Bioinformation Volume 5 *open access*

www.bioinformation.net Issue 10 **Hypothesis**

Genomic heterogeneity within conserved metabolic pathways of *Arthrobacter* species - a bioinformatic approach

Ayon Pal[1], Uttam Kumar Mondal[2], Subhasis Mukhopadhyay[3], Asim Kumar Bothra[2]*

[1]Department of Botany, Raiganj College (University College), Raiganj-733134, Uttar Dinajpur, West Bengal, India; [2]Cheminformatics Bioinformatics Laboratory, Department of Chemistry, Raiganj College (University College), Raiganj-733134, Uttar Dinajpur, West Bengal, India; [3]Department of Bio-Physics, Molecular Biology and Bioinformatics, University of Calcutta, 92, APC Road, Kolkata-700009, West Bengal, India; Asim Kumar Bothra – Email: asimbothra@gmail.com; Phone: +91 9474441570; Fax: +91 3523 242580; *Corresponding author

Received November 11, 2010; Accepted January 05, 2011; Published February 15, 2011

Abstract:
A comparative genomic analysis of three species of the soil bacterium *Arthrobacter* was undertaken with specific emphasis on genes involved in important and core energy metabolism pathways like glycolysis and amino acid metabolism. During the course of this study, it was revealed that codon bias of a particular species, namely *Arthrobacter aurescens* TC1, is significantly lower than that of the other two species *A. chlorophenolicus* A6 and *Arthrobacter* sp. FB24. The codon bias was also found to be negatively correlated with gene expression level which is determined by computing codon adaptation index of the genes. Uniformity in codon usage pattern among three species is evident in terms of genes which has high codon bias and multifunctional nature. Further, it was observed that this trend is present amongst the genes of important metabolic pathways, such as glycolysis and amino acid metabolism. The evolutionary divergence of the pathway gene sequences was calculated and was found to be equivalent in nature in the case of *Arthrobacter* sp. FB24 and *Arthrobacter chlorophenolicus* A6, but turned out to be dissimilar in the case of *Arthrobacter aurescens* TC1. A strong correlation between synonymous substitution rate and effective codon number or Nc was also observed. These observations clearly point out that the genes having low bias, in *Arthrobacter aurescens* TC1, and even of those that are part of highly conserved metabolic pathways like glycolysis and amino acid ensemble pathways have undergone a different type of evolution and might be subjected to positive selection pressure in comparison with *Arthrobacter* sp. FB24 and *Arthrobacter chlorophenolicus* A6.

Keywords: *Arthrobacter aurescens* TC1, *Arthrobacter chlorophenolicus* A6, *Arthrobacter* sp. FB24, Actinomycetales, Atrazine, Metabolic pathways, KEGG, MetaCyc, Glycolysis, Amino acid metabolism, Correspondence analysis, RSCU, Nc, Effective codon number, GC3, Codon adaptation index, Ks, Ka, Positive selection.

Background:
Arthrobacter is a ubiquitous soil-dwelling eubacterium **[1]**. They are also found in extreme environments, such as deep subsurface soils, arctic sea and radioactive waste tanks. They display exceptional nutritional versatility, and are proficient in using a wide range of compounds as carbon source and many of the species are potential bioremediation agents. They are competent enough to biodegrade serious environmental pollutants like endothal, nicotine, 2, 4-D, nitroglycerine, phenolic mixtures and atrazine containing herbicides **[2]**. The genus *Arthrobacter* is characterized by aerobic, high GC, gram positive bacteria categorized into the class Actinobacteria, order Actinomycetales, and family Micrococcaceae **[1]**. Of the many species of *Arthrobacter*, the complete genome sequence of three species had been published. These rich biological pools of data offer an important opportunity for comparing *Arthrobacter* species, both at the genomic as well as genic level, as an entirely new insight into organismal biology and gene expression is presented by comparative studies. In recent years, the depth and width of pathway molecular network interaction data made available through pathway informatics databases like KEGG **[3]**, BioCyc, MetaCyc **[4]** and other such endeavours have paved the way for comparing organisms, not only at the whole genome level but also across the core biochemical and biophysical metabolic pathways. Metabolic pathways contain important information on the function of organisms. In-depth analysis of gene sequences involved in major metabolic pathways provides an insight about the evolutionary process and comparative analysis of metabolic pathways among species is an effective means of obtaining information about the functional relation of organisms **[5]**. A striking feature of metabolism is the similarity of the basic metabolic pathways even among vastly different species **[6]**. These striking similarities in metabolism are most likely the result of the high efficiency of these pathways, and of their early appearance in evolutionary history **[7] [8]**. Two such important pathways are the hexose breakdown via EMP pathway of glycolysis and metabolism of amino acids.

Glycolysis is thought to be the archetype of a universal metabolic pathway that occurs, with variations, in nearly all organisms, aerobic and anaerobic alike. The wide occurrence of glycolysis indicates that it is one of the earliest known metabolic pathways **[9]**. The metabolism of amino acids also follows a common outline in large number of organisms, the differences being in their mode of regulation. Amino acid metabolism, in fact, is an assemblage of multitude of pathways that are primarily concerned with the biosynthesis and degradation of the twenty important amino acids participating universally in protein biosynthesis. A

comparative account of these two important metabolic pathways, namely glycolysis and amino acid metabolism, alongside whole genome comparison of an organism is supposed to provide us with a better understanding about the characteristics of closely related organisms.

The three strains of the genus *Arthrobacter* which are included in our study are *Arthrobacter aurescens* TC1, *Arthrobacter chlorophenolicus* A6 and *Arthrobacter* sp. FB24. *Arthrobacter aurescens* TC1 is an aerobic Gram-positive originally isolated from soil at a South Dakota atrazine herbicide spill site. It has the ability to grow on a wide variety of carbon compounds and to catabolize a variety of xenobiotics, such as glyphosate, methyl tert-butyl ether, 2, 4- dichlorophenoxyacetate (2, 4-D), nicotine, 4-nitrophenol, dimethyl silanediol, fluorene, phthalate, nitroglycerine and various s-triazine compounds **[2] [10]**.

Arthrobacter chlorophenolicus A6 is an aerobic gram positive isolated from soil at Fort Collins, Colorado that can degrade phenolic mixtures containing phenol, chlorophenol and nitrophenol **[11]**. *Arthrobacter* sp. FB24 is a non-sporulating, aerobic, mesophile isolated from chromate and xylene enriched soils capable of degrading hydrocarbons and is radiation resistant **[12].** A comparative genomic study of these three species at the level of metabolic pathways was performed to obtain a better insight about the similarities and dissimilarities in characteristics such as codon usage, gene expression, evolutionary divergence and nature of selection in these closely related species of *Arthrobacter*.

