# Gene Ontology and KEGG Orthology Mappings for 10 Strains of *Pseudomonas stutzeri*

**Jin Xing Lim[1] and Maurice HT Ling[2,3]***

[1]*School of Applied Sciences, Temasek Polytechnic, Republic of Singapore*

[2]*Colossus Technologies LLP, Republic of Singapore*

[3]*HOHY PTE LTD, Republic of Singapore*

**\*Corresponding Author**: Maurice HT Ling, Colossus Technologies LLP, Republic of Singapore.

## Abstract

Gene Ontology (GO) and KEGG Orthology (KO) are controlled vocabularies for annotating gene and protein functions, and mapping functions onto pathways; which enables metagenomic analysis. *Pseudomonas stutzeri* is an environmental bacterium with potential for biotechnology applications in the environment, despite being an opportunistic pathogen. However, there has been no GO nor KO annotations for *P. stutzeri*. This study presents the first GO and KO mapping for 10 strains of *P. stutzeri* for further studies into P. stutzeri. Of the 42764 peptides in 10 strains of *P. stutzeri*, 30435 (71.17%) peptides were annotated with one or more GO terms and 25034 (58.54%) of peptides were annotated with KO terms. The annotation files and sequences can be downloaded at https://tinyurl.com/GO-KO-Pstutzeri.

*Keywords: Gene Ontology; KEGG Orthology Mappings; Pseudomonas stutzeri*

## Introduction

Gene Ontology (GO) is a common set of defined and controlled vocabulary describing the roles of genes and gene products in any organism [1,2]. This makes GO a useful tool for comparative genomics, which requires analysis of gene and protein functions across different species [3-6]. GO had been used in many studies [7-9]. For example, GO had been used to analyse gene lists for biological functions [10-13], identifying candidate proteins for drug development [14], identifying homologous proteins [15] and functional similarities [16], identifying potential protein-protein interactions [17,18], and examining evolution across different phyla [19].

Although these studies demonstrated the usefulness of GO, a major drawback of GO is its lack of direct association with pathways [20]. KEGG Orthology (KO) aims to supplement this deficiency of GO [20], which allows for mapping functional analysis results onto pathway diagrams [21,22]. Similar to GO, KO had been used in a number of studies [22,23]. At the same time, both GO and KO had been used together in the same studies [24,25]; thus, implying the usefulness of KO and GO. However, the presence of GO and/or KO annotations for a specific organism cannot be assumed.

*Pseudomonas stutzeri* is an environmental bacterium with many metabolic feature; such as, denitrification and natural transforming abilities [26]; and had been isolated as a multi-drug resistant human opportunistic pathogen [27]. *P. stutzeri* had been suggested to be a host for expressing membrane proteins [28] and being an environmental isolate, it has the potential for environmentally relevant applications [29]. However, there has been no Gene Ontology nor KEGG Orthology annotations for *P. stutzeri*. This study presents the first Gene Ontology and KEGG Orthology mapping for 10 strains of *P. stutzeri*.

## Materials and Methods

**Strain Sequences:** The 10 strains of *P. stutzeri* were (a) *P. stutzeri* 19SMN4 (Accession number: CP007509.1), (b) *P. stutzeri* 28a24 (Accession number: CP007441.1), (c) *P. stutzeri* 273 (Accession number: CP015641.1), (d) *P. stutzeri* A1501 (Accession number: CP000304.1), (e) *P. stutzeri* CCUG 29243 (Accession number: CP003677.1), (f) *P. stutzeri* CGMCC 1.1803 (Accession number: CP002881.1), (g) *P. stutzer-i*DSM 4166 (Accession number: CP002622.1), (h) *P. stutzeri* DSM 10701 (Accession number: CP003725.1), (i) *P. stutzeri* RCH2 (Accession number: CP003071.1), and (j) *P. stutzeri* SLG510A3-8 (Accession number: CP011854.1).

