Journal of Theoretical Biology 435 (2017) 229–237

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

iPHLoc-ES: Identification of bacteriophage protein locations using evolutionary and structural features

Swakkhar Shatabda[a,*], Sanjay Saha[a], Alok Sharma[b,c,e], Abdollah Dehzangi[d]

[a] Department of Computer Science and Engineering, United International University, House 80, Road 8A, Dhanmondi, Dhaka-1209, Bangladesh
[b] Institute for Integrated and Intelligent Systems, Griffith University, Australia
[c] School of Engineering and Physics, University of the South Pacific, Fiji
[d] Department of Computer Science, School of Computer, Mathematical, and Natural Sciences, Morgan State University, United States
[e] RIKEN Center for Integrative Medical Sciences, Japan

ARTICLE INFO

Article history:
Received 20 July 2017
Revised 18 September 2017
Accepted 20 September 2017
Available online 21 September 2017

MSC:
00-01
99-00

Keywords:
Proteins
Locations
Phage
Classification
Feature selection

ABSTRACT

Bacteriophage proteins are viruses that can significantly impact on the functioning of bacteria and can be used in phage based therapy. The functioning of Bacteriophage in the host bacteria depends on its location in those host cells. It is very important to know the subcellular location of the phage proteins in a host cell in order to understand their working mechanism. In this paper, we propose iPHLoc-ES, a prediction method for subcellular localization of bacteriophage proteins. We aim to solve two problems: discriminating between host located and non-host located phage proteins and discriminating between the locations of host located protein in a host cell (membrane or cytoplasm). To do this, we extract sets of evolutionary and structural features of phage protein and employ Support Vector Machine (SVM) as our classifier. We also use recursive feature elimination (RFE) to reduce the number of features for effective prediction. On standard dataset using standard evaluation criteria, our method significantly outperforms the state-of-the-art predictor. iPHLoc-ES is readily available to use as a standalone tool from: https://github.com/swakkhar/iPHLoc-ES/ and as a web application from: http://brl.uiu.ac.bd/iPHLoc-ES/.

1. Introduction

The term 'bacteriophage' means 'bacteria eaters' in Latin. Bacteriophage or informally called phage proteins are viruses that can kill the bacteria by infection and replication. History of phage goes back 100 years back in 1910s when phages were used to cure dysentery (Keen, 2012; Lederberg, 1996). With the emergence of antibiotics, phage therapy somehow lost its popularity (Keen, 2012). However, in recent years due to continuous abuse of anti-bacterial drug by inappropriate prescription practices and poor drug access control (Liljeqvist et al., 2012) and evolving capability of the microbes, the commercial viability of new antibiotics is in decline (Hughes, 2011). The overuse of antibiotics have also been detrimental to the communities of beneficial bacteria (Buffie et al., 2012). In contrast, the phages are very precise in nature and the scientists are again looking back to these bacteriophages to treat the intractable bacterial infections (Deresinski, 2009; Sorokulova et al., 2014).

An injected bacteriophage transcribed by host cell polymerase typically has two life cycles: lytic and lysogenic. In lysogenic or temperate phase, the phage continues replication along with the host cell. However, lysis instigated typically by enzymes breaks open the host cell membrane and destroys it (Sass and Bierbaum, 2007). Phage proteins are either extra-cellular or not located in host cells or located in host cells. Extra cellular phages often take help of receptor for adsorption whose location are pivotal among other factors (Rakhuba et al., 2010). Subcellular localization of phage proteins are mostly distributed in host membrane or in host cytoplasm. Knowledge of the location of bacteriophage proteins are fundamental to the understanding of the mechanism of the virion and development of anti-bacterial therapy. Electron microscopy is generally used to find the locations of phage proteins in host cell (Altman et al., 1985; Casjens and Hendrix, 1988). However, the experimental methods are still time consuming and expensive.

Many computational methods have been developed to study and analyze phage proteins (Cheng et al., 2017a; 2017c; Chou and Shen, 2006; Ding et al., 2014; 2016a; 2016b; Khan et al., 2017; Seguritan et al., 2012; Shen and Chou, 2007a; 2007b; 2009; 2010a; 2010b; Wu et al., 2012; Xiao et al., 2011a; 2011b; Zhou et al., 2011). PHAST was introduced in Zhou et al. (2011) to identify and

* Corresponding author.
E-mail addresses: swakkhar@cse.uiu.ac.bd (S. Shatabda), sanjay@cse.uiu.ac.bd (S. Saha), alok.sharma@griffith.edu.au (A. Sharma), abdollah.dehzangi@morgan.edu (A. Dehzangi).

https://doi.org/10.1016/j.jtbi.2017.09.022
0022-5193/

annotate prophage sequences within bacterial genomes. Among other phage finding tools are PHASTER (Arndt et al., 2016), Phage_finder (Fouts, 2006). Another successful phage prediction tool was PhiSpy (Akhter et al., 2012) that used similarity and composition based strategies.

Several classification algorithms are used to predict phage or phage locations including Artificial Neural Network (ANN) (Galiez et al., 2015; Seguritan et al., 2012), Support Vector Machine (SVM) (Ding et al., 2016b), Random Forest (RF) (McNair et al., 2012) and Naive Bayesian Classifier (NBC) (Feng et al., 2013). Subcellular localization of proteins (Emanuelsson et al., 2000) and bacteriophages (Chou and Shen, 2007; Ding et al., 2014) are of interest for a long time in the research field. In a very recent work, a prediction methodology was proposed to identify phage locations in protein in Ding et al. (2016a) using feature selection method. They have used Support Vector Machine (SVM) classifier to solve two subcellular localization problems on a verified benchmark dataset.

