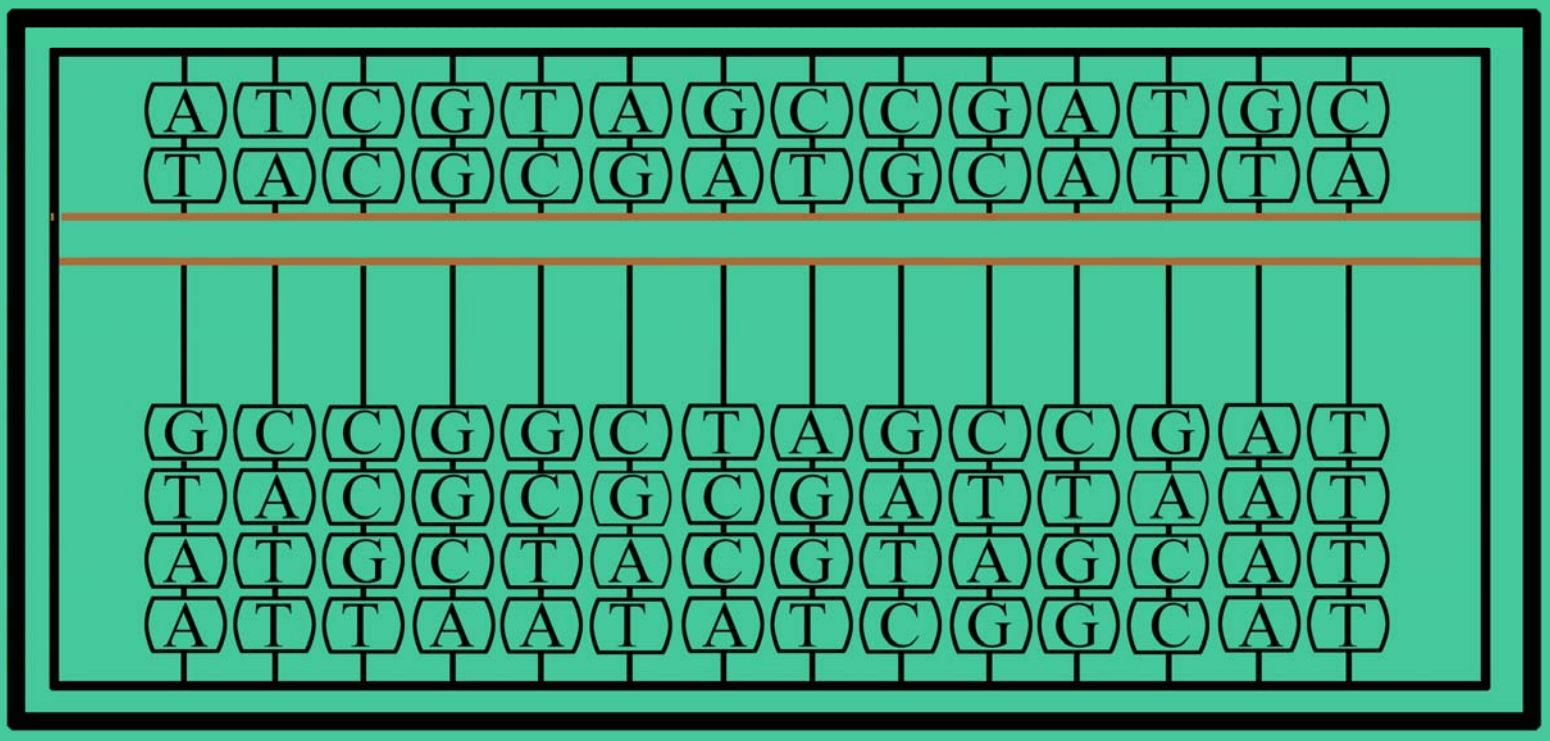
ISSN 1927-5587

BioPublisher©

Computational Molecular Biology

An integration of experimental molecular and genome biology with computational technology



ISSN 1927-5587
14004
9 771927 558004 >

2014
Vol.4

http://cmb.biopublisher.ca

Computational Molecular Biology (online), 2014, Vol. 4
ISSN 1927-5587
http://cmb.biopublisher.ca

Publisher

Sophia Publishing Group

Editor-in-Chief

Dr. Henry Smith

Edited by

Editorial Team of Computational Molecular Biology
Email: edit@cmb.biopublisher.ca
Website: http://cmb.biopublisher.ca

Address:
11388 Stevenston Hwy,
PO Box 96016,
Richmond, V7A 5J5, British Columbia
Canada

Computational Molecular Biology (Online, ISSN 1927-5587)), indexed by international database ProQuest, is an open access, peer reviewed journal publishing original research papers of general interest involving in the computational biology at the molecular level.

The Journal is publishing all the latest and outstanding research articles, letters, methods, and reviews in all areas of Computational Molecular Biology, covering new discoveries in molecular biology, from genes to genomes, using statistical, mathematical, and computational methods as well as new development of computational methods and databases in molecular and genome biology.

The papers published in the journal are expected to be of interests to computational scientists, biologists, and teachers/students/researchers engaged in biology, as well as are appropriate for R & D personnel and general readers interested in computational technology and biology.

Computational Molecular Biology is published independently by BioPublisher. It is committed to publishing and disseminating significant original achievements in the related research fields of molecular biology.