**Mapping to Gene Ontology:** Each of the 42764 peptides was scanned using HAMAP Scan [30] done on 27-Feb-2018[1] , and peptide blast against *Escherichia coli* K-12[2] and *Pseudomonas aeruginosa* PAO1[3] peptides using a purpose-generated BLAST database (Figure 1). HAMAP Scan [30] generated *P. stutzeri* to HAMAP IDs, which were then mapped to the corresponding GO IDs using HAMAP to GO map[4]. The peptide blasts generated *P. stutzeri* to *E. coli* K-12 blast result and *P. stutzeri* to *P. aeruginosa* PAO1 blast result respectively, which were then mapped to the corresponding GO IDs using existing *E. coli* to GO annotations[5] and *P. aeruginosa* to GO annotations[6] respectively for E-values below 1e-9. The result is 3 independent GO annotation files for *P. stutzeri*. These annotations files were then consolidated to yield a GO annotation where each *P. stutzeri* peptide was annotated by 3 methods.
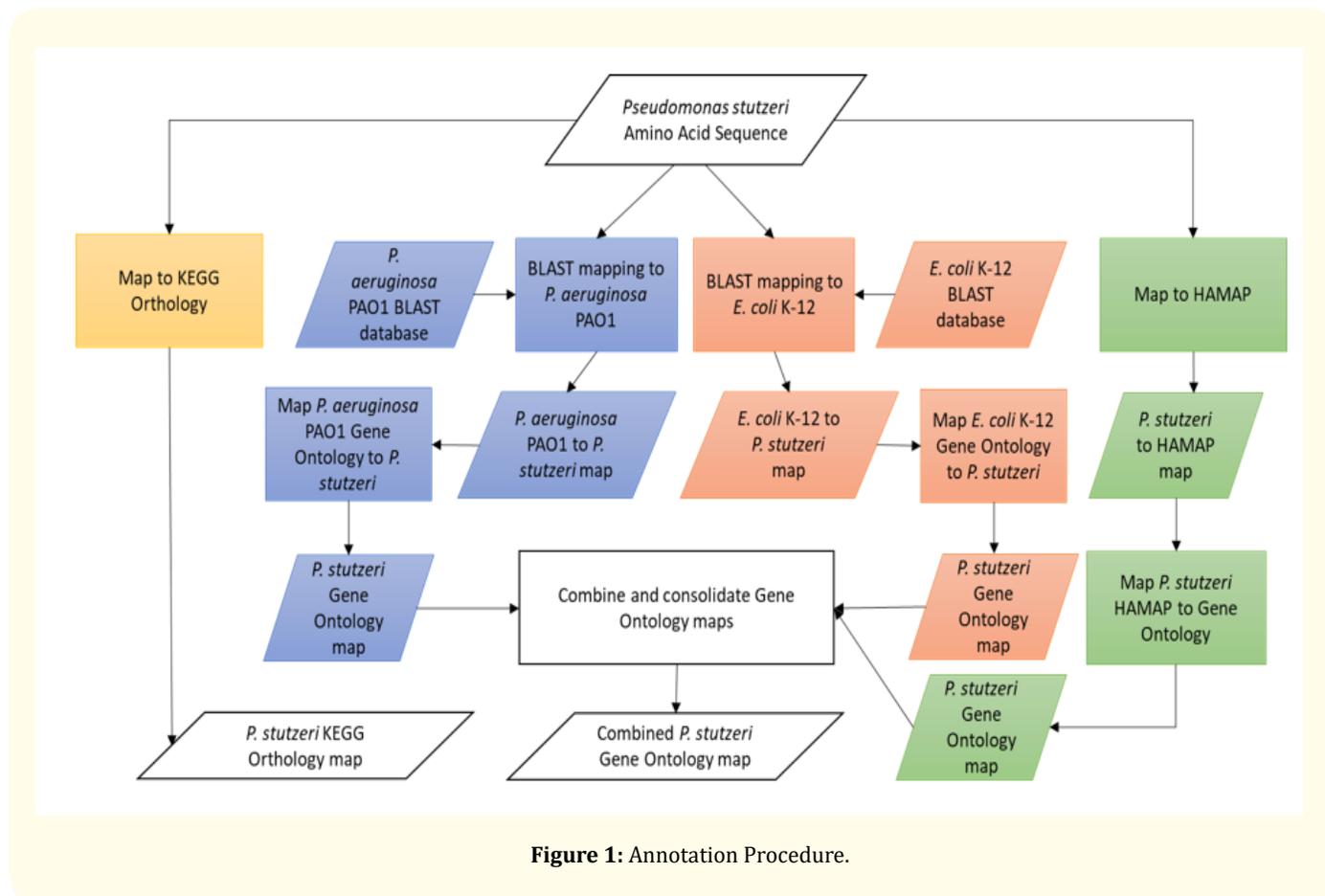


