004) only with the positions of the active sites of the proteins PurP and PurO.

Results

Comparative genomics

The comparative genomic analysis was carried out with a total of 1,405 completely sequenced genomes available in the NCBI database, representing 1,268 OTUs of the domain *Bacteria* and 137 OTUs of the domain *Archaea*. These

OTUs represent 26 out of the 33 described phyla of the domain *Bacteria* and all five phyla of the domain *Archaea* according to LPSN (Euzéby 1997). From the 1,405 analysed genomes, 135 did not have genes coding for PurH, PurO, purP, PurV and PurJ (fig. 2; supplementary table S1).

Fig. 2. Comparative genomic analysis of genes coding for the ninth and tenth steps of the purine biosynthetic pathway in 1,405 prokaryote complete genomes. Coloured boxes represent presence of the gene and smaller black boxes inside the coloured ones represent the cases in which purH is replaced by a combination of genes that are functionally equivalent to purH. The numbers below each taxonomical category indicate the number of genera and OTUs analysed. None indicates the number of OTUs that do not contain any of the genes encoding the last two steps of the PBP.

PurH-coding genes were found in genomes of 23 OTUs of the domain *Archaea*, all of which are in the class *Methanomicrobia* and in the families *Methanoregulaceae, Methanocorpusculaceae, Methanomicrobiaceae, Methanospirillaceae* and *Methanosarcinaceae* and in the class *Thermoplasmata*, families *Ferroplasmaceae, Picrophilaceae* and *Thermoplasmataceae* of the phylum *Euryarchaeota* (fig. 2; supplementary table S1). In contrast, PurH-coding genes were found in 1,081 OTUs of the domain *Bacteria* distributed in most phyla of this domain (fig. 2; supplementary table S1).

Genes that code for proteins homologous to the domains AICARFT and IMPCH of PurH were found in OTUs of the domains *Archaea* and *Bacteria* (fig. 2;

supplementary table S1). The name *purJ* was used to designate the domain IMPCH of the PurH of *Salmonella typhymurium*, when it was incorrectly identified as a gene (Gots et al. 1969). Therefore, from this point on we will use the name *purJ* for the gene encoding the domain IMPCH in accordance with its original nomenclature (Gots et al. 1969). For the domain AICARFT we propose the name *purV* (fig. 1). The majority of the *purV* and *purJ* were recovered from bacterial genomes: 109 *purVs*, 28 from the domain *Archaea* and 81 from domain *Bacteria*; and 55 *purJs*, 4 from the domain *Archaea* and 51 from domain *Bacteria* (table 1; supplementary table S1). The genes *purV* and/or *purJ* were found in 10 of the 26 bacterial phyla analysed, including gram-positive and gram-negative OTUs, indicating that they are widely distributed.

Table 1. Occurrence and co-occurrence of the genes encoding the last two steps of the purine biosynthetic pathway in Bacteria and Archaea and the genomic context they are found. Genes in genomic context are together with other genes of the PBP.

Occurrence per OTU				**Co-occurrence per OTU**			**In context with other *pur genes***		**Not in context with other *pur* genes**	
Genes	Archaea	Bacteria	Total	Genes	Archaea	Bacteria	Genes	Archaea	Bacteria	Archaea	Bacteria
purH	23	1081	1103	*purH/purV*	-	17	*purV*	28	26	-	55
purV	28	81	109	*purH/purJ*	-	4	*purJ*	-	16	4	34
purJ	4	50	54	*purH/purO*	-	2	*purO*	6	13	53	9
purO	59	22	81	*purV/purJ*	-	44	*purP I*	-	-	25	-
purP I	25	-	-	*purV/purO*	25	19	*purP II*	25	-	37	-
purP II	62	-	-	*purH/purV/purO*	-	1	*purP III*	32	-	29	-
purP III	61	-	-	*purP II/purP III*	37	-	*purP IV*	6	-	-	-
purP IV	6	-	-	*purP II/purP III/purJ*	4	-					
				purP I/purO	24	-					
				purP I/purP II/purP III/purO	1	-					
				purP II/ purP III/purV/purO	3	-					
				purP II/purP III/purH	10	-					
				purP II/purP III/purP IV/purO	6	-					

The gene *purO* was until now considered a signature of the domain *Archaea* (Graupner et al. 2002; Ownby et al. 2005; Zhang et al. 2008-I; Zhang et al. 2008-II; Armenta-Medina et al. 2014). Surprisingly, 22 homologs of *purO* were found in bacterial genomes, whereas 59 *purO*s were found in archaeal genomes (table 1; fig. 2). Genes that code for homologs of PurP were only found in genomes of OTUs of the domain *Archaea*, a total of 154, with the number of copies varying from one to four per genome (fig. 2). Homologs of PurH, PurO, PurV and PurJ were not found in genomes of the phyla *Crenarchaeota*, *Korarchaeota* and *Thaumarchaeota*, but genes coding for homologs of PurP were present.

The patterns of presence and absence of the genes *purH, purV, purJ* and *purO* show that the putative new bacterial genes of the purine biosynthetic pathway (PBP), in general, anticorrelate with *purH.* In other words, the majority of the OTUs that do not have *purH* possess *purV* and *purO* or *purV* and *purJ* in the genome (fig. 2). Similarly, archaeal genomes contained the combinations *purV*/*purO*, *purJ*/*purP* and *purP*/*purO*, but not *purV and purJ*, the most common in *Bacteria*, after *purH*. These results suggest that the combinations of genes mentioned earlier for bacteria and archaea are replacing *purH,* maintaining the last two steps of the PBP, as already proposed for the domain *Archaea* (Brown et al. 2011).

Only two assembly/annotation errors in the genes coding for proteins of the last two steps of the PBP were found in the 1,405 analysed genomes: two truncated PurH sequences (Supplementary table S1). These two sequences were included in our analyses. This remarkably low number of misannotations may be due to the fact that these genes are well represented in the biological databases.

In summary, genes coding for PurH were found in 85% of the bacterial genomes and the combinations *purV*/*purJ*, *purV*/*purO* or *purJ* alone were in 65 genomes, representing 5% of the total number of bacterial genomes analysed. The remaining 10% did not harbour any of the genes for the last two steps of the PBP. In Archaea, genes coding for PurH were in 17% of the genomes and the combinations *purV*/*purO*, *purJ*/*purP*, *purP*/*purO* or *purP* alone were found in 73% of the genomes and the remaining 10% of the genomes did not have the genes for the last two steps of the PBP (fig. 2; table 2; Supplementary table S1).

Table 2. Bacterial OTUs with purV, purJ and purO in genomic context with other genes of the purine biosynthetic pathway. Grey boxes indicate genes that do not participate in last two steps or are not part of the purine biosynthetic pathway.

Phylum	Class	Order	OTUs	Genomic Context					
Firmicutes	*Bacilli*	*Bacillales*	*Paenibacillus mucilaginosus* KNP414	*purF*	*purM*	*purN*	***purV***	***purO***	*purD*
			Thermobacillus composti KWC4	*purF*	*purM*	*purN*	***purV***	***purO***	*purD*
	Clostridia	*Clostridiales*	*Butyrivibrio proteoclasticus* B316	*Other*	***purO***	***purV***	*Other*		
			Cellulosilyticum lentocellum DSM 5427	*Other*	***purO***	***purV***	*Other*		
			Clostridium saccharolyticum WM1	*Other*	***purO***	***purV***	*Other*		
			Clostridium sp. SY8519	*Other*	***purO***	***purV***	*Other*		
			Eubacterium eligens ATCC 27750	*Other*	***purO***	***purV***	*Other*		
			Eubacterium rectale ATCC 33656	*Other*	***purO***	***purV***	*Other*		
			Roseburia hominis A2-183	*Other*	***purO***	***purV***	*Other*		
			Oscillibacter valericigenes Sjm18-20	*purF*	*purM*	*purN*	***purV***	*purD*	*purL*
	Negativicutes	*Selenomonadales*	*Acidaminococcus fermentans* DSM 20731	*purF*	*purM*	*purN*	***purJ***	*purD*	
			Acidaminococcus intestini RyC-MR95	*purF*	*purM*	*purN*	***purJ***	*purD*	
			Selenomonas ruminantium subsp. lactilytica TAM6421	*purF*	*purM*	*purN*	***purJ***	*purD*	
Actinobacteria	*Actinobacteria*	*Coriobacteriales*	*Adlercreutzia equolifaciens* DSM 19450	*Other*	***purO***	***purV***	*Other*		
			Olsenella uli DSM 7084	*Other*	***purO***	***purV***	*Other*		
			Slackia heliotrinireducens DSM 20476	*Other*	***purO***	***purV***	*Other*		
Thermotogae	*Thermotogae*	*Thermotogales*	*Fervidobacterium nodosum* Rt17-B1	*purF*	*purN*	***purV***	*purD*	*purM*	
			Fervidobacterium pennivorans DSM 9078	*purF*	*purN*	***purV***	*purD*	*purM*	
			Thermosipho africanus TCF52B	*purF*	*purN*	***purV***	*purD*	*purM*	
			Thermosipho melanesiensis BI429	*purF*	*purN*	***purV***	*purD*	*purM*	
Spirochaetes	*Spirochaetia*	*Spirochaetales*	*Sphaerochaeta globus* str. Buddy	*purF*	*purM*	*purN*	***purV***	*purD*	*purL*
			Sphaerochaeta pleomorpha str. Grapes	*purF*	*purM*	*purN*	***purV***	*purD*	*purL*
			Spirochaeta africana DSM 8902	*Other*	***purJ***	***purV***	*Other*		
			Spirochaeta smaragdinae DSM 11293	*Other*	*purB*	***purJ***	*Other*		
			Spirochaeta sp. L21-RPul-D2	*Other*	***purJ***	***purV***	*Other*		
			Spirochaeta thermophila DSM 6192	*Other*	***purJ***	***purV***	*Other*		
Proteobacteria	*Deltaproteobacteria*	*Desulfarculales*	*Desulfarculus baarsii* DSM 2075	*Other*	***purJ***	*purD*	*Other*		
		Desulfobacterales	*Desulfobacterium autotrophicum* HRM2	*Other*	***purJ***	***purV***	*Other*		
			Desulfobacula toluolica Tol2	*Other*	***purJ***	***purV***	*Other*		
			Desulfobulbus propionicus DSM 2032	*Other*	***purJ***	*purN*	*Other*		
			Desulfocapsa sulfexigens DSM 10523	*Other*	***purJ***	*purN*	*Other*		
			Desulfococcus oleovorans Hxd3	*Other*	***purJ***	***purV***	*Other*		
			Desulfotalea psychrophila LSv54	*Other*	***purJ***	*purN*	*Other*		
			Desulfurivibrio alkaliphilus AHT2	*Other*	***purJ***	*purN*	*Other*		
		Syntrophobacterales	*Syntrophobacter fumaroxidans* MPOB	*Other*	***purJ***	*purD*	*Other*		
	Epsilonproteobacteria	*Campylobacterales*	*Helicobacter felis* ATCC 49179	*purF*	*purM*	***purO***	***purV***	*purD*	*PurL*

Genomic context and taxonomical patterns

The results of these studies demonstrated that *purP* is present only in the domain *Archaea*, while *purO*, *purV* and *purJ* were found both in the domains *Bacteria* and *Archaea* (fig. 2, supplementary table S1). The genes *purP*/*purO* in *purV* and *purJ* are potentially replacing *purH* in *Bacteria* and *Archaea*, performing the last two steps of the PBP. Approximately 37% of the total number of the genes that are potentially replacing *purH* in *Bacteria* and *Archaea* is in genomic context with other genes of the PBP and the remaining 63% is not in context in *Bacteria* and *Archaea* (tables 1 and 2).

In some OTUs the putative new genes, *purO*, *purV* and *purJ* in *Bacteria* are in genomic context with themselves or with other genes of the PBP, from which the most frequent are *purD* and *purN* (table 2). The arrangements of these genes in *Bacteria* (table 2) indicate that they are putative operons that are co-expressed and have a role in the PBP.

The comparative genomic analysis shows that there are taxonomical patterns at the genus, family and class levels for the combinations *purP*, *purO*, *purV* and *purJ*, which are able to replace *purH* in *Archaea* and *Bacteria*. Thus, the presence of these genes is typical of specific higher taxa of prokaryotes. Some patterns are maintained at the genus, family or class level, but not at the phylum level (table 3).

Table 3. Taxonomic patterns of ocurrence of purPs, purO, purV and purJ. All OTUs of these taxa, including classes, families and genera included harbour the gene shown. The numbers between brackets are the amount of OTUs in each group.

Taxonomic Classification				Taxonomic Patterns
Archaea	*Crenarchaeota*	Family *Thermoproteaceae*	[12]	*purP II / purP III*
	Euryarchaeota	Class *Halobacteria*	[25]	*purV-N / purO*
		Family *Methanosaetaceae*	[03]	*purV / purO / purP II / purP III*
		Genus *Archaeoglobus*	[04]	*purJ / purP II / purP III*
Bacteria	*Firmicutes*	Family *Lachnospiraceae*	[08]	*purV / purJ*
	Bacteriodetes	Family *Prevotellaceae*	[05]	*purV / purJ*
	Thermotogae	Genus *Fervidobacterium*	[02]	*purV / purJ*
		Genus *Thermosipho*	[02]	*purV / purJ*
	Thermodesulfobacteria	Class *Thermodesulfobacteria*	[03]	*purV / purJ*
		Family *Desulfovibrionaceae*	[18]	*purV / purJ*
		Family *Desulfobulbaceae*	[04]	*purV / purJ*
		Family *Desulfobacteraceae*	[04]	*purV / purJ*

The evolutionary history of *purH*, *purV* and *purJ* is intimately related and these results will be presented elsewhere, whereas the evolution of *purO* and *PurP* will be presented in this publication.

Evolution of PurO

The maximum likelihood (ML) tree of PurO shows two distinct groups, one containing PurOs from OTUs of the domain *Archaea* and the other one with PurOs of the domain *Bacteria* (fig. 3a). This topology was obtained both with sequences of aminoacids and nucleotides (supplementary fig. S1) and is congruent with the current prokaryotic phylogeny (Woese and Fox 1977) and therefore the root of this tree was placed in the branch that connects these two groups. The topology of the

archaeal PurOs in the rooted tree agrees with the taxonomy of this domain, at least at the family level (fig. 3b).

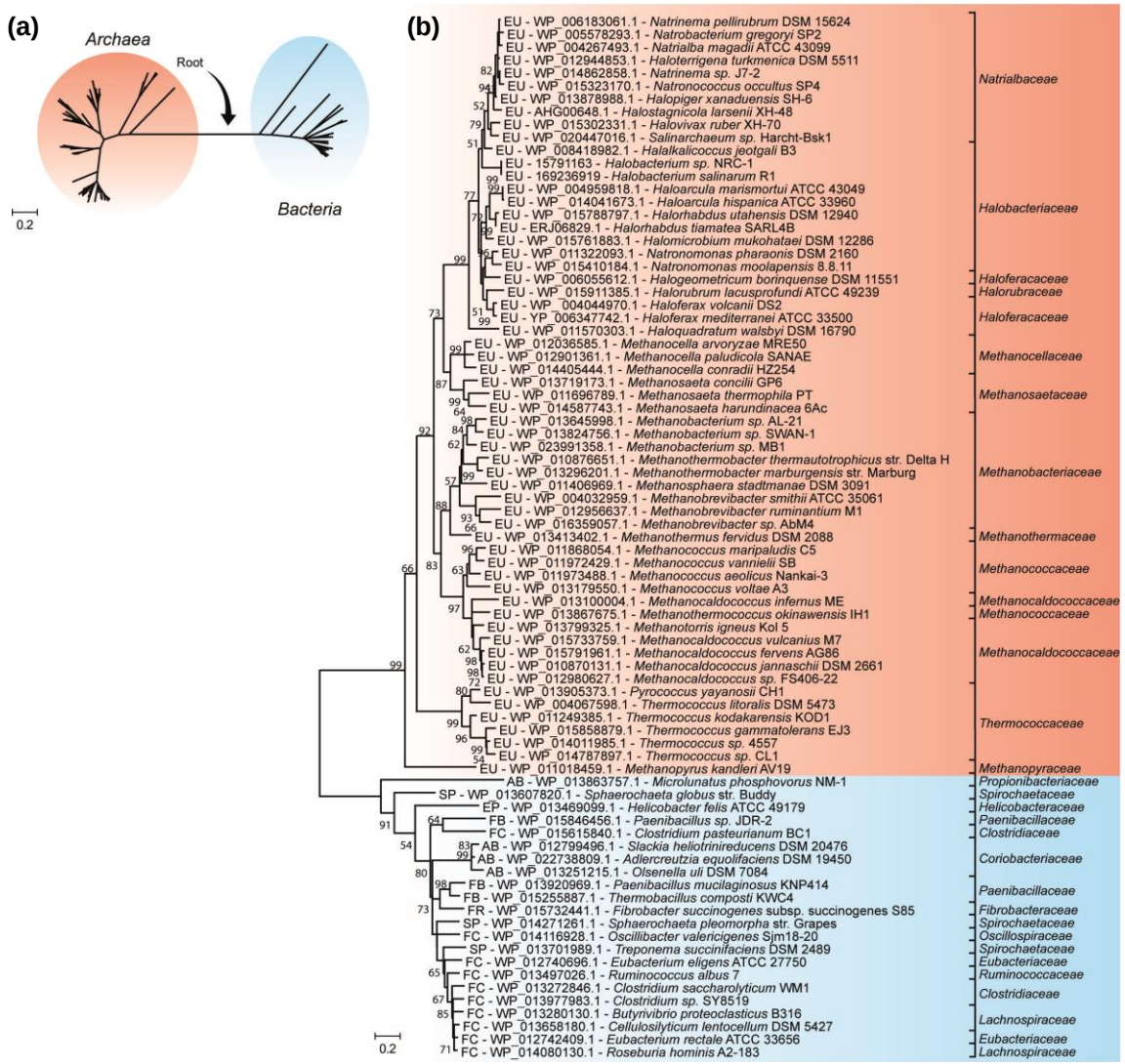


Fig. 3. Phylogenetic relationships among archaeal PurOs and their bacterial counterparts. (a) Maximum likelihood tree of PurOs showing the root placement at the branch that connects PurOs from Archaea and Bacteria. (b) Rooted maximum likelihood tree showing PurOs from Archaea and their homologs in Bacteria in distinct groups. The multiple alignment utilised in these reconstructions contained 81 sequences of aminoacids with 134 sites. The trees were constructed with the model LG+G+I and bootstrap was performed with 1,000 resamplings. The OTUs are identified by abbreviations that represent their phylum and class, accession number of the protein and species name. The abbreviations are: EU- phylum Euryarchaeota; FC- phylum Firmicutes class Clostridia; FB- phylum Firmicutes class Bacillales; AB- phylum Actinobacteria; SP- phylum Spirochaetes; FR- phylum Fibrobacteres; EP- phylum Proteobacteria classe Epsilonproteobacteria. The scale indicates the number of substitutions per site.

The average divergence of *purO* homologs in *Archaea* and *Bacteria* is 73%, indicating that they are highly divergent. However, the results of the genomic analyses previously shown (fig. 2; tables 1 and 2) suggest that the bacterial homologs of purOs are analogs of their archaeal counterparts. To test whether the PurOs in *Archaea* and *Bacteria* are indeed analogs, additional analyses were performed. The sliding window plot analysis showed that the conservation in the primary structure of the archaeal PurOs and their bacterial counterparts is similar. The most conserved regions are the same in both archaeal and bacterial PurOs (fig. 4a). These conserved aminoacids are located in both highly and poorly conserved regions of the primary structure (fig. 4a). The PurO from *Methanothermobacter thermoautotrophicus*, that was already functionally characterized, is included in our analyses. This enzyme is monomeric and its active site comprises 12 aminoacid residues (Kang et al. 2007). The logos constructed with the positions of the multiple alignments of purOs from *Archaea* and their bacterial homologs containing the aminoacids of the active sites of PurO from *M. thermoautotrophicus* show that the aminoacids in these positions are identical or chemically equivalent in both archaeal and bacterial homologs (fig. 4b). The only exceptions are Ser24 replaced by Gly and Asn54 replaced by Asp or Val in the bacterial homologs (fig. 4b). These results are congruent with the hypothesis that purOs in *Archaea* and *Bacteria* are analogous.

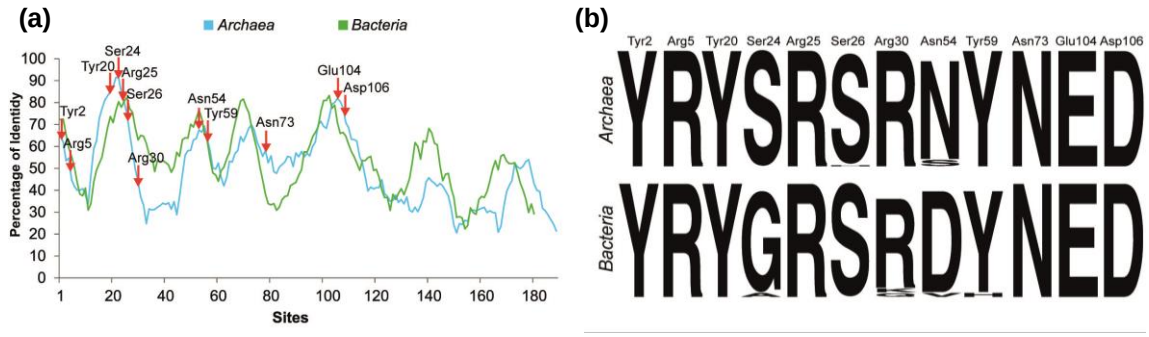


Fig. 4. Analysis of the primary structures of PurOs from Archaea and their homologs in Bacteria. The PurO from Methanothermobacter thermoautotrophicus was used as reference. (a) Sliding window plot analysis of the multiple alignment of PurOs in Archaea and Bacteria show that the conservation in the primary structure is similar in these domains. The arrows indicate the positions of the active sites of PurO from M. thermoautotrophicus. (b) Sequence logos of the positions in the multiple alignment that correspond to the active sites of PurO from

M. thermoautotrophicus showing that most aminoacids are conserved in Archaea and Bacteria.

Evolution of PurP

The number of copies of *purP* in OTUs of the domain *Archaea* that contain this gene varies from one to three. The ML phylogenetic tree of PurP shows four distinct clades (fig. 5a) that were enumerated according to Zhang et al. (2008-II). The indels in the PurP alignment (fig. 5b) were used to root the phylogenetic tree (fig. 5a) because indels are rarely fixed and are highly conserved evolutionary events that do not revert easily as compared to nucleotide substitutions. The PurPs I, II and IV share two indels (fig. 5b), indicating that they are phylogenetically more related with each other than with PurP III. These indels were utilised as a non-arbitrary criterion to root the phylogenetic tree of PurP in the branch that connects PurP III with the other groups (fig. 6). The topology of the rooted tree indicates that PurPs I and II are themselves more related than they are with PurP IV (fig. 6).

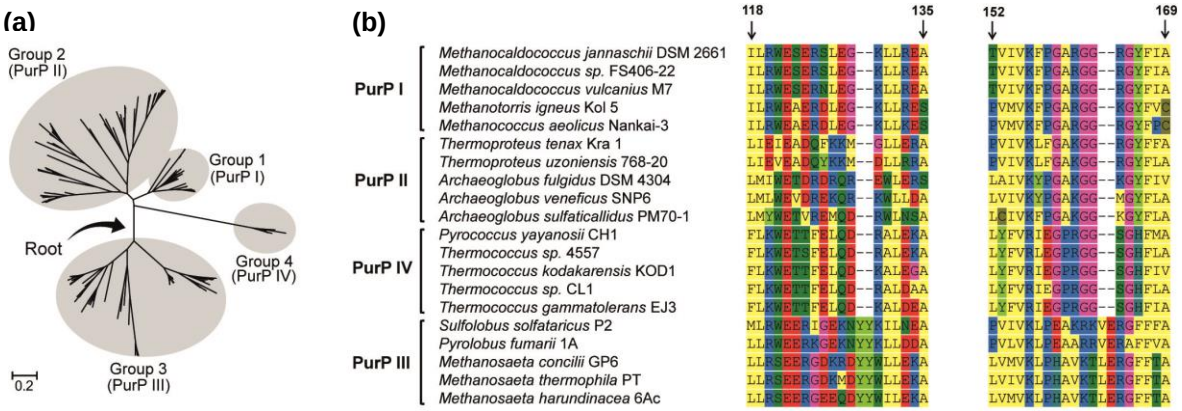


Fig. 5. Groups of PurP defined by phylogenetics and detail of the multiple alignment. (a) Unrooted maximum likelihood tree of PurP constructed with 154 aminoacid sequences containing 399 sites, the model LG+G+I and bootstrap with 1,000 resamplings. Root placement is indicated. (b) Indels shared by PurPs I, II and IV that were used to root the phylogenetic tree of PurP shown in (a). Five PurPs of each group are represented in the alignment.

