Chemometrics and Intelligent Laboratory Systems 180 (2018) 64–69

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Prediction of bacteriophage proteins located in the host cell using hybrid features

Jing-Hui Cheng [a], Hui Yang [a], Meng-Lu Liu [a], Wei Su [a], Peng-Mian Feng [b], Hui Ding [a,**], Wei Chen [a,c,***], Hao Lin [a,*]

[a] Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China
[b] Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry, School of Public Health, North China University of Science and Technology, Tangshan, China
[c] Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan, 063000, China

ARTICLE INFO

Keywords:
Phage proteins
Host cell
Subcellular location
Feature selection
Webserver

ABSTRACT

The identification of bacteriophage proteins in the host subcellular localization could provide important clues for understanding the interaction between phage and host bacteria as well as antibacterial drug design. To date, computational methods have been reported to identify bacteriophage proteins located in the host cell. However, there is still space for improving the prediction accuracy. The existing methods considering the sequence order correlation and the physicochemical property of protein provide us insights to construct an integrated descriptor based on sequence for phage proteins. Meanwhile, we proposed a feature selection technique to obtain the optimal features. In the jackknife test, the prediction accuracies are 86.7% and 97.9%, respectively for discrimination between PH proteins and non-PH proteins as well as PHM proteins and PHC proteins. Based on our model, we updated the web server PHPred to version 2.0 which can be freely accessed from http://lin-group.cn/server/PHPred2.0.

1. Introduction

Genome duplication is the most fundamental and orchestrated step. A bacteriophage, i.e. phage, is the virus that infects and proliferates within a bacterium. It can also kill host bacterium. In recent years, as more and more bacteria display the multi-drug-resistance, phages can be used as antibacterial agents [1].

Like other viruses, bacteriophage is parasitic to the host cell by injecting viral genetic materials (RNA or DNA) into the bacterial cell [2]. Based on the physiological process in the infected bacteria, there are two types of phages: temperate phage and intemperate phage. The former integrates its DNA (RNA) to the chromosome of host cell to replicate prophages, which is called lysogenic cycle. The later can produce daughter phages by controlling the expression system of bacterium and kill the host to infect other bacteria, which is called lytic cycle [3]. But the temperate phage could turn to lytic cycle induced by physicochemical and biological factors [4].

Phage proteins located in the host cell (PH proteins) play a key role in physiological processes. Thus, it is important to identify whether a phage protein locates in host bacterial cellular or not. In facts, the subcellular location of PH proteins in host cell often correlates with its special function. Specifically, phage proteins located in the host cell membrane (PHM proteins) may be the enzymes of lysis, such as hydrolases and lyases [5], which is pivotal for daughter phage to depart from the host bacterium [6]. And phage proteins located in the host cell cytoplasm (PHC proteins) may be the capsulate proteins [7] or the regulators [8] of the gene expression. Therefore, it is necessary to identify the subcellular location of PH proteins in host bacterial cell.

Previous researchers have successfully developed many computational methods dealing with phages and phage proteins, such as identifying the prophages [9], classifying the viral structural proteins [10], predicting the phage virion proteins [11,12]. The research about PH

* Corresponding author.
** Corresponding author.
*** Corresponding author. Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China.
E-mail addresses: hding@uestc.edu.cn (H. Ding), chenweiimu@gmail.com (W. Chen), hlin@uestc.edu.cn (H. Lin).

https://doi.org/10.1016/j.chemolab.2018.07.006
Received 8 June 2018; Received in revised form 7 July 2018; Accepted 13 July 2018
Available online 17 July 2018
0169-7439/

proteins was first developed by Ding et al. [13], in which they proposed a *g*-gap dipeptide composition descriptor and obtained an encouraging result [13]. Later on, Shatabda et al. proposed a new descriptor based on the structural and evolutionary information [14]. Although high accuracies were obtained, the evaluated results were not objective because of independent structural and evolutionary information. Thus, the correct way to design a powerful predictor is only based on sequence. However, to the best of our knowledge, no such descriptor based on sequence information can reach a wonderful prediction result for identifying PH proteins.

In this paper, we introduced an integrated descriptor based on the sequence composition and the basic property of amino acid to identifying PH proteins and their locations in host cell. The feature selection technique was used to obtain the optimal features. For the convenience of experimental scientists, an online web server called PHPred 2.0 was developed according to the proposed method.

2. Materials and methods

This work comprises four major steps: (i) constructing the benchmark dataset, (ii) formulating protein samples with feature extraction methods, (iii) selecting and obtaining optimal features, (iv) constructing and evaluating the model. The workflow diagram for constructing the prediction model can be found in Fig. 1.