Computational Molecular Biology (online), 2014, Vol. 4
ISSN 1927-5587
http://cmb.biopublisher.ca

Latest Content

PlantSecKB: the Plant Secretome and Subcellular Proteome KnowledgeBase

Gengkon Lum, John Meinken, Jessica Orr, Stephanie Frazier, Xiang Min

Computational Molecular Biology, 2014, Vol.4, No.1

GC2 Biology Dictates Gene Expressivity in *Camellia sinensis*

Supriyo Chakraborty, Prosenjit Paul

Computational Molecular Biology, 2014, Vol.4, No.2

Association Rules for Diagnosis of Hiv-Aids

Anubha Dubey

Computational Molecular Biology, 2014, Vol.4, No.3

In Silico Proteomic Functional Re-annotation of *Escherichia coli* K-12 using Dynamic Biological Data Fusion Strategy

Ramesh Gopal, Subazini Thankaswamy Kosalai, Rajadurai Chinnasamy Perumal, Palani Kannan Kandavel

Computational Molecular Biology, 2014, Vol.4, No.4

Computational Molecular Biology

2014, Vol.4, No.4, 34-43 http://cmb.biopublisher.ca

BioPublisher©

Research Report Open Access

In silico Proteomic Functional Re-annotation of *Escherichia coli* K-12 Using Dynamic Biological Data Fusion Strategy

Gopal Ramesh Kumar, Thankaswamy Kosalai Subazini, Chinnasamy Perumal Rajadurai, Kandavel Palani Kannan

Bioinformatics Lab, AU-KBC Research Centre, M.I.T Campus of Anna University, Chromepet, Chennai 600044, India

Corresponding Author email: gramesh@au-kbc.org; Author

Computational Molecular Biology, 2014, Vol.4, No.4 doi: 10.5376/cmb.2014.04.0004

Abstract *Escherichia coli*, one of the favorite model organisms, was initially annotated in 1997 and re-annotated in 2007. Although years of intensive research is being carried out on *E. coli* genome, still complete and accurate functional information of this organism is not available. In *E. coli*, about 40% of the protein sequences have been annotated as hypothetical proteins, because of lack of information. Hence, such sequences require advanced computational strategies and derive clues on their biological role. Herein, we have carried out re-annotation of the complete proteome of *E. coli* K-12 using "Dynamic biological data fusion method". It is a computational strategy we typically applied for combining the heterogeneous biological data sources to maximize knowledge sharing and generating the intersection of data sets. Functional re-annotation results reported in this paper help us to present high quality data on complete proteome of *E. coli* K-12. We have updated all the protein coding genes from previous annotation work and tried to assign new or more precise functions, wherever possible. About 29% of the protein sequences of *E. coli* which have been previously annotated as unclear/unknown (hypothetical; without functions) have now been assigned with clear/known functions. Further, the analysis also resulted in the revision of the protein sequences that have been found to be false positive or poorly annotated. Information from this work is made available as a database, "REC-DB", which will remain a useful repository with accurate and updated functional information. Availability: REC-DB is publicly available at http://recdb.bioinfo.au-kbc.org.in/recdb/

Keywords *E. coli*; Re-annotation; Hypothetical proteins; Confidence level; Phylogenetics; Motif

Background

The field of genomics has been expanding at a rapid pace since the annotated Escherichia coli K-12 genome was published in 1997 (Blattner et al., 1997). This has led to exponential growth of sequence information and related biological databases (Serres et al, 2001). Despite decades of intense research on the E. coli genome with the attributions through the biochemical experimentations, complete and accurate functional information of this model organism is still not available. Globally several genome projects have been completed and many of them are enduring. There are incomplete functional annotation results based on obsolete data or inappropriate sequence models. Moreover, this information is not updated for years and such a poor annotation will lead to the significant gaps in our genome knowledge (Salzberg, 2007). This incomplete annotation process causes an extensive occurrence of unknown proteins in their genomes that have not yet been characterized. The traditional functional enrichment is by BLAST analysis carried out for the entire genome and in most of the case it is incomplete because functions are not updated frequently. So it is necessary to frequently update the genome function through re-annotation, otherwise the information provided will be obsolete. The occurrence of several uncharacterized proteins in the genomes such as hypothetical or conserved hypothetical proteins and functions with negative terms such as possible, probable, etc leading to uncertainty. In general for any microbial genome it will be around 30~40%. The functions of unknown proteins like

Preferred citation for this article:

Kumar et al., 2014, In silico Proteomic Functional Re-annotation of Escherichia coli K-12 Using Dynamic Biological Data Fusion Strategy, Computational Molecular Biology, Vol.4, No.4 34-43 (doi: 10.5376/cmb.2014.04.0004)

Received: 15 Apr., 2014 | Accepted: 10 May, 2014 | Published: 05 Jul., 2014

hypothetical and conserved hypothetical proteins must be predicted as they might play a vital role in cellular physiology of microorganisms. Hypothetical proteins are proteins of unknown function with no homology or experimental evidence and conserved hypothetical proteins are unknown proteins with phylogenetic distribution and homology (Wood et al., 2001; Riley et al., 2006). These uncharacterized proteins might be involved in regulation of gene expression, cell signal transduction, host–parasite interaction and complex secondary metabolism (including antibiotic and biologically active compounds synthesis) and therefore biochemically investigation of conserved hypothetical proteins makes possible to discover new biomolecules with pharmacological and biotechnological significance (Galperin and Koonin, 2010; Roberts et al., 2001). The method of predicting protein function using different bioinformatics tools makes the annotation process easy and more efficient (Altschul et al., 1999).