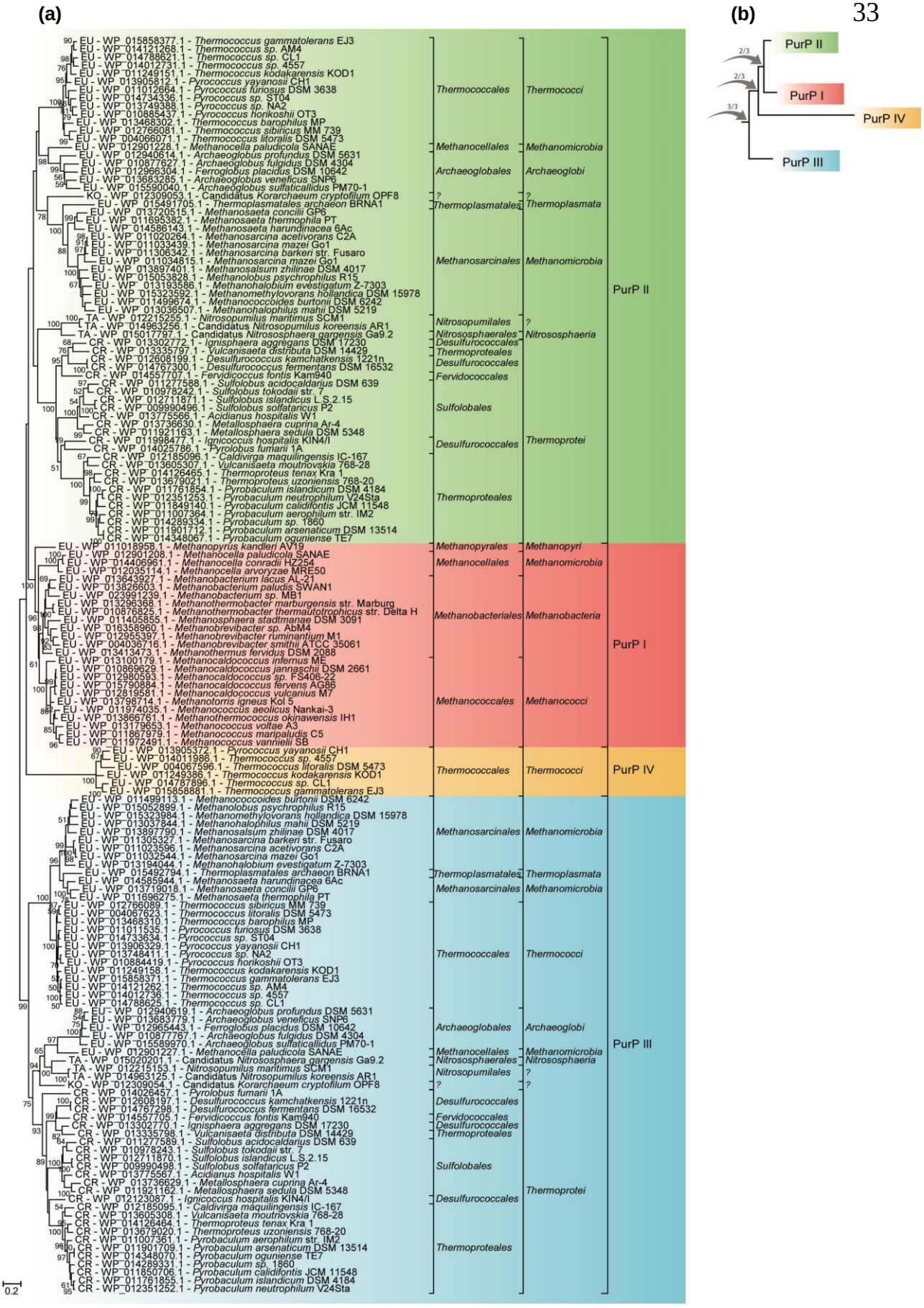


Fig. 6. Phylogenetic trees of the PurPs and the proposed events of duplication that gave rise to the different groups. (a) Phylogenetic tree rooted on the branch that

connects PurP III with the other PurPs. This tree is shown unrooted in figure 5a. (b) Proposed evolutionary history of PurPs. Arrows indicate duplication events and the numbers on the arrows indicate the number of times the branch is recovered in phylogenetic analyses performed with different settings (supplementary fig. S2). The abbreviations indicate the phylum of each sequence: EU - phylum *Euryarchaeota*; CR - phylum *Crenarchaeota*; TA - phylum *Thaumarchaeota*; KO - phylum *Korarchaeota*. The scale indicates the number of substitutions per site.

The topology of the PurP tree suggests that clades I, II, III and IV are paralogs that originated in three events of duplication that occurred in the ancestral of the *Archaea* domain (fig. 6b). The first duplication event originated the PurP III and two subsequent duplications originated the PurPs IV, I and II. These isoforms of PurP were selectively maintained or lost during the taxonomic diversification of the domain *Archaea*. However, the internal nodes representing the duplication events that originated the paralogs I, II and IV do not have enough bootstrap support in the protein ML tree (fig. 6). Therefore, we constructed other ML trees with the multiple alignment of nucleotides containing only unambiguous sites as determined in the Guidance server (Penn et al. 2010). The alternative ML trees were inferred with different nucleotide positions (1^{st} and 2^{nd} or all positions of the codon) and with the aminoacid positions of the protein multiple alignment. These reconstructions yielded alternative evolutionary histories (supplementary fig. S2) that were not always similar to the tree proposed for the PurPs in this study (fig. 6).

Group 1 (PurP 1) is composed of OTUs in four classes of methanogenic *Archaea*: *Methanobacteria*, families *Methanobacteriaceae*, *Methanothermaceae*; *Methanococci*, families *Methanocaldococcaceae* and *Methanococcaceae*; *Methanopyri*, família *Methanopyraceae*, with only 1 OTUs; and *Methanomicrobia*, family *Methanocellaceae*. All OTUs from these classes contain only one copy of *purP* (supplementary table S2). Although *Methanocella paludicola* SANAE presents three copies of the gene, only one copy is classified in group 1. This group includes the PurP from *Methanocaldococcus jannaschii*, the only PurP that had its FAICAR synthase activity functionally characterized (Zhang et al. 2008-II).

PurPs of groups 2 and 3 (PurP II and PurP III) are in the same 61 OTUs of the phyla *Crenarchaeota*, *Euryarchaeota*, *Thaumarchaeota* and *Korarchaeota*

(supplementary table S2). Group 2 is composed of 62 proteins, one in each OTUs, except for *Methanosarcina mazei* Go1 that harbours two PurP II. Group III is composed of 61 proteins, one copy in each OTUs. In some OTUs the genes that encode the PurP II and III are in genomic context with other *pur* genes, what suggests that they are functionally related with the PBP (supplementary table S3). The fourth group is composed of PurP homologs in the genome of six of thirteen OTUs in the family *Thermococcaceae* (supplementary table S2). These six OTUs also harbour the PurPs II and III (supplementary table S2). The genes coding for the PurP IV are in the same genomic context with other *pur* genes (supplementary table S3). The PurPs IV are being described for the first time in this study as a novel putative isoform of PurP in six of thirteen analysed OTUs of the *Thermococcaceae* family (supplementary table S2).

Discussion

The purine biosynthetic pathway (PBP) is ancient and its derivatives are involved in the synthesis of nucleic acid precursors, carbohydrate metabolism and in several cellular signalling pathways. Information on the diversity of the genes encoding the enzymes of the PBP is important to many biotechnological applications, such as the development of anticancer and antimicrobial drugs. In this study, we expand the scientific knowledge on the diversity in the last two steps of the PBP in prokaryotic lineages. To accomplish this, an extensive genome analysis was performed with 1,405 completely sequenced and annotated prokaryotic genomes. Comparative genomics and genomic context analyses showed that the diversity of PBP in the domain *Bacteria* is higher than previously reported. For example, the genes *purV* and *purJ*, initially reported only from *Archaea* and the gene *purO*, considered a signature of the domain *Archaea*, were found in this study to be relatively common in *Bacteria*.

The occurrence of genes coding for PurH in *Bacteria* and *Archaea* was previously reported (Brown et al. 2011; Armenta-Medina et al. 2014). However, our results showed that contrary to what occurs in the domain *Archaea*, the enzyme PurH seems to be the preferred evolutionary alternative selected by most bacterial phyla to catalyse the last two steps in the purine biosynthesis. Genes

encoding PurH were found in 17% of the archaeal genomes analysed and in 85% of the bacterial genomes (fig. 2). The genomic analyses performed in this study, as well as the taxonomical patterns of occurrence indicate that *purO*, *purV* and *purJ* are able to replace *purH* in members of the domain *Bacteria* that lack this gene, similar to what was proposed for *Archaea* (Brown et al. 2011). Genes that replace *purH* occurred in 73% of the archaeal genomes and in 5% of the bacterial genomes included in our analyses (fig. 2). Approximately 37% of the *purO*, *purV* and *purJ* were in genomic context with other genes of the PBP in both bacterial and archaeal genomes and 63% were not in genomic context (table 1). The fact that part of the genes coding for enzymes in the PBP are in genomic context is expected because genes functionally related tend to be in the same operon in prokaryotic genomes (Korbel et al. 2004). The genomic context is commonly used in the prediction of the functional interactions among genes (Huynen et al. 2000).

No homologs of genes that code for PurH, PurO, PurV and PurJ were found in genomes studied of OTUs in the phyla *Crenarchaeota*, *Korarchaeota* and *Thaumarchaeota*, but genes that code for homologs of PurP were found and they may be involved in the PBP, as will be discussed below. A total of 135 genomes did not harbour any of the genes coding for the last two steps of the PBP nor their homologs. These OTUs, 14 *Archaea* and 121 *Bacteria*, representing aproximately 10% of each domain are not able to synthesize purines *de novo*. Many parasites, such as members of the family *Rickettsiaceae*, *Helicobacter pylori*, *Nanoarchaeum equitans* and symbionts such as *Borrelia burgdorferi* and *Serratia symbiotica* acquire purines by recycling pre-existing purines found in the growth substrate through the salvage purine pathway (Waters et al., 2003; Jewett et al., 2009; Liechti and Goldberg, 2011). A number of OTUs in the domains *Archaea* and *Bacteria* lacking these genes are free-living, indicating that they are able to synthesize purines, most certainly through the salvage pathway. Another possibility is that these organisms synthesize purines *de novo* by enzymes that are still unknown to participate in the last two steps of the PBP, a hypothesis that awaits for confirmation.

According to Xu et al. (2007) the fusion of the genes that code for the domains AICARFT and IMPCH originating the PurH was favoured during evolution once FAICAR, product of the AICARFT domain of PurH, is spontaneously

converted into its precursor, AICAR. So, the origin of PurH contributed to accelerate the conversion of FAICAR into IMP by the domain IMPCH. The hypothesis proposed by Xu et al. (2007) for the origin of PurH, implies that the domains AICARFT and IMPCH coded by two distinct genes was the ancestral condition of the PBP and therefore, organisms with the PurH in their genomes would have an adaptive advantage. Zang et al. (2008-I) speculated that although there is no tunnel or channel between the domains AICARFT and IMPCH of PurH, there is a predominance of positive charges between them, what favours the transfer of FAICAR from the domain AICARFT to the domain IMPCH, corroborating the suggestions made by Xu et al. (2007) for the origin of PurH. However, our results show that *purV* and *purJ* are functionally involved in the PBP in bacteria, like was proposed by Brown et al. (2011) for *Archaea*. The domains AICARFT and IMPCH of the human PurH produce peptides able to catalyse their respective enzymatic reactions (Rayl et al. 1996). These results indicate that AICARFT and IMPCH do not need to be fused to perform their catalytic reactions. Therefore, the fact that the domains AICARFT and IMPCH are coded by distinct genes in many living OTUs of the domains *Archaea* and *Bacteria* do not implicate that these two genes cannot catalyse the last two steps of the PBP. The experimental characterization of the peptides encoded by these genes will certainly bring more information on their catalytic mechanisms and evolution.

According to the molecular phylogeny of PurO, it is possible to infer that the ancestral form of this enzyme was present in the common ancestor of *Archaea* and *Bacteria* domains. On the basis of this hypothesis, the absence of *purO* in the majority of the analysed bacterial genomes would be the result of losses of this gene during species diversification in this domain. Taking these results into consideration, *purO* can no longer be considered a signature of the domain *Archaea*, as previously suggested (Graupner et al. 2002; Ownby et al. 2005; Zhang et al. 2008-I; Zhang et al. 2008-II; Armenta-Medina et al. 2014).

Although the *purO* homologs in *Bacteria* and *Archaea* appear to be highly distinct, the combinations of bacterial homologs of *purO* and *purV* genes anticorrelated with *purH* (fig. 2) and part of the bacterial homologs of *purO* are in genomic context with *purV* and other genes of the PBP (table 1). The similarity of the aminoacids in the active sites of archaeal and bacterial PurOs suggests that

they are functionally related (fig. 4). The variation in the conservation of the primary structure of the PurO of *Archaea* and *Bacteria* are coincident (fig 4a), indicating that the selection pressure was similar on the primary structure of these proteins after the divergence of the domains *Archaea* and *Bacteria*. In addition, the conservation of the aminoacid residues of the active site is not related to the conservation of the region of the primary structure in which they are located, even at positions where amino acids residues vary (fig.4a). It suggests that they were conserved in the PurOs of archaeas and their bacterial counterparts despite the evolutionary divergence that results from speciation events. Taken together, these results strongly indicate that the bacterial homologs of *purO* are analogs of archaeal PurOs.

The PurPs are only present in archaeal genomes and the phylogenetic tree of these enzymes shows a division in four groups. Rooting the PurP phylogenetic tree with the shared indels is justifiable because these insertions or deletions of nucleotides or aminoacids have a high importance as phylogenetic markers. These events are rarely fixed and hardly reversed when compared to sequence substitutions (Rokas and Holland 2000). Therefore, genes or proteins that share indels are considered phylogenetically more related (Chan et al. 2007). Indels are used as molecular markers in studies as diverse as protein evolution and taxonomy of microorganisms (Rokas and Holland 2000; Chan et al. 2007; Ajawatanawong and Baldauf 2013; Naushad et al. 2014). Due to the different evolutionary histories recovered in phylogenetic trees constructed with different codon positions in the nucleotides of *purP* or with aminoacid positions with high levels of reliability, the relationships among these PurP paralogs proposed here must be viewed with caution. Perhaps, phylogenetic analyses including a greater number of PurP sequences or other methodological approaches, such as the use of complex networks, can help to solve this problem (Andrade et al. 2011; Carvalho et al. 2015).

The collective analyses of PurPs in archaeal genomes indicate that PurP IV is a new isoform, distinct from the ones previously described in the PBP (Zhang et al. 2008-II). The tertiary structure of the PurP IV from *Thermococcus kodakarensis* was resolved, however, it was not enzymatically characterized (Zhang et al. 2008-II; Brown et al. 2011). The topology of the clades PurP I, II, III and IV is

approximately congruent with the taxonomy of the domain *Archaea*. The incongruences between the phylogeny inferred with the PurPs in this study and the phylogeny of the domain *Archaea* may be the result of both the absence of PurPs in the genome of some OTUs (e.g. in the classes *Halobacteria*, *Thermococci* and *Thermoplasmata;* supplementary text) or events of horizontal gene transfer. For example, transference of PurP I among unrelated methanogenic *Archaea* (supplementary text). The OTUs with PurPs II and III are the same and are in the phyla *Crenarchaeota*, *Euryarchaeota*, *Thaumarchaeota* and *Korarchaeota*. The topology of this clade is approximately congruent with the taxonomy of the domain *Archaea* (fig. 6a; Spang et al. 2010; Podar et al. 2013 Petitjean et al. 2015). The exceptions are the PurP II of *Candidatus Korarchaeum cryptofilum* OPF8, phylogenetically related to the PurP II of phylum *Euryarchaeota* and the clade containing the PurPs III of the families *Archaeoglobaceae* and of *Methanocella paludicola* SANAE, that are phylogenetically related to the PurPs III of OTUs from the phyla *Thaumarchaeota* and *Korarchaeota* (fig. 6a). Considering that the phylum *Euryarchaeota* is monophyletic (Wolf et al. 2012; Petitjean et al. 2015), this incongruences indicate that the PurPs from *Candidatus Korarchaeum cryptofilum* OPF8 and the clade that contains the PurPs III of the family and *Archaeoglobaceae* and of *Methanocella paludicola* SANAE were acquired by horizontal gene transfer.

The PurP was recruited after the divergence of *Archaea* and *Bacteria*, its origin occurred in the ancestor of *Crenarchaeota* and *Euryarchaeota*. Its isoforms originated with a first event of duplication that occurred in the ancestor of *Crenarchaeota* and *Euryarchaeota* giving rise to PurP III and the ancestor of the PurPs I, II and IV followed by a second duplication event that originated the PurP IV and the ancestor of the PurPs I and II and the last duplication originated the PurPs I and II. The PurPs II and III were maintained in the *Crenarchaeota* and *Euryarchaeota* and the other isoforms only in *Euryarchaeota*.

There are evidences suggesting that the PurPs II, III and IV are functionally linked to the PBP. For example, most *Archaea* included in this study that harbour PurPs II and III or II, III and IV do not contain the PurP I or analogous enzymes such as PurH or PurV encoded in their genomes (supplementary table S1). In general, these archaeal OTUs are free-living (supplementary table S2) and the

PurPs II, III and IV of several phylogenetically unrelated *Archaea* are in genomic context with other genes of the PBP or in putative operons (supplementary table S3). These OTUs are widespread in the phyla *Crenarchaeota, Euryarchaeota, Thaumarchaeota* and *Korarchaeota* and represent more than one third of the *Archaea* analysed in this study. In some cases the conservation of the genomic context of the *purP*s extends to OTUs of different genera, such as the *purP*s from *Desulfurococcus kamchatkensis* 1221n and *Ignisphaera aggregans* DSM 17230 or to all OTUs of one order, such as in *Sulfolobales* (supplementary table S3). These are ecological and genomic evidences of the involvement of the PurPs II, III and IV in the PBP.

The PurP I of *Methanocaldococcus jannaschii* was shown to have FAICAR synthase activity, which is the ninth step of the PBP (Ownby et al. 2005). In contrast, neither *Pyrococcus furiosus* PurP II and PurP III have showed any detectable FAICAR synthase activity (Zhang et al. 2008-II). However, the tertiary and quaternary structures of PurP I of *M. jannaschii* and PurP II of *P. furiosus* revealed that their active sites were highly conserved (Zhang et al. 2008-II). Additionally, the PurP II of *P. furiosus* binds to both ATP and AICAR (Zhang et al. 2008-II). Based on these findings, Zhang et al. (2008-II) speculated that the PurP II of *P. furiosus* could utilize an alternative source of formyl to catalyse the conversion of AICAR to FAICAR while the PurP III would have a distinct catalytic activity or would have no catalytic function once it is highly divergent from PurP I and II. The PurP I of *M. jannaschii* and PurP II of *P. furiosus* have an hexameric quaternary structure formed by the interaction between two trimers. However, PurP I has a more compact structure, with approximately 2.5X more buried surface area between the two trimers than PurP II (Zhang et al. 2008-II). This weaker interaction between the trimers of PurP II of *P. furiosus* was attributed to a possible crystallization artefact rather than biologically relevant (Zhang et al. 2008-II). However, all *Archaea* analysed in this study that contain a PurP II also has a PurP III, except for some OTUs of the class *Thermococci*, which harbour the PurPs II, III and IV. Therefore, it is possible that this weaker interaction between the trimers of PurP II of *P. furiosus* is because in its biologically active form, the PurP II and III form heterohexamers composed of trimers with the same isoform.

So, we speculate that in this arrangement the PurPs II and III would be able to catalyse the ninth reaction of the PBP, the conversion of AICAR into FAICAR.

The results of this study contribute to a better understanding of the diversity of the PBP in prokaryotes. The genes *purV*, *purJ* and *purO*, previously reported only in the domain *Archaea* were also found in *Bacteria*, indicating a higher diversity of the PBP in this domain than previously described. In light of these results, *purO* cannot be considered a signature of the domain *Archaea*, as previously reported. Bacterial PurOs were inferred to have catalytic activity due to the conservation of aminoacids in its active site and to participate in the ninth step of the PBP.

According to the hypothesis of Horowitz (1945), enzymes of the last stages of biosynthetic pathways were the first to be recruited during their origin. The last two steps of the PBP show the highest variation when compared to the other steps of this pathway and indeed, this variation is found in both *Archaea* and *Bacteria*. Woese (1998) proposed that the Cenancestor was a diverse community of cells that survived and evolved as a biological unit rather than a discrete entity. This diversity observed in the first recruited enzymes probably resulted from the functional redundancy in ancient times, in accordance with Woese's proposition. Therefore, our results combined with the propositions presented above, are congruent with the hypothesis that the last steps of PBP evolved in the Cenancestor.

Literature cited

Aiba A, Mizobuchi K. 1989. Nucleotide sequence analysis of genes purH and purD involved in the de novo purine nucleotide biosynthesis of Escherichia coli. J Biol Chem. 264(35):21239-46.

Ajawatanawong P, Baldauf SL. 2013. Evolution of protein indels in plants, animals and fungi. BMC Evol Biol. 13:140.

Andrade RFS, et al. 2011. Detecting network communities: an application to phylogenetic analysis. PLoS Comput Biol. 7(5): e1001131.

Armenta-Medina D, Segovia L, Perez-Rueda E. 2014. Comparative genomics of nucleotide metabolism: a tour to the past of the three cellular domains of life. BMC Genomics. 15:800.

Brown AM, Hoopes SL, White RH, Sarisky CA. 2011. Purine biosynthesis in archaea: variations on a theme. Biol Direct. 6:63.

Buchanan JM, Hartman SC. 1959. Enzymatic reactions in the synthesis of purines. Adv Enzymol. 21:199–261.

Caetano-Anollés G, Kim SK, Mittenthal JE. 2007. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. Proc Natl Acad Sci USA. 104:9358-9363.

Carvalho DS, et al. 2015. What are the evolutionary origins of mitochondria? a complex network approach. PLoS One. 10(9): e0134988.

Chan SK, Hsing M, Hormozdiari F, Cherkasov A. 2007. Relationship between insertion/deletion (indel) frequency of proteins and essentiality. BMC Bioinformatics. 28:227.

Cheng YS, et al. 1990. Glycinamide ribonucleotide synthetase from Escherichia coli: cloning, overproduction, sequencing, isolation, and characterization. Biochemistry. 29(1):218-27.

Chopra AK, Peterson JW, Prasad R. 1991, Nucleotide sequence analysis of purH and purD genes from Salmonella typhimurium. Biochim Biophys Acta. 1090(3):351-4.

Crooks GE, Hon G, Chandonia JM, Brenner, SE. 2004. WebLogo: a sequence logo generator. Genome Res. 14:1188-1190.

Euzéby JP. 1997. List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. Int J Syst Bacteriol. 47:590-592.

Gots JS, Dalal FR, Shumas SR. 1969. Genetic separation of the inosinic acid cyclohydrolase-transformylase complex of Salmonella typhimurium. J Bacteriol. 99:441-449.

Graupner M, Xu H, White RH. 2002. New class of IMP cyclohydrolase in Methanococcus jannaschii. J Bacteriol. 184(5):1471-1473.

Gu ZM, Martindale DW, Lee BH. 1992. Isolation and complete sequence of the purL gene encoding FGAM synthase II in Lactobacillus casei. Gene. 119(1):123-6.

Han MV, Zmasek CM. 2009. phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics. 10:356.

He B, Smith JM, Zalkin H. 1992. Escherichia coli purB gene: cloning, nucleotide sequence, and regulation by purR. J Bacteriol. 174(1):130-6.

Horowitz NH. 1945. On the evolution of biochemical syntheses. Proc Natl Acad Sci USA. 31(6):153-157.

Hoskins AA, Anand R, Ealick SE, Stubbe J. 2004. The formylglycinamide ribonucleotide amidotransferase complex from Bacillus subtilis: metabolite-mediated complex formation. Biochemistry. 43(32):10314-27.

Huynen M, Snel B, Lathe W, Bork P. 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. 10(8):1204-1210.

Inglese J, Johnson DL, Shiau A, Smith JM, Benkovic SJ. 1990. Subcloning, characterization, and affinity labeling of Escherichia coli glycinamide ribonucleotide transformylase. Biochemistry. 29(6):1436-43.

Jewett JW, et al. 2009. GuaA and GuaB are essential for Borrelia burgdorferi survival in the tick-mouse infection cycle. J. Bacteriol. 191: 6231-6241.

Kang YN, Tran A, White RH, Ealick SE. 2007. A novel function for the NTN hydrolase fold demonstrated by the structure of an archaeal inosine monophosphate cyclohydrolase. Biochemistry. 46(17):5050–5062.

Kirsch DR, Whitney RR. 1991. Pathogenicity of Candida albicans auxotrophic mutants in experimental infections. Infect Immun. 59(9):3297-3300.

Korbel JO, Jensen LJ, Mering CV, Bork P. 2004. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat Biotechnol. 22(7):911-7.

Liechti G, Goldberg JB. 2011. Helicobacter pylori relies primarily on the purine salvage pathway for purine nucleotide biosynthesis. J Bacteriol. 194:839-854.

Liu Y, et al. 2014. Modification in de novo purine pathway for adenosine accumulation by Bacillus subtilis. Wei Sheng Wu Xue Bao. 54(6):641-7.

Marolewski A, Smith JM, Benkovic SJ. 1994. Cloning and characterization of a new purine biosynthetic enzyme: a non-folate glycinamide ribonucleotide transformylase from E. coli. Biochemistry. 33(9):2531-7.

Naushad HS, Lee B, Gupta RS. 2014. Conserved signature indels and signature proteins as novel tools for understanding microbial phylogeny and systematics: identification of molecular signatures that are specific for the phytopathogenic genera Dickeya, Pectobacterium and Brenneria. Int J Syst Evol Microbiol. 64:366–383.

Ni L, Guan K, Zalkin H, Dixon JE. 1991. De novo purine nucleotide biosynthesis: cloning, sequencing and expression of a chicken PurH cDNA encoding 5-aminoimidazole-4-carboxamide-ribonucleotide transformylase-IMP cyclohydrolase. Gene. 106(2):197-205.

Nilsson D, Kilstrup M. 1998. Cloning and expression of the Lactococcus lactis purDEK genes, required for growth in milk. Appl Environ Microbiol. 64(11):4321-7.

Ownby K, Xu H, White RH. 2005. A Methanocaldococcus jannaschii archaeal signature gene encodes for a 5-Formaminoimidazole-4-carboxamide-1--D-ribofuranosyl 5-Monophosphate synthetase: a new enzyme in purine biosynthesis. J Biol Chem. 280:10881-10887.

Parker J. 1984. Identification of the purC gene product of Escherichia coli. J Bacteriol. 157(3):712-7.

Peltonen T, Mäntsälä P. 1999. Isolation and characterization of a purC(orf)QLF operon from Lactococcus [correction of Lactobacillus] lactis MG1614. Mol Gen Genet. 261(1):31-41.

Penn O, et al. 2010. GUIDANCE: a web server for assessing alignment confidence scores. Nucleic acids research. 38:W23–W28.

Petitjean C, Deschamps P, Lopez-Garcia P, Moreira D, Brochier-Armanet C. 2015. Extending the conserved phylogenetic core of Archaea disentangles the evolution of the third domain of life. Mol Biol Evol. 32:1242–1254.

Podar M, et al. 2013. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone national park. Biol Direct. 8:9.

Pride DT. 2000. Swaap: a tool for analyzing substitutions and similarity in multiple alignments. Distributed by the author.

Rayl EA, Moroson BA, Beardsley GP. 1996. The Human purH gene product, 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase: cloning, sequencing, expression, purification, kinetic analysis, and domain mapping. J Biol Chem. 271(4):2225–2233.

Rokas A, Holland PW. 2000. Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol. 15(11):454-459.