Methodology:
The complete genome sequence of the three strains of the genus *Arthrobacter*, namely, *Arthrobacter* sp. FB24 (NCBI/RefSeq: NC_008537; NC_008538; NC_008539; NC_008541), *Arthrobacter chlorophenolicus* A6 (NCBI/RefSeq: NC_011879; NC_011881; NC_011886) and *Arthrobacter aurescens* TC1 (NCBI/RefSeq: NC_008711; NC_008712; NC_008713) were downloaded from the Integrated Microbial Genomes website http://www.img.jgi.doe.gov **[13]**. The nucleotide sequences encoding the information for the production of proteins and enzymes involved in amino acid metabolism pathways and hexose degradation by EMP pathway of glycolysis were sorted out using references from pathway metabolic information database KEGG **[3]** and MetaCyc **[4]**. The genome sequence of all the three *Arthrobacter* strains was scanned to isolate the different nucleotide sequences coding for the different tRNAs, proteins and enzymes involved in the biosynthesis and degradation of the standard amino acids along with the enzymes involved in the breakdown of hexose via EMP pathway.

We started off with basic local alignment of the individual gene sequences using the web based local alignment tool BLASTn **[14]** to find out the percentage of identity of the concerned gene sequences in the three *Arthrobacter* species. We then calculated the effective number of codons (ENc or Nc) as per Wright **[15]**, which is a measure of synonymous codon usage bias, for each nucleotide sequence encoding enzymes of the amino acid metabolism and EMP pathway. For calculating Nc at first (F caret) was calculated for each synonymous group: **(See Supplementary material 1)** where p represents the proportion of usage of a codon i within its synonymous group of size j, and the total usage of the synonymous group. The average for synonymous group of same size (i.e., 2, 4 and 6), and Nc is calculated as: **(See Supplementary material 2)**

Further, the frequency of guanine and cytosine at the synonymous third position of codon, known as GC3 content was calculated using CodonW. Correspondence analysis was also carried out for nucleotide sequences to investigate the major trend in relative synonymous codon usage (RSCU). A commonly used and well-accepted measure for calculating the expression levels of gene sequences known as the Codon Adaptation Index or CAI was then calculated using the method proposed by Sharp and Li **[16]** but with an improved implementation proposed by Xia **[17]**. In order to obtain an insight about the evolutionary divergence of the gene sequences of the three different *Arthrobacter* species together with the nature of the selection forces acting on them, we calculated the synonymous and non-synonymous substitution rates of the glycolysis and amino acid metabolism pathway gene sequences. Multiple sequence alignment was carried out using ClustalX **[18]** for all the genes involved in glycolysis, together with a number of genes involved in amino acid metabolism and a few randomly selected genes from the genome of the three different species of *Arthrobacter*. The curated alignment data files were used as input for the program DnaSP **[19]** to calculate the DNA sequence variation at synonymous (Ks) and nonsynonymous sites (Ka) **[20].**

Results and Discussion:
An important property of the genetic code is its degeneracy due to the presence of synonymous codons; some synonymous codons are used more abundantly or 'preferred'. A well accepted parameter for studying codon bias is the Nc index **[15]**. It is a simple measure of overall codon bias and ranges from twenty to sixty one where 20 is the value obtained when only one codon is used for each amino acid (i.e., the codon bias is maximum) and 61 is the value obtained when all synonymous codon for each amino acid are equally used (i.e., no codon bias). The whole genome of *Arthrobacter aurescens* TC1, *Arthrobacter chlorophenolicus* A6 and *Arthrobacter* sp. FB24 contains 4794, 4745 and 4606 gene sequences respectively. The Nc score of all the genes in the genome of *Arthrobacter aurescens* TC1, *Arthrobacter chlorophenolicus* A6 and *Arthrobacter* sp. FB24 ranges between 21.81-61, 20-61 and 20-61 respectively **(Table 1 see Supplementary material)**. An overwhelming 71% of *Arthrobacter aurescens* TC1 genes have effective codon numbers or Nc well above 40, whereas in comparison *Arthrobacter chlorophenolicus* A6 and *Arthrobacter* sp. FB24 have only 21% and 37% of its genes with effective codon numbers above 40. The mean whole genome Nc of *Arthrobacter aurescens* TC1 is substantially elevated at 43.54 compared to that of *Arthrobacter chlorophenolicus* A6 and *Arthrobacter* sp. FB24, which is 37.62 and 39.16 respectively. This signifies that the codon degeneracy is significantly high or alternatively, codon bias is minimal in *Arthrobacter aurescens* TC1. When we consider the Nc of the genes involved in amino acid metabolism of *Arthrobacter aurescens* TC1, the Nc scale ranges from 26.7 to 58.8 with an average of 40.43 which is quite high compared to that of the remaining two strains – *Arthrobacter chlorophenolicus* A6 and *Arthrobacter* sp. FB24. In the case of *Arthrobacter chlorophenolicus* A6, the Nc of the genes involved in amino acid metabolism ranges between 25.7-50.2 with an average of 34.33, which is the lowest among the three strains. *Arthrobacter* sp. FB24 has a mean Nc of 35.58 for the genes responsible for amino acid metabolism with the scale ranging from 25.2 to 54.1. Thus, *Arthrobacter chlorophenolicus* A6 genes exhibit higher codon bias both in terms of the genome and the metabolic pathways of amino acid metabolism.

The study of the codon usage pattern of the genes concerned with glycolysis shows comparable results where *Arthrobacter chlorophenolicus* A6 genes have a higher codon bias compared to the other two strains. In *Arthrobacter chlorophenolicus* A6 the genes coding for the different enzymes of the glycolytic pathway have Nc ranging between 27-39.8 with an average of 33.11. *Arthrobacter aurescens* TC1, on the contrary, has the least codon bias with Nc of the glycolysis pathway genes ranging between 29.1 to 54.7 with a mean of 40.15. *Arthrobacter* sp. FB24 glycolytic pathway genes reveal moderate codon bias with Nc ranging between 28.4-45.7 with an average of 36.51. *Arthrobacter chlorophenolicus* A6, thus, consistently exhibited a higher codon bias not only in terms of the whole genome but also in terms of the two important energy metabolic pathways of amino acid metabolism and glycolysis.

A plot of Nc values versus GC3 known as Nc plot **[15] (Figure 1)** was constructed. This plot clearly indicates the anti-correlation between Nc and GC3 for all the three species. *Arthrobacter aurescens* TC1 genes were found to be scattered in comparison to the other two species which forms a common cluster.