The wealth of biological information on *E. coli* is increasing rapidly (Serres et al., 2001) and is contributing to a better understanding of this organism as well as functions encoded in other organisms (Karp et al., 2007). It is therefore important that the most up-to-date and accurate information on *E. coli* functions are made available for the use of scientific community. Functional re-annotation, a process of annotating a previously annotated genome, would support in providing deeper insight into the genome (Rajadurai et al., 2011). This process generally involves a variety of computational techniques for functional prediction. Such functional assignments could also be achieved using more advanced high throughput technologies employed and however such techniques are highly laborious and expensive (Valencia, 2005). Hence in silico functional re-analysis would assist in making quick and reliable functional predictions. The functional re-annotation can potentially provide answers regarding higher levels of cellular processes, such as metabolism, transport, pathogenicity and regulation, thereby facilitating the elucidation of individual protein in a proteome (Zheng et al., 2002). Moreover, the results of re-annotation would also be helpful in understanding the dynamic interactions of the proteins and the underlying mechanism of metabolic processes since, all the processes are accomplished by large ordered complexes or cascading proteins. It also helps in identifying new protein functions that offer real promise of new therapies for the communicable diseases and genetic diseases. Previously, the genome of several organisms including *Mycoplasma pneumoniae* (Dandekar et al. 2000), *Mycobacterium tuberculosis H37Rv* (Camus et al., 2002), *Campylobacter jejuni* (Gundogdu et al., 2007), *Geobacter sulfurreducens* (Ashok et al., 2014) and *Saccharomyces cerevisiae* (Wood et al., 2001) were successfully re-annotated using various computational strategies and now they serve as useful pieces of information in the biological research.

Similarly, the genome analysis work has managed to analyze all the available annotations of *E. coli* and provide a snapshot of their functional information. However, at the end of their analyses, they reported ~14% of unknown sequences. They reported the results of their analyses as text files and excel sheets. Moreover, these analyses were carried out in 1997 and 2005, and by this time, a fresh set of annotations have appeared (Blattner et al. 1997; Riley et al., 2006). Hence, the aim of this work is to perform the manual proteomic re-annotation of the complete proteome sequences of *E. coli* K-12 strain using a multiple and dynamic biological data fusion strategy, where information from several protein databases are carefully compared and analyzed before assigning functions to the genome and make it available as a public database that can be useful for the scientific community dealing *E. coli* research. In this re-annotation work, a dynamic biological data fusion strategy has been implemented to perform sequence functional prediction. This strategy generally deals with the ability to dynamically form integrated data sets from the data sources by combining the heterogeneous data from database to maximize knowledge sharing (Elmore et al., 2003).

In recent years, the accumulation of complete genome sequences and related protein databases (Pearson and Lipman, 1988) provide useful comparisons with the close relatives among other organisms and facilitates

powerful re-annotation. In silico re-annotation permits uniform quality control, systemic updates, easy data parsing and more comprehensive comparative analysis, providing a valuable resource to the whole research community (Rajadurai et al., 2011). Though, several genome and proteome database are publicly available to facilitate life science research, there are a number of pitfalls associated with each database. The most common problems in these biological sequence databases are lack of reliability, redundancy, erroneous annotation etc. Recently, the misannotation levels in the four popularly used public protein sequence databases, UniProtKB/Swiss-Prot, GenBank NR, UniProtKB/TrEMBL, and KEGG have been investigated and identified the misannotated enzyme superfamilies that remains a larger problem during annotation process (Schnoes et al. 2009).

To alleviate the problem of erroneous annotation and redundancy, re-annotation of genes and proteins using a set of common, controlled context to describe a gene or protein function is necessary. Thus, deciphering the precise functions encoded by all gene products of this genome remains a great challenge in this genomic era (Bock and Gough, 2004; Altschul et al., 1990). Hence, to overcome such challenges, the re-annotation of the *E. coli* K-12 proteome has been carried out using a strategy known as “dynamic biological data fusion” in which the biological data from various available databases are integrated into a unique information source. Further, confidence level has been carried out to assign functions of unknown proteins and thereby facilitating more accurate functional information to the research society.

1 Results

The original sequence annotations of *E. coli* K-12 strain downloaded from EcoCyc database and it was identified to possess 4,290 protein sequences. Of these sequences, 2,560 sequences had clear annotations and the remaining 1,730 sequences were found to be uncharacterized with hypothetical, unknown, predicted, conserved and putative functions (Table 1, Figure 1a). Following the in silico functional re-annotation, several categories of changes were made in the previously annotated E. coli genome that includes i) Assigning functions to uncharacterized proteins ii) Transfer of functions (revision of already annotated function) and iii) Updating the functions.

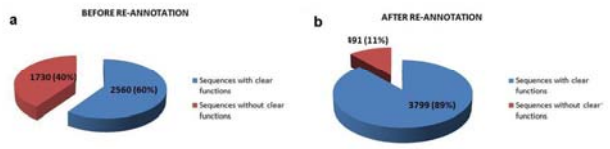


Figure 1 EcoCyc Genome Data

Note: a. A pie chart describing the percentage of known and unknown sequences in the original data downloaded from EcoCyc. b. Genome data after re-annotation. A pie chart representing the percentage of known and unknown sequences after re-annotation (Rec-DB data)

Table 1 Genome features of *E. coli* K-12

Sequence Category	Number of Sequences	Percentage (%)
Total No. of protein Sequences	4290	100
i. Sequences with unknown functions:		
Predicted	1033	24.08
Conserved	401	9.35
Putative	212	4.94
Conserved + Hypothetical	18	0.42
Hypothetical	59	1.38
Conserved + Predicted	2	0.05
Conserved + Putative	1	0.02
Hypothetical + Predicted	2	0.05
Putative + Predicted	2	0.05
Total Sequences with unknown functions	1730	40
ii. Sequences with clear functions	2560	60