Sampei G, et al. 2010. Crystal structures of glycinamide ribonucleotide synthetase, PurD, from thermophilic eubacteria. J Biochem. 148(4):429-38.

Schrimsher JL, Schendel FJ, Stubbe J, Smith JM. 1986. Purification and characterization of aminoimidazole ribonucleotide synthetase from Escherichia coli. Biochemistry 25(15):4366-71.

Smith PMC, Atkins CA. 2002. Purine biosynthesis: big in cell division, even bigger in nitrogen assimilation. Plant Physiol. 128:793–802.

Spang A, et al. 2010. Distinct gene set in two diferente lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. Trends Microbiol. 18:331-340.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 30(12):2725–9.

Watanabe W, Sampei G, Aiba A, Mizobuchi K. 1989. Identification and sequence analysis of Escherichia coli purE and purK genes encoding 5'-phosphoribosyl-5-amino-4-imidazole carboxylase for de novo purine biosynthesis. J Bacteriol. 171(1):198-204.

Waters E, et al. 2003. The genome of Nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. Proc. Natl. Acad. Sci. USA. 100: 12984-1298.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci USA. 74(11):5088-5090.

Woese CR. 1998. The universal ancestor. Proc Natl Acad Sci USA. 95(12):6854-6859.

Wolf YI, Makarova KS, Yutin N, Koonin EV. 2012. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol Direct. 7:46.

Xu L, et al. 2007. Structure-based design, synthesis, evaluation, and crystal structures of transition state analogue inhibitors of inosine monophosphate cyclohydrolase. J Biol Chem. 282(17):13033–13046.

Zalkin H. 1983. Structure, function, and regulation of amidophosphoribosyltransferase from prokaryotes. Adv Enzyme Regul. 21:225-37.

Zhang Y, Morar M, Ealick SE. 2008-I. Structural biology of the purine biosynthetic pathway. Cell Mol Life Sci. 65(23):3699-724.

Zhang Y, White RH, Ealick SE. 2008-II. Crystal structure and function of 5-formaminoimidazole-4-carboxamide-1-β-d-ribofuranosyl 5′-monophosphate synthetase from Methanocaldococcus jannaschii. Biochemistry. 47(1):205-217.

ANEXOS

(Figuras, Tabelas e Texto Suplementares)

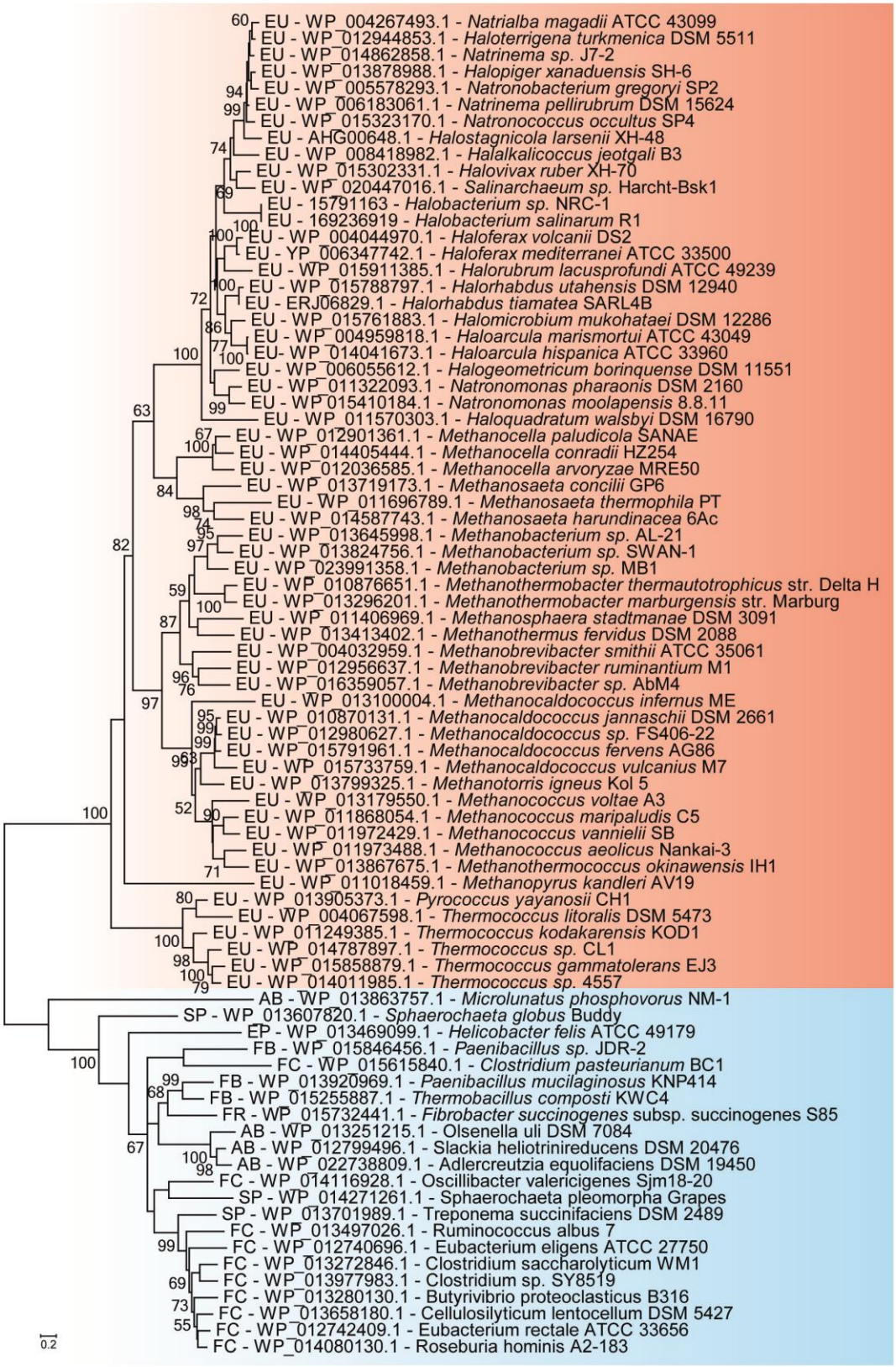
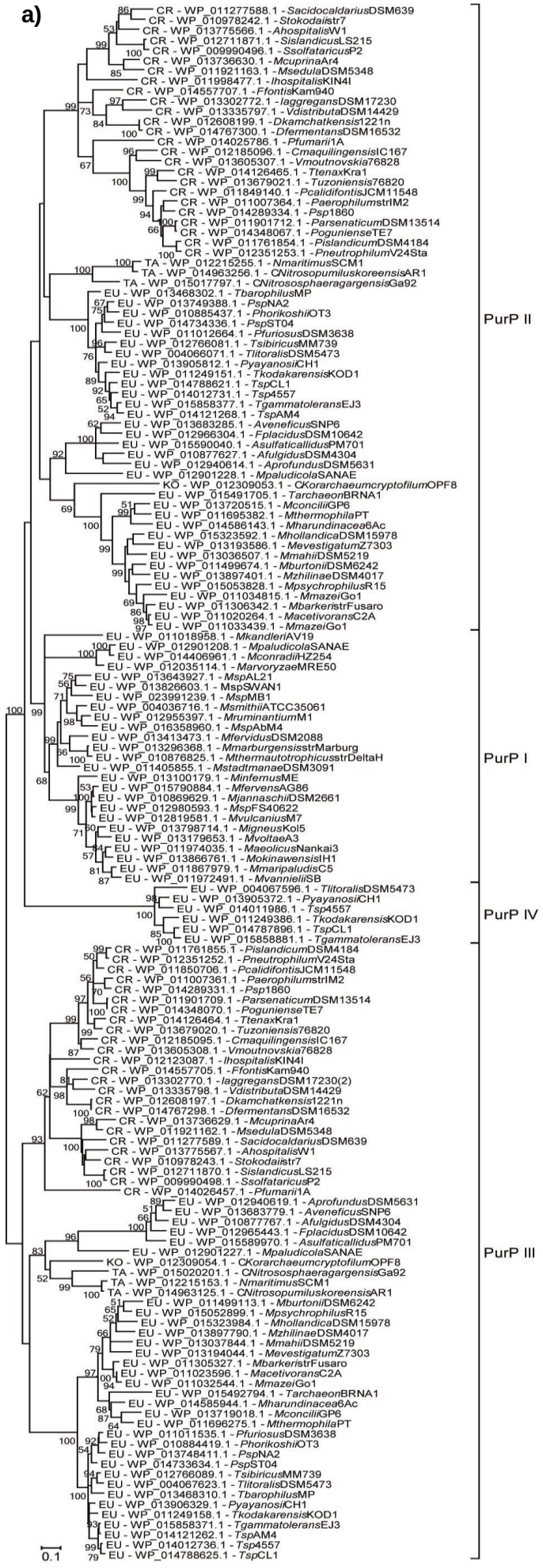
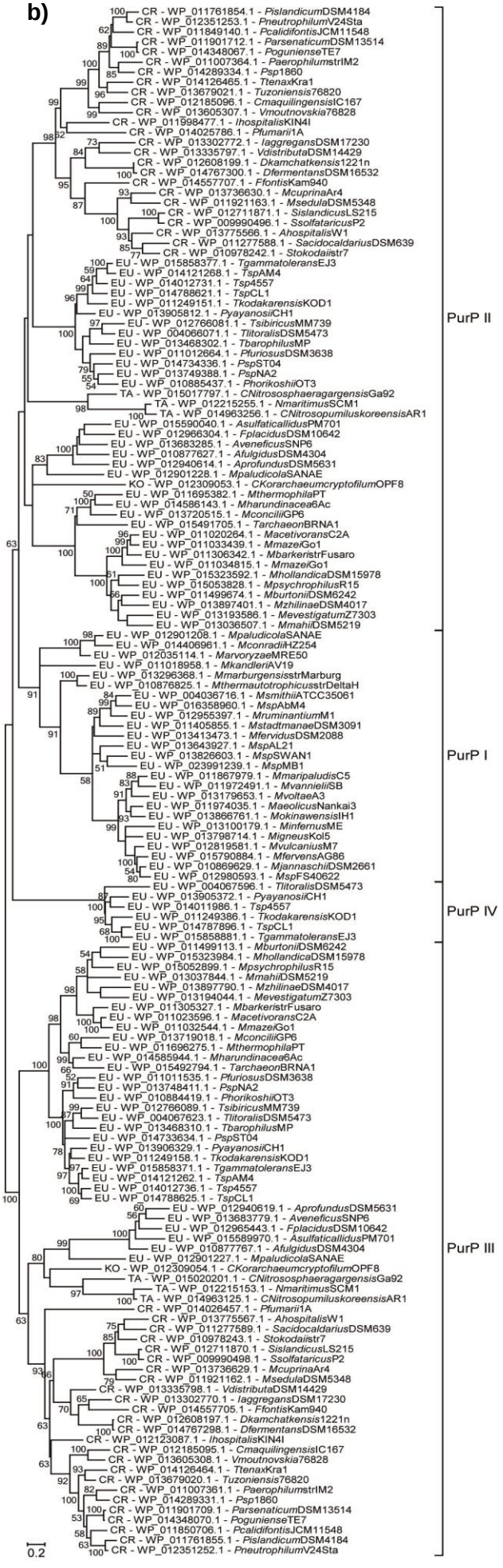


Fig. S1. Phylogenetic tree of 81 nucleotide sequences of purOs from Archaea and Bacteria reconstructed with the maximum likelihood method. The multiple alignment contained 386 sites and the evolutionary model GTR+G+I was used with 1,000 bootstrap resamplings. The scale indicates the number of substitutions per site.


a)
CR - WP_011277588.1 - SacidocaldariusDSM639
CR - WP_010978242.1 - Stokodaiistr7
CR - WP_013775566.1 - AhospitalisW1
CR - WP_012711871.1 - SislandicusLS215
CR - WP_009990496.1 - SsolfataricusP2
CR - WP_013736630.1 - McuprinaAr4
CR - WP_011921163.1 - MsedulaDSM5348
CR - WP_011998477.1 - IhospitalisKIN4I
CR - WP_014557707.1 - FfontisKam940
CR - WP_013302772.1 - IaggregansDSM17230
CR - WP_013335797.1 - VdistributaDSM14429
CR - WP_012608199.1 - Dkamchatkensis1221n
CR - WP_014767300.1 - DfermentansDSM16532
CR - WP_014025786.1 - Pfumarii1A
CR - WP_012185096.1 - CmaquilingensisIC167
CR - WP_013605307.1 - Vmoutnovskia76828
CR - WP_014126465.1 - TtenaxKra1
CR - WP_013679021.1 - Tuzoniensis76820
CR - WP_011849140.1 - PcalidifontisJCM11548
CR - WP_011007364.1 - PaerophilumstrIM2
CR - WP_014289334.1 - Psp1860
CR - WP_011901712.1 - ParsenaticumDSM13514
CR - WP_014348067.1 - PogunienseTE7
CR - WP_011761854.1 - PislandicumDSM4184
CR - WP_012351253.1 - PneutrophilumV24Sta
TA - WP_012215255.1 - NmaritimusSCM1
TA - WP_014963256.1 - CNitrosopumiluskoreensisAR1
TA - WP_015017797.1 - CNitrososphaeragargensisGa92
EU - WP_013468302.1 - TbarophilusMP
EU - WP_013749388.1 - PspNA2
EU - WP_010885437.1 - PhorikoshiiOT3
EU - WP_014734336.1 - PspST04
EU - WP_011012664.1 - PfuriosusDSM3638
EU - WP_012766081.1 - TsibiricusMM739
EU - WP_004066071.1 - TlitoralisDSM5473
EU - WP_013905812.1 - PyayanosiiCH1
EU - WP_011249151.1 - TkodakarensisKOD1
EU - WP_014788621.1 - TspCL1
EU - WP_014012731.1 - Tsp4557
EU - WP_015858377.1 - TgammatoleransEJ3
EU - WP_014121268.1 - TspAM4
EU - WP_013683285.1 - AveneficusSNP6
EU - WP_012966304.1 - FplacidusDSM10642
EU - WP_015590040.1 - AsulfataallidusPM701
EU - WP_010877627.1 - AfulgidusDSM4304
EU - WP_012940614.1 - AprofundusDSM5631
EU - WP_012901228.1 - MpaludicolaSANAE
KO - WP_012309053.1 - CKorarchaeumcryptofilumOPF8
EU - WP_015491705.1 - TarchaeonBRNA1
EU - WP_013720515.1 - MconciliiGP6
EU - WP_011695382.1 - MthermophilaPT
EU - WP_014586143.1 - MharundinaceaA6Ac
EU - WP_015323592.1 - MhollandicaDSM15978
EU - WP_013193586.1 - MevestigatumZ7303
EU - WP_013036507.1 - MmahiiDSM5219
EU - WP_011499674.1 - MburtoniiDSM6242
EU - WP_013897401.1 - MzhilinaeDSM4017
EU - WP_015053828.1 - MpsychrophilusR15
EU - WP_011034815.1 - MmazeiGo1
EU - WP_011306342.1 - MbarkeristrFusaro
EU - WP_011020264.1 - MacetivoransC2A
EU - WP_011033439.1 - MmazeiGo1
EU - WP_011018958.1 - MkandleriAV19
EU - WP_012901208.1 - MpaludicolaSANAE
EU - WP_014406961.1 - MconradiiHZ254
EU - WP_012035114.1 - MarvoryzaeMRE50
EU - WP_013643927.1 - MspAL21
EU - WP_013826603.1 - MspSWAN1
EU - WP_023991239.1 - MspMB1
EU - WP_004036716.1 - MsmithiiATCC35061
EU - WP_012955397.1 - MruminantiumM1
EU - WP_016358960.1 - MspAbM4
EU - WP_013413473.1 - MfervidusDSM2088
EU - WP_013296368.1 - MmarburgensisstrMarburg
EU - WP_010876825.1 - MthermautotrophicusstrDeltaH
EU - WP_011405855.1 - MstadtmanaeDSM3091
EU - WP_013100179.1 - MinfernusME
EU - WP_015790884.1 - MfervensAG86
EU - WP_010869629.1 - MjannaschiiDSM2661
EU - WP_012980593.1 - MspFS40622
EU - WP_012819581.1 - MvulcaniusM7
EU - WP_013798714.1 - MigneusKol5
EU - WP_013179653.1 - MvoltaeA3
EU - WP_011974035.1 - MaeolicusNankai3
EU - WP_013866761.1 - MokinawensisIH1
EU - WP_011867979.1 - MmaripaludisC5
EU - WP_011972491.1 - MvanniellSB
EU - WP_004067596.1 - TlitoralisDSM5473
EU - WP_013905372.1 - PyayanosiiCH1
EU - WP_014011986.1 - Tsp4557
EU - WP_011249386.1 - TkodakarensisKOD1
EU - WP_014787896.1 - TspCL1
EU - WP_015858881.1 - TgammatoleransEJ3
CR - WP_011761855.1 - PislandicumDSM4184
CR - WP_012351252.1 - PneutrophilumV24Sta
CR - WP_011850706.1 - PcalidifontisJCM11548
CR - WP_011007361.1 - PaerophilumstrIM2
CR - WP_014289331.1 - Psp1860
CR - WP_011901709.1 - ParsenaticumDSM13514
CR - WP_014348070.1 - PogunienseTE7
CR - WP_014126464.1 - TtenaxKra1
CR - WP_013679020.1 - Tuzoniensis76820
CR - WP_012185095.1 - CmaquilingensisIC167
CR - WP_013605308.1 - Vmoutnovskia76828
CR - WP_012123087.1 - IhospitalisKIN4I
CR - WP_014557705.1 - FfontisKam940
CR - WP_013302770.1 - IaggregansDSM17230(2)
CR - WP_013335798.1 - VdistributaDSM14429
CR - WP_012608197.1 - Dkamchatkensis1221n
CR - WP_014767298.1 - DfermentansDSM16532
CR - WP_013736629.1 - McuprinaAr4
CR - WP_011921162.1 - MsedulaDSM5348
CR - WP_011277589.1 - SacidocaldariusDSM639
CR - WP_013775567.1 - AhospitalisW1
CR - WP_010978243.1 - Stokodaiistr7
CR - WP_012711870.1 - SislandicusLS215
CR - WP_009990498.1 - SsolfataricusP2
CR - WP_014026457.1 - Pfumarii1A
EU - WP_012940619.1 - AprofundusDSM5631
EU - WP_013683779.1 - AveneficusSNP6
EU - WP_010877767.1 - AfulgidusDSM4304
EU - WP_012965443.1 - FplacidusDSM10642
EU - WP_015589970.1 - AsulfataallidusPM701
EU - WP_012901227.1 - MpaludicolaSANAE
KO - WP_012309054.1 - CKorarchaeumcryptofilumOPF8
TA - WP_015020201.1 - CNitrososphaeragargensisGa92
TA - WP_012215153.1 - NmaritimusSCM1
TA - WP_014963125.1 - CNitrosopumiluskoreensisAR1
EU - WP_011499113.1 - MburtoniiDSM6242
EU - WP_015052899.1 - MpsychrophilusR15
EU - WP_015323984.1 - MhollandicaDSM15978
EU - WP_013897790.1 - MzhilinaeDSM4017
EU - WP_013037844.1 - MmahiiDSM5219
EU - WP_013194044.1 - MevestigatumZ7303
EU - WP_011305327.1 - MbarkeristrFusaro
EU - WP_011023596.1 - MacetivoransC2A
EU - WP_011032544.1 - MmazeiGo1
EU - WP_015492794.1 - TarchaeonBRNA1
EU - WP_014585944.1 - MharundinaceaA6Ac
EU - WP_013719018.1 - MconciliiGP6
EU - WP_011696275.1 - MthermophilaPT
EU - WP_011011535.1 - PfuriosusDSM3638
EU - WP_010884419.1 - PhorikoshiiOT3
EU - WP_013748411.1 - PspNA2
EU - WP_014733634.1 - PspST04
EU - WP_012766089.1 - TsibiricusMM739
EU - WP_004067623.1 - TlitoralisDSM5473
EU - WP_013468310.1 - TbarophilusMP
EU - WP_013906329.1 - PyayanosiiCH1
EU - WP_011249158.1 - TkodakarensisKOD1
EU - WP_015858371.1 - TgammatoleransEJ3
EU - WP_014121262.1 - TspAM4
EU - WP_014012736.1 - Tsp4557
EU - WP_014788625.1 - TspCL1
0.1
PurP II
PurP I
PurP IV
PurP III


b)
CR - WP_011761854.1 - PislandicumDSM4184
CR - WP_012351253.1 - PneutrophilumV24Sta
CR - WP_011849140.1 - PcalidifontisJCM11548
CR - WP_011901712.1 - ParsenaticumDSM13514
CR - WP_014348067.1 - Poguniense TE7
CR - WP_011007364.1 - PaerophilumstrIM2
CR - WP_014289334.1 - Psp1860
CR - WP_014126465.1 - TtenaxKra1
CR - WP_013679021.1 - Tuzoniensis76820
CR - WP_012185096.1 - CmaquilingensisIC167
CR - WP_013605307.1 - Vmoutnovskia76828
CR - WP_011998477.1 - IhospitalisKIN4I
CR - WP_014025786.1 - PfumariiA
CR - WP_013302772.1 - IaggregansDSM17230
CR - WP_013335797.1 - VdistributaDSM14429
CR - WP_012608199.1 - DkamchatkensisA1221n
CR - WP_014767300.1 - DfermentansDSM16532
CR - WP_014557707.1 - FfontisKam940
CR - WP_013736630.1 - McuprinaAr4
CR - WP_011921163.1 - MsedulaDSM5348
CR - WP_012711871.1 - SislandicusLS215
CR - WP_009990496.1 - SsolfataricusP2
CR - WP_013775566.1 - AhospitalisW1
CR - WP_011277588.1 - SacidocaldariusDSM639
CR - WP_010978242.1 - StokodaiistrA7
EU - WP_015858377.1 - TgammatoleransEJ3
EU - WP_014121268.1 - TspAM4
EU - WP_014012731.1 - Tsp4557
EU - WP_014788621.1 - TspCL1
EU - WP_011249151.1 - TkodakarensisKOD1
EU - WP_013905812.1 - PyayanosiiCH1
EU - WP_012766081.1 - TsibiricusMM739
EU - WP_004066071.1 - TlitoralisDSM5473
EU - WP_013468302.1 - TbarophilusMP
EU - WP_011012664.1 - PfuriosusDSM3638
EU - WP_014734336.1 - PspST04
EU - WP_013749388.1 - PspNA2
EU - WP_010885437.1 - PhorikoshiiOT3
TA - WP_015017797.1 - CNitrososphaeragargensisGa92
TA - WP_012215255.1 - NmaritimusSCM1
TA - WP_014963256.1 - CNitrosopumiluskoreensisAR1
EU - WP_015590040.1 - AsulfataciallidusPM701
EU - WP_012966304.1 - FplacidusDSM10642
EU - WP_013683285.1 - AveneficusSNP6
EU - WP_010877627.1 - AfulgidusDSM4304
EU - WP_012940614.1 - AprofundusDSM5631
EU - WP_012901228.1 - MpaludicolaSANAE
KO - WP_012309053.1 - CKorarchaeumcryptofilumOPF8
EU - WP_011695382.1 - MthermophilaPT
EU - WP_014586143.1 - Mharundinacea6Ac
EU - WP_013720515.1 - MconciliiGP6
EU - WP_015491705.1 - TarchaeonBRNA1
EU - WP_011020264.1 - MacetivoransC2A
EU - WP_011033439.1 - MmazeiGo1
EU - WP_011306342.1 - MbarkeristrFusaro
EU - WP_011034815.1 - MmazeiGo1
EU - WP_015323592.1 - MhollandicaDSM15978
EU - WP_015053828.1 - MpsychrophilusR15
EU - WP_011499674.1 - MburtoniiDSM6242
EU - WP_013897401.1 - MzhilinaeDSM4017
EU - WP_013193586.1 - MevestigatumZ7303
EU - WP_013036507.1 - MmahiiDSM5219
EU - WP_012901208.1 - MpaludicolaSANAE
EU - WP_014406961.1 - MconradiiHZ254
EU - WP_012035114.1 - MarvoryzaeMRE50
EU - WP_011018958.1 - MkandleriAV19
EU - WP_013296368.1 - MmarburgensisstrMarburg
EU - WP_010876825.1 - MthermautotrophicusstrDeltaH
EU - WP_004036716.1 - MsmithiiATCC35061
EU - WP_016358960.1 - MspAbM4
EU - WP_012955397.1 - MruminantiumM1
EU - WP_011405855.1 - MstadtmanaeDSM3091
EU - WP_013413473.1 - MfervidusDSM2088
EU - WP_013643927.1 - MspAL21
EU - WP_013826603.1 - MspSWAN1
EU - WP_023991239.1 - MspMB1
EU - WP_011867979.1 - MmaripaludisC5
EU - WP_011972491.1 - MvanniellSB
EU - WP_013179653.1 - MvolteaA3
EU - WP_011974035.1 - MaeolicusNankai3
EU - WP_013866761.1 - MokinawensisIH1
EU - WP_013100179.1 - MinfernusME
EU - WP_013798714.1 - MigneusKol5
EU - WP_012819581.1 - MvulcaniusM7
EU - WP_015790884.1 - MfervensAG86
EU - WP_010869629.1 - MjannaschiiDSM2661
EU - WP_012980593.1 - MspFS40622
EU - WP_004067596.1 - TlitoralisDSM5473
EU - WP_013905372.1 - PyayanosiiCH1
EU - WP_014011986.1 - Tsp4557
EU - WP_011249386.1 - TkodakarensisKOD1
EU - WP_014787896.1 - TspCL1
EU - WP_015858881.1 - TgammatoleransEJ3
EU - WP_011499113.1 - MburtoniiDSM6242
EU - WP_015323984.1 - MhollandicaDSM15978
EU - WP_015052899.1 - MpsychrophilusR15
EU - WP_013037844.1 - MmahiiDSM5219
EU - WP_013897790.1 - MzhilinaeDSM4017
EU - WP_013194044.1 - MevestigatumZ7303
EU - WP_011305327.1 - MbarkeristrFusaro
EU - WP_011023596.1 - MacetivoransC2A
EU - WP_011032544.1 - MmazeiGo1
EU - WP_013719018.1 - MconciliiGP6
EU - WP_011696275.1 - MthermophilaPT
EU - WP_014585944.1 - Mharundinacea6Ac
EU - WP_015492794.1 - TarchaeonBRNA1
EU - WP_011011535.1 - PfuriosusDSM3638
EU - WP_013748411.1 - PspNA2
EU - WP_010884419.1 - PhorikoshiiOT3
EU - WP_012766089.1 - TsibiricusMM739
EU - WP_004067623.1 - TlitoralisDSM5473
EU - WP_013468310.1 - TbarophilusMP
EU - WP_014733634.1 - PspST04
EU - WP_013906329.1 - PyayanosiiCH1
EU - WP_011249158.1 - TkodakarensisKOD1
EU - WP_015858371.1 - TgammatoleransEJ3
EU - WP_014121262.1 - TspAM4
EU - WP_014012736.1 - Tsp4557
EU - WP_014788625.1 - TspCL1
EU - WP_012940619.1 - AprofundusDSM5631
EU - WP_013683779.1 - AveneficusSNP6
EU - WP_012965443.1 - FplacidusDSM10642
EU - WP_015589970.1 - AsulfataciallidusPM701
EU - WP_010877767.1 - AfulgidusDSM4304
EU - WP_012901227.1 - MpaludicolaSANAE
KO - WP_012309054.1 - CKorarchaeumcryptofilumOPF8
TA - WP_015020201.1 - CNitrososphaeragargensisGa92
TA - WP_012215153.1 - NmaritimusSCM1
TA - WP_014963125.1 - CNitrosopumiluskoreensisAR1
CR - WP_014026457.1 - PfumariiA
CR - WP_013775567.1 - AhospitalisW1
CR - WP_011277589.1 - SacidocaldariusDSM639
CR - WP_010978243.1 - StokodaiistrA7
CR - WP_012711870.1 - SislandicusLS215
CR - WP_009990498.1 - SsolfataricusP2
CR - WP_013736629.1 - McuprinaAr4
CR - WP_011921162.1 - MsedulaDSM5348
CR - WP_013335798.1 - VdistributaDSM14429
CR - WP_013302770.1 - IaggregansDSM17230
CR - WP_014557705.1 - FfontisKam940
CR - WP_012608197.1 - DkamchatkensisA1221n
CR - WP_014767298.1 - DfermentansDSM16532
CR - WP_012123087.1 - IhospitalisKIN4I
CR - WP_012185095.1 - CmaquilingensisIC167
CR - WP_013605308.1 - Vmoutnovskia76828
CR - WP_014126464.1 - TtenaxKra1
CR - WP_013679020.1 - Tuzoniensis76820
CR - WP_011007361.1 - PaerophilumstrIM2
CR - WP_014289331.1 - Psp1860
CR - WP_011901709.1 - ParsenaticumDSM13514
CR - WP_014348070.1 - Poguniense TE7
CR - WP_011850706.1 - PcalidifontisJCM11548
CR - WP_011761855.1 - PislandicumDSM4184
CR - WP_012351252.1 - PneutrophilumV24Sta
0.2
PurP II
PurP I
PurP IV
PurP III