As for the synonymous codon usage bias, the Codon Adaptation Index (CAI) was estimated for the gene sequences related to glycolysis pathways, the amino acid metabolism pathways and a few randomly selected genes from the genome. Codon Adaptation Index is a well-accepted parameter

for studying the expressivity of a gene and assesses the extent to which the selection has been effective in moulding the pattern of codon usage. The Codon Adaptation Index ranges from 0 to 1.0, with higher CAI values signifying that the gene of concern has a higher degree of expressivity **[16]**. Codon adaptation index has also been successfully used in predicting m-RNA expression in different microorganisms **[21] [22]**. Our study reveals that *Arthrobacter chlorophenolicus* A6 has the highest average CAI values both in terms of glycolysis and amino acid metabolism, whereas *Arthrobacter aurescens* TC1 and *Arthrobacter* sp. FB24 have quite similar codon adaptation index values with respect to each other.

A comprehensive and comparative study of the ten glycolytic pathway genes from the three different *Arthrobacter* species **(Table 3 see Supplementary material)** shows that for all the gene sequences starting from glucokinase (EC 2.7.1.2), the enzyme which intiates glycolysis by converting glucose to glucose-6-phosphate, till pyruvate kinase (EC 2.7.1.40), the enzyme which catalyzes the last step of glycolysis converting phosphoenol pyruvate to pyruvate, *Arthrobacter aurescens* TC1 consistently displays a high Nc value in comparison with *Arthrobacter* sp. FB24 and *Arthrobacter chlorophenolicus* A6. In the case of the enzymes glucokinase, glucose-6-phosphate isomerase, triose phosphate isomerase, phosphoglycerate kinase and phosphoglycerate mutase, *Arthrobacter aurescens* TC1 has Nc value well above 40, whereas at the same time the other two species have Nc values well below 40, pointing towards some degree of codon bias. For the remaining enzymes too, the Nc values for *Arthrobacter aurescens* TC1 are comparatively higher than the other two species. The codon adaptation index (CAI) values were found to be negatively correlated with the Nc value where enzyme gene sequences with higher Nc has relatively low CAI values ($r = -0.8514$ for glycolytic pathway genes and $r = -0.9$ for amino acid metabolism pathway genes; $p>>0.01$). We did not observe any sort of deviation from this consistent trend and for each individual cluster of enzymes, gene sequences with the highest codon bias or low Nc has the highest expression level or CAI value. The CAI values for all the gene sequences with Nc values below 30 was significantly high zeroing near to 0.8 to 0.9. *Arthrobacter chlorophenolicus* A6 exhibited an overall elevated range of CAI values for all the glycolytic enzyme genes. Apart from glucokinase, one or a few of all the remaining enzyme gene sequences exhibited CAI values above 0.8. *Arthrobacter aurescens* TC1 and *Arthrobacter* sp. FB24 on the other hand, had moderate CAI values. The multifunctional enzyme enolase (EC 4.2.1.11) exhibited significantly high CAI values and low Nc values (27 – 29.2) in all the three species concerned.

A comprehensive and comparative study of the gene sequences involved in the amino acid metabolism pathways reveals the participation of more than three hundred gene sequences (data not shown). *Arthrobacter aurescens* TC1, *Arthrobacter* sp. FB24 and *Arthrobacter chlorophenolicus* A6 have 364, 324 and 308 genes respectively coding for the different enzymes and RNA involved in the whole host of synthesis and breakdown pathways collectively termed as amino acid metabolism. The codon usage index Nc for amino acid metabolism pathway gene sequences range between 26.7-58.8, 25.7-50.2 and 25.2-54.1 for *Arthrobacter aurescens* TC1, *Arthrobacter chlorophenolicus* A6 and *Arthrobacter* sp. FB24 respectively. A thorough analysis reveals that in *Arthrobacter aurescens* TC1, a very high percentage (51.10%) of all the genes involved in amino acid metabolism have Nc values in excess of 40. In contrast, only 7.76% and 13.37% of *Arthrobacter chlorophenolicus* A6 and *Arthrobacter* sp. FB24 amino acid metabolism genes respectively have Nc values above 40. The relation between Nc and CAI of the genes involved in amino acid metabolism given in Table 2 shows that gene sequences with Nc score less than 40 uniformly return a higher average CAI values compared to the sequences with Nc score of more than 40 in all the three species in question.

In this study, our observation confirms that the Nc value had an inverse relationship with the CAI value. A high Nc value indicates low codon bias and vice versa ($r = -0.8514$ for glycolytic pathway genes and $r = -0.9$ for amino acid metabolism pathway genes; $p>>0.01$). We observed that the gene sequences with high Nc score displayed lower expression level (CAI values).

We have picked up 13 different gene sequences **(Table 4 see Supplementary material)** involved in the various aspects of amino acid metabolism, like alanine and aspartate metabolism, aromatic amino acid biosynthesis, histidine and methionine metabolism and lysine biosynthesis, to name a few. Based on their functional similarity and comparable lengths of the sequences from the three concerned species we clustered them. Table 4 displays a picture similar to what we have observed previously in the case of glycolysis. *Arthrobacter aurescens* TC1 have higher Nc values for all its gene sequences compared to the other two species. *Arthrobacter chlorophenolicus* A6 on the other hand, have Nc values that are quite low. In this respect mention must be made of the gene sequence coding for methionyl-tRNA formyltranferase (EC 2.1.2.9) in *Arthrobacter aurescens* TC1, since it has an overwhelming high Nc score of nearly 52. Relatively higher Nc values were also observed in the case of nucleotide sequences coding for phosphoserine phosphatase SerB, chorismate synthase and arginosuccinate lyase in *Arthrobacter aurescens* TC1. In all these cases, the other two organisms *Arthrobacter* sp. FB24 and *Arthrobacter chlorophenolicus* lagged way behind in terms of their Nc scores. Those gene sequences of *Arthrobacter aurescens* TC1 which have Nc values less than 40 showed somewhat similarity with the Nc score of the same gene sequences found in the other two organisms. To substantiate this point we might mention that argininosuccinate synthase, 4-aminobutyrate aminotransferase apoenzyme, cysteine synthase, cystathionine gamma lyase and dihydropicolinate reductase displayed Nc values less than 40.

