Feature Reduction Using a Topic Model for the Prediction of Type III Secreted Effectors

Sihui Qi, Yang Yang*, and Anjun Song

Department of Computer Science and Engineering, Information Engineering College,
Shanghai Maritime University, 1550 Haigang Ave., Shanghai 201306, China
yangyang@shmtu.edu.cn

Abstract. The type III secretion system (T3SS) is a specialized protein delivery system that plays a key role in pathogenic bacteria. Until now, the secretion mechanism has not been fully understood yet. Recently, a lot of emphasis has been put on identifying type III secreted effectors (T3SE) in order to uncover the signal and principle that guide the secretion process. However, the amino acid sequences of T3SEs have great sequence diversity through fast evolution and many T3SEs have no homolog in the public databases at all. Therefore, it is notoriously challenging to recognize T3SEs. In this paper, we use amino acid sequence features to predict T3SEs, and conduct feature reduction using a topic model. The experimental results on *Pseudomonas syringae* data set demonstrate that the proposed method can effectively reduce the features and improve the prediction accuracy at the same time.

Keywords: type III secretion system, type III secreted effector, topic model, feature reduction.

1 Introduction

The type III secretion system (T3SS) is one of the six types of secretion systems that have been discovered in gram-negative bacteria. T3SS is an essential component for a large variety of pathogens, such as *Pseudomonas*, *Erwinia*, *Xanthomonas*, *Ralstonia*, *Salmonella*, *Yersinia*, *Shigella*, *Escherichia*, *etc* [1], which can cause devastating diseases on plants, animals and human beings. T3SS plays an important role in developing the diseases by injecting virulence proteins into the host cells.

Researchers have been exploring the working principle and mechanism of T3SS for over a decade. The detailed structure of T3SS has been identified, including a needle-like structure and bases embedded in the inner and outer bacterial membranes [1]. The virulence proteins, called type III secreted effectors (T3SEs), are secreted directly from the bacterial cell into the host cell through the needle. Although the structure of T3SS apparatus has been uncovered, the precise mechanism of the secretion process has not been fully understood. In recent years,

* Corresponding author.

B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part I, LNCS 7062, pp. 155–163, 2011.
© Springer-Verlag Berlin Heidelberg 2011

more and more effort has been put into the studies of T3SEs, because the characteristics determining what kind of proteins could be secreted have not been discovered yet. A lot of questions remain unresolved, such as how the T3SEs are recognized, and how they are transported into host cells. Once we know the answers, we know much better about how the T3SS works.

Although the structure of T3SS is conserved, T3SEs are highly variable even among different strains of the same bacterial species. This is because they evolve fast in order to adapt to different hosts and respond to the resistance from the host immune systems. Therefore, it is notoriously challenging to recognize T3SEs. Some wet-bench methods have been used to verify T3SEs, *e.g.*, functional screen and protein secretion assay [2]. These methods are time and labor consuming, and cannot deal with high-throughput screening, while computational tools can save the laborious work in wet-bench experiments and help biologists find the T3SE candidates more quickly. Therefore, bioinformatics approaches are in great demand for the study of T3SS.

There is very little domain knowledge could be used for identifying T3SEs. Actually, many T3SEs were hypothetical proteins before they were verified. As the sequencing techniques have gained breakthrough for the past decade, a large number of sequenced genomes for plant and animal pathogens became available, thus the genome sequences and amino acid sequences are widely used to discriminate effectors and non-effectors. Researchers have detected amino acid composition biases in T3SEs, especially in the N-termini. For example, Guttman *et al.* [2] reported that the first 50 amino acids of *P. syringae* effectors have a high proportion of Ser and a low proportion of Asp residues. A conserved regulatory motif on promoters was also found in some T3SEs [3]. However, these features are not accurate enough to identify new effectors because some effectors do not possess these features at all.

Recently, some machine learning methods have been proposed for the prediction of T3SEs. Arnold *et al.* [4] used the frequencies of amino acids as well as the frequencies from two reduced alphabets, *i.e.*, they mapped amino acids to groups according to the amino acid properties. They also computed the frequencies of di- and tri-peptides from each of the alphabets. Löwer and Schneider [5] used sliding-window technique to extract features. The sliding window procedure divides a sequence in a number of overlapping segments. Each segment is encoded by a bit string containing $W \times 20$ bits (W is the size of the window). Yang *et al.* used amino acid composition, K-mer composition, as well as SSE-ACC method (amino acid composition in terms of different secondary structures and solvent accessibility states) [6]. Wang *et al.* [7] proposed a position-specific feature extraction. The position-specific occurrence time of each amino acid is recorded, and then the profile is analyzed to compose features.

These methods mainly utilize sequence features, like amino acid composition and position information, but they do not consider the most discriminating residues or peptides. In this paper, we regard the protein sequences as text written in a certain kind of biological language. The residues and peptides, *i.e.*, K-mers (K-tuple amino acid sequences), are the words composing the text.