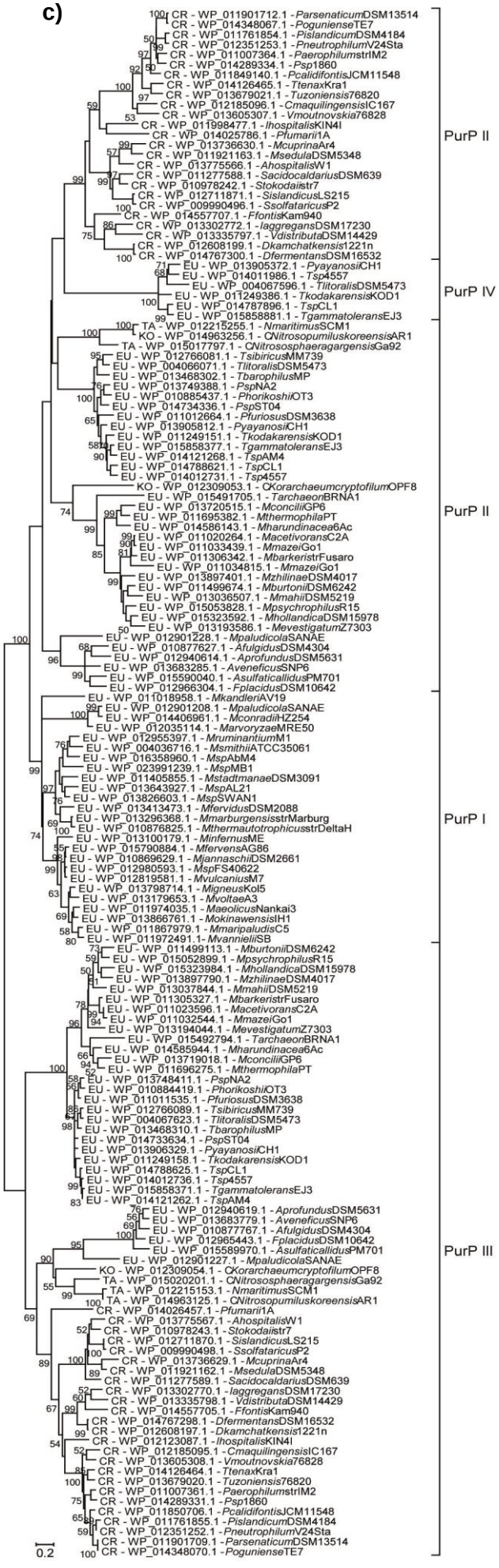

c)
CR - WP_011901712.1 - ParsenaticumDSM13514
CR - WP_014348067.1 - PogunienseTE7
CR - WP_011761854.1 - PislandicumDSM4184
CR - WP_012351253.1 - PneutrophilumV24Sta
CR - WP_011007364.1 - PaerophilumstrIM2
CR - WP_014289334.1 - Psp1860
CR - WP_011849140.1 - PcalidifontisJCM11548
CR - WP_014126465.1 - TtenaxKra1
CR - WP_013679021.1 - Tuzoniensis76820
CR - WP_012185096.1 - CmaquilingensisIC167
CR - WP_013605307.1 - Vmoutnovskia76828
CR - WP_011998477.1 - IhospitalisKIN4I
CR - WP_014025786.1 - PfumariiA
CR - WP_013736630.1 - McuprinaAr4
CR - WP_011921163.1 - MsedulaDSM5348
CR - WP_013775566.1 - AhospitalisW1
CR - WP_011277588.1 - SacidocaldariusDSM639
CR - WP_010978242.1 - Stokodaiistr7
CR - WP_012711871.1 - SislandicusLS215
CR - WP_009990496.1 - SsolfataricusP2
CR - WP_014557707.1 - FfontisKam940
CR - WP_013302772.1 - IaggregansDSM17230
CR - WP_013335797.1 - VdistributaDSM14429
CR - WP_012608199.1 - DkamchatkensisA1221n
CR - WP_014767300.1 - DfermentansDSM16532
EU - WP_013905372.1 - PyayanosiCH1
EU - WP_014011986.1 - Tsp4557
EU - WP_004067596.1 - TlitoralisDSM5473
EU - WP_011249386.1 - TkodakarensisKOD1
EU - WP_014787896.1 - TspCL1
EU - WP_015858881.1 - TgammatoleransEJ3
TA - WP_012215255.1 - NmaritimusSCM1
KO - WP_014963256.1 - CNitrosopumiluskoreensisAR1
TA - WP_015017797.1 - CNitrososphaeragargensisGa92
EU - WP_012766081.1 - TsibiricusMM739
EU - WP_004066071.1 - TlitoralisDSM5473
EU - WP_013468302.1 - TbarophilusMP
EU - WP_013749388.1 - PspNA2
EU - WP_010885437.1 - PhorikoshiiOT3
EU - WP_014734336.1 - PspST04
EU - WP_011012664.1 - PfuriosusDSM3638
EU - WP_013905812.1 - PyayanosiCH1
EU - WP_011249151.1 - TkodakarensisKOD1
EU - WP_015858377.1 - TgammatoleransEJ3
EU - WP_014121268.1 - TspAM4
EU - WP_014788621.1 - TspCL1
EU - WP_014012731.1 - Tsp4557
KO - WP_012309053.1 - CKorarchaeumcryptofilumOPF8
EU - WP_015491705.1 - TarchaeonBRNA1
EU - WP_013720515.1 - MconciliiGP6
EU - WP_011695382.1 - MthermophilaPT
EU - WP_014586143.1 - MharundinaceaA6Ac
EU - WP_011020264.1 - MacetivoransC2A
EU - WP_011033439.1 - MmazeiGo1
EU - WP_011306342.1 - MbarkeristrFusaro
EU - WP_011034815.1 - MmazeiGo1
EU - WP_013897401.1 - MzhilinaeDSM4017
EU - WP_011499674.1 - MburtoniiDSM6242
EU - WP_013036507.1 - MmahiiDSM5219
EU - WP_015053828.1 - MpsychrophilusR15
EU - WP_015323592.1 - MhollandicaDSM15978
EU - WP_013193586.1 - MevestigatumZ7303
EU - WP_012901228.1 - MpaludicolaSANAE
EU - WP_010877627.1 - AfulgidusDSM4304
EU - WP_012940614.1 - AprofundusDSM5631
EU - WP_013683285.1 - AveneficusSNP6
EU - WP_015590040.1 - AsulfatallidusPM701
EU - WP_012966304.1 - FplacidusDSM10642
EU - WP_011018958.1 - MkandleriAV19
EU - WP_012901208.1 - MpaludicolaSANAE
EU - WP_014406961.1 - MconradiiHZ254
EU - WP_012035114.1 - MarvoryzaeMRE50
EU - WP_012955397.1 - MruminantiumM1
EU - WP_00403676.1 - MsmithiiATCC35061
EU - WP_016358960.1 - MspAbM4
EU - WP_023991239.1 - MspMB1
EU - WP_011405855.1 - MstadtmanaeDSM3091
EU - WP_013643927.1 - MspAL21
EU - WP_013826603.1 - MspSWAN1
EU - WP_013413473.1 - MfervidusDSM2088
EU - WP_013296368.1 - MmarburgensisstrMarburg
EU - WP_010876825.1 - MthermautotrophicusstrDeltaH
EU - WP_013100179.1 - MinfernusME
EU - WP_015790884.1 - MfervensAG86
EU - WP_010869629.1 - MjannaschiiDSM2661
EU - WP_012980593.1 - MspFS40622
EU - WP_012819581.1 - MvulcaniusM7
EU - WP_013798714.1 - MigneusKol5
EU - WP_013179653.1 - MvoltaeA3
EU - WP_011974035.1 - MaeolicusNankai3
EU - WP_013866761.1 - MokinawensisIH1
EU - WP_011867979.1 - MmaripaludisC5
EU - WP_011972491.1 - MvanniellSB
EU - WP_011499113.1 - MburtoniiDSM6242
EU - WP_015052899.1 - MpsychrophilusR15
EU - WP_015323984.1 - MhollandicaDSM15978
EU - WP_013897790.1 - MzhilinaeDSM4017
EU - WP_013037844.1 - MmahiiDSM5219
EU - WP_011305327.1 - MbarkeristrFusaro
EU - WP_011023596.1 - MacetivoransC2A
EU - WP_011032544.1 - MmazeiGo1
EU - WP_013194044.1 - MevestigatumZ7303
EU - WP_015492794.1 - TarchaeonBRNA1
EU - WP_014585944.1 - MharundinaceaA6Ac
EU - WP_013719018.1 - MconciliiGP6
EU - WP_011696275.1 - MthermophilaPT
EU - WP_013748411.1 - PspNA2
EU - WP_010884419.1 - PhorikoshiiOT3
EU - WP_011011535.1 - PfuriosusDSM3638
EU - WP_012766089.1 - TsibiricusMM739
EU - WP_004067623.1 - TlitoralisDSM5473
EU - WP_013468310.1 - TbarophilusMP
EU - WP_014733634.1 - PspST04
EU - WP_013906329.1 - PyayanosiCH1
EU - WP_011249158.1 - TkodakarensisKOD1
EU - WP_014788625.1 - TspCL1
EU - WP_014012736.1 - Tsp4557
EU - WP_015858371.1 - TgammatoleransEJ3
EU - WP_014121262.1 - TspAM4
EU - WP_012940619.1 - AprofundusDSM5631
EU - WP_013683779.1 - AveneficusSNP6
EU - WP_010877767.1 - AfulgidusDSM4304
EU - WP_012965443.1 - FplacidusDSM10642
EU - WP_015589970.1 - AsulfatallidusPM701
EU - WP_012901227.1 - MpaludicolaSANAE
KO - WP_012309054.1 - CKorarchaeumcryptofilumOPF8
TA - WP_015020201.1 - CNitrososphaeragargensisGa92
TA - WP_012215153.1 - NmaritimusSCM1
TA - WP_014963125.1 - CNitrosopumiluskoreensisAR1
CR - WP_014026457.1 - PfumariiA
CR - WP_013775567.1 - AhospitalisW1
CR - WP_010978243.1 - Stokodaiistr7
CR - WP_012711870.1 - SislandicusLS215
CR - WP_009990498.1 - SsolfataricusP2
CR - WP_013736629.1 - McuprinaAr4
CR - WP_011921162.1 - MsedulaDSM5348
CR - WP_011277589.1 - SacidocaldariusDSM639
CR - WP_013302770.1 - IaggregansDSM17230
CR - WP_013335798.1 - VdistributaDSM14429
CR - WP_014557705.1 - FfontisKam940
CR - WP_014767298.1 - DfermentansDSM16532
CR - WP_012608197.1 - DkamchatkensisA1221n
CR - WP_012123087.1 - IhospitalisKIN4I
CR - WP_012185095.1 - CmaquilingensisIC167
CR - WP_013605308.1 - Vmoutnovskia76828
CR - WP_014126464.1 - TtenaxKra1
CR - WP_013679020.1 - Tuzoniensis76820
CR - WP_011007361.1 - PaerophilumstrIM2
CR - WP_014289331.1 - Psp1860
CR - WP_011850706.1 - PcalidifontisJCM11548
CR - WP_011761855.1 - PislandicumDSM4184
CR - WP_012351252.1 - PneutrophilumV24Sta
CR - WP_011901709.1 - ParsenaticumDSM13514
CR - WP_014348070.1 - PogunienseTE7
0.2
PurP II
PurP IV
PurP II
PurP I
PurP III

Fig. S2. Maximum likelihood trees of the PurPs reconstructed with different settings. (a) Tree constructed with nucleotides utilising the first and second positions of each codon. (b) Tree constructed with aligned nucleotides of the first, second and third positions of the codon. (c) Tree constructed with aligned aminoacids. Initially two multiple alignments of nucleotides of purP homologs were obtained in the Guidance Server using different pairwise alignment methods: Localpair and Globalpair. The multiple alignments were visually checked and sites aligned differently in these alignments were considered ambiguous and were excluded. The final multiple alignments contained 666 sites, 1,4% with a Guidance confidence score of 0,75 and 98.6% with a Guidance confidence score higher than 0.9. The model utilised for aligned nucleotide trees was GTR+G+I and the matrix for aminoacids was LG+G+I and branch confidence was calculated for all trees with 1,000 bootstrap resamplings. The scale indicates the number of substitutions per site.

Table S2. Archaeal taxa with OTUs contaning PurPs in the genome. The black cicle represents the only archaeal species that is not free-living.

Phylum	Class	Order	Family	OTUs	PurP I	PurP II	PurP III	PurPIV
Crenarchaeota	*Thermoprotei*	*Desulfurococcales*	*Desulfurococcaceae*	*Desulfurococcus fermentans* DSM 16532		+	+	
				Desulfurococcus kamchatkensis 1221n		+	+	
				Ignicoccus hospitalis KIN4/I		+	+	
				Ignisphaera aggregans DSM 17230		+	+	
			Pyrodictiaceae	*Pyrolobus fumarii* 1A		+	+	
		Fervidicoccales	*Fervidicoccaceae*	*Fervidicoccus fontis* Kam940		+	+	
		Sulfolobales	*Sulfolobaceae*	*Acidianus hospitalis* W1		+	+	
				Metallosphaera cuprina Ar-4		+	+	
				Metallosphaera sedula DSM 5348		+	+	
				Sulfolobus acidocaldarius DSM 639		+	+	
				Sulfolobus islandicus L.S.2.15		+	+	
				Sulfolobus solfataricus P2		+	+	
				Sulfolobus tokodaii str. 7		+	+	
		Thermoproteales	*Thermoproteaceae*	*Caldivirga maquilingensis* IC-167		+	+	
				Pyrobaculum aerophilum str. IM2		+	+	
				Pyrobaculum arsenaticum DSM 13514		+	+	
				Pyrobaculum calidifontis JCM 11548		+	+	
				Pyrobaculum islandicum DSM 4184		+	+	
				Pyrobaculum oguniense TE7		+	+	
				Pyrobaculum sp. 1860		+	+	
				Pyrobaculum neutrophilum V24Sta		+	+	
				Thermoproteus tenax Kra 1		+	+	
				Thermoproteus uzoniensis 768-20		+	+	
				Vulcanisaeta distributa DSM 14429		+	+	
				Vulcanisaeta moutnovskia 768-28		+	+	
Euryarchaeota	*Archaeoglobi*	*Archaeoglobales*	*Archaeoglobaceae*	*Archaeoglobus fulgidus* DSM 4304		+	+	
				Archaeoglobus profundus DSM 5631		+	+	
				Archaeoglobus sulfaticallidus PM70-1		+	+	
				Archaeoglobus veneficus SNP6		+	+	
				Ferroglobus placidus DSM 10642		+	+	
	Methanobacteria	*Methanobacteriales*	*Methanobacteriaceae*	*Methanobacterium lacus* AL-21	+			
				Methanobacterium sp. MB1	+			
				Methanobacterium paludis SWAN1	+			
				Methanobrevibacter sp. AbM4	+			
				Methanobrevibacter ruminantium M1	+			
				Methanobrevibacter smithii ATCC 35061	+			
				Methanosphaera stadtmanae DSM 3091	+			
				Methanothermobacter marburgensis str. Marburg	+			
				Methanothermobacter thermautotrophicus str. Delta H	+			
			Methanothermaceae	*Methanothermus fervidus* DSM 2088	+			
	Methanococci	*Methanococcales*	*Methanocaldococcaceae*	*Methanocaldococcus fervens* AG86	+			
				Methanocaldococcus infernus ME	+			
				Methanocaldococcus jannaschii DSM 2661	+			
				Methanocaldococcus sp. FS406-22	+			
				Methanocaldococcus vulcanius M7	+			
				Methanotorris igneus Kol 5	+			
			Methanococcaceae	*Methanococcus aeolicus* Nankai-3	+			
				Methanococcus maripaludis C5	+			
				Methanococcus vannielii SB	+			
				Methanococcus voltae A3	+			
				Methanothermococcus okinawensis IH1	+			
	Methanomicrobia	*Methanocellales*	*Methanocellaceae*	*Methanocella arvoryzae* MRE50	+			
				Methanocella conradii HZ254	+			
				Methanocella paludicola SANAE	+	+	+	
		Methanosarcinales	*Methanosaetaceae*	*Methanosaeta concilii* GP6		+	+	
				Methanosaeta harundinacea 6Ac		+	+	
				Methanosaeta thermophila PT		+	+	
			Methanosarcinaceae	*Methanococcoides burtonii DSM* 6242		+	+	
				Methanohalobium evestigatum Z-7303		+	+	
				Methanohalophilus mahii DSM 5219		+	+	
				Methanolobus psychrophilus R15		+	+	
				Methanomethylovorans hollandica DSM 15978		+	+	
				Methanosalsum zhilinae DSM 4017		+	+	
				Methanosarcina acetivorans C2A		+	+	
				Methanosarcina barkeri str. Fusaro		+	+	
				Methanosarcina mazei Go1		+	+	
	Methanopyri	*Methanopyrales*	*Methanopyraceae*	*Methanopyrus kandleri* AV19	+			
	Thermococci	*Thermococcales*	*Thermococcaceae*	*Pyrococcus furiosus* DSM 3638		+	+	
				Pyrococcus horikoshii OT3		+	+	
				Pyrococcus sp. NA2		+	+	
				Pyrococcus sp. ST04		+	+	
				Pyrococcus yayanosii CH1		+	+	+
				Thermococcus barophilus MP		+	+	
				Thermococcus gammatolerans EJ3		+	+	+
				Thermococcus kodakarensis KOD1		+	+	+
				Thermococcus litoralis DSM 5473		+	+	+
				Thermococcus sibiricus MM 739		+	+	
				Thermococcus sp. 4557		+	+	+
				Thermococcus sp. AM4		+	+	
				Thermococcus sp. CL1		+	+	+
	Thermoplasmata	*Thermoplasmatales*	*?*	*Thermoplasmatales archaeon* BRNA1 ●		+	+	
Korarchaeota	*?*	*?*	*?*	Candidatus *Korarchaeum cryptofilum* OPF8		+	+	
Thaumarchaeota	*?*	*Nitrosopumilales*	*Nitrosopumilaceae*	Candidatus *Nitrosopumilus koreensis* AR1		+	+	
				Nitrosopumilus maritimus SCM1		+	+	
	?	*Nitrososphaerales*	*Nitrososphaeraceae*	Candidatus *Nitrososphaera gargensis* Ga9.2		+	+	

Table S3. OTUs in which the genes encoding PurPs II, III and IV are in genomic context with other genes of the purine biosynthetic pathway.

Phylum	Class	Order	OTUs	Genomic Context
Crenarchaeota	Thermoprotei	Desulfurococcales	*Desulfurococcus fermentans* DSM 16532	purF purD purT purE purM **purP III** purB **purP II**
			Desulfurococcus kamchatkensis 1221n	purC purS purQ purL purF purF purD purT purE purM **purP III** purB **purP II**
			Ignisphaera aggregans DSM 17230	purC purS purQ purL purF purF purD purT purE purM **purP III** purB **purP II**
		Fervidicoccales	*Fervidicoccus fontis* Kam940	purC purF purD purT purE purS purQ purL purF purM **purP III** purB **purP II**
		Sulfolobales	*Acidianus hospitalis* W1	purB **purP II** **purP III** purA
			Metallosphaera cuprina Ar-4	purB **purP II** **purP III** purA
			Metallosphaera sedula DSM 5348	purB **purP II** **purP III** purA
			Sulfolobus acidocaldarius DSM 639	purB **purP II** **purP III** purA
			Sulfolobus islandicus L.S.2.15	purB **purP II** **purP III** purA
			Sulfolobus solfataricus P2	purB **purP II** **purP III** purA
			Sulfolobus tokodaii str. 7	purB **purP II** **purP III** purA
		Thermoproteales	*Caldivirga maquilingensis* IC-167	purA purB **purP III** **purP II**
			Pyrobaculum islandicum DSM 4184	**purP III** **purP II**
			Thermoproteus tenax Kra 1	**purP III** **purP II**
			Thermoproteus uzoniensis 768-20	**purP III** **purP II**
			Vulcanisaeta distributa DSM 14429	**purP II** **purP III**
			Vulcanisaeta moutnovskia 768-28	**purP III** **purP II** purS purQ purT purD purC
Euryarchaeota	Methanomicrobia	Methanocellales	*Methanocella paludicola* SANAE	**purP III** **purP II**
	Thermococci	Thermococcales	*Pyrococcus furiosus* DSM 3638	purD **purP III**
			Pyrococcus horikoshii OT3	purD **purP III**
			Pyrococcus sp. NA2	purD **purP III**
			Pyrococcus sp. ST04	purD **purP III**
			Pyrococcus yayanosii CH1	purD **purP III** / purC **purP IV** purO
			Thermococcus barophilus MP	purM purC purF **purP II** / purT purE purD **purP III** Other purS purQ purL
			Thermococcus gammatolerans EJ3	purS purQ purL **purP II** / purD **purP III** / purC **purP IV** purO
			Thermococcus kodakarensis KOD1	purL **purP II** / purD **purP III** purS purQ / purC **purP IV** purO
			Thermococcus litoralis DSM 5473	purM purT purE purD **purP III** / purC **purP IV** purO
			Thermococcus sibiricus MM 739	purD **purP III** purS purQ purL
			Thermococcus sp. 4557	purM purT purE purD **purP III** purS purQ Other purL **purP II** / purC **purP IV** purO
			Thermococcus sp. AM4	purS purQ purL **purP II** / purD **purP III**
			Thermococcus sp. CL1	purM purT purE purD **purP III** purS purQ purL **purP II** / purC **purP IV** purO
Korarchaeota	?	?	Candidatus *Korarchaeum cryptofilum* OPF8	**purP III** **purP II**

Supplementary text:

The phylum *Euryarchaeota* is monophyletic, however methanogenic *Archaea* are not monophyletic and they are divided in two unrelated classes (Yutin et al. 2012; Wolf et al. 2012; Petitjean et al. 2015). The methanogenic class I includes OTUs from the orders *Methanobacteriales*, *Methanococcales* and *Methanopyrales* that are from the classes *Methanobacteria*, *Methanococci* and *Methanopyri*, respectively. These classes are phylogenetically related. The methanogenic class II includes OTUs from *Methanocellales*, *Methanosarcinales* and *Methanomicrobiales*, all orders from the class *Methanomicrobia*.

The clade PurP I contains methanogenic OTUs from the class I (orders *Methanobacteriales*, *Methanococcales* and *Methanopyrales*) and from order *Methanocellales*, a class II methanogenic higher taxa (fig. 6a). The evolutionary relations between the PurP I of methanogenic *Archaea* class I is congruent with the phylogeny of the *Archaea* domain (Wolf et al. 2012; Petitjean et al. 2015). The exception is the close phylogenetic relationship between the PurP I from *Methanocellales* with the PurPs from the order *Methanobacteriales* (fig. 6a). This suggests that the former ones were acquired in a horizontal gene transfer event from an ancestor of a methanogenic *Archaea* class I. *Methanosarcina mazei* Go1 is the only OTUs from *Methanosarcinales* (methanogenic *Archaea* class II) that harbours three PurPs, two in group II and the other one in group III. All other OTUs analysed from this order possess two PurPs, one in group II and the other one in group III. The topology of the tree indicates that one of the copies of PurP II of *M. mazei* Go1 was acquired from *Methanosarcina acetivorans* C2A in an event of horizontal gene transfer. One alternative hypothesis is that the PurPs II from *M. mazei* Go1 originated in a duplication event in the ancestor of the species of the genus *Methanosarcina* and one of the copies was lost in the other OTUs of the genus included in this study (fig. 6a).