1.1 Assigning functions to uncharacterized proteins

Out of 1,730 unknown protein sequences, from the previous annotations, 1,239 sequences have now been assigned with clear functions. From the re-annotation results, it was found that about 156 proteins functions were not clearly defined. For example (Table 2), the re-annotation of the protein sequence ec2389 gave different functions from different tools viz., 'metallo-beta lactomase superfamily' from Pfam, 'Zinc dependent Hydrolases' from COG, 'Probable Hydrolase' from ProDom and No Hits from BLAST and ScanProsite. Of these Pfam and ProDom produced similar function but Pfam gave a different function. Note that ProDom gave a function with the negative term "Probable", highlighted in Table 2, and hence not considered for the analysis. Similarly, for Protein Sequence 4267, Pfam and ScanProsite produced "Thiolase", as function, but other tools produced "Acetyl Transferases" as function. For some proteins, though different functions, few functional contexts were seemed to be synonymous, and others were not clearly defined. In such cases, it has practically become difficult to make a decision upon the functions. To decide properly among them, further advanced analysis strategies must be devised or the biochemical experimentations must be carried out. The two sample problem sets of sequences with different functions received from different tools that are made by manual annotation is shown in Table 2.

Table 2 Sample problem sets after re-annotation

Results	Sequence ID 2389[a]	Sequence ID 4267[a]
PFAM	Metallo-beta lactomase[b]	Thiolase[b]
COG	Zinc dependent Hydrolases[b]	Acetyl CoA Acetyl transferases[b]
SCANPROSITE	No Hits	Thiolase enzymes[b]
BLAST	No hits	Acetyl CoA Acetyl transferases[b]
PRODOM	Probable Hydrolase[c]	Probable Acetyl Transferase[c]

Note: [a] Sequence ID in Rec-DB database; [b] Different functions from different tools; [c] Functions predicted with a negative term "Probable"

For example, the conserved hypothetical protein encoded by the sequence ec1270 has now been predicted as Endoribonuclease and the complete list of data is now available at REC-DB database. Before annotation, there were 1730 (40% of sequences), unknown sequences and as a result of re-annotation it was reduced to 491 protein sequences (Figure 2b). This shows that only 11% of the proteins were left unknown/without function. Thus an overall outcome of re-annotation was found to be 29% efficient in analyzing unknown sequences of *E .coli*.

1.2 Transfer of functions (Revised functions)

Re-analysis of incompletely annotated sequences resulted in transferring the already available functions to the new and accurate functions. For example (Supplementary Table S1), protein sequence ec1034 was originally been annotated as MEND-MONOMER MenD (complement (2377281-2375611)) but now it was reassigned as 2-succinyl-6-hydroxy-2, 4-cyclohexadiene-1-carboxylate synthase. Hence, such types of functional transfers will be much useful and facilitate the scientific community to work with the more precise and reliable functions. The complete list of data is available in REC-DB database (http://recdb.bioinfo.au-kbc.org.in/recdb/).

1.3 Updates of protein functions

Re-annotation of *E. coli* has also resulted in updating the previously annotated functions based on the available biological information in the databases. In original annotation, a few proteins were designated with certain general function that was not considered to be adequate for understanding their actual biological roles. Our re-annotation has helped in adding more elaborate functions to such types of protein sequences. For example (Supplementary Table S1), protein sequence ec1915 was previously given a function, 'reductase', which is not sufficient for the functional annotation. But our re-annotation work helped us to update this function as 'Oxidoreductase molybdopterin binding domain'. Full list of data of

such types of proteins can be viewed in REC-DB database. The updated functional information of the proteins, in turn, will help the researchers to develop deeper insights into the molecular systems. The complete re-annotated functional information is available as supplementary Table S2. From REC-DB search, we could able to retrieve the gene information with clearly annotated function (Supplementary Table S2, A) which was obtained from the re-annotation results. Next, hypothetical gene hits can be obtained which means that there were no functional predictions obtained for the unknown gene sequences (available as Supplementary Table S2, B) and the functional hits with only predicted functions which need to be further annotated (Supplementary Table S2, C).

1.4 Inconvenient Outcomes

Although our re-annotation was found to be efficient in updating the functional information of the *E. coli* genome, there were a few inconveniences that had occurred in the results and these difficulties are listed below.

1.4.1 Transitive Catastrophe

Transitive catastrophe is a phenomenon whereby a function is transferred to another on the basis of sequence similarity searches but the original name is incorrect (Salzberg, 2007). As more genomes are annotated and several BLAST searches are carried out, the functional representation of some protein sequences gets transferred from one to another function. It is well known that in the genomic data resources thousands of such transitive errors have propagated through sequence databases. Thus, in the case of such incorrectly annotated information being propagated through the sequence databases using which re-annotation was carried out, then transitive catastrophes, leading to false positive functional predictions could have appeared. These inconveniences remained difficult to handle and were unable to make critical decisions upon them. Hence, such hassles are left open to the scientific community or any expert for handling and suggestions.