In this paper we tackle two types of localization problems. The first problem we denote as *PH vs non-PH* discrimination problem, where the aim is to classify whether a given phage protein is a host located phage (PH) or a extra-cellular phage (non-PH). The second problem is denoted by *PHM vs PHC* classification where the aim is to classify between two types of host located phages, whether they are located in cell membrane (PHM) or in cell cytoplasm (PHC). We propose iPHLoc-ES for prediction of subcellular locations of phage proteins. iPHLoc-ES is also able to discriminate between host located phages and extra-cellular phages. Our predictor is based on extracting a set of evolutionary and structural features and using a Support Vector Machine (SVM) classifier along with recursive feature elimination (RFE) as feature selection technique. On the standard benchmark dataset of phage proteins our method significantly outperforms the state-of-the-art predictor. We have also made iPHLoc-ES available as a stand-alone tool that is freely available to use (https://github.com/swakkhar/iPHLoc-ES/). We have also made it available as a web application from: http://brl.uiu.ac.bd/iPHLoc-ES/.

In this paper, we follow the guidelines in compliance with Chou's 5-step rule (Chou, 2011) to establish a useful statistical predictor for a biological system. The rest of the paper is organized accordingly: (a) description of the benchmark dataset and construction of train and test sets for the predictor; (b) mathematical formulation of the biological sequence samples that can reflect their intrinsic correlation with the target to be predicted; (c) a powerful model for feature selection and classification algorithm; (d) proper experimentation with cross-validation tests; (e) a user-friendly web-server for the predictor that is accessible to the public.

2. Materials and methods

In this section, we describe the materials and methods required to develop iPHLoc-ES. We call our system **i**dentification of bacterio**PH**age protein **Loc**ations using **E**volutionary and **S**tructural Features (iPHLoc-ES). A system flow-chart of our prediction model is given in Fig. 1.

Phage protein sequences from the benchmark dataset are first fed to PSI-BLAST (Altschul et al., 1997) and SPIDER2 (Heffernan et al., 2015; Yang et al., 2017). PSI-BLAST produces a position specific scoring matrix (PSSM) file and SPIDER2 predicts structural information and generates a SPD file that is used by the feature generation module to generate a set of features. Features are generated belonging to three different groups: composition based evolutionary features, PSSM based evolutionary features and SPD based structural features. After the feature generation a feature selection method selects only a small subset of features to train the dataset. With the help of this selected small set of features the original dataset is transformed and trained using a classification model. We used Support Vector Machine (SVM) (Cortes and Vapnik, 1995) in this paper due to superiority over other methods (Ding et al., 2016b). The trained model is saved for prediction phase. Whenever a new sequence is given, it goes through the same process and given the instance with selected features, the trained model predicts its label. For both of the problems (PH vs non-PH and PHM vs PHC), we follow the same procedure.

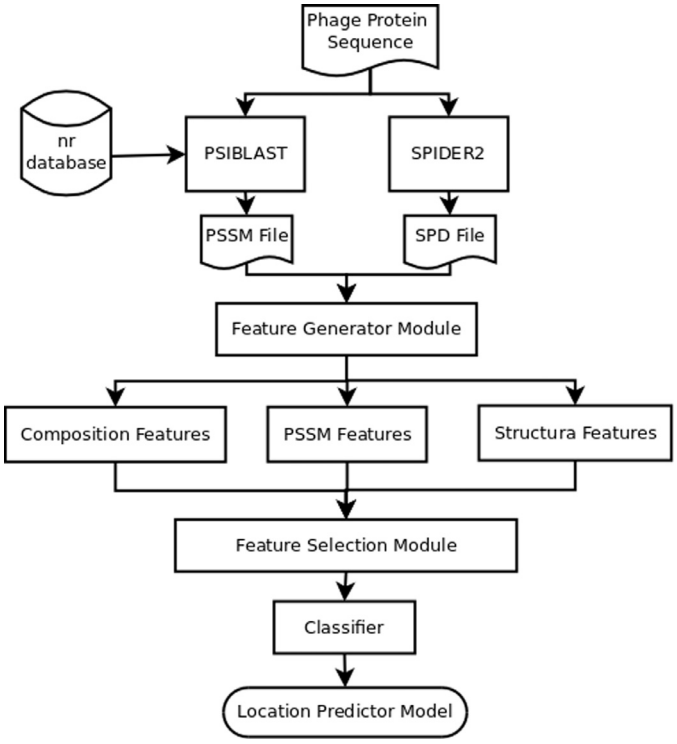

Fig. 1. System flowchart of iPHLoc-ES.

Table 1
Summary of bacteriophage protein dataset for pH vs non-PH prediction.

Phage Type	Number of Samples
Host-Located Proteins (PH)	144
Extra-Cellular Proteins (non-PH)	134

2.1. Benchmark dataset

The description of the datasets used in this paper for pH vs non-PH problem is given in Table 1. There are total 278 instances out of which 144 are positive instances or host-located proteins and 134 are extra-cellular proteins or negative samples. This dataset is similar to the one used in Ding et al. (2016a). All the protein sequences are collected from UniProt Database (Consortium, 2014). All these subcellular locations are experimentally validated. Subphages that are part of other phage proteins or the phages with non-standard amino-acids were discarded to generate the dataset. This dataset excludes the redundant sequences with similarity threshold set to 30%.

From the host located protein dataset, a second dataset was derived for PHC vs PHM problem. The description is given in Table 2. In total, 68 phages are location in cell membrane and 76 phages are located in cell cytoplasm.