Methanomicrobiales is the only order of class *Methanomicrobia* (methanogenic *Archaea* class II) that does not possess PurP-coding genes in their genomes and all OTUs of this order contain the gene *purH*, where the AICARFT domain is analogous to PurP (supplementary table S1). Some studies did not recover the monophyly of class *Methanomicrobia* and show that the order *Methanomicrobiales* is phylogentically related to the order *Halobacteriales* (class

Halobacteria) (Brochier-Armanet et al. 2008; Podar et al. 2013; Petitjean et al. 2015). The OTUs of *Halobacteriales* also do not contain purP-coding genes, but they harbour *purV*, which is a putative analog of PurP (supplementary table S1). The absence of *purP* in the genome of OTUs of the orders *Methanomicrobiales* and *Halobacteriales* might be a shared characteristic due to their close phylogenetic relationship. This is maybe an additional evidence that the class *Methanomicrobia* is not monophyletic, a question still under debate.

Additional literature cited:

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. 2008. Mesophilic crenarchaeota: proposal for a third archaeal phylum, the *Thaumarchaeota*. Nat Rev Microbiol. 6:245-252.

Petitjean C, Deschamps P, Lopez-Garcia P, Moreira D, Brochier-Armanet C. 2015. Extending the conserved phylogenetic core of Archaea disentangles the evolution of the third domain of life. Mol Biol Evol. 32:1242–1254.

Podar M, et al. 2013. Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone national park. Biol Direct. 8:9.

Wolf YI, Makarova KS, Yutin N, Koonin EV. 2012. Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer. Biol Direct. 7:46.

Yutin N, Puigbo P, Koonin EV, Wolf YI. 2012. Phylogenomics of prokaryotic ribosomal proteins. PLoS One. 7:1-10.

CAPÍTULO 2

Diversidade e evolução de *purV* e *purJ*: Prováveis novos genes da via biossintética de purinas

Artigo a ser submetido à revista Genome Biology and Evolution (GBE)

RESUMO

Cruz, DCB. Diversidade e evolução de *purV* e *purJ*: Prováveis novos genes da via biossintética de purinas

O estudo da via biossintética de purinas (VBP) tem contribuído de forma significativa no desenvolvimento de fármacos anticâncer, antimicrobianos, na área da biotecnologia proporcionando a produção de compostos de valor comercial a partir de microrganismos, e também na área agrícola, uma vez que suas enzimas estão diretamente relacionadas a fatores como crescimento e virulência de fitopatógenos. O fosforribosil-pirofosfato (PRPP) é o precursor da via de purinas, ele é convertido a inosina-monofosfato (IMP) através de 10 etapas enzimáticas. O nono e décimo passos da VBP é originalmente realizado pela PurH nas Bacterias e pela PurP e PurO nas Archaeas, respectivamente. No capítulo 1 relatamos a existência da PurO, PurV e PurJ no Domínio Bacteria, novas enzimas da VBP que provavelmente também estão relacionadas com o nono e décimo passo da via. Como discutido anteriormente a PurV e PurJ são homólogos aos domínios AICARFT e IMPCH da PurH entretanto ainda não se sabe qual teria sido a origem delas, se foram originados a partir dos domínios ancestrais que deram origem a PurH ou a partir de quebras do gene da PurH, devido a isso o objetivo desse trabalho foi primeiramente compreender melhor a diversidade dos genes envolvidos com o nono e décimo passos da VBP nos genomas procarióticos, buscar evidencias que nos permitam inferir que PurV e PurJ são ativas na via e entender a relação evolutiva existente entre elas e a PurH. Foi realizada uma genômica comparativa com 2735 genomas procarióticos completamente sequenciados, buscando pelos genes *purH, purV, purJ, purP* e *purO*. Apesar do número de genomas ser quase o dobro do capítulo 1 os resultados da genômica comparativa foram basicamente os mesmos. Foi realizada uma análise dos aminoácidos do sítio ativo dos AICARFTs e IMPCHs de todas as PurHs recuperadas e PurVs e PurJs recuperados, com isso foi possível observar que apesar de algumas variações existe conservação de aminoácidos do sítio ativo de PurV e PurJ inclusive de aminoácidos que foram descritos como essenciais para

a atividade da enzima. O padrão de conservação da estrutura primária de AICARFTs/PurVs e IMPCH/PurJ é similar, eles apresentam basicamente as mesmas regiões conservadas. Na árvore da PurH foi possível observar uma clara divisão entre Gram positivas e Gram negativas. Dos 32 filos amostrados apenas 8 foram monofiléticos para PurH, entretando a monofilia foi observada também em taxas inferiores como família, ordem ou classe. As PurHs de Archaea ficaram dispersas e provavelmente teriam sido adquiridas a partir de transferência lateral de genes. A topologia que a árvore das PurJs com os domínios IMPCHs das PurHs assumiu, reflete que os *purJs* não tiveram uma única origem e provavelmente surgiram a partir de quebras de *purH,* uma vez que parte deles agruparam com os IMPCHs do mesmo filo. A PurV por outro lado formou grupos distintos na árvore e não estão relacionados com os AICARFTs das PurHs, indicando que provavelmente tenham sido originadas a partir do domínio ancestral.

Palavras-chave: Inosina, IMP, AICARFT, IMPCH, PurH.

ABSTRACT

Cruz, DCB. Diversity and evolution of *purV* and *purJ*: Probable new genes of the purine biosynthetic pathway

The study of the purine biosynthetic pathway (PBP) has contributed significantly to the development of anticancer and , antimicrobial drugs, , providing the production of compounds of commercial value from microorganisms, and also in the agricultural area, since their enzymes are directly related to factors such as growth and virulence of phytopathogens. Phosphoribosyl-pyrophosphate (PRPP) is the precursor of the purine pathway, it is converted to inosine monophosphate (IMP) through 10 enzymatic steps. The ninth and tenth steps of PBP are originally performed by PurH in Bacteria and by PurP and PurO in Archaea, respectively. In Chapter 1 we report the existence of PurO, PurV and PurJ in the Bacterial Domain, new enzymes of the PBP that are probably also related to the ninth and tenth passage of the pathway. As discussed earlier, PurV and PurJ are homologous to the AICARFT and IMPCH domains of PurH, however, it is not yet known what their origin would have been if they originated from the ancestral domains that gave rise to PurH or from the PurH gene , due to this the objective of this work was first to better understand the diversity of the genes involved with the ninth and tenth steps of the PBP in the prokaryotic genomes, to find evidence that allow us to infer that PurV and PurJ are active in the pathway and to understand the evolutionary relation between they are the PurH. A comparative genomic was performed with 2735 completely sequenced prokaryotic genomes, searching for the purH, purV, purJ, purP and purO genes. Although the number of genomes is almost double that of chapter 1 the results of comparative genomics were basically the same. An analysis of the amino acids of the active site of the AICARFTs and IMPCHs of all the recovered PurHs and PurVs and PurJs recovered was performed, with it being possible to observe that despite some variations there is conservation of amino acids from the PurV and PurJ active site including amino acids that were described as essential for the activity of the enzyme. The conservation pattern of the primary structure of AICARFTs/PurVs and IMPCH/PurJ is similar, they present basically the same conserved regions. In the PurH tree it

was possible to observe a clear division between Gram positive and Gram negative. Of the 32 phyla sampled only 8 were monophyletic for PurH, while monophyly was also observed at lower rates such as family, order or class. The Archaea PurHs were dispersed and probably would have been acquired from lateral gene transfer. The topology that the PurJs tree with the PurHs IMPCH domains took, reflects that the purJs did not have a single origin and probably arose from purH breaks, since some of them clustered with the IMPCHs of the same edge. The PurV on the other hand formed distinct groups in the tree and are not related to the PurHs AICARFTs, indicating that they probably originated from the ancestral domain.

Key words: Inosine, IMP, AICARFT, IMPCH, PurH.

Introdução

A via biossintética de purinas (VBP) é uma das mais antigas, logo, o sua evolução está intimamente relacionada com a evolução e diversificação dos seres vivos (Caetano-Anolles et al. 2007). Além disso, o estudo da VBP tem contribuído de forma importante na saúde, com o desenvolvimento de drogas anticâncer e antimicrobianas (Xu et al. 2007; Firestine et al. 2009; Nours et al. 2011; Tranchimand et al. 2011; Baggott & Tamura, 2015). A inativação das enzimas da VBP está relacionada com fenótipos agronomicamente importantes, como a perda de virulência de espécies fitopatogênicas redução na eficiência de associação simbiótica microrganismo-planta e também interferência no processo de conidiação de fungos que exercem controle biológico (Xie et al. 2005; Park et al. 2007; Qin et al. 2011; Yuan et al. 2013).

Na maioria dos organismos, existem dois caminhos para a produção das purinas, a via biossintética de novo, onde a síntese ocorre a partir do fosforibosilpirofosfato (PRPP), e a via de salvamento, onde as purinas são recuperadas dos produtos resultantes da degradação dos ácidos nucleicos ou coenzimas (Xu et al. 2007; Nours et al. 2011). A via biossintética de purinas de novo é um processo celular caro, pois pode consumir até 10 moléculas de ATP para produzir 1 molécula de Inosina monofosfato (IMP) que é o produto final da via (Peifer et al. 2012), entretanto é a forma mais rápida de obtenção de purinas pela célula (Xu et al. 2007). Já a via de salvamento converte nucleobases extracelulares, ou degrada os compostos de purinas até os nucleotídeos ou nucleosídeos correspondente, gastando menos ATPs que a via de novo e assim sendo mais econômico para a célula (Peifer et al. 2012). Na via de novo, o IMP é produzido a partir de 10 passos enzimáticos, ele é a primeira purina, e a partir dele passos enzimáticos adicionais irão realizar a conversão nas demais purinas, como a adenina, guanina, xantosina, hypoxantina entre outras (Peifer et al. 2012).

A quantidade de etapas enzimáticas e de enzimas envolvidas na via de novo pode variar entre os grupos taxonômicos. Nos animais, por exemplo, o sexto passo da via é realizado pela enzima PurE II, enquanto em procariotos, fungos e plantas esse passo é realizado pelas enzimas PurK e PurE I. O terceiro passo

pode ser realizado pela PurN ou pela PurT, os eucariotos não tem PurT, e apesar dela estar presente em procariotos, o mais comum nas archaeas e bacterias é a PurN (Zhang et al. 2008).

O nono e décimo passo da VBP também apresentam variações, em bactérias e eucariotos essas etapas são realizadas pela enzima PurH, enquanto nas archaeas, esses passos são catalisados principalmente pela PurP e PurO, respectivamente. A PurH é uma enzima bifuncional, que apresenta dois domínios, o domínio AICARFT realiza o nono passo da via, onde o intermediário AICAR é convertido em FAICAR, e o domínio IMPCH que realiza o decimo passo, convertendo FAICAR em IMP (Zhang et al. 2008). Brown e Colaboradores (2011), descreveram a existência da PurH em algumas espécies do domínio *Archaea*, assim como a ocorrência dos seus domínios em uma forma não fusionada, como genes independentes. Em um trabalho anterior (Dados ainda não publicados), nós encontramos os domínios da PurH como genes independentes no Domínio *Bacteria*, os quais nomeamos *purV* (domínio AICARFT) e *purJ* (domínio IMPCH), encontramos também homólogos bacterianos de *purO,* algo inédito até o momento.

No nosso trabalho anterior mostramos que a distribuição de *purV* e *purJ* nos Domínios *Archaea* e *Bacteria* é significativa, dos 31 Filos analisados 25 apresentam *purV* ou *purJ*, entretanto ainda não estudamos qual seria a origem desses genes e qual a relação evolutiva deles com *purH.* Podemos inferir que existem duas hipóteses para a origem dos *purVs* e *purJs* encontrados atualmente em linhagens de *Archaea* e *Bacteria*, esses genes podem ter sido originados a partir dos domínios ancestrais que originaram a PurH, ou ser resultado de eventos de quebra do gene da PurH.

Devido a isso o objetivo do presente trabalho foi, primeiramente expandir o nosso conhecimento sobre a distribuição e diversidade das enzimas do nono e décimo passos da VBP, buscar novas evidencias que nos permitam associar os novos genes *purV* e *purJ* funcionalmente à VBP e principalmente conhecer a história evolutiva deles, a fim de saber se foram originados a partir do domínio ancestral ou da quebra do *purH*.

Material e Métodos

Busca pelos genes do nono e décimo passo da VBP nos genomas procarióticos

Inicialmente foi criada uma lista de espécies com genoma completamente sequenciado, depositados nas bases de dados do NCBI (*National Center for Biothecnology Information*) até o mês de Março de 2017. Estas espécies aqui foram tratadas como Unidades Taxonômicas Operacionais (OTUs). Esta lista foi organizada taxonomicamente, e teve um total de 2735 OTUs, englobando os Domínios *Archaea* e *Bacteria*.

As buscas foram realizadas para todos os genes envolvidos com a nona e decima etapa da VBP, *purH, purV, purJ, purP* e *purO,* entretanto apenas as sequências de nucleotídeos homologas de *purH*, *purV* e *purJ* foram recuperadas para a realização das análises seguintes. Para isso foram utilizados os algoritmos Blastp e tBlastn do programa BLAST disponível no NCBI, utilizando as bases de dados “Non-redundant protein sequences (nr)” e “Nucleotide collection (nr/nt)” respectivamente.

A sequência de aminoácidos da PurH (NP_418434.1) de *Escherichia coli* foi utilizada como *query* para as buscas por homólogos da PurH. A região que codifica o domínio AICARFT dessa mesma PurH foi utilizada como *query* para as buscas de *purV* e a região que codifica o domínio IMPCH foi utilizada como *query* para as buscas de *purJ.* Para as buscas por PurP e PurO foram utilizadas como *query* as proteínas homólogas de *Methanocaldococcus jannaschii* (WP_010869629.1 e Q58043.2 respectivamente). Essas proteínas foram escolhidas por já terem sido descritas e caracterizadas funcionalmente.

A presença e ausência de *purH, purV, purJ, purP* e *purO* e o número de cópias em cada genoma analisado foi anotado, assim como o contexto genômico em que os novos genes *purV* e *purJ* estão inseridos. A análise do contexto genômico dos genes foi realizada através da ferramenta de visualização de contexto genômico *Gene Context Tool NG* (Martinez-Guerrero et al. 2008) e

também com o programa Artemis (Rutherford et al. 2000) utilizando os dados genômicos disponíveis no NCBI.

Análise da estrutura primária e secundária de PurV e PurJ

A conservação dos aminoácidos da estrutura primária das proteínas foi estudada através da análise de *Sliding window plot* com o programa Geneious 11.1.5 (Kearse et al. 2012) a partir do alinhamento múltiplo das proteínas. Utilizando uma janela de tamanho 1, o programa calculou o percentual de identidade de cada posição do alinhamento múltiplo de proteínas. Para essa análise foram considerados apenas os sítios conservados entre AICARFTs e PurV (242 sítios), e IMPCHs e PurJ (144 sítios), os mesmos utilizados para as análises filogenéticas.

Para analisar a conservação dos aminoácidos do sítio ativo das PurVs e PurJs, em relação aos domínios AICARFT e IMPCH da PurH, foram construidos LOGOS, comparando-se os aminoácidos dos sítios ativos dos domínios AICARFT e IMPCH de todas as PurHs recuperadas com todas as PurVs e PurJs recuperadas, respectivamente, e também com a PurH humana (NP_004035.2) que foi utilizada como referência (Nours et al. 2011; Axelrod et al. 2008). Os LOGOS foram criados com o programa WEB LOGO (Crooks et al. 2004). A análise da estrutura secundária de PurV e PurJ foi baseada na predição de estrutura secundaria da PurH humana (1P4R) disponível no Protein Data Bank (PDB).

Alinhamentos múltiplos e filogenia de PurH, PurV e PurJ

As sequências de aminoácidos da PurH, PurV e PurJ, foram alinhadas no programa MAFFT (Katoh et al. 2017) e posteriormente editadas no programa MEGA 6 (Tamura et al. 2013).

O alinhamento das PurHs foi separado nas partes correspondentes aos domínios AICARFTs e IMPCHs, estes foram alinhados e editados separadamente e depois unidos para realização da filogenia da PurH. A parte correspondente aos

AICARFTs também foi unida às PurVs e o mesmo foi realizado com os IMPCHs e PurJs, gerando assim 3 alinhamentos. As árvores filogenéticas foram realizadas com o FastTree (Price et al. 2009).

Resultados

Análise de genômica comparativa

A busca pelos genes *purH*, *purV*, *purJ*, *purP* e *purO* foi realizada em 2735 OTUs com genomas completamente sequenciados, 205 do Domínios *Archaea* e 2530 do Domínio *Bacteria*. Foram recuperadas 2257 sequencias homólogas de *purH*, 42 no Domínio *Archaea* e 2215 no Domínio *Bacteria*. Além disso, foram recuperadas 4 homólogos de *purJ* em *Archaea* e 70 em *Bacteria*, assim como, 37 homólogos de *purV* em *Archaea* e 120 em *Bacteria.* O gene da PurO foi encontrado em 85 OTUs do Domínio *Archaea* e 37 do Domínio *Bacteria*, apenas uma cópia por OTU enquanto a PurP foi encontrada em 132 OTUs de *Archaea*, variando de uma a quatro cópias, totalizando 248 homólogos de *purP* (tabela 1).

Tabela 1. Relação de OTUs e genes recuperados por filo.

Domínio	Filo	OTUs	*purH*	*purV*	*purJ*	*purP*	*purO*	Nenhum dos genes
Archaea	*Crenarchaeota*	44	-	-	-	58	-	15
	Euryarchaeota	145	42	37	4	167	85	2
	Korarchaeota	1	-	-	-	2	-	-
	Nanoarchaeota	2	-	-	-	-	-	2
	Thaumarchaeota	13	-	-	-	21	-	3
Bacteria	*Acidobacteria*	9	9	-	-	-	-	-
	Actinobacteria	364	353	13	1	-	6	6
	Aquificae	15	15	-	-	-	-	-
	Bacteriodetes	173	159	16	9	-	-	5
	Chlamydiae	17	1	-	-	-	-	16
	Chloroflexi	18	17	-	-	-	-	1
	Chrysiogenetes	1	1	-	-	-	-	-
	Clorobi	12	12	-	-	-	-	-
	Cyanobacteria	80	79	-	-	-	-	1
	Deferribacteres	4	4	-	-	-	-	-
	Deinococcus - Thermus	23	23	-	-	-	-	-
	Dictyoglomi	2	2	-	-	-	-	-
	Elusimicrobia	3	3	-	-	-	-	-
	Fibrobacteres	1	-	1	-	-	1	-
	Firmicutes	441	386	39	8	-	24	23
	Fusobacteria	9	8	-	-	-	-	1
	Gemmatimonadetes	3	3	-	-	-	-	-
	Ignavibacteriae	2	2	-	-	-	-	-
	Nitrospirae	7	7	-	-	-	-	-
	Planctomycetes	14	13	1	1	-	-	-
	Proteobacteria	1171	1080	33	28	-	3	60
	Spirochaetes	43	18	7	5	-	3	18
	Synergistetes	5	5	-	-	-	-	-
	Tenericutes	73	4	-	-	-	-	69
	Thermodesulfobacteria	4	-	4	4	-	-	-
	Thermotogae	27	18	6	14	-	-	1
	Verrucomicrobia	9	9	-	-	-	-	-
Total		2735	2273	157	74	248	122	223

Aproximadamente 95% das OTUs do Domínio *Bacteria*, que não possuem *purH*, possuem *purV* e *purJ* ou *purV* e *purO,* combinações gênicas que são equivalentes funcionais de *purH* (fig. 1). Essas combinações foram frequentemente encontradas inseridas no mesmo contexto genômico (tabela 2).

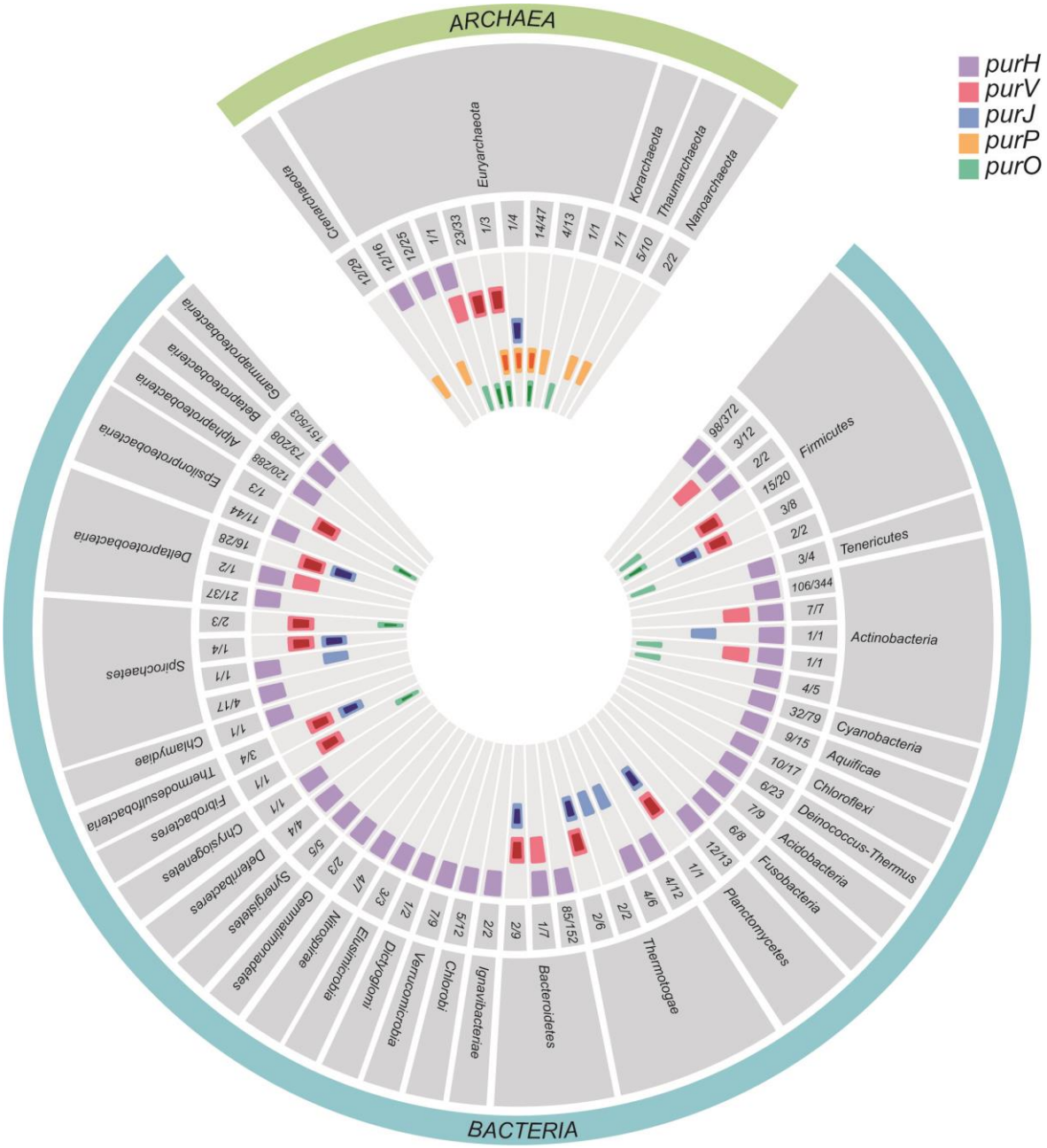
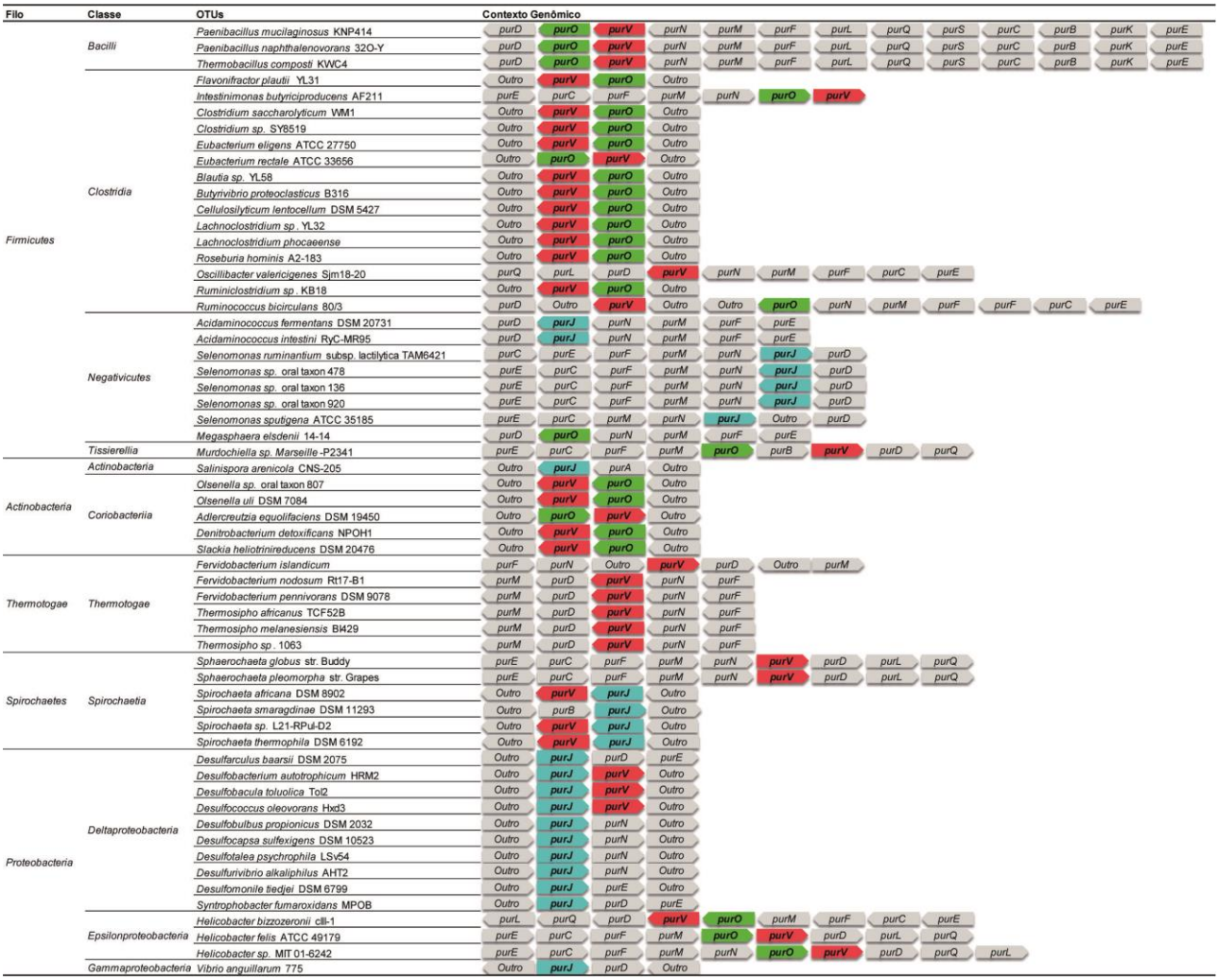


Fig. 1. Genômica comparativa com 2735 genomas completamente sequenciados dos Domínios Archaea e Bacteria. Caixas coloridas indicam presença do gene, caixas coloridas escuras indicam os casos em que na ausência da PurH ocorrem combinações gênicas que são equivalentes funcionais da PurH. Os números representam quantidade de gênero/OTUs que apresentam o gene em questão.