1.5 REC-DB

The outcome of this research work has been published online as a public database named “REC-DB – A Re-annotated *Escherichia coli* Database”. Several enhanced features have been incorporated within this database for searching functions. In this database, user can able to retrieve the re-annotated *E. coli* genome data by querying REC-DB accession number (eg. ec001), by choosing GenBank id (GI.No. 90111633) or by giving Gene id (Gene-ID. 948195). While querying, user may find “Null”, “No GI” and “No Gene id” in search option which actually means that there are no REC-DB function if it is queried as “Null” and there are no GenBank id, if it is searched as “No GI” and no gene id occurs in REC-DB, if it is search as “No Gene id”.

2 Discussion

Although genome projects have the potential to provide a better understanding of the organisms, the lack of updated and accurate functional annotation for the genome hampers the ability to exploit these data for any further research on the organism. Hence in this in silico functional proteomic re-annotation, an attempt has been made to substantially update the functions of the entire sequences of *E. coli* K-12, incorporating a vast amount of research information performed since the original annotation in 1997. Much knowledge has been gained about the molecular functions encoded by the *E. coli* K-12 genome.

Analyzing a single sequence using a regular BLAST program (http://www.ebi.ac.uk/Tools/BLAST/), will itself generate large amount of results in terms of hits accompanied with varied parameters such as E-value, Percentage of Identity, Percentage of Similarity, BLAST score and sequence length. The results obtained from BLAST with a maximum alignment score and optimal E-value of 1×10-6 up to 1×10 -52 can be obtained as a result hit (Gabriel et al., 2008). This requires a lot of human interventions to interpret and choose the best positive hit. Thus, analyzing the entire proteome of *E. coli* using a regular BLAST program will be tedious (Aravindhan et al., 2009; Hulo et al., 2004). AIM-BLAST with a well structured and in a concise manner, supported us greatly in performing sequence comparisons of the complete genome of *E. coli* efficiently and in a very short span of time (Aravindhan et al. 2009).

Each of the *E. coli* K-12 protein sequences previously predicted and annotated has been manually re-analyzed based on the diverse approaches such as similarity based search approach (BLASTP), Pattern based search (ScanProsite), Phylogenetic classification based search (COG), Domain based search (ProDom) and Proteins Family based search (Pfam). Each approach has different emphasis and collects different sets of information related to the function of the gene products. Further, since each of these databases has been designed for specific problems and therefore have their own inherent strengths and weaknesses (Rust et al., 2002).

Our re-annotation strategy helped in postulating featured functions to almost 29% of the protein sequences. For example (Supplementary Table S1), the hypothetical protein encoded by the sequence ec0903 has now been newly assigned as Formate dehydrogenase. Importantly, ec0903 whose Refseq id YP_001165334.1 was previously reported as hypothetical protein and they used for vaccine prediction for infectious disease in human for which assigned a new function (Xiang and He, 2009). Re-annotation also helped in revising the previously annotated proteins. In several cases, the original annotation was very wide and less precise ones. While analyzing such sequences, we found that they were poorly annotated and were consigned with a false positive function (Aravindhan et al., 2009).

As a result of our manual re-annotation, almost 29% of the genes, whose functions were not determined earlier, are now assigned with a known function. The data presented in REC-DB should be useful for analysis of *E. coli* gene products as well as gene products encoded by other genomes. Hence, we believe that our re-annotation should be useful for the scientific community in *E. coli* research.

3 Material and Methods

3.1 Protein Sequences

The complete protein sequences of *E. coli* K-12 organism were downloaded from the EcoCyc Database (Keseler et al., 2005) and analyzed. The previous functional genomics analysis on *E. coli* showed that there were totally 4,290 protein sequences of which only 60% of the sequences were found to have clear/known (with functions) functions. The remaining 40% of sequence functions were left as unclear/unknown genes (without functions). The protein functions of unknown genes such as hypothetical and conserved hypothetical proteins must be predicted as they might play a significant role in cellular physiology of microorganisms. Hypothetical proteins are proteins of unknown function with no homology or experimental evidence and conserved hypothetical proteins are unknown proteins with phylogenetic distribution and homology (Tao et al., 1999).

3.2 Functional annotation of Sequences

The complete genome sequence comprises of totally 4,290 protein sequences of which 2,560 protein sequences have clear/known function and 1,730 sequence functions are unclear/unknown (Table 1). Computers play a significant role in sequence analysis and re-annotation as it reduces the analysis time taken for the processing of large amounts of data and through the integration of several approaches (Nascimento and Bazzan, 2005). Here, the functional re-annotation of entire *E. coli* proteome was carried out using the advanced re-annotation strategies in which several sequence analysis methods were incorporated into a coherent and an efficient annotation schema (Figure 2).

3.2.1 Similarity Search Approach using AIM BLAST

Proteins that are evolutionarily related are commonly referred to as homologues and close homologues often have similar functions (Ofran et al., 2005). Based on this concept, homology-based or similarity based transfer of functional annotations remain a native prediction method to assign functions of unknown proteins that have not been previously annotated. In turn, due to the serious and quicker accumulation of the fresh biological information in the protein databases, this similarity search approach would also help in revising or updating the previously annotated functions.

BLAST, Basic Local Alignment Search Tool is one of the most favorite and widely used Bioinformatics program for identifying the similarity between the

biological sequences (Altschul et al., 1990). Since, this tool remains computationally intensive and time consuming as they employ a voluminous data transfer and requires manual intervention for parsing BLAST sequence analysis result. The E-value threshold for BLAST search is 1*10-6 to 1*10-52 (Gabriel et al. 2008). To overcome such difficulties, we have developed a program, AIM-BLAST that has been interfaced with AJAX and SOAP services of EBI (European Bioinformatics Institute) to support multiple sequence searches at a stretch during re-annotation. Further, AIM-BLAST has enhanced features for performing automated parsing of the huge blast results of individual sequences and presenting them as "one sequence-one function" manner with manual curation.

