Journal of Infection and Public Health (2015) 8, 526—532

http://www.elsevier.com/locate/jiph

In silico functional annotation of a hypothetical protein from *Staphylococcus aureus*

P. Bharat Siva Varma[a], **Yesu B. Adimulam**[b,*], **Subrahmaniam Kodukula**[c]

[a] *Department of Computer Science & Engineering, SRKR Engineering College, Bhimavaram, Andhra Pradesh, India*
[b] *Department of Computer Science & Engineering, Sir C.R. Reddy College of Engineering, Eluru, Andhra Pradesh, India*
[c] *Department of Computer Science and Systems Engineering, K.L. University, Vaddeswaram, Vijayawada, Andhra Pradesh, India*

Received 27 December 2014; received in revised form 1 March 2015; accepted 28 March 2015

KEYWORDS
Hypothetical protein;
Blast;
HHpred;
Domains of unknown function

Summary Unknown proteins or hypothetical proteins exist but have not been characterized or linked to known genes. Domains of unknown function are experimentally identified proteins with no known functional or structural domain. In this paper, the investigation and characterization of the likely functional aspects of a hypothetical protein, YP_001317347.1, from *Staphylococcus aureus* was performed using various computational methods and tools. Based on the analysis, the protein has a YbbR domain and is expected to bind ribosomal subunits. The analysis reported here helps in understanding the importance of YbbR domains and will aid in the development of novel antibacterial agents.

Introduction

A large portion of mammalian proteomes is represented by hypothetical proteins (HP), which are proteins predicted from nucleic acid sequences only and protein sequences with unknown function [1]. Several approaches have been developed by scientists with the aid of various computational tools to predict protein function. This has been achieved from information derived from sequence similarity, phylogenetic analysis, protein-protein interaction, protein—ligand interactions, active site residue similarity, conserved domains, motifs, phosphorylation regions and gene expression profiles. However,

* Corresponding author. Tel.: +91 7032739277.
E-mail address: yesubabuadimulam9@gmail.com (Y.B. Adimulam).

http://dx.doi.org/10.1016/j.jiph.2015.03.007

the classical method of inferring function is based on sequence similarity using programs such as BLAST, FASTA [2] and PSI-BLAST [3]. HPs are predicted proteins from nucleic acid sequences that have no experimental protein chemical evidence for their existence. Moreover, these proteins are characterized by a low identity to known, annotated proteins [1]. Few HPs are conserved and are found in organisms from several phylogenetic lineages. HPs represent a large fraction of genes in sequenced microbial genomes; however, they have not been functionally characterized and described at the protein chemical level [4]. Two classes of HPs exist. One class is the uncharacterized protein families (UPFs) and the other class is the domains of unknown function (DUFs). Unknown proteins have experimental structures that have been shown to exist but have not been characterized or linked to a known gene. DUFs are experimentally identified proteins; however, they have no known functional or structural domains. They may contain coiled-coil structures or transmembrane regions that do not allow for the assignment of function.

Analyzing the function of proteins with no known function offers many advantages, such as determining new conformational orientations of 3-dimensional structures, which makes it possible to evaluate new domains and motifs as well as reveals additional protein pathways and cascades. These new domains might offer potential pharmacological targets.

Moreover, function prediction can be inferred from the phylogenetic profiling of proteins in multiple genomes [5], and high throughout methods, such as protein complex identification by mass spectrometry, microarray gene expression profiles [6] and systematic synthetic lethal analysis [7], are useful. Clustering gene-expression profiles is a general widely implemented approach that is used to predict function based on the assumption that genes with similar functions are likely to be co-expressed [8]. Schwikowski et al. [9] used the neighbor-counting method to predict function. They assigned a function to an unknown protein based on the frequencies of its neighbors with certain functions. Instead of searching for a simple consensus among the functions of the interacting partners, Deng et al. used the Bayesian approach to assign a probability for a hypothetical protein to display the annotated function.

Many protein domains have unknown functions; however, these domains participate in the metabolic pathways of organisms and can cause adverse effects. Sometimes the function of the protein may change due to mutations, such as insertions, deletions and substitutions. The main objective of the study is to identify a protein domain of unknown function and determine its classification using bioinformatics tools.

Materials and methods

Selection of the hypothetical protein

Hypothetical proteins were searched in the protein database of NCBI using the keyword, ''hypothetical protein,'' and the resultant hits were randomly selected to study the near relatives using blast programs. To predict the function of the query protein, a similarity search was performed using NCBI blast tools to identify proteins that may have structural similarity with that of the hypothetical protein [10].

Physicochemical characterization of the hypothetical protein

The hypothetical protein in raw sequence format was evaluated for physicochemical properties using the ProtParam tool in the ExPASy server [11]. The parameters computed by the program and reported here include the molecular weight, theoretical pI, amino acid composition, total number of positive and negative residues, extinction coefficient, instability index, aliphatic index and grand average of hydropathicity (GRAVY). The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. The instability index provides an estimate of the stability of a protein in a test tube. An instability index <40 is predicted to be stable, and a value >40 is predicted to be unstable. The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chain amino acids. The GRAVY value for a peptide or protein is calculated as the sum of the hydropathy values of all of the amino acids divided by the number of residues in the sequence [12].

Sequence analysis

The Basic Local Alignment Search Tool (BLAST) [13] is the most frequently used tool for calculating sequence similarity. The FASTA sequence of the YP_001317347.1protein was the query sequence, and similar proteins in different databases were searched for using the BLASTP program. BLASTP is used to identify a query amino acid sequence and to find similar sequences in protein databases.

HHPred model generation

Conventional sequence search methods examine sequence databases, such as UniProt or