**Figure 1:** Annotation Procedure.

[1]https://hamap.expasy.org/hamap_scan.html

[2]http://www.uniprot.org/uniprot/?query=proteome: UP000000625

[3]https://www.ncbi.nlm.nih.gov/nuccore/AE004091.2

[4]http://geneontology.org/external2go/hamap2go

[5]http://geneontology.org/gene-associations/gene_association.ecocyc.gz

[6]http://geneontology.org/gene-associations/gene_association.pseudocap.gz

**Mapping to KEGG Orthology:** Mapping of *P. stutzeri* peptides to KEGG Orthology (Figure 1) was performed using BlastKOALA [31].

## Results and Discussion

Gene Ontology (GO) and KEGG Orthology (KO) enable metagenomic analysis of gene and protein functions and had been used in a number of studies [22,23]. However, there has been no GO nor KO annotations for *P. stutzeri*. This study presents the first GO and KO mapping for 10 strains of *P. stutzeri*.

**Different number of CDS in *P. stutzeri* strains:** The number of coding sequences (CDS) between *P. stutzeri* strains varies substantially (Figure 2), ranging from 3851 CDS in DSM 10701 to 4595 CDS in CGMCC 1.1803; a difference of 744 CDS or between 19.32% to 16.19%. The average number of CDS is 4276 with standard deviation of 209. A possible reason might be horizontal gene transfer (HGT), which *P. stutzeri* is known for [26] and a complete repertoire of genes necessary for HGT had been found in several *P. stutzeri* strains [32]. There are several recent studies on HGT in *P. stutzeri*. For example, Saha., *et al.* [33] found evidence of HGT of dinitrogen reductase gene. Venieraki., *et al.* [34] analysed nitrogen-fixation islands across several pseudomonas strains including *P. stutzeri* and found highly similar sequences from geologically distant *pseudomonas* strains, suggesting HGT events early in evolution of *Pseudomonas*. Dougherty., *et al.* [35] demonstrated an HGT event of a 1Mb megaplasmid from *Pseudomonas syringae* MAFF301305 to *P. stutzeri* 23a24.
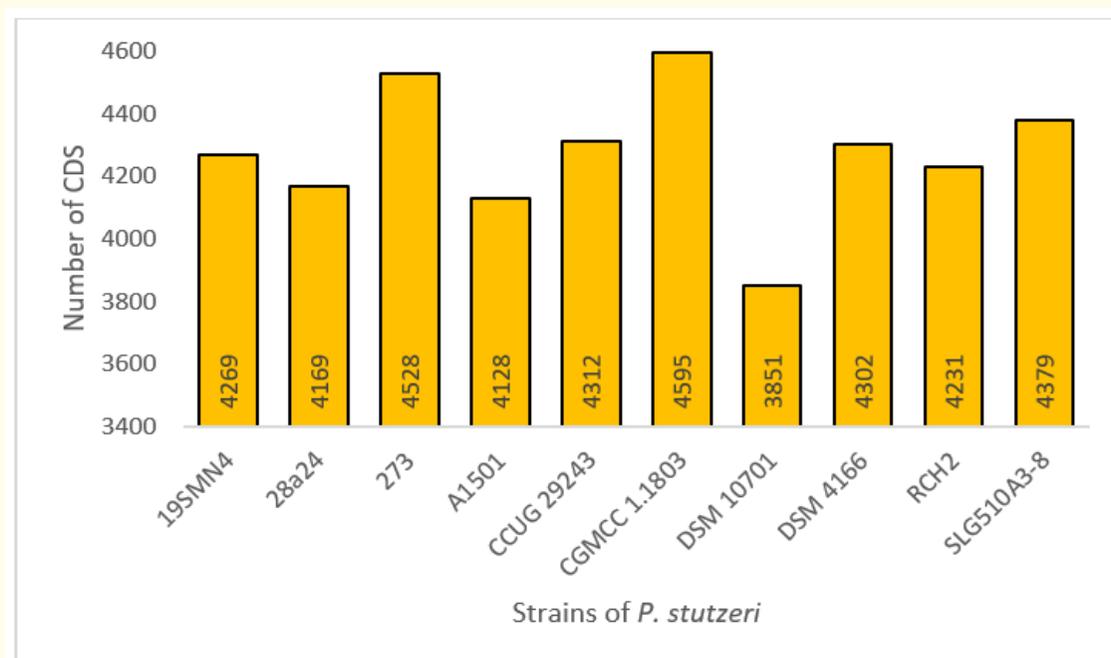