Tabela 2. OTUs em que purV, purJ e purO bacterianos estão em contexto genômico com outros genes da via biossintética de purinas. Seta cinza escrito outro, indicam outros genes que não tem relação com a via.



Além de estarem no mesmo contexto genômico, os genes bacterianos *purV*, *purJ* e *purO*, estão inseridos também no mesmo contexto que outros genes da VBP (tabela 2), possivelmente compondo operons. Em 3 OTUs da Classe *Bacilli* os genes *purV* e *purO* estão agrupados com todos os outros genes da VBP, formando um operon completo. No Domínio *Archaea*, apenas as *purVs* das OTUs da Classe *Halobacteria* foram encontrados em contexto com *purB* (tabela 3).

Tabela 3. OTUs em que *purVs* de *Archaea* estão em contexto com outros genes da via biossintetica de purinas.

Filo	Classe	Ordem	Família	OTUs	Contexto Genômico			
Euryarchaeota	*Halobacteria*	*Halobacteriales*	*Haloarculaceae*	*Natronomonas pharaonis* DSM 2160	*purB*	*purV/N*	*Outro*	
			Halobacteriaceae	*Halalkalicoccus jeotgali* B3	*purB*	*Outro*	*purV/N*	*Outro*
				Halobacterium salinarum NRC-1	*purB*	*purV/N*	*Outro*	
		Haloferacales	*Haloferacaceae*	*Haloferax mediterranei* ATCC 33500	*purB*	*purV/N*	*Outro*	
				Haloferax volcanii DS2	*purB*	*purV/N*	*Outro*	
				Halogeometricum borinquense DSM 11551	*purB*	*purV/N*	*Outro*	
				Haloquadratum walsbyi DSM 16790	*purB*	*purV/N*	*purA*	
			Halorubraceae	*Halorubrum lacusprofundi* ATCC 49239	*purB*	*purV/N*	*Outro*	
		Natrialbales	*Natrialbaceae*	*Halopiger xanaduensis* SH-6	*purB*	*purV/N*	*Outro*	
				Haloterrigena turkmenica DSM 5511	*purB*	*purV/N*	*Outro*	
				Halovivax ruber XH-70	*purB*	*purV/N*	*Outro*	
				Natrialba magadii ATCC 43099	*purB*	*purV/N*	*Outro*	
				Natrinema pellirubrum DSM 15624	*purB*	*purV/N*	*Outro*	
				Natrinema sp. J7-2	*purB*	*purV/N*	*Outro*	
				Natronobacterium gregoryi SP2	*purB*	*purV/N*	*Outro*	
				Natronococcus occultus SP4	*purB*	*purV/N*	*Outro*	

Estrutura primária das PurVs e PurJs

A análise dos logos mostrou que existe variação nos aminoácidos do AICARFT e IMPCH das PurHs, das PurVs e das PurJs homólogos aos aminoácidos do sítio ativo dos domínios AICARFT e IMPCH da PurH humana (fig. 2 e 3). O sítio ativo do domínio AICARFT da PurH humana, possui 13 resíduos de aminoácidos (Cheong et al. 2004, Zhang, et al. 2008-a), destes, 7 são 100% de conservados nos domínios AICARFTs das PurHs e PurVs recuperadas, (Lys266, His267, G316, Asn431, Arg451, Asp546, Arg588 da PurHs humana) (fig. 2b). Nas demais posições que apresentam variações os aminoácidos de maior ocorrência geralmente são os mesmos nos domínios AICARFTs das PurHs e nas PurVs.

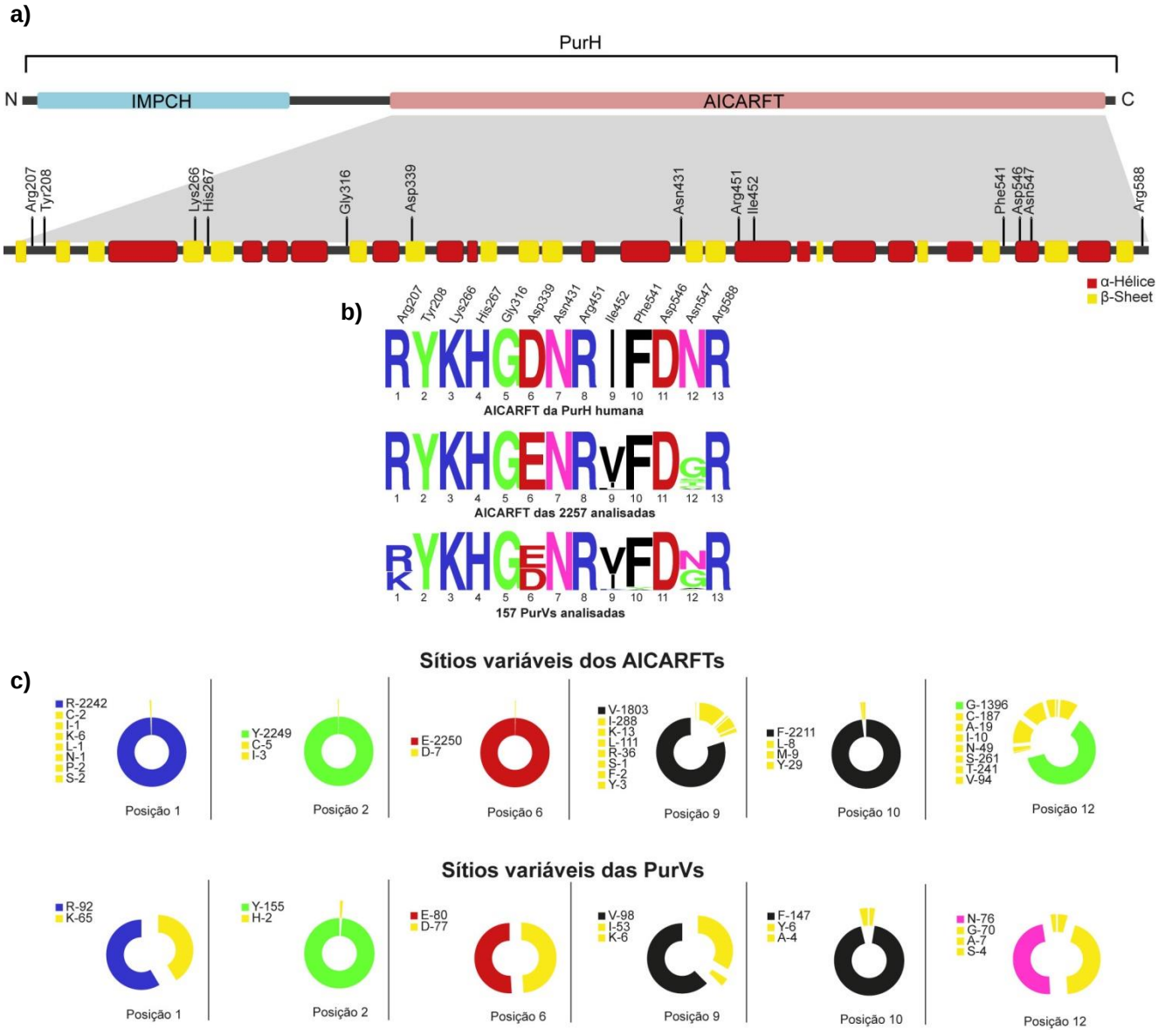


Fig. 2. Variações dos aminoácidos do sítio ativo dos AICARFTs e PurVs. a- Estrutura secundária do AICARFT da PurH com indicação das posições dos aminoácidos do sítio ativo. b-LOGOs dos aminoácidos do sítio ativo dos AICARFTs da PurH humana, 2257 PurHs recuperadas e PurVs recuperadas. c- Gráficos ilustrando variações encontradas nos aminoácidos do sítio ativo dos AICARFTs e PurVs.

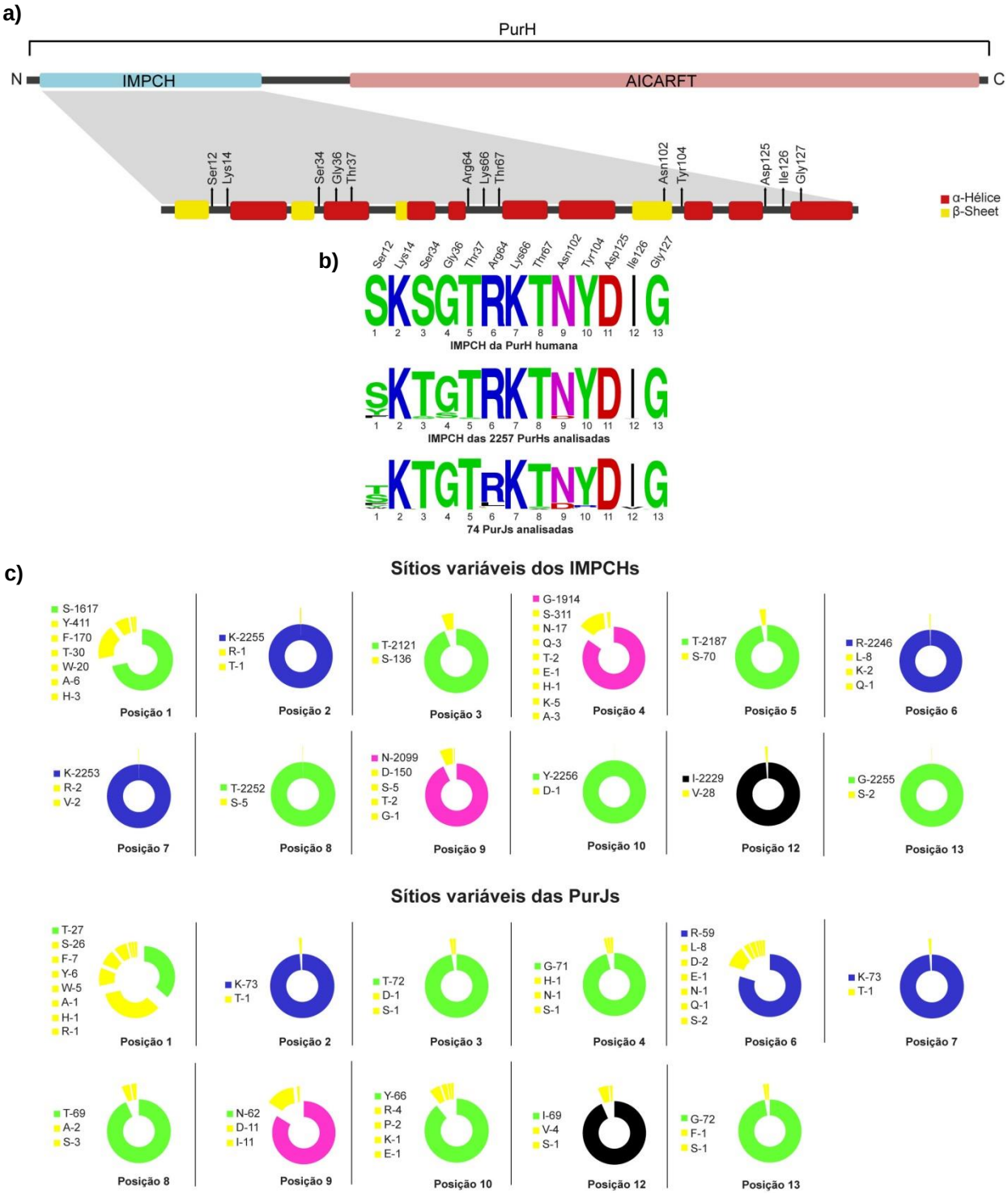


Fig. 3. Variações dos aminoácidos do sítio ativo dos IMPCHs e PurJs. a) Estrutura secundária do IMPCH da PurH com indicação das posições dos aminoácidos do sítio ativo. b-LOGOs dos aminoácidos do sítio ativo dos IMPCHs da PurH humana, 2257 PurHs recuperadas e PurJs recuperadas. c-Gráficos ilustrando variações encontradas nos aminoácidos do sítio ativo dos IMPCH e PurJs.

As posições 1, 9 e 12, foram as mais variáveis nos domínios AICARFTs, oito aminoácidos distintos foram encontrados em cada uma dessas posições. Os aminoácidos que ocorreram nessas posições dos AICARFTs também ocorreram nas mesmas posições das PurVs, entretanto a variação encontrada nas PurVs foi menor (fig. 2). Os aminoácidos predominantes em cada uma dessas posições foram os mesmos tanto nos domínios AICARFTs como nas PurVs, com exceção da glicina na posição 12 que predominou nos AICARFTs, porém foi a segunda mais frequente nas PurVs. A posição 12 foi a que apresentou maior variação nas PurVs, quatro aminoácidos diferentes, nos demais as variações não passaram de três aminoácidos distintos (fig. 2).

O sítio ativo do domínio IMPCH também possui 13 aminoácidos, mas apenas um é 100% conservado no IMPCHs das PurHs (Asp125) (fig. 3b) e dois nas PurJs recuperadas (Thr37 e Asp125)(fig. 3b). Nos demais sítios que apresentaram variações, os aminoácidos de maior ocorrência geralmente são os mesmos (fig. 3c). As posições 1, 4 e 9 foram mais variáveis nos domínios IMPCHs enquanto que as posições 1, 6 e 10 foram as mais variáveis nas PurJs. Foram encontrados 7 aminoácidos distintos na posição 1 dos IMPCHs e 8 nas PurJs, com exceção da arginina encontrada em uma PurJ nessa posição, todos os outros aminoácidos encontrados nos IMPCHs também foram encontrados nas PurJs. Na posição 4 dos IMPCHs foram encontrados 9 aminoácidos distintos entretanto a predominância foi da glicina. 5 aminoácidos diferentes foram encontrados na posição 9 dos IMPCHs entre eles a asparagina o aminoácido de maior ocorrência nessa posição. As posições 6 e 10 das PurJs apresentaram 7 e 5 aminoácidos distintos respectivamente, com predominância de Arginina para a primeira e tirosina para a segunda. Com exceção da lisina encontrada na posição 6 e do ácido aspártico encontrado na posição 10 dos IMPCHs, os aminoácidos restantes encontrados nessas posições foram correspondentes entre IMPCHs e PurJ. Na posição 5 das PurJs a treonina é 100% conservada e também foi a que mais ocorreu nos IMPCHs.

Algumas dessas variações observadas, tanto para AICARFTs, PurVs, IMPCHs e PurJ, estão em regiões de estrutura secundária do tipo α-hélice ou β-sheet (fig. 2a e 3a). Para os AICARFTs e PurVs, as variações nas posições 6, 9 e 12 ocorrem em estrutura secundária (fig. 2a e c), enquanto que para os IMPCHs e

PurJs as variações que ocorreram dentro de estrutura secundária foram nas posições 4, 9 e 13 e apenas para os IMPCHs também na posição 5 (fig. 3a e c).Das 6 posições que variaram nos AICARFTs e PurVs, apenas 3 estão em estrutura secundária. Das 12 posições que apresentam variações nos IMPCHs, apenas 4 estão inseridas em estrutura secundária, e das 11 encontradas para PurJs, apenas 3 estão em estrutura secundária. Os IMPCHs e PurJ apresentaram mais variações que os AICARFTs e PurVs os quais aparentemente são mais conservados.

A análise de *sliding window plot* mostrou que a variação na conservação da estrutura primária das PurVs e os domínios AICARFTs das PurHs e das PurJs e os domínios IMPCHs das PurHs é similar (fig. 4a e b). Geralmente, as regiões mais conservadas na estrutura primaria dos AICARFTs e dos IMPCHs, coincidem com as regiões mais conservadas na estrutura primária das PurVs e PurJs, respectivamente. Doze dos treze aminoácidos dos sítios ativos do AICARFTs e IMPCHs bem como seus homólogos nas PurVs e PurJs ficaram em uma região com identidade média acima de 66%, apenas Asn547 do sítio ativo do AICARFT e Ser12 do sítio ativo IMPCH estão em posições com identidade média abaixo de 55%.

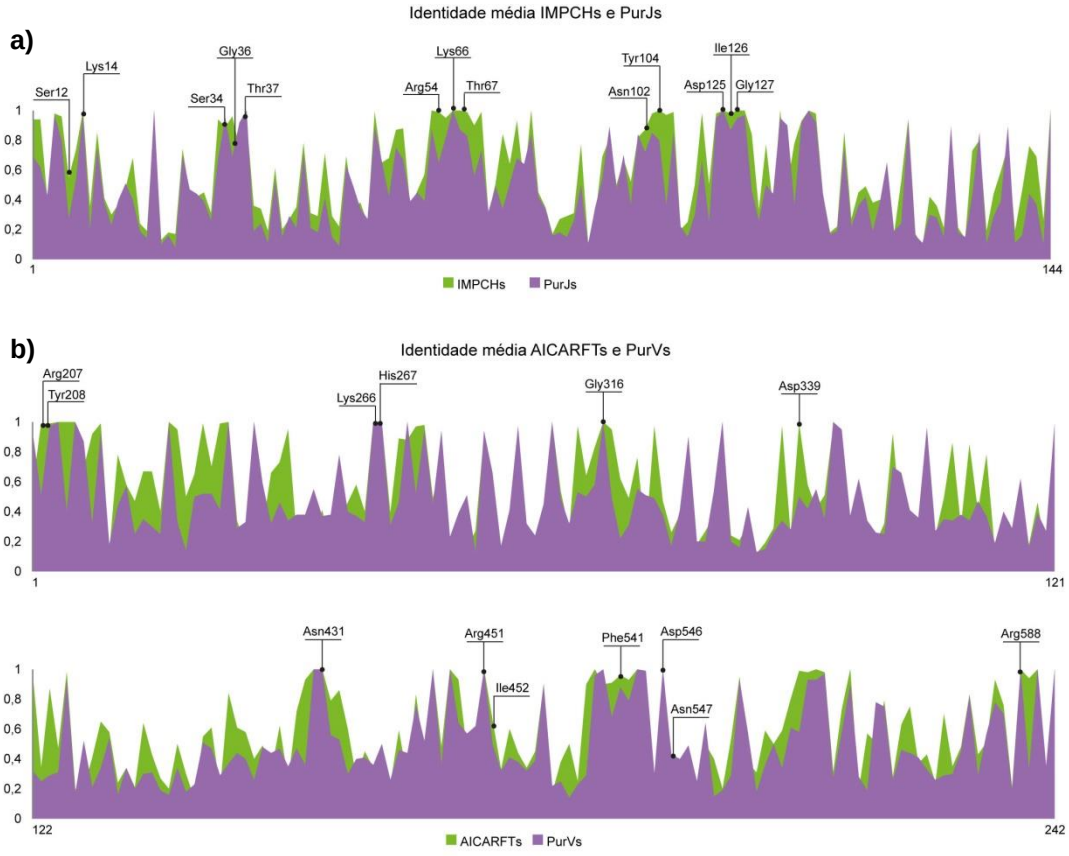


Fig. 4. Análise de *sliding window plot*. Gráficos ilustrando o padrão de conservação da estrutura primaria dos AICARFTs, IMPCHs, PurV e PurJs. A estrutura primária de AICARFTs/PurVs e IMPCHs/PurJs apresentam conservação similar, as regiões mais conservadas são basicamente as mesmas. a) Gráfico de identidade da estrutura primária IMPCHs/PurJs; b) Gráfico de identidade da estrutura primária AICARFTs/PurVs.

Relações filogenéticas entre PurHs procarióticas

A Topologia da arvore filogenética da PurH reflete em grande parte a taxonomia procariótica, sendo possível observar uma divisão entre Gram positivas e Gram negativas, mesmo nos casos em que as PurHs dos dois grupos taxonômicos estão em um mesmo grupo da árvore filogenética (fig. 5). As PurHs do Domínio *Archaea* ficaram dispersas em 5 grupos distintos na árvore filogenética e a maioria filogeneticamente relacionadas com PurHs de bactérias Gram negativas (fig. 5), Das 42 PurHs recuperadas no Domínio *Archaea*, 5 são de OTUs da Classe *Thermoplasmata*, e são filogeneticamente relacionadas com PurHs do Filo *Firmicutes*, as únicas PurHs desse Domínio que agruparam com Gram positivas, entretanto esse grupo emerge dentro de um grupo maior constituído em sua maioria por PurHs de bactérias Gram negativas. Vinte das PurHs de *Archaea*, todas pertencentes a Família *Methanosarcinaceae* agruparam com PurHs dos Filos *Bacteriodetes*, *Deltaproteobacteria* e com a única PurH encontrada para o Filo *Chlamydiae*. Cinco PurHs pertencentes a OTUs da Ordem *Thermoplasmatales* agruparam com PurHs do Filo *Thermotogae*. Duas PurHs pertencentes a OTUs do Filo *Euryarchaeota* ainda sem táxons inferiores definidos, agruparam com PurHs de *Spirochaetes* e *Chloroflexi*. Nove pertencentes a OTUs da Ordem *Methanomicrobiales* ficaram próximas a um grupo predominantemente de *Deltaproteobacterias* e 1 agrupou com proteobacterias da Classe *Gammaproteobacteria*.

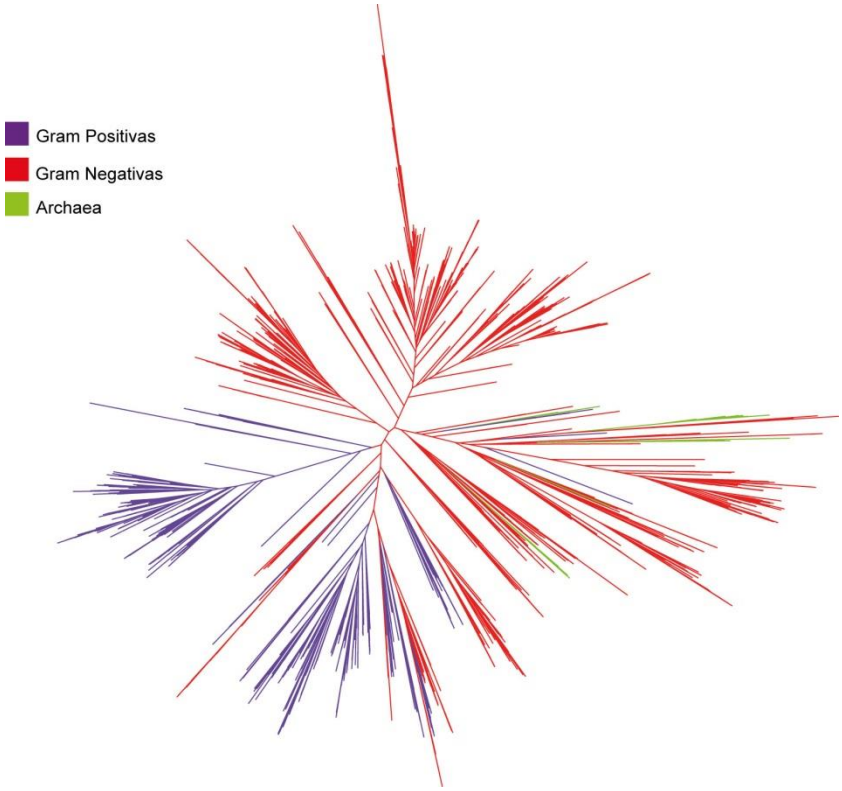


Fig. 5. Árvore filogenética de *maximum likelihood* com a sequencia de proteína das 2257 PurH recuperadas. A árvore foi reconstruida pelo método LG+G+I, com 1000 repetições de bootstrap. Ramos roxos representam as PurHs pertencentes a OTUs de Gram positivas; vermelho representam as PurHs pertencentes a OTUs de Gram negativas; verde representam as PurHs pertencentes a OTUs de Archaea.

Dos 32 Filos procarióticos representados nesse trabalho, 8 são monofiléticos para PurH (*Clorobi, Cyanobacteria, Deferribacteres, Dictyoglomi, Gemmatimonadetes, Ignavibacteriae, Synergistetes,* e *Verrucomicrobia*), e 16 não são monofiléticos, sendo que em 7 deles apenas uma PurH ficou fora do grupo contendo as PurHs das demais OTUs do Filo (tabela 4). Os 8 Filos restantes não apresentam PurH, ou apresentam apenas 1 PurH. Táxons monofiléticos abaixo de Filo foram observados em 29 Filos, apenas os Filos *Elusimicrobia*, *Fusobacteria* e *Nitrospirae* não apresentaram táxons monofiléticos em níveis taxonômico inferiores, como Classe, Ordem ou Família (tabela 4).

Tabela 4. Grupos taxonômicos que são monofiléticos para PurH.