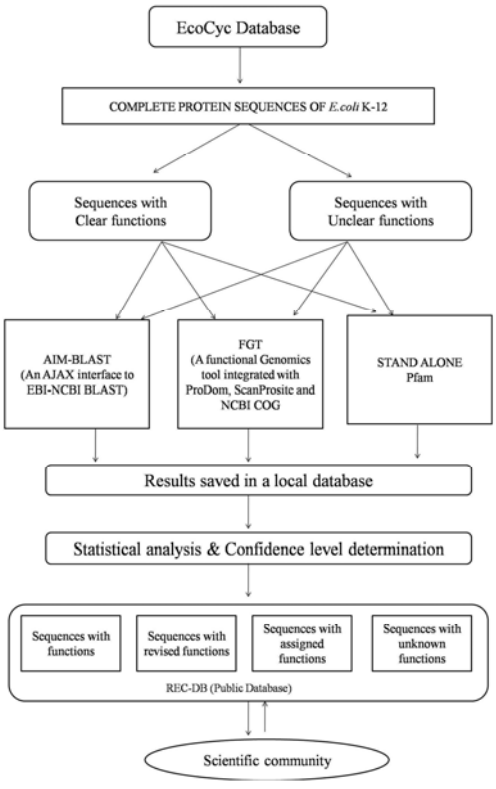


Figure 2 Schematic representation of the Re-annotation strategy "Dynamic biological data fusion"
This flow chart describes the step by step procedures for carrying out the re-annotation of *E.coli* K12 strain.

3.2.2 Sequence motifs and patterns search for controlled annotation

Genome annotations describe that the function of sequences are important to researchers during laboratory investigations and when making computational inferences. Re-annotation based on the BLAST analysis alone would not be helpful in making accurate functional re-annotation and predictions, as in some cases these annotations may be inconsistent, incomplete and erroneous (Karp et al., 2007). Hence, sequence analysis based on the several approaches such as motifs and pattern searches, phylogeny based search, domain based search and family based search were carried out to efficiently annotate the *E. coli* K-12 proteome. When comparing the protein sequences, although they might not seem to be very identical possess a short region of sequence "motif" in common that is explicit to specific functions. Thus, identifying such distinctive motif patterns in the protein sequences could help in predicting the functions on un-annotated proteins that contain similar motifs. A few databases are dedicated to identify such motifs patterns are available, of which ScanProsite (http://expasy.org/tools/scanprosite/) was chosen for this work to search for hits by specific motifs in the protein databases. This tool makes use of ProRules-context-dependent annotation templates to discover functional and structural intra-domain residues by scanning the protein sequences for the occurrence of possible motifs and predicts their function (Hulo et al., 2004).

3.2.3 Phylogenetic Classification based approach

The database of Clusters of Orthologous Groups of proteins (COGs) consists of information on the classifications of the proteins sequences based on their phylogenies (Tatusov et al., 2000). This COGs database serves as a best portal for carrying out functional annotation of the proteins sequences based on their genome evolution. To facilitate functional studies, the COGs have been classified into 17 broad functional categories, including a class for which only a general functional prediction, usually that of biochemical activity, was feasible and a class of uncharacterized COGs. Additionally, COGs with known functions are organized to represent specific cellular systems and biochemical pathways. Thus, sequence analysis using the COGNITOR program of COG database would produce deeper insights to the

protein functions based on their genome evolution.

3.2.4 Domain based search

Domains are the structural, functional and evolutionary units of proteins. Domains of common lineage are clustered into superfamilies. Thus, functional annotation based on such domain superfamilies will support enhanced knowledge on the protein functions. ProDom, an exclusive database of protein domain families has also been used that support our analysis based on the domain arrangements of proteins (Servant et al., 2002). ProDom can be useful in providing functional information of the proteins by carrying out a global comparison of the submitted sequence against all the available protein sequences. For ProDom a cutoff default E-value of 0.01 is kept and searched using ncbi-blastp program and multiple alignment method.

3.2.5 Protein family search using Pfam

Family based classification also remains as an important means of providing functional annotation for the biological sequences. Pfam (Finn et al., 2008) is a collection of multiple sequences alignments and profile Hidden Markov Models (HMM) that represent protein families. Information on the protein functions can thus be realized by comparing the sequences against the Pfam library of HMMs (Wu et al., 2003). Using Pfam database the complete proteome of *E. coli* was analyzed and the functions predicted based on these families. Although Pfam was integrated with the FGT program, Pfam analyses were carried out separately. The cutoff E-value is set as default 0.001. This is because, Pfam analysis generally consumes more time comparatively and running Pfam for a large number of protein sequences in FGT would affect its performance and slow down the overall process. But, at the same time, running each and every sequence separately in the Pfam server directly would also be a monotonous process. Hence, the Pfam FTP files were downloaded and installed into a local system and a stand alone Pfam was devised to support large-scale sequence analysis during the re-annotation.

3.3 Hectic process of annotation

ScanProsite, COG and ProDom were selected for re-analyzing the complete *E. coli* proteome, because they operate on different strategies to explore the biological roles of proteins. But there are some inconveniences prevailing with these tools. These tools do not allow multiple sequence searches at an instance. Also, for every single sequence search, these tools produce many hits and the users have to carefully interpret them and choose the best hit. After choosing the best hit, the users have to copy the appropriate function to a local database for final interpretation. Further, users have to simultaneously open and deal with multiple browsers when analyzing with these tools at a given time. This in turn consumes much of the man power and man hour. Hence, it would be an immense tiresome process for performing searches for the entire *E. coli* proteome using all these tools.