**Figure 2:** Number of Coding Sequences (CDS) in Each Strain.

***P. stutzeri* Proteome Map to 3867 Unique Gene Ontology Terms:** 42764 *P. stutzeri* peptides were annotated with GO terms using 3 methods; BLAST and mapping via *E. coli* K-12 annotations and *P. aeruginosa* PAO1 annotations, and scanned with HAMAP scan [30] and mapped via HAMAP-GO annotations. Mapping via HAMAP yields 1239 unique GO terms. Mapping via *E. coli* K-12 annotations and *P. aeruginosa* PAO1 annotations yield 2900 and 2021 unique GO terms respectively. In total, 42764 *P. stutzeri* peptides were mapped to 3867 unique GO terms.

A total of 30435 (71.17%) *P. stutzeri* peptides were annotated with one or more GO terms, generating 589229 GO term mappings (Figure 3). Of which, 7231 (16.90%) out of 42764 peptides were found in all 3 mappings, yielding 14694 GO term mappings. 11811 GO term mappings were common in HAMAP and *E. coli* K-12 only, which accounts for 6003 (14.03%) out of 42764 peptides. 8166 GO term mappings were common in HAMAP and *P. aeruginosa* PAO1 only, which accounts for 4578 (10.71%) out of 42764 peptides. 62712 GO mappings were common in *E. coli* K-12 and *P. aeruginosa* PAO1 only, which accounts for 12228 (28.59%) out of 42764 peptides.
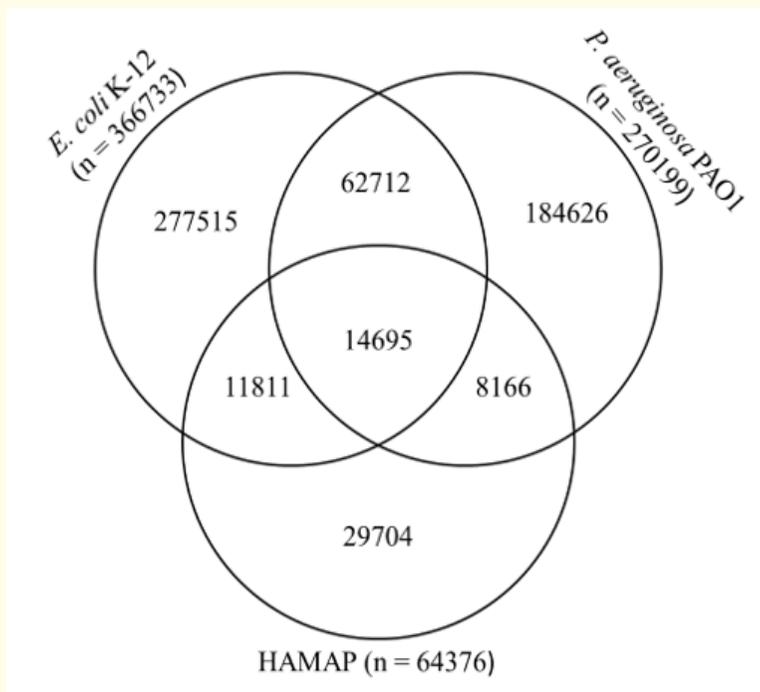
**Figure 3:** Number of GO Terms Mapped.

