

In silico Functional Proteomic Re-annotation of Escherichia coli K12

Aravindhan Ganesan^{1*}
aravind_garveend@yahoo.co.in

Ramesh Kumar Gopal¹
gramesh@au-kbc.org

Sathish Kumar Radhakrishnan²
rsathish@au-kbc.org

¹ Bioinformatics Division, Life Sciences Department, AU-KBC Research Centre, MIT Campus, Anna University, Chrompet, Chennai – 600044. India

² NRCFOSS, AU-KBC Research Centre, MIT Campus, Anna University, Chrompet, Chennai-600044. India

Abstract

Background:

Escherichia coli is one of the most favorite model organisms of the scientific community. Since the complete genome of *E.coli K12* was initially annotated in 1997, additional information based on biological characterization and functions of sequence similar proteins have become available. Although, five decades of intense research is being carried out on *E.coli* genome, still complete and accurate functional information of this organism is not available. Thus, in silico functional re-annotation of the *E.coli K12* becomes highly vital that will help us to update our knowledge and develop a deeper insight on this important model organism ^[4]. Further, past decade had seen the accumulation of multiple complete genome sequences and related protein databases which provide useful comparisons with the close relatives among other organisms and facilitated powerful re-annotation ^[1, 2, 3]. Hence, on the basis of this newly available database and software resources, we have attempted to carry out the functional proteomic re-annotation of the complete sequences of *E.coli K12*.

Methods:

The *E.coli K12* genome, downloaded from Ecocyc database, represented a total number of 4,290 proteins of which 40% of the sequences were found to be without clear functions. Hence, we have manually carried out the functional proteomic annotation of this organism using the advanced re-annotation strategies, in which we have incorporated several sequence analysis methods into a coherent and an efficient prediction schema. Each of the *E.coli K12* protein sequences previously predicted and annotated has been manually re-analyzed based on the diverse approaches such as similarity search approach (BLASTP), Family based search (PFAM) ^[6], Orthologous Group search (COG) ^[7], Pattern based search (SCANPROSITE) ^[8] and Domain based search (PRODOM). But a regular BLAST search will allow only a single search at a time and will also consume a lot of bandwidth and time for analyzing the entire genome. To overcome this difficulty, we have developed a simple program, AIM-BLAST ^[5] as an AJAX interface to the SOAP services of EBI (European Bioinformatics Institute) to support multiple sequence searches at a stretch during re-annotation.

Results:

This functional re-annotation helped to produce a high quality, consistently named proteome of *E.coli K12*. As a result of this work, 89% of the *E.coli* sequences have been manually assigned with functions (Figure 1b). We have updated all the protein coding genes previously identified in 1997(Figure 1a) and tried to assign new or more precise functions when possible. 29% of the protein sequences of *E.coli* which have been previously uncharacterized with hypothetical,

predicted, putative or unknown annotation have now been assigned with know functions(Figure 1c). Further, this analysis also resulted in the revision of the protein sequences that have been found to be false positive or poorly annotated. Hence this re-annotation methodology has been found to be more efficient and has helped us to achieve sensitivity, specificity and accurate biological information content to the *E.coli K12*. The information from this research work was incorporated into a database named “**REC-DB: a Re-annotated *E.coli* Database**” that will remain a useful repository with accurate and updated functional information of *E.coli k12* genome. Also, this database offers a BLAST facility for the users to compare any protein sequences from any organisms against REC-DB so as to ensure significant new information to the research society.

Conclusion:

In summary, the functional re-annotation of the *E.coli K12* protein sequences has led to the substantial updates across the functions of the entire genome. The data presented here should be constructive for the any future analysis of *E.coli* gene products as well as gene products encoded by other genomes.

Availability:

REC-DB is publicly available at <http://122.165.25.137/bioinfo/recdb/>

Keywords: re-annotation, *E.coli*, genome, proteome.

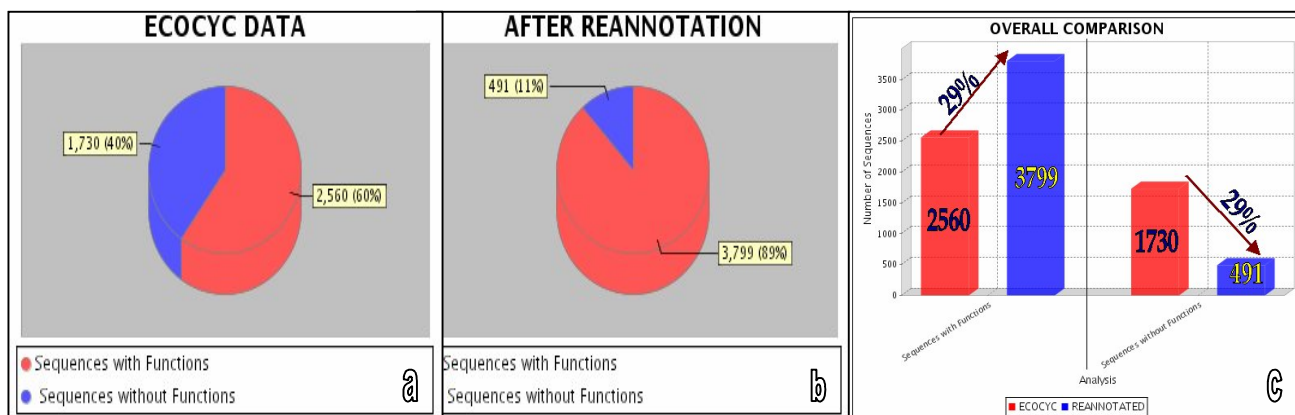


Figure 1: Overall Efficiency of the Functional Re-annotation of *E.Coli K12*. a) Before annotation, *E.coli* genome consisted of 60% of sequences with functions and 40% of sequence with unknown functions. B) After annotation, 89% of sequences were assigned with clear functions and only 11% of sequences were unknown. C) The overall comparison shows that the re-annotation was 29% efficient.

Bibliography:

- [1] Jean-Christophe Camus, Melinda J. Pryor and Claudine Médiq, Re-annotation of the genome sequence of *Mycobacterium tuberculosis H37Rv*, *Microbiology*, 148, 2967-2973, 2002.
- [2] Stéphanie Bocs, Antoine Danchin and Claudine Médiq, Re-annotation of genome microbial CoDing-Sequences: finding genes and inaccurately annotated genes, *BMC Bioinformatics*, 3:5, 2002.
- [3] Steven L Salzberg, Genome re-annotation: a wiki solution? , *Genome Biology*, 8:102, 2007.
- [4] Yu Zheng, Richard J Roberts and Simon Kasif, Genomic functional annotation using co-evolution profiles of gene clusters, *genome biology*, 3(11), 2002.
- [5] <http://biotool.nrcfosshelpline.in/aimblast/>
- [6] <http://pfam.jouy.inra.fr/hmmsearch.shtml>
- [7] <http://expasy.org/tools/scanprosite/>
- [8] <http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html>