3.4 In-house Functional Genomics Tool

Understanding the complicatedness of dealing with many tools simultaneously, a simple and but novel system, Bioinfotracker (a Functional Genomics Tool-FGT), was developed for performing controlled annotation of the protein sequences locally, by concurrently using different online functional prediction tools (Kumar et al., 2009). FGT is a well-structured, flexible and a highly systematic functional analysis program developed by us to carry out large-scale protein annotation. Different online tools, operating on the diverse research strategies, such as ScanProsite, ProDom, COG and Pfam have been integrated in this tool. Once a sequence is submitted to this tool, the sequence is forwarded and submitted to the different servers of the tools integrated and the process is carried out at the individual servers in tandem. In ScanProsite the first option of scanning against PRSOTE collection of motif was used. For ProDom a cutoff default E-value of 0.01 is kept. The cutoff E-values for COG, Pfam are mentioned as 0.001. On Completion and when the results are available this tool will perform an automated parsing of the results to choose the best function, fetch them from the corresponding servers. Further, the results of the submitted sequences are provided as a simple table format that will be easier to interpret. Hence, FGT tool was very much supportive for carrying out the re-annotation of the complete

proteome sequences of *E. coli* K-12 organism.

Here, in silico functional proteomic re-annotation of the *E. coli* K-12 was carried out using a well-organized and proficient annotation procedure that involves dynamic fusion of biological data from various databases (Figure 2).The complete genome of *E. coli* K-12 was downloaded from the EcoCyc database and an initial analysis on the total number of protein sequences with clear functions and the sequences without clear functions were carried out. Then, all the sequences were submitted to the AIM BLAST for BLAST analyses, FGT for ProDom, ScanProsite and COG analyses and stand alone Pfam for family based analyses. The results of individual sequences from all the five tools were stored in a local database for further analyses. When the results of all the sequences were available, a statistical analysis was undertaken to determine the confidence level of predicted functions.

Authors' contributions

GRK did Overall data management,software development and writing paper. TKS carried out software development and writing paper. CPR did data analysis and interpretation. KPK involved in software development and data analysis. The authors have read and approved the manuscript.

Acknowledgements

Authors would like to thank Aravindan Ganesan, Kalyanamoorthy Subha and R Sathish Kumar for their constructive comments which helped in preparation of this manuscript.

References

Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J., 1990, Basic local alignment search tool. J. Mol. Biol., 215: 403-410 http://dx.doi.org/10.1016/S0022-2836(05)80360-2

Ashok S., Venil S., Nupoor C., and Kumar G.R., 2014, Prediction and classification of ABC transporters in *Geobacter sulfurreducens* PCA using computational approaches. Current Bioinformatics, 9(2): 166-172 http://dx.doi.org/10.2174/15748936089991401091132 36

Aravindhan G., Kumar G.R., Kumar R.S., and Subha K., 2009, AJAX Interface: A Breakthrough in Bioinformatics Web Applications. Proteomics Insights, 2: 1-7

Aravindhan G., Kumar R.S., Subha K., Subazini T.K., Dey A., Kant K., and Kumar G.R., 2009, AIM-BLAST-AJAX Interfaced Multisequence Blast. Proteomics Insights, *2:* 9-13

Blattner F.R., Plunkett G., Bloch C.A., Perna N.T., Burland V., and Riley M. et al., 1997, The complete genome sequence of *Escherichia coli* K-12. Science, 277: 1453-1474 http://dx.doi.org/10.1126/science.277.5331.1453

Bock J.R., and Gough D.A., 2004, In silico biological function attribution: a different perspective. Drug Discov Today Biosilico, 2: 30-37 http://dx.doi.org/10.1016/S1741-8364(04)02381-9

Camus J.C., Pryor M.J., Médigue C., and Cole S.T., 2002, Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. Microbiology, 148: 2967-2973

Dandekar T., Huynen M., Regula J.T., Ueberle B., Zimmermann C.U., and Andrade M.A., 2000, Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. Nucl. Acids Res., 28: 3278-3288 http://dx.doi.org/10.1093/nar/28.17.3278

Elmore M.T., Potok T.E., and Sheldon F.T., 2003, Dynamic Data Fusion Using An Ontology-Based Software Agent System. Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics, 1-6

Finn R.D., Tate J., Mistry J., Coggill P.C., Sammut S.J., Hotz H.R, Ceric G., Forslund K., Eddy S.R., Sonnhammer E.L, and Bateman A., 2008,The Pfam protein families database. Nucl. Acids Res., 36: 281-288 http://dx.doi.org/10.1093/nar/gkm960

Gabriel M.H., and Kristen L., 2008, Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics, 24: 319-324 http://dx.doi.org/10.1093/bioinformatics/btm585

Galperin M.Y., and Koonin E.V., 2010, From complete genome sequence to 'complete' understanding? Trends Biotechnol., 28: 398-406 http://dx.doi.org/10.1016/j.tibtech.2010.05.006

Gundogdu O., Bentley S.D., Holden M.T., Parkhill J., Dorrell N., and Wren B.W., 2007, Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. BMC Genomics, 8: 162-170 http://dx.doi.org/10.1186/1471-2164-8-162