In terms of KEGG Orthology, only 25034 (58.54%) of the 42764 *P. stutzeri* peptides were annotated with KO terms by BlastKOALA [31], yielding 2419 unique KO terms. In total, 22706 (53.09%) of the 42764 *P. stutzeri* peptides were mapped to both a KO term and a GO term via at least one of the three mappings.

## Conclusion

In summary, both GO and KO had been used in significant number of studies [36-38]. However, there has been no GO annotations nor KO annotations for *P. stutzeri*, despite its potential use in environmental biotechnology. Hence, this study presents both GO and KO annotations for *P. stutzeri* as a new resource supporting further *P. stutzeri* studies.

### Annotation and Sequence File

The annotation files and sequences can be downloaded at https://tinyurl.com/GO-KO-Pstutzeri.

## Conflict of Interest

The authors declare no conflict of interest.

## Bibliography

1.  Ashburner M., *et al*. "Gene ontology: tool for the unification of biology". *Nature Genetics* 25.1 (2000): 25-29.

2.  Gene Ontology Consortium. "Creating the gene ontology resource: design and implementation". *Genome Research* 11.8 (2001): 1425-1433.

3.  Schlicker A., *et al*. "A new measure for functional similarity of gene products based on Gene Ontology". *BMC Bioinformatics* 7 (2006): 302.

4.  Cai Z., *et al*. "Genome comparison using Gene Ontology (GO) with statistical testing". *BMC Bioinformatics* 7 (2006): 374.

5.  Zwaenepoel A., *et al*. "MorphDB: Prioritizing Genes for Specialized Metabolism Pathways and Gene Ontology Categories in Plants". *Frontiers in Plant Science* 9 (2018): 352.

6.  Chibucos MC., *et al*. "Describing commonalities in microbial effector delivery using the Gene Ontology". *Trends in Microbiology* 17.7 (2009): 312-319.

7.  McCarthy FM., *et al*. "Understanding animal viruses using the Gene Ontology". *Trends in Microbiology* 17.7 (2009): 328-335.

8.  Primmer CR., *et al*. "Annotated genes and nonannotated genomes: cross-species use of Gene Ontology in ecology and evolution research". *Molecular Ecology* 22.12 (2013): 3216-3241.

9.  Torto-Alalibo T., *et al*. "Genetic resources for advanced biofuel production described with the Gene Ontology". *Frontiers in Microbiology* 5 (2014): 528.

10. Heng SSJ., *et al*. "Glucan Biosynthesis Protein G (mdoG) is a suitable reference gene in Escherichia coli K-12". *ISRN Microbiology* (2011): 469053.

11. Keng BM., *et al*. "Transcriptome analysis of Spermophilus lateralis and Spermophilus tridecemlineatus liver does not suggest the presence of Spermophilus-liver-specific reference genes". *ISRN Bioinformatics* (2013): 361321.

12. Too IHK and Ling MHT. "Signal Peptidase Complex Subunit 1 and Hydroxyacyl-CoA Dehydrogenase Beta Subunit are suitable reference genes in human lungs". *ISRN Bioinformatics* (2012): 790452.

13. Fruzangohar M., *et al*. "A novel hypothesis-unbiased method for Gene Ontology enrichment based on transcriptome data". *PLoS ONE* 12.2 (2017): e0170486.

14. Passi A., *et al*. "RepTB: a gene ontology based drug repurposing approach for tuberculosis". *Journal of Cheminformatics* 10.1 (2018): 24.

15. Ikram N., *et al*. "Investigating Correlation between Protein Sequence Similarity and Semantic Similarity Using Gene Ontology Annotations". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 15.3 (2018): 905-912.

16. Peng J., *et al*. "Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach". *BMC Systems Biology* 12.2 (2018): 18.

17. Yim S., *et al*. "Annotating activation/inhibition relationships to protein-protein interactions using gene ontology relations". *BMC Systems Biology* 12.1 (2018): 9.