2.1. Benchmark dataset

Ding's dataset [13], which could be obtained from http://lin-group.cn/server/PHPr/data, was used in this work. According to the description in Ding et al.'s work [13], the phage proteins in the benchmark dataset was extracted from the UniProt [15] database according to the following steps:

Firstly, only phage proteins whose subcellular locations are experimentally confirmed were selected. Secondly, only phage proteins which are not the fragments of other proteins were selected. Thirdly, only phage proteins whose sequences do not contain nonstandard letters ('B′, 'U′, 'X′ or 'Z′) were selected. Finally, phage proteins with sequence identity greater than 0.3 were removed by using the software CD-HIT [16]. After performing these rules, they obtained 278 phage proteins, of which 144 were located in host cell, 134 were not located in host cell. Based on these proteins, a benchmark dataset $\mathbb{S}$ is formulated as:

$$\mathbb{S} = \mathbb{S}_{PH} \cup \mathbb{S}_{non-PH} \tag{1}$$

where $\mathbb{S}_{PH}$ contains 144 proteins located in host cell (PH proteins), $\mathbb{S}_{non-PH}$ contains 134 proteins that do not locate in host cell (non-PH proteins). The PH proteins can be further classified into two classes, i.e. the phage proteins that located in membrane of host cell (PHM proteins) and the phage proteins that located in host cell cytoplasm (PHC proteins), which can be described as:

$$\mathbb{S}_{PH} = \mathbb{S}_{PHM} \cup \mathbb{S}_{PHC} \tag{2}$$

where $\mathbb{S}_{PHM}$ contains the 68 PHM proteins and $\mathbb{S}_{PHC}$ contains the 76 PHC proteins, respectively.

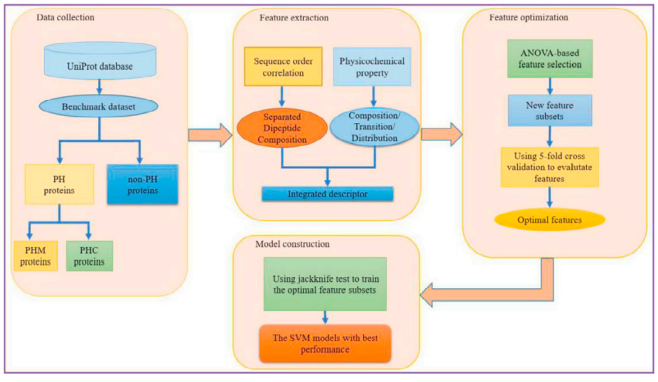

Fig. 1. The workflow of this work.

2.2. Feature vector construction

After constructing the objective and strict benchmark dataset, we should formulate each protein sample with a mathematical descriptor. However, the lengths of proteins are different. Therefore, it's necessary to convert them to vectors that can be handled by the existing machine-learning algorithms. In fact, many efficient descriptors have been proposed and applied for this aim, such as the amino acid composition (AAC) [17] and dipeptide composition (DC) [18]. To consider both sequence order correlation and amino acid composition of the protein, Chou proposed a pseudo amino acid composition (PseAAC) [19] to formulate proteins. Here, we applied three kinds of higher dimensional descriptors described as follows.

(I) *g*-gap dipeptide composition (*g*-gap DC)

Suppose a protein sequence **P** with the length of L, denoted as follows:

$$P = R_1R_2R_3R_4\cdots R_i\cdots R_{L-1}R_L \tag{3}$$

where R_i means the i-th residue of the protein **P**.

In order to contain the long-range correlation information of residues, the interval of *g*-gap residues extended from dipeptide composition [11] was used in work. Then, the protein **P** can be expressed as:

$$P = [f_1^g, f_2^g, \cdots, f_i^g, \cdots, f_{400}^g]^T \tag{4}$$

where f_i^g is the normalized frequency of the *i*-th ($i = 1, 2, \cdots, 400$) *g*-gap dipeptide [13] and is calculated by

$$f_i^g = \frac{n_i^g}{\sum_{k=1}^{400} n_k^g} = \frac{n_i^g}{L - g - 1} \tag{5}$$

where n_i^g means the occurrence number of the *i*-th *g*-gap dipeptide, *L* denotes the length of protein **P**.

(II) Separated dipeptide compositions (SDC)

With the avalanche of protein sequences generated in the post-genomic era, the lengths of protein sequences vary widely. To extract important information from one protein, some researchers have spited a sequence to different fragments [20,21], which could highlight the properties of head or tail or special part of protein. Considering the lengths of proteins in the benchmark dataset are from 32 to 1825 residues, we segmented each protein sequence into two parts: the first 30 residues and the rest part. Then we calculated the *g*-gap dipeptide composition for each part based on Eq. (4) and Eq. (5). Finally, the SDC was obtained by combining the *g*-gap dipeptide composition of the two parts. Thus, the protein P can be expressed as a 800-D vector.

(III) composition/Transition/distribution (CTD)

Although SDCs contain much more sequence order correlation, the physicochemical properties are still lost [22]. The previous researchers have successfully developed some reasonable approaches [19,23] to extract the physicochemical property. Here, we chose the lower dimensional but more integrative physicochemical property descriptor—CTD (Composition/Transition/Distribution) to encode protein sequences.

CTD was first proposed to predict the protein folding class by Dubchak et al. [24] and has also been used to predict other protein cellular attributes. This paper also used the CTD to formulate protein samples. In the CTD feature, C represents the global composition of the given property in a protein sequence, T denotes the frequencies of the property changed along the protein sequence, and D is the distribution pattern for