Hulo N., Sigrist C.J., Le Saux V., Langendijk-Genevaux P.S., Bordoli L., Gattiker A., De Castro E., Bucher P., and Bairoch A., 2004, Recent improvements to the PROSITE database. Nucl. Acids Res., 32: 134-137 http://dx.doi.org/10.1093/nar/gkh044

Karp P.D., Keseler I.M., Shearer A., Latendresse M.,

Krummenacker M., Paley S.M., and Paulsen I., 2007, Multidimensional annotation of the *Escherichia coli* K-12 genome. Nucl. Acids Res., 35: 7577-90
http://dx.doi.org/10.1093/nar/gkm740

Keseler I.M., Collado-Vides J., Gama-Castro S., Ingraham J., Paley S., Paulsen I.T., Peralta-Gil M., and Karp P.D., 2005, EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucl. Acids Res., 33: 334-337
http://dx.doi.org/10.1093/nar/gki108

Kumar G.R., Aravindhan G., Subazini T.K., and Kumar R.S., 2009, Bioinfotracker: A novel system for advanced genome functional insight. Journal of Bioinformatics and Sequence Analysis, 1(3): 046-049

Nascimento L.V., and Bazzan A.L., 2005, An agent-based system for re-annotation of genomes. Genet. Mol. Res., 4: 571-580

Ofran Y., Punta M., Schneider R, and Rost B., 205, Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. Drug Discov. Today Biosilico., 10: 1475-1482
http://dx.doi.org/10.1016/S1359-6446(05)03621-4

Pearson W.R, and Lipman D.J., 1988, Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA, 85: 2444-2448
http://dx.doi.org/10.1073/pnas.85.8.2444

Rajadurai C.P., Subazini T.K., and Kumar G.R., 2011, An integrated re-annotation approach for functional predictions of hypothetical proteins in microbial genomes. Current Bioinformatics, 6(4): 450-461
http://dx.doi.org/10.2174/157489311798072954

Riley M., Abe T., Arnaud M.B., Berlyn M.K., Blattner F.R., Chaudhuri R.R., and Glasner J.D., 2006, *Escherichia coli* K-12: a cooperatively developed annotation snapshot. Nucl. Acids Res., 34: 1-9
http://dx.doi.org/10.1093/nar/gkj405

Roberts R.J., Chang Y.C., and Hu Z. Rachlin J.N., Anton B.P., Pokrzywa R.M., Choi H.P., Faller L.L., Guleria J., Housman G., Klitgord N., Mazumdar V., McGettrick M.G., Osmani L., Swaminathan R., Tao K.R., Letovsky S., Vitkup D., Segrè D., Salzberg S.L., Delisi C., Steffen M., Kasif S., 2011, COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. Nucl. Acids Res., 39: D11-D14
http://dx.doi.org/10.1093/nar/gkq1168

Rust A.G., Mongin E., and Birney E., 2002, Genome annotation techniques: new approaches and challenges. Drug Discov. Today Biosilico., 7: 70-76
http://dx.doi.org/10.1016/S1359-6446(02)02289-4

Salzberg S.L., 2007, Genome re-annotation: a wiki solution? Genome Biol., 8: 1-5
http://dx.doi.org/10.1186/gb-2007-8-1-r1

Schnoes A.M., Brown S.D., Dodevski I., and Babbitt P.C., 2009, Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies, PLoS Comput. Biol., 5: e1000605
http://dx.doi.org/10.1371/journal.pcbi.1000605

Serres M.H., Gopal S., Nahum L.A., Liang P., Gaasterland T., and Riley M.A., 2001, functional update of the *Escherichia coli* K-12 genome. Genome Biol., 2: 1-7
http://dx.doi.org/10.1186/gb-2001-2-9-research0035

Servant F., Bru C., Carrère S., Courcelle E., Gouzy J., Peyruc D., and Kahn D., 2002, ProDom: Automated clustering of homologous domains. Briefings in Bioinformatics, 3: 246-251
http://dx.doi.org/10.1093/bib/3.3.246

Tao H., Bausch C., Richmond C., Blattner F.R., and Conway T., 1999, Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. J. Bacteriol., 181: 6425-6440

Tatusov R.L., Natale D.A., Garkavtsev I.V., Tatusova T.A., Shankavaram U.T., Rao B.S., Kiryutin B., Galperin M.Y., Fedorova N.D., and Koonin E.V., 2001, The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res., 129: 22-28
http://dx.doi.org/10.1093/nar/29.1.22

Valencia A., 2005, Automatic annotation of protein function. Curr. Op. Struct. Biol., 15: 267-274
http://dx.doi.org/10.1016/j.sbi.2005.05.010

Wood V., Rutherford K.M., Ivens A., Rajandream M.A., and Barrell B., 2001, A re-annotation of the *Saccharomyces cerevisiae* genome. Comp. Funct. Genomics, 2: 143-154
http://dx.doi.org/10.1002/cfg.86

Wu C.H., Huang C.H., Yeh L.S., and Barker W.C., 2003, Protein family classification and functional annotation. Comput. Biol. Chem.,, 27: 37-47
http://dx.doi.org/10.1016/S1476-9271(02)00098-1

Xiang Z., and He Y., 2009, Vaxign: a web-based vaccine target design program for reverse vaccinology. Procedia in Vaccinology, 1: 23-29
http://dx.doi.org/10.1016/j.provac.2009.07.005

Zheng Y., Roberts., and Kasif S., 2002, Genomic functional annotation using co-evolution profiles of gene clusters. Genome Biol., 3: 1-9
http://dx.doi.org/10.1186/gb-2002-3-11-research0060