Domínio	Filo	Monofilético	Classes		Ordens		Familias	
			Total	Monofiléticas	Total	Monofiléticas	Total	Monofiléticas
Archaea	*Crenarchaeota*	-	-	-	-	-	-	-
	Euryarchaeota	Não	8	0	9	2	24	4
	Korarchaeota	-	-	-	-	-	-	-
	Nanoarchaeota	-	-	-	-	-	-	-
	Thaumarchaeota	-	-	-	-	-	-	-
Bacteria	*Acidobacteria*	Não[+]	3	1	2	1	2	1
	Actinobacteria	Não[+]	4	1	22	7	41	13
	Aquificae	Não	1	0	2	1	3	3
	Bacteriodetes	Não	5	2	7	2	20	6
	Chlamydiae	[a]	[a]	[a]	[a]	[a]	[a]	[a]
	Chloroflexi	Não	5	2	7	2	8	3
	Chrysiogenetes	[a]	[a]	[a]	[a]	[a]	[a]	[a]
	Clorobi	Sim	1	1	1	1	1	1
	Cyanobacteria	Sim	-	-	7	2	22	4
	Deferribacteres	Sim	1	1	1	1	1	1
	Deinococcus - Thermus	Não[+]	1	0	2	0	3	2
	Dictyoglomi	Sim	1	1	1	1	1	1
	Elusimicrobia	Não[+]	1	0	1	0	1	0
	Fibrobacteres	-	-	-	-	-	-	-
	Firmicutes	Não	6	1	10	3	33	9
	Fusobacteria	Não	1	0	1	0	2	0
	Gemmatimonadetes	Sim	1	1	1	1	1	1
	Ignavibacteriae	Sim	1	1	1	1	2	-
	Nitrospirae	Não[+]	1	0	1	0	1	0
	Planctomycetes	Não[+]	2	1	2	1	4	0
	Proteobacteria	Não	5	1	45	14	104	34
	Spirochaetes	Não	1	0	2	1	4	1
	Synergistetes	Sim	1	1	1	1	1	1
	Tenericutes	Não[+]	1	0	3	1	4	1
	Thermodesulfobacteria	-	-	-	-	-	-	-
	Thermotogae	Não	1	0	3	1	4	1
	Verrucomicrobia	Sim	4	1	4	2	4	2

[+] Apenas a PurH de 1 OTU agrupou fora, o que provavelmente é um caso de transferencia horizontal
- Não tem PurH
[a] Tem apenas 1 PurH

Relações filogenéticas entre as PurVs e o domínio AICARFT das PurHs procarióticas

Na árvore filogenética obtida a partir do alinhamento múltiplo contendo os domínios AICARFTs das PurHs e as PurVs, as últimas formaram um grande grupo contendo 113 PurVs bacterianas que não está relacionado com um grupo específico de AICARFTs de PurHs procarióticas (fig. 6, tabela 5). Este grupo contém dois subgrupos menores, um contendo predominantemente PurVs de OTUs Gram positivas, incluindo OTUs High GC e Low GC, e o outro contendo predominantemente PurVs de OTUs da Classe *Deltaproteobacteria* (fig. 6). Das outras PurVs restantes, seis do Filo *Thermotogae* agruparam com AICARFTs de PurHs de OTUs do mesmo Filo. As três PurVs do Filo *Euryarchaeota*, pertencentes a OTUs da Família *Methanosaetaceae*, agruparam com o domínio

AICARFT de PurHs de OTUs do Filo *Verrucomicrobia* (fig. 6). Por fim, a única PurV encontrada em OTUs do Filo *Fibrobacteres* agrupou com AICARFTs de PurHs de OTUs do Domínio *Archaea* todos pertencentes à Família *Methanosarcinaceae* da Ordem *Methanosarcinales* e Classe *Methanomicrobia*. (fig. 6). O domínio AICARFT foi encontrado fusionado à PurN em todas as trinta e quatro OTUs da Classe *Halobacteria*, Domínio *Archaea*, que foram analisadas neste trabalho (tabela 3). Todos eles ficam em um único grupo na filogenia dos domínios AICARFT das PurHs e das PurVs (fig. 6).

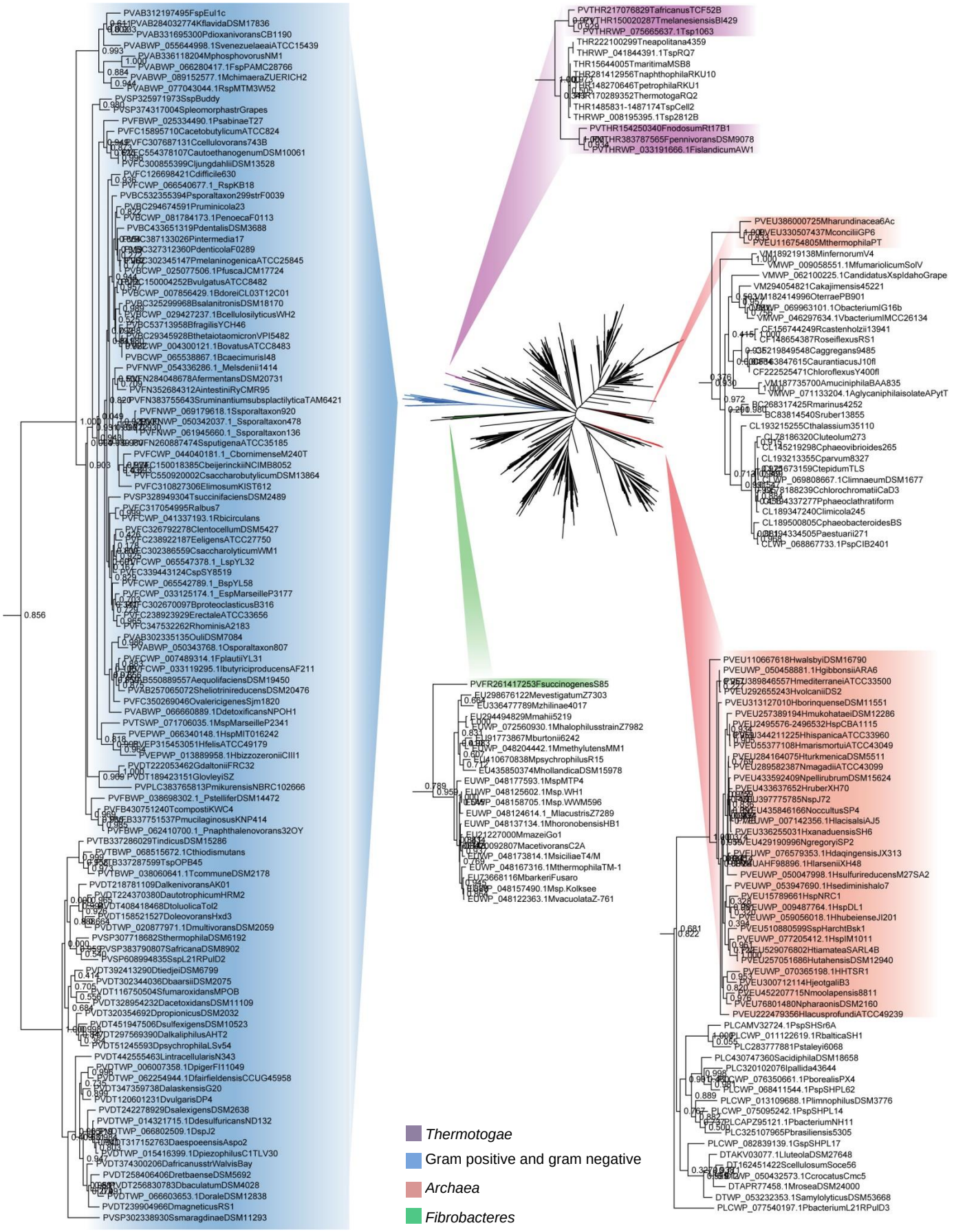

Thermotogae
Gram positive and gram negative
Archaea
Fibrobacteres

Fig. 6. Recortes da árvore filogenética de *maximum likelihood* dos AICARFTs das 2257 PurHs recuperadas com as 157 PurVs recuperadas. A árvore foi realizada pelo método LG+G+I, com 1000 repetições de bootstrap. Marcações coloridas indicam os grupos formados pelas PurVs. Azul, grupo com 113 PurVs de Gram positivas e Gram negativas; vermelho, PurVs de *Archaea,* roxo, PurVs do Filo *Thermotogae;* e verde, PurV do Filo *Fibrobacteres.*

Relações filogenéticas entre as PurJs e o domínio AICARFT das PurHs procarióticas

Na árvore filogenética obtida a partir do alinhamento múltiplo dos IMPCHs e PurJs, as PurJs ficaram distribuídas em 12 grupos. Em geral esses grupos contém PurJs pertencentes a OTUs do mesmo Filo, e não estão diretamente relacionados com IMPCHs, indicando que elas são filogeneticamente mais relacionadas entre si (tabela 5). Aproximadamente 39% das PurJs agruparam com IMPCHs de OTUs do mesmo Filo (PurJs de *Euryarchaeota, Firmicutes, Planctomycetes, Thermotogae, Bacteriodetes, e Spirochaetes*) (fig. 7 e tabela 5).

Tabela 5. Relação de Filos em que as PurV e PurJ agruparam com AICARFTs e IMPCHs

Filo com PurV e PurJ	PurV	Relacionada com AICARFTs do mesmo filo	PurJ	Relacionada com IMPCHs do mesmo filo
Euryarchaeota	37	-	4	4
Firmicutes	39	-	8	8
Actinobacteria	13	-	1	-
Planctomycetes	1	-	1	1
Thermotogae	6	6	14	6
Bacteriodetes	16	-	9	9
Fibrobacteres	1	-	-	-
Thermodesulfobacteria	4	-	4	-
Spirochaetes	7	-	5	1
Proteobacteria	33	-	28	-

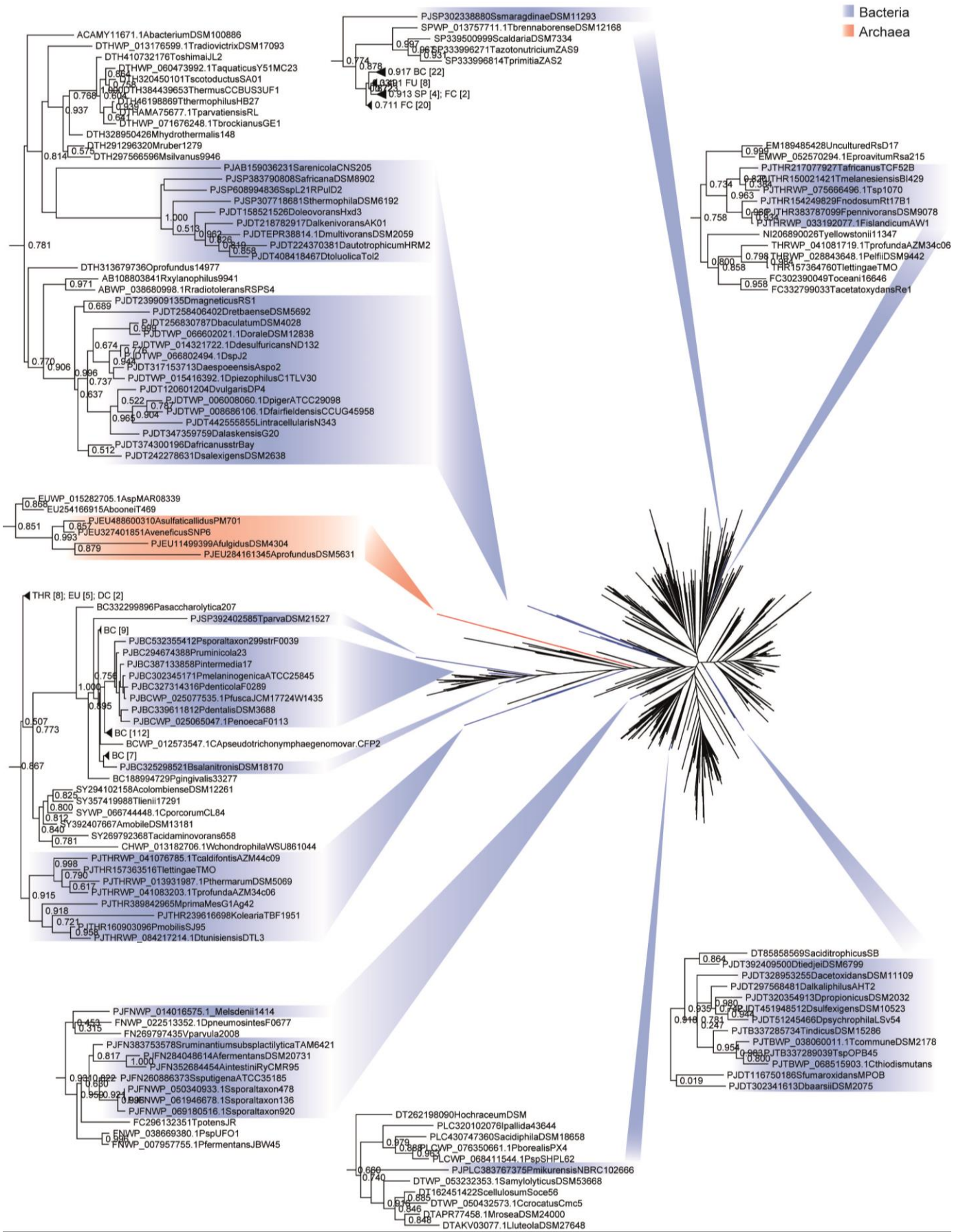

Bacteria
Archaea
PJSP302338880SsmaragdinaeDSM11293
SPWP_013757711.1TbrennaborenseDSM12168
SP339500999ScaldariaDSM7334
SP333996814TprimitiaZAS2
0.917 BC [22]
0.913 SP [4]; FC [2]
0.711 FC [20]
ACAMY11671.1AbacteriumDSM100886
DTHWP_013176599.1TradiovictrixDSM17093
DTH410732176ToshimaiJL2
DTHWP_060473992.1TaquaticusY51MC23
DTH320450101TscotoductusSA01
DTH384439653ThermusCCBUS3UF1
DTH46198869TthermophilusHB27
DTHAMA75677.1TparvatiensisRL
DTHWP_071676248.1TbrockianusGE1
DTH328950426Mhydrothermalis148
DTH291296320Mruber1279
DTH297566596Msilvanus9946
PJAB159036231SarenicolaCNS205
PJSP383790808SafricanaDSM8902
PJSP608994836SspL21RPuiD2
PJSP307718681SthermophilaDSM6192
PJDT158521526DoleovoransHxd3
PJDT218782917DalkenivoransAK01
PJDTEPR38814.1DmultivoransDSM2059
PJDT224370381DautotrophicumHRM2
PJDT408418467DtoluolicaTol2
DTH3136797360profundus14977
AB108803841Rxylanophilus9941
ABWP_038680998.1RradiotoleransRSPS4
PJDT239909135DmagneticusRS1
PJDT258406402DretbaenseDSM5692
PJDT256830787DbaculatumDSM4028
PJDTWP_066602021.1DoraleDSM12838
PJDTWP_014321722.1DdesulfuricansND132
PJDTWP_066802494.1DspJ2
PJDT317153713DaespoeensisAspo2
PJDTWP_015416392.1DpiezophilusC1TLV30
PJDT120601204DvulgarisDP4
PJDTWP_006008060.1DpigerATCC29098
PJDTWP_008686106.1DfairfieldensisCCUG45958
PJDT442555855LintracellularisN343
PJDT347359759DalaskensisG20
PJDT374300196DafricanusstrBay
PJDT242278631DsalexigensDSM2638
EM189485428UnculturedRsD17
EMWP_052570294.1EproavitumRsa215
PJTHR217077927TafricanusTCF52B
PJTHR150021421TmelanesiensisBI429
PJTHRWP_075666496.1Tsp1070
PJTHR154249829FnodosumRt17B1
PJTHRWP_033192077.1FislandicumAW1
NI206890026Tyellowstonii11347
THRWP_041081719.1TprofundaAZM34c06
THRWP_028843648.1PelfiiDSM9442
THR157364760TlettingaeTMO
FC302390049Toceani16646
FC332799033TacetatoxydansRe1
EUWP_015282705.1AspMAR08339
EU254166915AbooneiT469
PJEU488600310AsulfaticallidusPM701
PJEU327401851AveneficusSNP6
PJEU11499399AfulgidusDSM4304
PJEU284161345AprofundusDSM5631
THR [8]; EU [5]; DC [2]
BC332299896Pasaccharolytica207
PJSP392402585TparvaDSM21527
BC [9]
PJBC532355412Psporaltaxon299strF0039
PJBC294674388Pruminicola23
PJBC387133858Pintermedia17
PJBC302345171PmelaninogenicaATCC25845
PJBC327314316PdenticolaF0289
PJBCWP_025077535.1PfuscaJCM17724W1435
PJBC339611812PdentalisDSM3688
PJBCWP_025065047.1PenoecaF0113
BC [112]
BCWP_012573547.1CApseudotrichonymphaegenomovar.CFP2
BC [7]
PJBC325298521BsalanitronisDSM18170
BC188994729Pgingivalis33277
SY294102158AcolombienseDSM12261
SY357419988Tlienii17291
SYWP_066744448.1CporcorumCL84
SY392407667AmobileDSM13181
SY269792368Tacidaminovorans658
CHWP_013182706.1WchondrophilaWSU861044
PJTHRWP_041076785.1TcaldifontisAZM44c09
PJTHR157363516TlettingaeTMO
PJTHRWP_013931987.1PthermarumDSM5069
PJTHRWP_041083203.1TprofundaAZM34c06
PJTHR389842965MprimaMesG1Ag42
PJTHR239616698KolearíaTBF1951
PJTHR160903096PmobilisSJ95
PJTHRWP_084217214.1DtunisiensisDTL3
DT85858569SaciditrophicusSB
PJDT392409500DtiedjeiDSM6799
PJDT328953255DacetoxidansDSM11109
PJDT297568481DalkaliphilusAHT2
PJDT320354913DpropionicusDSM2032
PJDT451948512DsulfexigensDSM10523
PJDT51245466DpsychrophilaLSv54
PJTB337285734TindicusDSM15286
PJTBWP_038060011.1TcommuneDSM2178
PJTB337289039TspOPB45
PJTBWP_068515903.1Cthiodismutans
PJDT116750186SfumaroxidansMPOB
PJDT302341613DbaarsiiDSM2075
PJFNWP_014016575.1_Melsdenii1414
FNWP_022513352.1DpneumosintesF0677
FN269797435Vparvula2008
PJFN383753578SruminantiumsubsplactilyticaTAM6421
PJFN284048614AfermentansDSM20731
PJFN352684454AintestiniRyCMR95
PJFN260886373SsputigenaATCC35185
PJFNWP_050340933.1Ssporaltaxon478
PJFNWP_061946678.1Ssporaltaxon136
PJFNWP_069180516.1Ssporaltaxon920
FC296132351TpotensJR
FNWP_038669380.1PspUFO1
FNWP_007957755.1PfermentansJBW45
DT262198090HochraceumDSM
PLC320102076Ipallida43644
PLC430747360SacidiphilaDSM18658
PLCWP_076350661.1PborealisPX4
PLCWP_068411544.1PspSHPL62
PJPLC383767375PmikurensisNBRC102666
DTWP_053232353.1SamylolyticusDSM53668
DT162451422ScellulosumSoce56
DTWP_050432573.1CcrocatusCmc5
DTAPR77458.1MroseaDSM24000
DTAKV03077.1LluteolaDSM27648

Fig. 7. Recortes da árvore filogenética de *maximum likelihood* dos IMPCHs das 2257 PurHs recuperadas com as 74 PurJs recuperadas. A árvore foi realizada pelo método LG+G+I, com 1000 repetições de bootstrap. Marcações coloridas indicam os grupos formados pelas PurJs. Azul, PurJ de *Bacteria;* vermelho, PurJs de *Archaea*.

Discussão

Análise de genômica comparativa

Nesse trabalho foi analisado quase o dobro de genomas em comparação com o trabalho anterior (97% a mais), e os resultados encontrados não diferem dos resultados obtidos previamente: A PurH é a escolha evolutiva das bactérias para desempenhar os dois últimos passos da VBP, enquanto as archaeas além da PurH, recorrem a PurP e a PurO. A PurP continua sendo assinatura do Domínio *Archaea*, e a PurO está presente em um número maior de genomas bacterianos do que o que foi mostrado em nosso trabalho anterior (Capítulo 1).

Devido a preferência das bactérias pela PurH, a quantidade de PurV, PurJ e PurO encontradas nos genomas procarióticos analisados foi surpreendente, esse fato chama atenção para a possibilidade de uma diversidade enzimática ainda maior para a VBP, que ainda é desconhecida, principalmente para as espécies de *Archaea* do Filo *Crenarchaeota*, que não possuem nenhumas das outras enzimas do nono e décimo passos da VBP, apenas a PurP.

Até então não existe indícios de qualquer outra enzima que realize a função da PurH no nono e decimo passo da VBP no Domínio *Bacteria*, dessa forma a correlação negativa das combinações gênicas *purV*/*purJ* ou *purV*/*purO* com *purH* (fig. 1) pode indicar que esses novos genes estão funcionalmente ativos na via, uma vez que essas combinações são equivalentes funcionais de *purH* e foram encontrados em 95% dos genomas que não tem *purH*.

A inserção de *purV* e *purJ* bacterianos no mesmo contexto genômico que outros genes da VBP (tabela 1), é um dos indícios mais relevantes demonstrados nesse trabalho, que nos permite atribuir relação funcional desses genes com a

VBP. Isso porque é comum nos genomas procarióticos as vias metabólicas serem codificadas por agrupamentos de genes ou operons, eles são pistas relevantes para a atribuição funcional (Korbel et al. 2004; Rocha 2008; Zhao et al. 2013). Isso devido ao fato de que genes para uma determinada característica importante são comumente agrupados dentro do genoma de procariotos (Rocha 2008; Martí-Arbona et al. 2014) e tendem a ser conservados devido às pressões de seleção (Santos et al. 2010, Yurkov et al. 1994).

Boa parte dos *purVs* encontrados no Domínio *Archaea* estão inseridos no mesmo contexto genômico que *purB* (tabela 2), porém, em direções opostas. Isso nos permite reforçar a associação dos *purVs* à VBP nesse Domínio, uma vez que a relação funcional pode ser estabelecida quando os genes estão posicionados em uma mesma direção, ou não, pois, genes posicionados divergentemente também são fortemente co-regulados e podem ser explorados para fazer predições funcionais precisas (Korbel et al. 2004).

É valido salientar também, que a PurB produz o intermediário que será utilizado pela PurV, e além disso, esses genes *purVs* estão todos fusionados à *purN.* Além da conservação de grupos gênicos, e a co-ocorrência de genes a fusão gênica também é uma características genômicas associada a funcionalidade (Berger et al. 2008; Moreno-Hagelsieb & Santoyo 2015).

Análise da estrutura primária das PurVs e PurJs

A maioria das variações que ocorreram nos resíduos de aminoácidos dos sítios ativos das PurVs e PurJs, também foram encontradas nos domínios AICARFT e IMPCH das PurHs analisadas (fig. 2c e 3c), o que possivelmente pode ser uma característica própria dessas enzimas procarióticas. Muitas das variações ocorreram entre aminoácidos com as mesmas características físico-químicas, como por exemplo, mudança de Asp para Glu (fig. 2b posição 6), ambos aminoácidos polares e ácidos, mudanças de Ile para Val (fig. 2b posição 9) que são aminoácidos apolares com cadeia ramificada, ou ainda mudanças entre Ser e Thr (fig. 3b posição 3) que são aminoácidos polares sem carga.

A mudança entre aminoácidos físico-quimicamente semelhantes tem uma grande chance de ter equivalência funcional e estrutural, uma vez que, de acordo com Malkov e Colaboradores (2008), possivelmente são as características físicas e químicas de cada aminoácido que definem como ele vai interagir e quais estruturas secundárias ele vai formar dentro de uma proteína. Sendo assim, quando um aminoácido é substituído por outro de características iguais, as consequências disso pode ser irrelevante.

Segundo a literatura existe uma propensão dos aminoácidos em formar α-helices, β-Sheet, ou alças. Os aminoácidos Arg, Glu, Lys, Leu, Met, Ala e Gln, preferencialmente estão presente nas α-helices, Val, Ile, Thr, Phe, Tyr e Trp têm propensão em formar β-Sheets, já a Ser, Asp, Asn, Gly e Pro geralmente formam as alças. Os aminoácidos His e Cys podem está relacionados com qualquer uma das estruturas secundárias (Malkov et al. 2008).

De acordo com Fujiwara e Colaboradores (2012) as propensões dos aminoácidos que formam α-hélice são frequentemente os mesmos, entretanto para β-Sheet isso pode variar muito de acordo com o contexto, e a frequência dos aminoácidos na cadeia. Isso porque, segundo Chemmama e Colaboradores (2015), a propensão de um aminoácido assumir uma determinada estrutura secundária é diretamente influenciada pelos aminoácidos que se encontram antes e depois dele na estrutura primária.

Sendo assim, por vezes, os aminoácidos relacionados à formação de β-Sheet podem estar também presentes em estrutura secundária do tipo α-hélice, assim como aqueles com propensão a formar alças, que também podem compor α-hélices e β-Sheet (Malkov et al. 2008). Por esse motivo, as alterações encontradas nos aminoácidos do sítio ativo inseridos em estrutura secundária, podem ou não ter alguma influencia na atividade da enzima,

De acordo com a literatura, as variações de aminoácidos encontradas nos aminoácidos do sítio ativo AICARFTs/PurVs e IMPCH/PurJs são estruturalmente equivalentes, e não interfeririam na formação das α-hélices e β-Sheets. Com exceção apenas da mudança de Asp para Glu (fig. 2c posição 6), entretanto esses dois aminoácidos são físico-quimicamente semelhantes, e podem ser estruturalmente equivalentes, uma vez que a propensão dos aminoácidos em

formar certa estrutura secundária pode está relacionada às suas propriedades físico-quimicas (Malkov et al. 2008).

Através de mutações sítio dirigidas, Verma e Colaboradores (2017) observaram que Lys255 (Lys 266 para PurH humana) e His256 (His 267 para a PurH humana) na PurH de *Staphylococcus lugdunensis* são resíduos de aminoácidos essenciais para a atividade catalítica do domínio AICARFT, quando substituídos por outro aminoácido o domínio perdeu sua atividade catalítica. Em todas as PurVs analisadas neste trabalho os aminoácidos correspondentes a Lys255 e His256 de *Staphylococcus lugdunensis* são 100% conservados (fig. 2b, posições 3 e 4). A conservação desses aminoácidos mostra que existe uma pressão de seleção atuando sobre as PurVs, fazendo com que esses aminoácidos essenciais à atividade da enzima sejam mantidos, reforçando assim a participação e importância da PurV na VBP. Além disso, o Logo indica que outros aminoácidos também são 100% conservados e devem ser essenciais para a atividade enzimática do domínio AICARFT.

Apesar das variações existentes entre os aminoácidos dos sítios ativos, a similaridade entre a estrutura primaria das PurVs e PurJs e os domínios AICARFT e IMPCH que foi observada na análise de *sliding window plot*, sugere que existe uma pressão de seleção atuando sobre PurV e PurJ, e ela agiu de forma similar para elas e para a PurH. Essa conservação da estrutura primária de PurV e PurJ, também pode indicar atividade catalítica funcional, uma vez que os domínios AICARFT e IMPCH são cataliticamente independentes (Rayl et al. 1996; Zhang et al. 2008; Verma et al. 2017).

PurV e PurJ podem ser funcionalmente ativas?