To further understand the implications of codon usage variation and to detect and quantify synonymous codon usage pattern, we performed a multivariate statistical analysis (correspondence analysis or CA) as codon usage intrinsically is multivariate in nature. CA is a statistical technique widely employed to detect and quantify synonymous codon usage pattern. In this technique high dimensional data are reduced to a limited number of variables or axes and the most prominent axes contributing to the codon usage variation among the gene sequences is considered **[23]**. Correspondence analysis on relative synonymous codon usage (RSCU) of the genes involved in glycolysis and amino acid metabolism was performed. We also performed correspondence analysis on RSCU of the entire genomes of the three *Arthrobacter* species. Whole genome correspondence analysis of the three *Arthrobacter* species involving 14142 genes revealed a major trend of codon usage variation and in comparison to all the axes generated, the first and second axis accounted respectively for 13.34% and 4.91% of the total variation. They turn out to be highest among all the axes. The position of the genes along the first and second major axes when plotted, a significant positive correlation between the positions of the genes on the first axis with the level of expression was observed. Whereas the highly expressed genes appeared on the right hand side, the lowly expressed genes were found to be present on the left hand side of the plot. It was also observed that *Arthrobacter aurescens* genes formed a large cluster towards the left hand side of the plot compared to the other two species. This would suggest that a relatively larger number of *Arthrobacter aurescens* genes are lowly expressed. This result is consistent with what we have observed previously in the case of glycolysis and amino acid metabolism pathways. These low expression levels of individual genes may be well compensated for by the presence of a large number of gene duplicates or paralogs which might additively contribute to the expression level of a gene product in *Arthrobacter aurescens* TC1.
Correspondence analysis on RSCU of the genes involved in glycolysis of the three *Arthrobacter* species involving 56 genes detected a single major trend of codon usage variation and in comparison to all the axes generated the first and second axis accounted for 23.15% and 10.72% of the total variation respectively, which are highest among all the axes. Plotting the first and second axes data for glycolysis in Fig. 2 we find that the first axis significantly correlates with the expression level of the gene ($r = -0.85$). We observed that the sequences on the left hand side of the plot are highly expressed compared to the sequences on the right hand side of the plot. Consistent with our findings related to Nc, we also observed that *Arthrobacter chlorophenolicus* has the maximum number of genes which

Bioinformation Volume 5 *open access*

www.bioinformation.net Issue 10 **Hypothesis**

are potentially highly expressed. A majority of the gene sequences of *Arthrobacter* sp. FB24 and *Arthrobacter chlorophenolicus* A6 have a tendency to cluster or stay together whether they are highly expressed or not showing similarity in terms of expression level; but in the case of *Arthrobacter aurescens* TC1, it was seen that only those genes that have low Nc or higher expression level clustered along with the genes of the other two species. The remaining genes of *Arthrobacter aurescens* that are relatively lowly expressed and have higher Nc were found in a rather dispersed manner altogether in a separate quadrant **(Figure 2)**. This observation further validates our previous finding that among the three species of *Arthrobacter* included in our study, *Arthrobacter aurescens* TC1 has a codon usage pattern quite distinct from its other two relatives and that a large number of *Arthrobacter aurescens* TC1 orthologs have low codon bias.

Correspondence analysis of the 998 genes involved in amino acid metabolism in the three *Arthrobacter* species shows highest variation along the first axis (17.48%) and the variation along the second axis is 7.27%. Major trend of codon usage variation for the three species in question is observed in **Figure 3**. The first axis corresponds to the expression level of the genes and gene sequences on the right hand side of the plot are potentially highly expressed compared to the sequences on the left hand side (r = -0.9). *Arthrobacter aurescens* TC1 have comparatively lesser number of genes on the right hand side of the plot compared to *Arthrobacter chlorophenolicus* A6 and *Arthrobacter* sp. FB24. Here also we find that the CA on RSCU pattern of *Arthrobacter aurescens* TC1 is rather dissimilar in comparison to its other two relatives.

So it is quite evident that rather unusual codon usage pattern of *Arthrobacter aurescens* TC1 has percolated all the way to very important and primordial metabolic pathways like glycolysis and amino acid metabolism. It is generally agreed upon that metabolic pathways which are universal and core to an organism's survival would normally have tight and conserved gene sequences even in distantly related individuals. We observe that *Arthrobacter aurescens* TC1 which is supposed to be very close to the other two *Arthrobacter* strains reveal significant dissimilarities at the level of glycolysis and amino acid metabolism gene sequences.

To study the evolutionary divergence of the gene sequences involved in amino acid biosynthesis, the synonymous (Ks) and non-synonymous (Ka) nucleotide substitution rates were taken into account. Mutation and selection are known to have varied effects on synonymous and non-synonymous substitution rates, hence estimation of these rates are crucial in deciphering the mechanisms of molecular sequence evolution **[24]**. A substitution is defined as synonymous if it does not change the amino-acid sequence, otherwise it is called non-synonymous. A change in an amino acid due to a non-synonymous nucleotide change is called a replacement. Software tool DnaSP was employed to estimate Ka (the number of non-synonymous substitutions per non-synonymous site), and Ks (the number of synonymous substitutions per synonymous site) for all the gene sequences involved in glycolysis, 13 genes from amino acid metabolic pathways and nearly 20 genes were selected randomly spanning the entire genome of all the three species of *Arthrobacter*. The selected gene sequences were involved in energy metabolism, DNA replication, cell wall biosynthesis and xenobiotic degradation to name a few. The Ks or synonymous substitution rates was found to be directly correlated with codon usage pattern signifying the fact that gene sequences with low codon bias or high Nc score show higher rate of synonymous or silent substitution (r=0.7). The gene sequences representing the genome cluster concur to this fact. *Arthrobacter aurescens* TC1 sequences were found to have significantly elevated synonymous substitution rates pointing to its low codon bias. When we consider the glycolytic pathway, apart from enolase, all the other nucleotide sequences coding for the enzymes of glycolysis in *Arthrobacter aurescens* TC1 have relatively higher rate of synonymous substitution. This too is consistent with what we have observed in the codon usage nature of the glycolytic genes in the three different *Arthrobacter* species. Finally, we studied the previously selected bunch of gene sequences involved in different synthetic and catabolic pathways constituting the amino acid metabolism. *Arthrobacter aurescens* TC1 displayed high Ks values. We found that quite a few gene sequences have synonymous substitution rates well above unity. The Ks versus Nc plots **(Figure 4)** clearly depict the positive correlation between Nc and synonymous nucleotide substitution rate in the three *Arthrobacter* species.

The synonymous substitution rate of *Arthrobacter aurescens* TC1 certainly reveals the organism's low codon bias to cope with the continued selective pressure of nature. In contrast, *A. chlorophenolicus* A6 genes are more subtle with respect to their Ks rates which corroborate the higher expression level of majority of the gene sequences.

The non-synonymous nucleotide substitution rate (Ka) was found to be low for gene sequences with effective codon numbers or Nc below 40. But sequences with Nc scores beyond 40 have high Ka values suggesting amino acid altering changes **(Figure 5)**. With a large number of gene sequences having Nc beyond 40, *Arthrobacter aurescens* TC1 have higher Ks to counter significant changes at the protein level. The presence of a large number of gene duplicates or paralogs in the genome of *Arthrobacter aurescens* TC1 compared to the other strains ensure maintenance of form and functionality in the extreme environments where *Arthrobacter aurescens* TC1 is known to thrive.

