Comparative Genomics of Arginine Biosynthesis Pathways and Regulons in Human Microbiome.

Olga M. Sigalova¹, Dmitry A. Rodionov²

¹A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia.
olga.sigalova88@gmail.com

²A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia.
rodionov@burnham.org

Abstract. The aim of the Human Microbiome Project (HMP) is to infer metabolic interconnections between diet and microbiome composition. This study was devoted to the systematic analysis of arginine de novo biosynthesis, transport and regulation in HMP bacteria with significance to human health. We used a subsystems-based comparative genomics approach to reconstruct arginine biosynthesis and salvage pathways and ArgR transcriptional regulons in diverse bacterial genomes from the HMP project. In the Bacteroidetes phylum we predicted a novel isoform of N-acetylglutamate synthase, which is non-orthologous to known enzymes catalyzing acetylation of L-glutamate on the first step of arginine biosynthesis. As result, for each analyzed HMP microorganism we assigned the arginine prototrophic or auxotrophic phenotype and predicted arginine uptake transporters and ArgR regulon composition in many of them. The results of this study can be useful for understanding metabolic interactions between the members of human microbiota.

Keywords: arginine biosynthesis, arginine transporter, phenotype prediction, arginine repressor, bacteria

1 Introduction

Biosynthesis of arginine is remarkable considering variability of pathway variants and multiplicity of non-orthologous enzymes that can catalyze some of its reaction steps. Systematic analysis of the variable pathway enzymes across bacterial lineages and prediction of prototrophic and auxotrophic phenotypes for arginine synthesis is an essential step in understanding...
metabolic interactions within bacterial communities such as human gut microbiome.

At least four distinct variants of arginine biosynthesis pathways have been described in prokaryotes. First, the arginine biosynthesis can proceed either through a linear or cyclical (via acetyl cycle) pathway. In *Escherichia coli*, arginine is produced from L-glutamate in eight sequential reaction steps catalyzed by enzymes ArgA to ArgH [5]. On the first step, glutamate is acetylated at the N-α position using N-acetylglutamate synthase (ArgA). The flow of acetylated precursors then proceeds until the fifth step of the pathway when N-acetyl-ornithine (NAG) is deacetylated by NAG deacetylase (ArgE) to form L-ornithine. L-ornithine is subsequently transcarbamoylated by ornithine carbamoyltransferase (ArgF) to form L-citrulline, which is converted to L-arginine in two more reaction steps. This pathway is now considered to be neither the most widespread nor the ancestral variant of arginine production. In a more energy efficient pathway, formation of L-ornithine is catalyzed by ornithine acetyltransferase (ArgJ), which recycles the acetyl group back on glutamate. In this cyclical pathway, some portion of N-acetyl-ornithine is still produced from glutamate to replenish loss of NAG during bacterial life cycle [2].

Next, the pathways may differ by the point, at which arginine precursors are deacetylated. In the ‘classical’ pathway in *E. coli*, N-acetyl-L-ornithine is deacetylated to L-ornithine by ornithine deacetylase (ArgF1) on the fifth step with its subsequent transcarbamoylation to form L-citrulline. In another variant described for some *Bacteroidetes* [29] and *Proteobacteria* [21], on the fifth step of the pathway, N-acetyl-L-ornithine is transcarbamoylated by N-acetylnornithine carbamoyltransferase (ArgF2) resulting in N-acetyl-L-citrulline, which is then deacetylated into L-citrulline by acetylnornithine deacetylase (ArgE). The ArgF2-containing microorganisms described so far can synthesize arginine using the linear pathway, however, variants with the cyclical pathway may also exist.

Finally, there are several mechanisms of preventing spontaneous cyclization of glutamate derivatives and therefore preventing the reactions from proceeding to proline biosynthesis instead of arginine. In all pathways
described above, cyclization is prevented by acetylation of L-glutamate at the N-position in the first step of the pathway [2]. However, it has been recently shown that some archaea use a dedicated carrier protein (L-2-amino adipate carrier protein, LysW) to protect the early intermediates of arginine biosynthesis [23].

Moreover, formation of acetyl-glutamate from L-glutamate and acetyl-coenzyme A (acetyl-CoA) on the first step of the pathway can be catalyzed by a variety of non-orthologous enzymes (for early review see [37]). Classical ArgA protein from the Gcn5-related N-acetyltransferase (GNAT) superfamily was first described in E. coli [5]. This 440 amino acid protein has N-Acetylglutamate synthase (NAGS) and acetylglutamate kinase (NAGK) domains. NAGK domain is present in another enzyme from the second step of arginine biosynthesis, ArgB, which supports a hypothesis that “classical” ArgA protein emerged in the course of evolution as a result of fusion of a one-domain GNAT acetyltransferase with ArgB.