Em 1996, Rayl e Colaboradores demonstraram que os domínios IMPCH e AICARFT da PurH humana, apresentam atividades enzimáticas independentes. Eles sintetizaram os domínios separadamente e estabeleceram a atividade enzimática de cada um, observando que a atividade de um domínio independe do outro domínio.

Recentemente Verma e Colaboradores (2017), chegaram a mesma conclusão, estudando a PurH de *Staphylococcus lugdunensis*, eles concluíram que os dois domínios dessa PurH também são cataliticamente ativos de forma independente, e os aminoácidos do sítio ativo de cada domínio não interagem entre si durante a reação enzimática dos domínios. Dessa forma PurV e PurJ são cataliticamente capazes de desempenhar o nono e decimo passos da via biossintetica de purinas.

O que se sabe até então é que os domínios da PurH necessitam formar um dímero para que tenha atividade funcional (Greasley et al. 2001; Cheong et al. 2004; Axelrod et al. 2008), principalmente o domínio AICARFT, uma vez que seu sítio ativo é composto por aminoácidos dos dois monômeros que formam o dímero da PurH, diferente do domínio IMPCH, que os aminoácidos do sitio ativo estão todos no mesmo monômero (Zhang et al. 2008-a).

É proposto que para ter uma atividade funcional eficiente o domínio AICARFT da PurH necessita de um acoplamento próximo ao domínio IMPCH, o que favoreceria a conversão do FAICAR em IMP (Bulock et al. 2002; Xu et al. 2007). Entretanto, não há evidencias da formação de um túnel que conecte os sítios ativos dos dois domínios (Bulock et al. 2002; Xu et al. 2007; Zhang et al. 2008; Nours et al. 2011), não justificando a necessidade dessa proximidade.

Verma e Colaboradores (2017), colocam que os domínios fusionados na forma da PurH, pode ser relevante para a aquisição da conformação ideal da enzima, entretanto ele assume como essencial para a atividade enzimática a formação de dímeros de cada domínio da enzima, e descreve que os domínios AICARFT e IMPCH separadamente podem se auto associar de forma independente e seguir o mesmo padrão da PurH, formando dímeros estáveis. Sendo assim a PurV e PurJ também são capazes de formar dímeros funcionais.

Diante do exposto nesse trabalho, nada impede que *purV* e *purJ* sejam funcionalmente ativos na VBP, uma vez que eles formam combinações gênicas que substituem o gene da PurH em alguns organismos, estão inseridos no mesmo contexto genômico que outros genes da VBP formando operons, apresentam sítio ativo aparentemente funcional e podem sim se associar de forma dimérica.

Relações filogenéticas entre PurHs procarióticas

A árvore da PurH tem sentido taxonômico, isso fica claro com a divisão de Gram Positivas e Gram Negativas e também pela formação de grupos monofiléticos, sejam eles a nível de Filo, Classe, Ordem ou Família. Nos casos em que Gram Positivas agruparam com Gram Negativas, é nítido a ocorrência de eventos de transferência lateral do gene da PurH de um grupo para o outro.

As archaeas preferencialmente utilizam a PurP e a PurO para realizar os últimos passos da VBP, sendo essas as enzimas que predominantemente ocorrem nesse Domínio da vida (Zhang et al. 2008-a). Devido a isso não foi surpreendente o fato das PurHs de archaeas estarem relacionadas com PurHs do Domínio *Bacteria*, evidenciando que elas foram adquiridas a partir de transferência horizontal. A distribuição das PurHs das archaeas na árvore, mostra também que elas não tiveram uma única origem dentro do Domínio *Bacteria*, como mostrado anteriormente, parte delas agrupou com Gram positivas e a maioria com Gram Negativas (fig. 5).

Sendo assim, podemos inferir que as PurHs recuperadas são derivadas de um único evento de fusão que ocorreu no ancestral das bacterias, uma vez que a topologia da árvore não sugere que exista mais de uma isoforma da PurH, e que elas possivelmente tenham sido originadas a partir de eventos de fusão distintos.

Relações filogenéticas entre as PurVs e o domínio AICARFT das PurHs procarióticas

Aparentemente, não existe relação filogenética próxima entre as PurVs e os domínios AICARFTs das PurHs, isso ficou evidente diante da topologia que a árvore assumiu, exceto para as PurVs do Filo *Thermotogae* que agruparam com AICARFTs do mesmo Filo. Essas PurVs provavelmente foram originadas a partir da quebra do *purH* desse grupo. O mesmo aconteceu com as 3 PurVs de *Archaea* que agruparam com Gram negativas e a PurV do Filo *Fibrobacteres* que agrupou com AICARFTs de OTUs de *Archaea*, entretanto esses últimos aparentemente foram adquiridos através de transferência horizontal.

As 113 PurVs que agruparam juntas estão amplamente distribuídas em Gram positivas e Gram negativas. Elas são muito distintas das demais PurVs e não estão relacionadas com os AICARFTs analisados, aparentemente estas PurVs são derivadas do domínio ancestral e surgiram no ancestral das bacterias.

A fusão da PurV com a PurN nas *Halobacterias*, provavelmente é decorrente de um evento de fusão que ocorreu no ancestral desse grupo, possivelmente similar ao evento de fusão que deu origem à PurH. Na árvore elas estão proximamente relacionadas com AICARFTs do Filo *Planctomycetes*.

Relações filogenéticas entre as PurJs e o domínio IMPCH das PurHs procarióticas

A árvore filogenética das PurVs e PurJs contam histórias diferentes. A topologia que a árvore das PurJs com os IMPCHs assumiu deixa claro que as PurJs apresentam origens distintas e possivelmente surgiram a partir de quebra do gene da PurH, uma vez que elas estão distribuídas de forma dispersa na árvore e em sua maioria (59%) agrupando com IMPCH sem relação taxonômica. Nos casos em que as PurJs agruparam com IMPCH do mesmo grupo taxonômico podemos inferir que essa quebra possivelmente tenha ocorrido no ancestral do grupo.

Diante dos resultados obtidos podemos inferir que, se as PurJs atuais fossem originadas a partir do domínio IMCPH ancestral que originou a PurH, elas formariam um grupo a parte na árvore, como ocorreu com a PurV, devido a divergência que existiria entre elas e os IMCPHs. Por isso, devido a distribuição das PurJs na árvore de forma dispersa, e em alguns casos agrupando com o IMCPH do mesmo grupo taxonômico, podemos inferir que as PurJs atuais foram originadas a partir de eventos de quebra da PurH, que ocorreram em momentos distintos da evolução do gene.

A topologia da árvore mostra também que houve casos de transferência lateral do gene, o que justifica a falta de relação taxonômica entre a maioria das PurJs e os IMPCHs (aproximadamente 59%).

Diante do exposto podemos concluir que os resultados desse estudo contribuem para compreender melhor a diversidade e distribuição das enzimas relacionadas com o nono e décimo passo da VBP, assim como esclarece um pouco da história evolutiva dos genes *purV*, *purJ,* e *purH*. Sugere também grupos taxonômicos em que a PurH pode ser utilizada para realizar filogenia. Além disso, os dados da genômica comparativa podem ser utilizados, por exemplo, para indicar os genes da VBP aqui estudados, que podem ser utilizados como assinaturas para um táxon específico.

REFERÊNCIAS

Aiba A, Mizobuchi K. 1989. Nucleotide sequence analysis of genes purH and purD involved in the de novo purine nucleotide biosynthesis of Escherichia coli. J Biol Chem. 264(35):21239-46.

Ajawatanawong P, Baldauf SL. 2013. Evolution of protein indels in plants, animals and fungi. BMC Evol Biol. 13:140.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Axelrod HL, et al. 2008. Crystal structure of AICAR transformylase IMP cyclohydrolase (TM1249) from Thermotoga maritima at 1.88 A˚ resolution. Proteins: Structure, Function, and Bioinformatics. DOI: 10.1002/prot.21967.

Andrade RFS, et al. 2011. Detecting network communities: an application to phylogenetic analysis. PLoS Comput Biol. 7(5): e1001131.

Armenta-Medina D, Segovia L, Perez-Rueda E. 2014. Comparative genomics of nucleotide metabolism: a tour to the past of the three cellular domains of life. BMC Genomics. 15:800.

Baggott JE, Tamura T. 2015. Folate-Dependent Purine Nucleotide Biosynthesis in Humans. Adv. Nutr. v.6, n.5, p.564-571.

Berger MF, et al. 2008. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. Cell. 133, p.1266–1276.

Brown AM, Hoopes SL, White RH, Sarisky CA. 2011. Purine biosynthesis in archaea: variations on a theme. Biology Direct. v.6, n.63. p.1-21.

Buchanan JM, Hartman SC. 1959. Enzymatic reactions in the synthesis of purines. Adv Enzymol. 21:199–261.

Bulock KG, Beardsley GP, Anderson KS. 2002. The Kinetic Mechanism of the Human Bifunctional Enzyme ATIC (5-Amino-4-imidazolecarboxamide Ribonucleotide Transformylase/Inosine 5′-Monophosphate Cyclohydrolase): A surprising lack of substrate channeling. J Biol Chem. v.21 n.277. p.22168-74.

Caetano-Anollés G, Kim SK, Mittenthal J E. 2007. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. PNAS. v. 104, p. 9358-9363.

Chemmama IE, Chapagain PP, Gerstman BS. 2015. Pairwise amino acid secondary structural propensities. Phys Rev E Stat Nonlin Soft Matter Phys. 91(4):042709.

Cheong CG, et al. 2004. Crystal structures of human bifunctional enzyme aminoimidazole-4-carboxamide ribonucleotide transformylase/IMP cyclohydrolase in complex with potent sulfonyl-containing antifolates. J. Biol. Chem. 279 p.18034-18045.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. Genome Research. 14:1188-1190.

Euzéby JP. 1997. List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. Int J Syst Bacteriol. 47:590-592.

Firestine SM, Paritala H, Mcdonnell JE, Thoden JB, Holden HM. 2009. Identification of inhibitors of N5-carboxyaminoimidazole ribonucleotide synthetase by high-throughput screening. Bio org Med Chem. v.17, n.9, p.3317-3323.

Fujiwara K, Toda H, Ikeguchi M. 2012. Dependence of α-helical and β-sheet amino acid propensities on the overall protein fold type. BMC Structural Biology.12:18.

Greasley SE, et al. 2001. Crystal structure of a bifunctional transformylase and cyclohydrolase enzyme in purine biosynthesis. Nature structural biology. v. 8 n. 5.

Katoh K, Rozewicki J, Yamada KD. 2017. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in Bioinformatics. doi: 10.1093/bib/bbx108.

Kearse M, et al. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28(12): 1647–1649.

Korbel JO, Jensen LJ, Mering CV, Bork P. 2004. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat Biotechnol. 7:911-7.

Malkov SN, Živković MV, Beljanski MV, Hall MB, Zarić SD. 2008. A reexamination of the propensities of amino acids towards a particular secondary structure: classification of amino acids based on their chemical structure. J Mol Model. 14:769–775.

Martí-Arbona R, Um F, Nowak-Lovato KL, Wren MS, Unkefer CJ. 2014. Automated genomic context analysis and experimental validation platform for discovery of prokaryote transcriptional regulator functions. BMC Genomics. 15:1142.

Martinez-Guerrero CE, Ciria R, Abreu-Goodger C, Moreno-Hagelsieb G, Merino E. 2008. GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways. Nucleic Acids Res. 36: W176–W180.

Moreno-Hagelsieb G, Santoyo G. 2015. Predicting Functional Interactions Among Genes in Prokaryotes by Genomic Context. Advances in Experimental Medicine and Biology. DOI: 10.1007/978-3-319-23603-2_5.

Ni L, Guan K, Zalkin H, Dixon JE. 1991. De novo purine nucleotide biosynthesis: cloning, sequencing and expression of a chicken PurH cDNA encoding 5-aminoimidazole-4-carboxamide-ribonucleotide transformylase-IMP cyclohydrolase. Gene. 106(2):197-205.

Nilsson D, Kilstrup M. 1998. Cloning and expression of the Lactococcus lactis purDEK genes, required for growth in milk. Appl Environ Microbiol. 64(11):4321-7.

Nours JL, et al. 2011. Structural analyses of a purine biosynthetic enzyme from Mycobacterium tuberculosis reveal a novel bound nucleotide. The Journal of Biological Chemistry. v.286, n.47, p.40706–40716.

Park Y, et al. 2007. Analysis of virulence and growth of a purine auxotrophic mutant of Xanthomonas oryzae pathovaroryzae. FEMS Microbiol Lett. 276, p. 55–59.

Peifer S, et al. 2012. Metabolic engineering of the purine biosynthetic pathway in Corynebacterium glutamicum results in increased intracellular pool sizes of IMP and hypoxanthine. Microbial Cell Factories. v.11, n.138, 2012.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 26(7):1641-50. doi: 10.1093/molbev/msp077. Epub 2009 Apr 17.

Qin L, et al. 2011. Phosphoribosylamidotransferase, the first enzyme for purine de novo synthesis, is required for conidiation in the sclerotial mycoparasite Coniothyrium minitans. Fungal Genetics and Biology. 48, p. 956–965.

Rayl EA, Moroson BA, Beardsley GP. 1996. The Human purH Gene Product, 5-Aminoimidazole-4-carboxamide Ribonucleotide Formyltransferase/IMP Cyclohydrolase: Cloning , sequencing , expression, purification, kinetic analysis, and domain mapping. The Journal of biological chemistry. 271(4):2225–2233.

Rocha EPC. 2008. The Organization of the Bacterial Genome. Annu. Rev. Genet. 42:211-233.

Rutherford K, et al. 2000. Artemis: sequence visualization and annotation. Bioinformatics. v.16, p.944–945, doi.org/10.1093/bioinformatics/16.10.944.

Santos MA, et al. 2010. Objective sequence-based subfamily classifications of mouse homeodomains reflect their in vitro DNA-binding preferences. Nucleic Acids Research 38.22, p.7927–7942.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular biology and evolution. 30(12):2725–9.

Tranchimand S, Starks CM, Mathews II, Hockings SC, Kappock TJ. 2011. Treponema denticola PurE Is a bacterial AIR carboxylase. Biochemistry. v.50, p.4623–4637.

Verma P, Kar B, Varshney, Roy P, Sharma AK. 2017. Characterization of AICAR transformylase/IMP cyclohydrolase (ATIC) from Staphylococcus lugdunensis. The FEBS Journal. doi: 10.1111/febs.14303.

Xie B, et al. 2005. Symbiotic abilities of Sinorhizobium fredii with modified expression of purL. Appl Microbiol Biotechnol 71, p. 505–514.

Xu L, et al. 2007. Structure-based design, synthesis, evaluation, and crystal structures of transition state analogue inhibitors of inosine monophosphate cyclohydrolase. The Journal of Biological Chemistry. v. 282, n.17, p.13033–13046.

Yuan Z, Wang L, Sun S, Wu Y, Qian W. 2013. Genetic and Proteomic Analyses of a Xanthomonas campestris pv. campestris purC Mutant Deficient in Purine Biosynthesis and Virulence. Journal of Genetics and Genomics, 40, p. 473-487.

Yurkov V, et al. 1994. Phylogenetic Positions of Novel Aerobic, Bacteriochlorophyll a-Containing Bacteria and Description of Roseococcus thiosulfatophilus gen. nov., sp. nov., Erythromicrobium ramosum gen. nov., sp. nov., and Erythrobacter litoralis sp. nov. International Journal of Systematic Bacteriology. v44, n3. p. 427434.

Zhang Y, Morar M, Ealick SE. 2008-a. Structural biology of the purine biosynthetic pathway. Cell Mol Life Sci. 65(23):3699-724.

Zhang Y, White RH, Ealick SE. 2008-b. Crystal structure and function of 5-formaminoimidazole-4-carboxamide-1-β-d-ribofuranosyl 5′-monophosphate synthetase from Methanocaldococcus jannaschii. Biochemistry 47(1):205-217.

Zhao S, et al. 2013. Discovery of new enzymes and metabolic pathways by using structure and genome contexto. Nature. 31;502(7473):698-702. doi: 10.1038/nature12576.

CONSIDERAÇÕES FINAIS

Os resultados aqui apresentados expandem o conhecimento sobre a diversidade e evolução das enzimas do nono e décimo passo da VBP, principalmente no Domínio Bacteria, algo que ainda não havia sido relatado. As enzimas PurO, PurV e PurJ, agora também associadas aos últimos passos da VBP em Bacteria, não podem mais ser considerados assinaturas do Domínio Archaea. A PurP surgiu após a divergência do Domínio Archaea e suas isoformas foram originadas a partir de eventos de duplicação. O trabalho possibilitou também entender a origem das PurVs e PurJs e a relação filogenética delas com as PurHs, que foram a escolha evolutiva das bactérias para os últimos passos da VBP e provavelmente surgiram a partir de um único evento de fusão no ancestral do Domínio Bacteria, e aqui sugerimos como uma enzima que pode ser utilizada para fazer filogenia de alguns táxons específicos.

ANEXOS

(Normas da revista Genome Biology and Evolution - GBE)

GBE
GENOME BIOLOGY AND EVOLUTION

Instructions to Authors

PREPARATION OF MANUSCRIPTS

General Format

Prepare your manuscript text using a Word processing package (save in .doc or .rtf format). Remember to number each page. Use double spacing (space between lines of type not less than 6 mm) throughout the manuscript and leave margins of 25 mm (1 inch) at the top, bottom and sides of each page. Please avoid footnotes. Type references in the correct order and in *GBE* style (see below). Type without hyphenation, except for compound words. Type headings in the style of the journal. Use the TAB key once for paragraph indents. Where possible use Times for the text font and Symbol for the Greek and special characters.

Use the word processing formatting features to indicate strong **Bold**, *Italic*, Greek, Maths,$^{\text{Superscript}}$, and $_{\text{Subscript}}$ characters. Clearly identify unusual symbols and Greek letters. Differentiate between the letter O and zero, and the letters I and l and the number 1. Mark the approximate position of each figure and table.

Check the final copy of your paper carefully, as any spelling mistakes and errors may be translated into the typeset version.

Sections of the Manuscript

- **Title**: The title should accurately advertise the paper's content and contain 150 characters or less including spaces.

- **Authors and affiliations:** Provide the name and institutional address of all authors, match addresses to names using superscript numbers.

- **Corresponding author:** The name of the author to whom all correspondence is to be addressed should be indicated with an asterisk in the author line and specified as follows:
  *Author for Correspondence: John Smith, Department of Science, University of Somewhere, Anytown, USA, telephone number, fax number, email address

- **Data deposition:** Supply all accession numbers for the relevant databases. New sequence data must be deposited in GenBank/DDBJ/EMBL. Any sequence alignments used must be made available as supplemental information or by the corresponding author upon request.

- **Abstract:** The first page of the manuscript should begin with the abstract, which should be a concise summary of the paper. Avoided reference citations in the Abstract; if mentioned, the full reference must be given. The Abstract should contain 250 words or less.

- **Key words:** Up to six key words should be given below the abstract. Key words facilitate retrieval of articles by search engines, web directories and indexes; therefore, terms that are too general should be avoided. The selected key words should not repeat words given in the title. The aim is to assist potential readers to find the article by clearly and specifically describing its subject matter, including aspects of methodology or the theoretical framework.

- **References:** Published articles and those in press (state the journal that has accepted them, provide a doi where possible) may be included. Do not include any reference cited *only* in Supplementary files.
  In the text citation, a reference should be cited by author and date. Do not place text other than the author and date within the parentheses. No more than two authors may be cited per text citation; if there are more than two authors, use et al. in the text (unless more are necessary to distinguish between references).
  In the reference list, list all authors if the author total is five authors or fewer; with

more than five list the first author (only) followed by et al. At the end of the manuscript, the references should be typed in alphabetical order, with the authors' names, year, paper title, journal, volume number, inclusive page numbers, and name and address of publisher (for books only). The name of the journal should be abbreviated according to the World List of Scientific Periodicals. References should therefore be listed as follows:

- Cagan RH, Rhein LD. 1980. Biochemical basis of recognition of taste and olfactory stimuli. In: van der Starre H, editor. *Olfaction and Taste VII.* Oxford: IRL Press. p. 35-44.
- Marshall DA, Moulton DG. 1981. Olfactory sensitivity to alpha-ionone in humans and dogs. Chem Senses. 6:53-61.
- van der Starre H, editor. 1980. *Olfaction and Taste VII*. Oxford: IRL Press.
- Avoid personal communications and mention of unpublished data.
- References to websites should be avoided, but if they are given, the references should give authors (if known), title of cited page, URL in full, and year of posting in parentheses.

- **Tables:** Tables should be prepared on separate sheets and numbered consecutively with Arabic numerals. They must be supplied in an editable format (.xlsx or .docx, for example) rather than an image. They should be self-explanatory and include a brief descriptive title. They should be of such a size that they fit easily onto a journal page, the type area of which is 234 (height) x 185 mm (double column width) or 89 mm (single column width). Footnotes to tables indicated by lower case letters are acceptable, but they should not include extensive experimental details.

- **Illustrations:** All illustrations (line drawings and photographs) must be referred to in the text (as Figure 1 etc.) and should be abbreviated to 'Fig. 1.' only in the figure legend. At online submission, you will be required to submit images electronically in one of the following formats: .jpg, .gif, .tif, .pdf or .eps.

  Each figure should be on a separate page and should be submitted at roughly final magnification. Use sans serif fonts such as Arial or Helvetica in figures. Use

uniform font size in each figure whenever possible and recall that labels should never smaller than 6 pt at final magnification.

- **Electronic submission of figures:** Save figures at a resolution of at least 300 pixels per inch at the final printed size for color figures and photographs, and 600 pixels per inch for black and white line drawings. Color art ***must*** be submitted in CMYK rather than RGB format. Authors should be satisfied with the colors in CMYK (both on screen and when printed) before submission. Please also keep in mind that colors can appear differently on different screens and printers. Failure to follow these guides could result in complications and delays. For useful information on preparing your figures for publication, click here.

- **Figure legends:** These should be included at the end of the manuscript text. Define all symbols and abbreviations used in the figure. Common abbreviations and others in the preceding text need not be redefined in the legend.

- **Color Figures:** All figures submitted to the journal in color will be published in color online at no cost to authors.

- **Permissions for Illustrations and Figures:** Permission to reproduce copyright material, for print and online publication in perpetuity, must be cleared and if necessary paid for by the author; this includes applications and payments to DACS, ARS, and similar licensing agencies where appropriate. Evidence in writing that such permissions have been secured from the rights-holder must be made available to the editors. It is also the author's responsibility to include acknowledgements as stipulated by the particular institutions. Oxford Journals can offer information and documentation to assist authors in securing print and online permissions: please see the Guidelines for Authors section. Information on permissions contacts for a number of main galleries and museums can also be provided. Should you require copies of this, please contact the editorial office of the journal in question or theOxford Journals Rights department.

- **Funding:** Authors who are NIH-funded will have their paper automatically deposited in PubMed Central. Details of all funding sources for the work should be given in the 'Acknowledgements' section. A full list of RIN-approved UK funding agencies may be found here.

  The following convention should be followed:

- The sentence should begin: 'This work was supported by …'
- The full official funding agency name should be given, i.e. 'the National Cancer Institute at the National Institutes of Health' or simply 'National Institutes of Health' not 'NCI' (one of the 27 subinstitutions) or 'NCI at NIH (full RIN-approved list of UK funding agencies)Grant numbers should be complete and accurate and provided in brackets as follows: '[grant number ABX CDXXXXXX]'
- Multiple grant numbers should be separated by a comma as follows: '[grant numbers ABX CDXXXXXX, EFX GHXXXXXX]'
- Agencies should be separated by a semi-colon (plus 'and' before the last funding agency)
- Where individuals need to be specified for certain sources of funding the following text should be added after the relevant agency or grant number 'to [author initials]'.
- Example: This work was supported by the National Institutes of Health [AP50 CA098252 and CA118790 to R.B.S.R.]; and the Education Research Council [hfygr667789].
- Oxford Journals will deposit all NIH-funded articles in PubMed Central. See this page for details. Authors must ensure that manuscripts are clearly indicated as NIH-funded using the guidelines above.

Other Information

- **Conventions:** In general, the journal follows the conventions of the *CSE Style Manual* (Council of Science Editors, Reston, VA, 2006, 7th ed.). Follow *Chemical Abstracts* and its indexes for chemical names. For guidance in the use of

biochemical terminology follow the recommendations issued by the IUPAC-IUB Commission on Biochemical Nomenclature, as given in *Biochemical Nomenclature and Related Documents*, published by the Biochemical Society, UK. For enzymes use the recommended name assigned by the IUPAC-IUB Commission on the Biochemical Nomenclature, 1978, as given in *Enzyme Nomenclature*, published by Academic Press, New York, 1980. Where possible, use the recommended SI (Systéme International) units.

Genotypes should be italicized; phenotypes should not be italicized.

- **Abbreviations:** Try to restrict the use of abbreviations to SI symbols and those recommended by the IUPAC-IUB. Abbreviations should be defined in parentheses after their first mention in the text. Standard units of measurements and chemical symbols of elements may be used without definition in the body of the paper.

- **Chemical Formulae and Mathematical Equations:** Wherever possible, write mathematical equations and chemical formulae on a single line. Submit complicated chemical structures as artwork.

- **Human and Animal Experiments:** The editors draw the authors' attention to the *Declaration of Helsinki* for Medical Research involving Human Subjectshttp://www.wma.net/e/policy/pdf/17c.pdf. In addition, when reporting experiments on animals, authors should indicate whether the institutional and national guidelines for the care and use of laboratory animals were followed.

- **Ethics Guidelines:** In order to guarantee a consistent policy of review and publication,*Genome Biology and Evolution* endorses the Ethics Guidelines offered by the Society for Neuroscience. These guidelines describe the responsibilities and expected conduct not only of authors of scientific articles, but also of the editors and reviewers. We encourage our readers to take a few minutes to download and look over these guidelines athttp://www.sfn.org/guidelines/.

- **Crossref Funding Data Registry** : In order to meet your funding requirements authors are required to name their funding sources, or state if there are none,

during the submission process. For further information on this process or to find out more about the CHORUS initiative please click here.

LANGUAGE EDITING

Language editing, if your first language is not English, to ensure that the academic content of your paper is fully understood by journal editors and reviewers is optional. Language editing does not guarantee that your manuscript will be accepted for publication. For further information on this service, please click here. Several specialist language editing companies offer similar services and you can also use any of these. Authors are liable for all costs associated with such services.