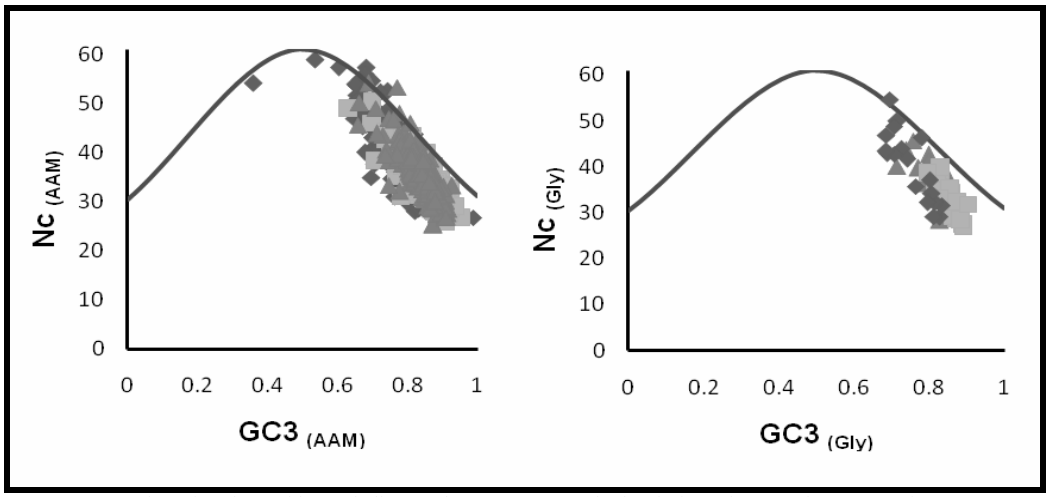

Figure 1: Nc plot of amino acid metabolism pathway (AAM) and glycolysis pathway (Gly) genes of *Arthrobacter aurescens* TC1 (◆), A. chlorophenolicus A6 (■) and *Arthrobacter* sp. FB24 (▲).

Bioinformation Volume 5 *open access*

www.bioinformation.net Issue 10 **Hypothesis**

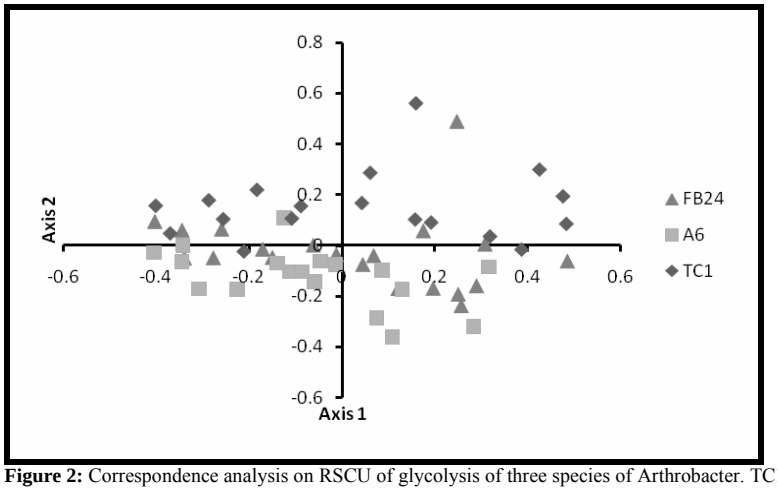

Figure 2: Correspondence analysis on RSCU of glycolysis of three species of Arthrobacter. TC1= *A. aurescens* TC1; A6 = *A. chlorophenolicus* A6; FB24 = *Arthrobacter* sp. FB24.

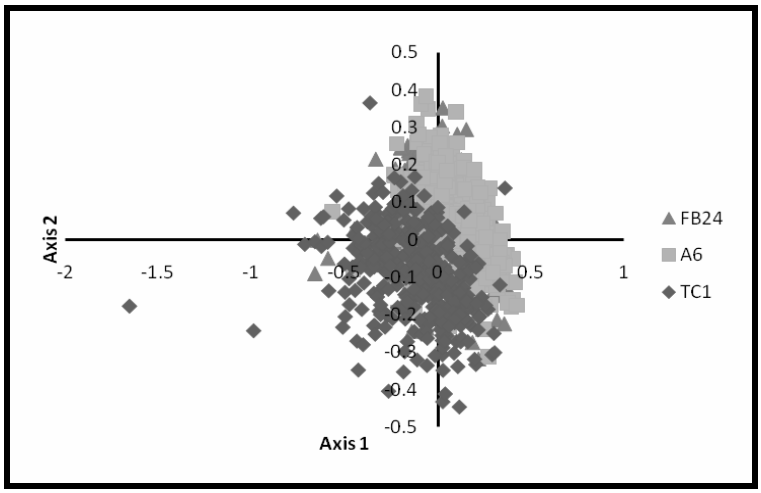

Figure 3: Correspondence analysis on RSCU of amino acid metabolism of three species of *Arthrobacter*. TC1= *A. aurescens* TC1; A6 = *A. chlorophenolicus* A6; FB24 = *Arthrobacter* sp. FB24.

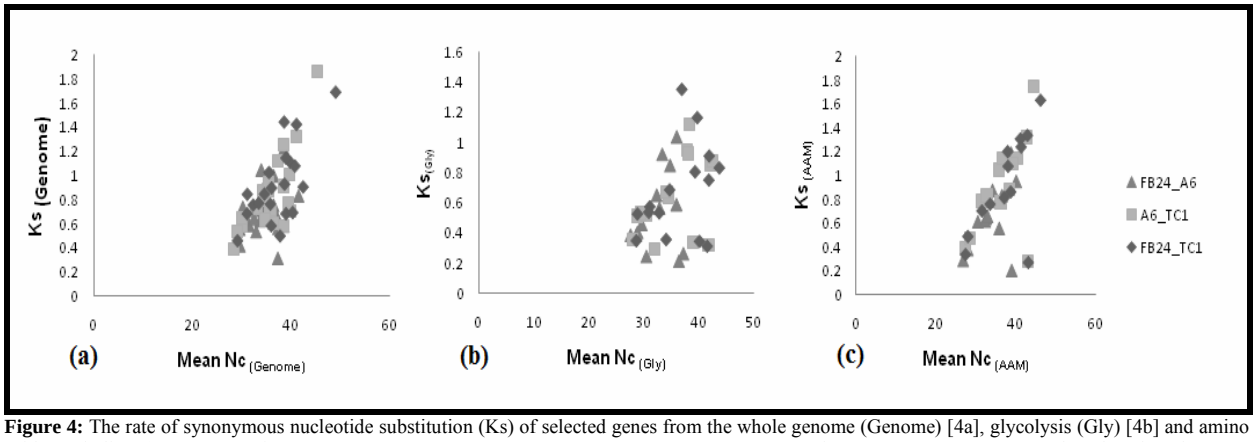

Figure 4: The rate of synonymous nucleotide substitution (Ks) of selected genes from the whole genome (Genome) [4a], glycolysis (Gly) [4b] and amino acid metabolism (AAM) [4c] of *Arthrobacter* sp. FB24 (=FB24), *A. chlorophenolicus* A6 (=A6) and *A. aurescens* TC1 (=TC1) plotted against the mean Nc of the respective pathways.

Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 10 **Hypothesis**

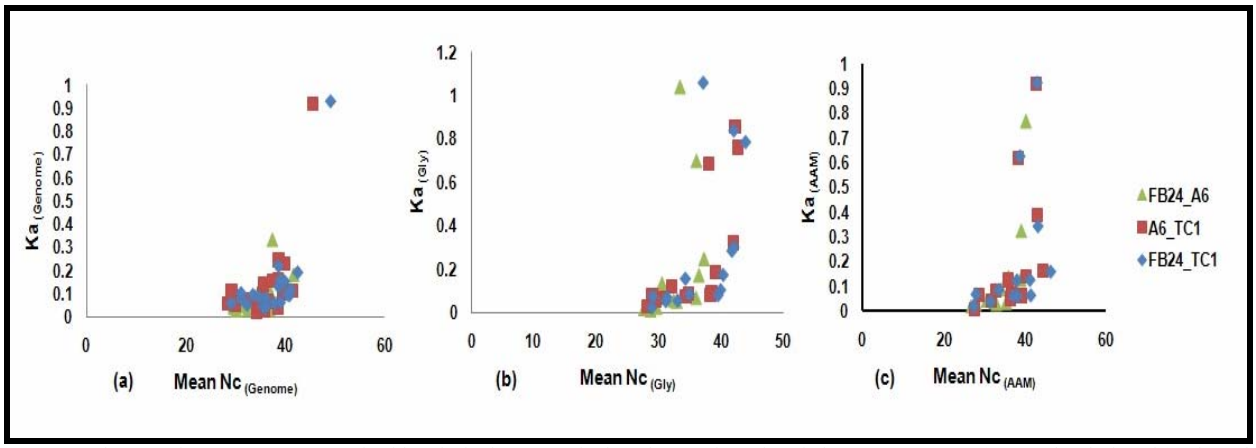

Figure 5: The rate of non-synonymous nucleotide substitution (Ka) of selected genes from the whole genome (Genome) [5a], glycolysis (Gly) [5b] and amino acid metabolism (AAM) [5c] of *Arthrobacter* sp. FB24 (=FB24), *A. chlorophenolicus* A6 (=A6) and *A. aurescens* TC1 (=TC1) plotted against the mean Nc of the respective pathways.

Conclusion:

A significantly different style of codon usage pattern was consistently observed in the case of *Arthrobacter aurescens* TC1. This was confirmed by our comprehensive study of parameters like Nc, codon adaptation index along with the synonymous and non-synonymous nucleotide substitution rate. The BLAST results (data not shown) of gene sequences of certain selected enzymes/proteins spanning the entire genome of the three *Arthrobacter* species in question comprehensively reveals the fact that the similarity between *Arthrobacter chlorophenolicus* A6 and *Arthrobacter* sp. FB24 is high. *Arthrobacter aurescens* TC1 on the other hand have low similarity with the other two species and the genomic heterogeneity has percolated to a certain degree even in conserved energy metabolism pathways like glycolysis and amino acid metabolism. The gene sequences of the enzymes/proteins included in the BLAST analysis were alcohol dehydrogenase GroES domain protein, amidase, aspartate ammonia-lyase, chorismate mutase, diacylglycerol kinase catalytic subunit, gluconate transporter, glutaminase, glycosyl transferase group 1, histidinol-phosphate aminotransferase, ketose-bisphosphate aldolase class-II and phosphoribosylglycinamide formyltransferase to name a few. From this, it may be finally concluded that *Arthrobacter aurescens* TC1 has evolved separately or due to selection pressure it is quite distant from the other two *Arthrobacter* species, where highly conserved pathways like glycolysis and amino acid metabolism ensemble shows distinct codon usage differences. *Arthrobacter chlorophenolicus* A6 and *Arthrobacter* sp. FB24 however share a lot of similarity in terms of their codon usage style and evolutionary divergence of nucleotide sequences.

Acknowledgements:

One of the authors SM would like to acknowledge the Department of Biotechnology, Government of India for its two grants BT/BI/004/93 and BT/BI/019/99.

References:

[1] http://genome.jgi-psf.org/art_f/art_f.home.html
[2] E Mongodin *et al. PLoS Genet.* (2006) **2**: 2094 [PMID: 17194220]
[3] M Kanehisa *et al. Nucleic Acids Res.* (2008) **36**: D480 [PMID: 18077471]
[4] P Karp *et al. Nucleic Acids Res.* (2002) **30**: 59 [PMID: 11752254]
[5] Y Ashida *et al. IPSJ Digit Cour.* (2008)**4**: 228
[6] E Smith & H Morowitz, *Proc Natl Acad Sci. USA* (2004) **101**: 13168 [PMID: 15340153]
[7] Ebenhöh, *Bull Math Biol*. (2001) **63**: 21 [PMID: 11146883]
[8] E Meléndez-Hevia *et al. J Mol Evol* (1996) **43**: 293 [PMID: 8703096]
[9] AH Romano & T Conway, *Res Microbiol*. (1996) **147**: 448 [PMID: 9084754]
[10] CL Strong *et al. Environ Microbiol*. (2002) **68**: 5973 [PMID: 12450818]
[11] M Unell *et al. Biodegradation* (2008) **19**: 495 [PMID: 17917705]
[12] http://www.genome.jp/dbget-bin/www_bget?refseq+NC_008539
[13] V Markowitz *et al. Nucleic Acids Res*. (2008) **36**: D528 [PMID: 17933782]
[14] SF Altschul *et al. J Mol Biol*. (1990) **62**: 403 [PMID: 2231712]
[15] F Wright, *Gene* (1990) **87**: 23 [PMID: 2110097]
[16] P Sharp & W Li, *Nucleic Acids Res.* (1987) **15**: 1281 [PMCID: PMC340524]
[17] X Xia, *Evol Bioinform Online.* (2007) **3**: 53 [PMCID: PMC2684136]
[18] J Thompson *et al. Nucleic Acids Res*. (1994) **22**: 4673 [PMCID: PMC308517]
[19] J Rozas *et al. BMC Bioinformatics* (2003) **19**: 2496
[20] M Nei & T Gojobori *Mol Biol Evol*. (1986) **3**: 418 [PMID: 3444411]
[21] RM Goetz & A Fuglsang, *Biochem Biophys Res Commun*. (2005) **327**: 4
[22] G Wu *et al. Microbiology* (2005) **151**: 2175
[23] S Basak *et al. J Biomol Struct Dyn* (2004) **22**: 205
[24] Z Yang, *Mol Biol Evol*. (1998) **15**: 568 [PMID: 9580986]