The bifunctional ArgA-B enzyme with both NAGS and NAGK activities was identified in the Xanthomonadales order of Gammaproteobacteria [26]. Interestingly, this ArgA-B gene has higher sequence similarity to mammalian NAGS than to other bacterial ArgA proteins, and therefore may be ancestral to mammalian NAGS.

Two-domain ArgA proteins are absent from many major divisions of Bacteria, including Thermotogales, Aquificales, Planctomycetes, and Bacteroidetes. On the other hand, “short” NAGS (S-NAGS), 160-200 amino acid long one-domain acetyltransferases from the GNAT superfamily, were found to catalyze glutamate acetylation from acetyl-CoA in a variety of bacteria from different phylogenetic branches, including Campylobacter jejuni [9], Mycobacterium tuberculosis [7], Corynebacterium glutamicum [25], as well as species from the Thermus, Streptomyces and Deinococcus genera [36]. In this work, we report a novel S-NAGS enzyme (ArgA2) from the GNAT superfamily, catalyzing the first step of arginine biosynthesis in the Bacteroidetes phylum and in closely related newly discovered phylum Caldithrix with two species described up to date [19,20]. The previously described S-NAGS enzymes and ArgA2 in the Bacteroidetes and Caldithrix
phyla are non-orthologous to each other as they belong to different proteins families within the GNAT superfamily.

Apart from the bifunctional ArgA-B enzymes, some GNAT acetyltransferases were found as domain fusions with some other enzymes from the arginine biosynthesis pathway. Two-domain ArgH-A proteins catalyzing the first and the last steps in the pathway were characterized in several Gammaproteobacteria including Moritella spp. [35], Pseudoalteromonas haloplanktis and Idiomarina loihiensis [36]. In some Bacteroidetes species, the newly characterized ArgA2 enzyme was found in a fusion with ArgG. Moreover, a few Pedobacter species possess two paralogous ArgA2, with one paralog fused to the peptidase M20 domain present in ArgE.

In addition to acetyltransferases from the GNAT superfamily, acetyl-glutamate can be produced by the ornithine acetyltransferase (OAT) superfamily protein ArgJ. As discussed before, in organisms with ArgJ, acetyl group is transferred from acetylinornithine to glutamate via the acetyl cycle. However, some bacteria including Geobacillus stearothermophilus possess bifunctional ArgJ. In these organisms, ArgJ can use both acetyl-CoA and acetylnornithine as acetyl group donor [18].

Variability in the arginine biosynthesis pathways goes beyond diversity of enzymes and methods of protection of glutamate derivatives from cyclization. Different bacteria regulate arginine production differently. First, the transcription of arginine biosynthesis genes is negatively controlled by the arginine-responsive regulators, first described in Esherichia coli [15]. Several ArgR-AhrC-family regulators were isolated and experimentally characterized in diverse bacteria including E. coli [11], Bacillus subtilis [4] and Salmonella typhimurium [12] and B. stearothermophilus [6]. Another transcription factor from the AraC/XylS family family (also named ArgR) that controls the arginine biosynthesis gene in Pseudomonas spp. was recently described [13]. Regulators from the ArgR-AhrC family are widely distributed among bacteria, most of which have a single copy of argR gene. However, in some Firmicutes such as Staphylococcus spp., two paralogs of arginine repressors (referred to as ArgR and AhrC) were shown to regulate the arginine metabolism genes differently and not in a fully complementary way.
Further, the arginine biosynthesis and uptake genes in several Firmicutes from the *Clostridia* genus are transcriptionally controlled via cis-regulatory RNA elements, called T-boxes, that respond to the concentration of specific uncharged tRNAs and thus measures the amino acid availability in the cell [32,33].

Second, the arginine biosynthesis is regulated on the enzyme level by the final product of the pathway, L-arginine, in different ways. In organisms with linear pathways and no ArgJ, arginine inhibits NAGS, the first step enzyme. When acetyl group is recycled and most acetylglutamate is produced by OAT in the acetyl cycle, NAGK (the second step enzyme ArgB) becomes regulated (for review, see [37]).