18. Bandyopadhyay S and Mallick K. "A New Feature Vector Based on Gene Ontology Terms for Protein-Protein Interaction Prediction". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 14.4 (2017): 762-770.

19. Koç I and Caetano-Anollés G. "The natural history of molecular functions inferred from an extensive phylogenomic analysis of gene ontology data". *PLoS ONE* 12.5 (2017): e0176129.

20. Mao X., *et al*. "Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary". *Bioinformatics* 121.19 (2005): 3787-3793.

21. Orakov AN., *et al*. "ASAR: visual analysis of metagenomes in R". *Bioinformatics* 34.8 (2018): 1404-1405.

22. Han B., *et al*. "Nutritional and reproductive signaling revealed by comparative gene expression analysis in Chrysopa pallens (Rambur) at different nutritional statuses". *PLoS ONE* 12.7 (2017): e0180373.

23. Berlanga M., *et al*. "Functional Stability and Community Dynamics during Spring and Autumn Seasons Over 3 Years in Camargue Microbial Mats". *Frontiers in Microbiology* 8 (2017): 2619.

24. Yang H., *et al*. "Transcriptome analysis in different developmental stages of Batocera horsfieldi (Coleoptera: Cerambycidae) and comparison of candidate olfactory genes". *PLoS ONE* 13.2 (2018): e0192730.

25. Huang CG., *et al*. "Identification of genes associated with castration-resistant prostate cancer by gene expression profile analysis". *Molecular Medicine Reports* 16.5 (2017): 6803-6813.

26. Lalucat J., *et al*. "Biology of Pseudomonas stutzeri". *Microbiology and Molecular Biology Reviews* 70.2 (2006): 510-547.

27. Bashar S., *et al*. "Emergence of IntI1 associated blaVIM-2 gene cassette-mediated carbapenem resistance in opportunistic pathogen Pseudomonas stutzeri". *Emerging Microbes and Infections* 6.5 (2017): e29.

28. Sommer M., *et al*. "Pseudomonas stutzeri as an alternative host for membrane proteins". *Microbial Cell Factories* 16 (2017): 157.

29. Nicolay T., *et al*. "Probing the applicability of autotransporter based surface display with the EstA autotransporter of Pseudomonas stutzeri A15". *Microbial Cell Factories* 11 (2012): 158.

30. Pedruzzi I., *et al*. "HAMAP in 2015: updates to the protein family classification and annotation system". *Nucleic Acids Research* 43.D1 (2015): D1064-D1070.

31. Kanehisa M., *et al*. "BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences". *Journal of Molecular Biology* 428.4 (2016): 726-731.

32. Brunet-Galmés I., *et al*. "Complete genome sequence of the naphthalene-degrading bacterium Pseudomonas stutzeri AN10 (CCUG 29243)". *Journal of Bacteriology* 194.23 (2012): 6642-6643.

33. Saha C., *et al*. "A consortium of non-rhizobial endophytic microbes from Typha angustifolia functions as probiotic in rice and improves nitrogen metabolism". *Plant Biology* 18.6 (2016): 938-946.

34. Venieraki A., *et al*. "The nitrogen-fixation island insertion site is conserved in diazotrophic Pseudomonas stutzeri and Pseudomonas sp. isolated from distal and close geographical regions". *PLoS ONE* 9.9 (2014): e105837.

35. Dougherty K., *et al*. "Multiple phenotypic changes associated with large-scale horizontal gene transfer". *PLoS ONE* 9.7 (2014): e102170.

36. Feuermann M., *et al*. "Large-scale inference of gene function through phylogenetic annotation of Gene Ontology terms: case study of the apoptosis and autophagy cellular processes". *Database (Oxford)* (2016).

37. Tian Z., *et al*. "An improved method for functional similarity analysis of genes based on Gene Ontology". *BMC Systems Biology* 10.4 (2016): 119.

38. Kotowski N., *et al*. "Improved orthologous databases to ease protozoan targets inference". *Parasites and Vectors* 8 (2015): 494.

**Volume 3 Issue 1 November 2019**