Edited by N Srinivasan

Citation: Pal *et al.* Bioinformation 5(10): 446-454 (2011)

Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 10 **Hypothesis**

Supplementary material:

Material 1:

$$\hat{F} = \frac{\left[n_{aa} \sum_{i=1}^{j} p^2\right] - 1}{n_{aa} - 1}$$

Material 2:

$$N_c = 2 + 9/\hat{F}^{av2} + 1/\hat{F}^{3} + 5/\hat{F}^{av4} + 3/\hat{F}^{av6}$$

Table 1: Mean Nc, GC3 and CAI of the three *Arthrobacter* species.

Organism	Whole Genome		Amino acid metabolism			Glycolysis		
	Nc	GC3	Nc	GC3	CAI	Nc	GC3	CAI
Arthrobacter aurescens TC1	43.54	0.73	40.43	0.77	0.721	40.15	0.75	0.721
Arthrobacter chlorophenolicus A6	37.62	0.81	34.33	0.84	0.806	33.11	0.86	0.819
Arthrobacter sp. FB 24	39.16	0.79	35.58	0.83	0.731	36.51	0.81	0.700

Table 2: A comparison of the CAI values of amino acid metabolism pathway genes with respect to the Nc values in the three *Arthrobacter* species.

Organism	CAI values of amino acid metabolism pathway genes	
	Nc value < 40	Nc value > 40
Arthrobacter aurescens TC1	0.76	0.68
Arthrobacter chlorophenolicus A6	0.81	0.73
Arthrobacter sp. FB24	0.74	0.64

Table 3: List of gene sequences involved in glycolysis pathway of *Arthrobacter aurescens* TC1, *A. chlorophenolicus* A6 and *Arthrobacter* sp. FB24. (Gene ids correspond to that in Integrated Microbial Genomes website http://www.img.jgi.doe.gov)

Gene Product	Length	%G+C(3)	Nc	CAI
639799654_glucokinase_Arthrobacter_aurescens	981	68.5	46.9	0.608
639800284_glucokinase_Arthrobacter_aurescens	894	68.8	42.9	0.669
639800499_glucokinase_Arthrobacter_aurescens	1059	77.9	46.3	0.67
643589959_glucokinase_Arthrobacter_chlorophenolicus_A6	1092	82.1	37.4	0.777
639689363_glucokinase_Arthrobacter_sp._FB24	1092	82.7	37	0.688
639800897_glucose-6-phosphate_isomerase_Arthrobacter_aurescens	1632	68.6	43.4	0.663
643590241_glucose-6-phosphate_isomerase_Arthrobacter_chlorophenolicus_A6	1632	82.9	33.3	0.797
639689903_glucose-6-phosphate_isomerase_Arthrobacter_sp._FB24	1638	79.3	36.3	0.695
639689619_glucose-6-phosphate_isomerase_[Arthrobacter_sp._FB24	591	77.2	39.7	0.665
639801760_6-phosphofructokinase_Arthrobacter_aurescens_TC1	1026	76.3	35.7	0.747
643591126_6-phosphofructokinase_Arthrobacter_chlorophenolicus_A6	1026	88.6	28.4	0.871
639690827_6-phosphofructokinase_Arthrobacter_sp._FB24	1026	83.3	32.7	0.743
639688033_6-phosphofructokinase_Arthrobacter_sp._FB24	963	81.9	38.4	0.679
639799393_fructose-bisphosphate_aldolase_Arthrobacter_aurescens_TC1	1020	82.6	29.1	0.819
643589056_fructose-bisphosphate_aldolase_Arthrobacter_chlorophenolicus_A6	1020	87.6	28.6	0.858
639688315_fructose-bisphosphate_aldolase_Arthrobacter_sp._FB24	1020	87.6	28.9	0.809
639800286_fructose-bisphosphate_aldolase_Athrobacter_aurescens_TC1	840	72.5	43.9	0.676
643588903_fructose-bisphosphate_aldolase_Arthrobacter_chlorophenolicus_A6	876	88.7	31.9	0.812
639689534_fructose-bisphosphate_aldolase_Arthrobacter_sp._FB24	840	71.4	40.1	0.618
639800892_triosephosphate_isomerase_Arthrobacter_aurescens	816	70.6	42.8	0.695
643590235_triosephosphate_isomerase_Arthrobacter_chlorophenolicus_A6	816	85.7	35.1	0.819
639689897_triosephosphate_isomerase_Arthrobacter_sp._FB24	816	81.6	37.7	0.73
639799937_triosephosphate_isomerase_Arthrobacter_aurescens	801	71.5	50	0.646