Prokaryotes can utilize different transporters to uptake arginine from the environment. Various arginine transporters have been experimentally characterized in diverse bacteria, including the ArgT-HisJQMP ABC transport system in *Salmonella typhimurium* and *E. coli*, the ArtPIQMJ ABC system in *E. coli* [34], the AotJQMP ABC transporter in *Pseudomonas aeruginosa* [22], high- and low-affinity arginine permeases in *Trypanosoma cruzi* [24],[3], putative arginine transporter ArgW from the COG3314 family in *Shewanella* spp. [27], the arginine periplasmic binding protein in *Thermotoga maritima* [1]. However, systematic comparison of these transporters is still missing.

Considering heterogeneity in arginine biosynthesis, regulation and transport in prokaryotes, the goal of this work was to reconstruct diverse arginine biosynthesis and uptake pathways and regulons in over thousand human microbiome bacteria sequenced as a part the Human Microbiome Project (HMP). This task was addressed by the methods of comparative genomics using the metabolic subsystems approach implemented in the SEED database and the regulon reconstruction approaches in the RegPredict tool.

## Results

In the first part of this project, the arginine biosynthesis pathways were reconstructed in a set of 1143 bacterial genomes from seven phyla out of those sequenced in the course of Human Microbiome Project. Results were
represented in the subsystem table available in the PubSEED database. The reconstructed subsystem is an extended spreadsheet with lines corresponding to genomes and columns denoting different arginine biosynthesis enzymes, transporters and regulators. Each genome in the subsystem table was assigned to a particular arginine pathway variant based on the following information: (1) the predicted arginine biosynthesis phenotype (auxotroph - A or prototroph - P); (2) the predicted ability to salvage arginine from the environment by the presence of arginine transporter (salvage - S); (3) the predicted transcriptional regulon for arginine metabolism genes (presence or absence of ArgR). In addition, we captured different variants of the arginine biosynthesis pathways, that are (1) enzymes used for acetylglutamate production (family or type of isoenzyme, e.g. classical NAGS, S-NAGS or OAT); (2) enzyme variant used for transcarbamoylation on the fifth step (ornithine carbamoyltransferase ArgF or N-acetylorlnithine carbamoyltransferase ArgF2). We further used the RegPredict comparative genomics tool to analyze transcriptional regulons for arginine biosynthesis and transport genes controlled by regulators from the ArgR family. Predicted DNA binding sites and reconstructed regulons for ArgR transcription factors in different bacterial lineages were uploaded into the RegPrecise database. The obtained DNA binding motifs of ArgR regulators in diverse bacterial phyla showed significant conservation.

The second part of this research was dedicated to the in-depth analysis of the arginine biosynthesis in bacteria from the Bacteroidetes phylum that play an important role in human gut microbiome. First, we predicted a putative GNAT acetyltransferase (ArgA2) catalyzing the first acetylation step in the arginine biosynthesis. Search against non-redundant protein database in NCBI revealed that ArgA2 is almost exclusively limited to the Bacteroidetes phylum. The only exceptions are two bacteria from the newly discovered Caldithrix phylum, which is confidently assigned to the Bacteroidetes-Chlorobi-Caldithrix superphylum (unpublished results), as well as two Betaproteobacteria and one Gammaproteobacterium that have probably acquired the argA2 gene through horizontal gene transfer. Moreover, no Bacteroidetes possess classical ArgA enzyme, while ArgJ is present only in a few bacteria.
ArgA2 gene was identified in 470 genomes of the Bacteroidetes phylum in the NCBI non-redundant database. In Candidatus Uzinura diaspicolor, Candidatus Sulcia muelleri and Blattabacterium species, ArgA2 was found to be merged to ArgG gene from the seventh step of the pathway. This two-domain protein apparently has dual activity in these organisms, because it is encoded by the arginine operon, with no other NAGS and ArgG homologs found in their genomes. Microscilla marina has the ArgA2 domain fused to an N-terminal short septidase family S49 domain. In Pedobacter heparinus, ArgA2 gene was duplicated, with one paralog fused to the peptidase M20 domain. Phylogenetic tree built on 282 ArgA2 sequences with less than 95% similarity in 191 conservative positions is generally consistent with the phylogenetic division within the Bacteroidetes phylum. Bifunctional ArgA-G proteins form a distinct branch on the tree, confirming that there was only one such domain fusion event.