639688696_triosephosphate_isomerase_Arthrobacter_sp._FB24	840	84.3	38.7	0.704
639800890_glyceraldehyde-3-phosphate_dehydrogenase_Arthrobacter_aurescens_TC1	1011	79.5	32.2	0.766
643590233_glyceraldehyde-3-phosphate_dehydrogenase_Arthrobacter_chlorophenolicus_A6	1011	86.6	29.1	0.842
639689895_glyceraldehyde-3-phosphate_dehydrogenase_Arthrobacter_sp._FB24	1011	86.4	30.3	0.763
639801205_glyceraldehyde-3-phosphate_dehydrogenase_Arthrobacter_aurescens_TC1	1488	80.2	37.1	0.772
643590593_glyceraldehyde-3-phosphate_dehydrogenase_Arthrobacter_chlorophenolicus_A6	1479	87.6	32.2	0.842
639690252_glyceraldehyde-3-phosphate_dehydrogenase_Arthrobacter_sp._FB24	1479	87.8	32.6	0.774
639800891_phosphoglycerate_kinase_Arthrobacter_aurescens_TC1	1227	74.1	41.7	0.702
643590234_phosphoglycerate_kinase_Arthrobacter_chlorophenolicus_A6	1227	81.9	34.7	0.801
639689896_phosphoglycerate_kinase_Arthrobacter_sp._FB24	1227	81.7	37.2	0.722
639799722_phosphoglycerate_mutase_Arthrobacter_aurescens_TC1	756	83.3	31.6	0.831
639799782_phosphoglycerate_mutase_Arthrobacter_aurescens_TC1	750	69.6	54.7	0.653
639802947_phosphoglycerate_mutase_Arthrobacter_aurescens_TC1	582	70.6	48.8	0.677
643588696_Phosphoglycerate_mutase_Arthrobacter_chlorophenolicus_A6	594	85.9	34.5	0.794
643589248_phosphoglycerate_mutase_Arthrobacter_chlorophenolicus_A6	747	89.2	27.4	0.894
643591446_Phosphoglycerate_mutase_Arthrobacter_chlorophenolicus_A6	777	90.3	31.8	0.819
643592237_Phosphoglycerate_mutase_Arthrobacter_chlorophenolicus_A6	675	83.1	39.8	0.757
643592332_Phosphoglycerate_mutase_Arthrobacter_chlorophenolicus_A6	585	79.5	38.5	0.77
639687915_Phosphoglycerate_mutase_Arthrobacter_sp._FB24	789	79.8	42.6	0.661
639688523_phosphoglycerate_mutase_Arthrobacter_sp._FB24	747	88	30.5	0.804
639691865_Phosphoglycerate_mutase_Arthrobacter_sp._FB24	672	79.9	41.2	0.665
639691987_Phosphoglycerate_mutase_Arthrobacter_sp._FB24	585	79	37	0.702
639800088_enolase_Arthrobacter_aurescens_TC1	1281	81.3	29.2	0.804
643589625_enolase_Arthrobacter_chlorophenolicus_A6	1281	89	27	0.885
639688953_enolase_Arthrobacter_sp._FB24	1281	82.9	28.4	0.769
639800652_pyruvate_kinase_Arthrobacter_aurescens_TC1	1494	80.7	34.3	0.797
643590103_pyruvate_kinase_Arthrobacter_chlorophenolicus_A6	1491	83.7	34.2	0.818
639689514_pyruvate_kinase_Arthrobacter_sp._FB24	1491	85.5	31.6	0.786

Table 4: Sequence analysis data of genes involved in amino acid metabolism of *Arthrobacter aurescens* TC1, *A. chlorophenolicus* A6 and *Arthrobacter* sp. FB24. (Gene ids correspond to that in Integrated Microbial Genomes website http://www.img.jgi.doe.gov)

Gene Product	Length	Nc	CAI
639800449_argininosuccinate_synthase_Arthrobacter_aurescens	1206	28	0.814
643589909_argininosuccinate_synthase_Arthrobacter_chlorophenolicus_A6	1206	27.1	0.87
639689313_argininosuccinate_synthase_Arthrobacter_sp._FB24	1206	26.7	0.85
639801849_4-aminobutyrate_aminotransferase_apoenzyme_Arthrobacter_aurescens_TC1	1371	28.5	0.824
643591212_4-aminobutyrate_aminotransferase_apoenz_Arthrobacter_chlorophenolicus_A6	1371	28.7	0.872
639690921_4-aminobutyrate_aminotransferase_apoenzyme_Arthrobacter_sp._FB24	1407	27.4	0.815
639801210_cysteine_synthase_Arthrobacter_aurescens_TC1	936	32.6	0.771
643590599_cysteine_synthase_Arthrobacter_chlorophenolicus_A6	936	30.8	0.845
639690260_cysteine_synthase_Arthrobacter_sp._FB24	936	30.4	0.805
639801818_cystathionine_gamma-lyase_Arthrobacter_aurescens_TC1	1164	33.6	0.78
643591174_cystathionine_gamma-lyase_Arthrobacter_chlorophenolicus_A6	1164	31.9	0.827
639690887_cystathionine_gamma-lyase_Arthrobacter_sp._FB24	1170	33.5	0.782
639800394_dihydrodipicolinate_reductase_Arthrobacter_aurescens_TC1	759	39.5	0.729
643589857_dihydrodipicolinate_reductase_Arthrobacter_chlorophenolicus_A6	771	32.1	0.828
639689259_dihydrodipicolinate_reductase_Arthrobacter_sp._FB24	759	36.5	0.708
639800450_argininosuccinate_lyase_Arthrobacter_aurescens_TC1	1476	41	0.733
643589910_argininosuccinate_lyase_Arthrobacter_chlorophenolicus_A6	1443	31.6	0.81
639689314_argininosuccinate_lyase_Arthrobacter_sp._FB24	1440	33.3	0.757
639802282_glycerate_kinase_Arthrobacter_aurescens	1167	41.1	0.727
643588636_Glycerate_kinase_Arthrobacter_chlorophenolicus_A6	1125	35.6	0.778
639687918_Glycerate_kinase_Arthrobacter_sp._FB24	1125	36.4	0.685
639801025_2-isopropylmalate_synthase_Arthrobacter_aurescens_TC1	1740	41.8	0.708
643590367_2-isopropylmalate_synthase_Arthrobacter_chlorophenolicus_A6	1740	31.7	0.84

639690035_2-isopropylmalate_synthase_Arthrobacter_sp._FB24	1740	34.4	0.755
639801401_homoserine_kinase_Arthrobacter_aurescens_TC1	975	44.8	0.656
643590760_homoserine_kinase_Arthrobacter_chlorophenolicus_A6	975	35.8	0.749
639690440_homoserine_kinase_Arthrobacter_sp._FB24	918	37.6	0.691
639801076_chorismate_synthase_Arthrobacter_aurescens_TC1	1263	45.2	0.648
643590415_chorismate_synthase_Arthrobacter_chlorophenolicus_A6	1200	33.3	0.798
639690083_chorismate_synthase_Arthrobacter_sp._FB24	1200	37.6	0.673
639803394_amidase_Arthrobacter_aurescens	1542	45.4	0.65
643589528_Amidase_Arthrobacter_chlorophenolicus_A6	1419	40	0.747
639687930_Amidase_Arthrobacter_sp._FB24	1413	40.6	0.67
639800926_phosphoserine_phosphatase_SerB_Arthrobacter_aurescens_TC1	894	47.1	0.683
643590278_phosphoserine_phosphatase_SerB_Arthrobacter_chlorophenolicus_A6	894	39	0.773
639689931_phosphoserine_phosphatase_Arthrobacter_sp._FB24	894	39.3	0.704
639800631_methionyl-tRNA_formyltransferase_Arthrobacter_aurescens_TC1	921	51.7	0.615
643590080_methionyl-tRNA_formyltransferase_Arthrobacter_chlorophenolicus_A6	921	37.1	0.758
639689492_methionyl-tRNA_formyltransferase_Arthrobacter_sp._FB24	921	40.9	0.656