Genome context analysis of the argA2 genes was performed using the subsystems approach in PubSEED. It was found that argA2 almost always lies at the beginning of the arginine operon before the argG gene, which explains fusion of these two genes in several bacterial species. In the cases when the arginine repressor gene argR is present in a genome, it typically lies upstream of argA2. Out of 380 Bacteroidetes genomes in PubSEED, 96 bacteria (25%) are arginine auxotrophs that lack the complete pathway. In the remaining 284 genomes, ArgA2 is present in 280 cases. However, two strains of Rhodothermus marinus and two strains of Salinibacter ruber have all genes required for the arginine biosynthesis except the first step NAGS. As a hypothesis, these two extremophilic bacterial species may synthesize arginine in the archaeal way, potentially using a non-orthologous isoform of LysW to stabilize glutamate derivatives. These organisms also encode a hypothetical protein with ATPase domain in arginine operon, while the glutamate-LysW ligase ArgX typically contains the ATP grasp domain. Anyway, there are currently no proofs for this hypothesis, and it requires more investigations. As noted before, the argJ gene was found in nine Bacteroidetes genomes from the subsystem. Interestingly, in all these bacteria there are two arginine operons with highly overlapping gene content, a shorter operon with the argJ and argBCD genes and a longer one
with the argA2 and argBCDEFGH genes, thus the ArgJ-encoding operon is redundant and presumably nonessential in these species. However, the ArgJ-encoding species belong to different genera and orders of the Bacteroidetes phylum, suggesting that the ArgJ-encoding operon was subject to several horizontal gene transfer events.

Activity of the predicted ArgA2 enzyme from Bacteroides thetaiotaomicron will be tested in the collaborating laboratory. The tertiary structure prediction for a novel enzyme was performed through the structural alignment of the ArgA2 proteins with other acetyltransferases from the GNAT and OAT superfamilies that have solved 3D structures.

Next, we looked at the occurrence of ArgF variants in the Bacteroidetes phylum. It was found that both pathways (through deacetylation of either N-acetyl-L-ornithine or N-acetyl-L-citrulline) are equally frequent across the phylum. Finally, for all bacteria in the PubSEED subsystem, which have the arginine repressor, we reconstructed phylogeny of ArgR proteins and assigned their DNA binding motifs.

3 Materials and methods

3.1 Multiple sequence alignment and phylogenetic analysis of ArgA2

On the first step, we searched the NCBI non-redundant protein database using BLAST and selected 474 protein sequences with more than 40% identity to the predicted ArgA2 protein from Bacteroides thetaiotaomicron, 470 being from the Bacteroidetes phylum and the remaining four proteins from Betaproteobacteria, Gammaproteobacteria and Caldithrix.

Multiple protein sequence alignment was performed using MEGA software with default parameters. 191 conservative amino acid positions were selected for phylogenetic analysis. Sequences with above 95% similarity in these 191 sites were filtered out, resulting in the list of 294 sequences from 282 species. Phylogeny of ArgA2 was reconstructed by maximum likelihood approach with 100 bootstraps using MEGA. Phylogenetic trees were visualized with Dendroscope software.
3.2 Genome context analysis

Metabolic reconstruction and genome context analysis were performed with the metabolic subsystems approach implemented in the PubSEED database. Functional roles of enzymes from the arginine biosynthesis, arginine transporters as well as the arginine repressor ArgR were added to the manually curated subsystems in PubSEED, one for the HMP genomes and another one for the Bacteroidetes phylum. These subsystems were filled with 1143 and 380 genomes, respectively. The HMP subset of 1143 bacterial genomes represented eight bacterial phyla, namely Actinobacteria, Bacteroidetes, Chlorobi, Firmicutes, Fusobacteria, Proteobacteria, Spirochaetes and Synergistetes.

3.3 ArgR binding sites search

Maximum likelihood phylogenetic tree of the ArgR regulators was constructed for proteins from various bacterial species (stains excluded) using MEGA. Distinct DNA binding motifs of were found for distinct branches of the ArgR tree. Motifs were identified using the MEME and SignalX motif search tools and further confirmed by phylogenomic footprinting using multiple DNA sequence alignments of upstream regions of ArgR-regulated genes. Positional weighted matrices (PWMs) were obtained for each identified set of ArgR binding sites in each studied taxonomic group. Genomes were further scanned with the obtained lineage-specific PWM profiles for additional candidate ArgR-binding sites using the GenomeExplorer software.

3.4 Structural alignment and 3D structure prediction

Multiple protein sequence and structure alignments of several ArgA2 proteins from distinct Bacteroidetes order and acetyltransferases with known 3D structure was performed with the PROMALS3D server. Alignments were visualized with the ESPript software. 3D structure of ArgA2 was predicted with the Modeller software.
4 Literature


