Computational identification and characterization of Type III secretion substrates

Eric Sakk
Department of Computer Science
Morgan State University
Baltimore, MD 21251
esakk@jewel.morgan.edu

David J. Schneider, Samuel W. Cartinhour
USDA Agricultural Research Service
Ithaca, NY 14853
djs30@cornell.edu
sc167@cornell.edu

Christopher R. Myers
Cornell Theory Center
Cornell University
Ithaca, NY 14853

Monica Vencato, Alan Collmer
Department of Plant Pathology
Cornell University
Ithaca, NY 14853

Abstract

Many bacterial pathogens employ a Type III secretion system (TTSS) to deliver specific proteins (or "substrates") into a host cytoplasm in order to interfere with defense responses and alter physiology. In this work, we present a computational formalism for characterizing the compositional properties of the Type III secretion signal. While various rule sets derived from empirical observations have been suggested, developing a consistent and comprehensive description of the TTSS signal is still of interest. This problem differs from typical signal peptide classification and identification problems (e.g. - nuclear, chloroplast, mitochondrial signal peptides) because known TTSS substrates lack the similarity expected from signal sequences involved in a similar function (e.g. - from a multiple alignment profile or signal consensus sequence). Using a training set derived from empirically verified substrate sequences in Pseudomonas syringae, we apply divergence measures derived from information theory in order to classify similar patterns and characterize the Type III signal. The TTSS characterization developed in this work leads to a diffuse targeting signal confined to the first 50 amino acids starting from the N-terminus. Finally, using the P. syringae training set, the method is applied to verify and predict substrate candidates in other organisms possessing a TTSS.

1 Introduction

The formulation presented here draws from research to identify and classify substrate proteins secreted by the type III secretion system (TTSS) [1, 2]. Research within this group has focused on *Pseudomonas syringae* pv. *tomato* DC3000 [1, 4]. It has been demonstrated that the first 50 amino acids in known substrates are necessary for identifying proteins that are translocated by the TTSS. While compositional properties have been found to be useful for detecting substrate sequences, no obvious consensus sequence or alignment appears to exist. Furthermore, cursory analyses by this group indicate that TTSS substrates do not fit into typical signal peptide models.

2 Sequence Composition Analysis

For this work, we present results based upon a training set of 35 sequences in $Pseudomonas\ syringae$ pv. $tomato$ DC3000 known to be translocated by the TTSS [1] and 5400 protein sequences from the same organism for background statistics. Let W be the window size, let M be the number of sequences and let N be the number of symbols in the encoding alphabet. In our case $N = 20$ in order to represent each amino acid. Consider an $M \times W$ block of symbols starting at sequence position m and ending at sequence position $m + W - 1$. We estimate the block symbol probability as $p_i = \frac{n_i}{MW}$ where n_i is the number of times the i^{th} symbol appears in the block, and make an estimate of the entropy as

$$H_e = -\sum_{i=1}^{N} p_i \log_2 p_i. \quad (1)$$

Figure 1 shows the results of this calculation at positions 2-47 using a sliding window of size $W = 3$. The dotted line represents the entropy estimate for the background set which is fairly constant at about 4.1 bits (approaching a limit of $log_2(20) \approx 4.3$ bits). Also, observe that, at several positions, the substrate set (dashed line) can differ in information content from the background set by more than .5 bits. It is this information difference that we wish to exploit in our classification algorithm.

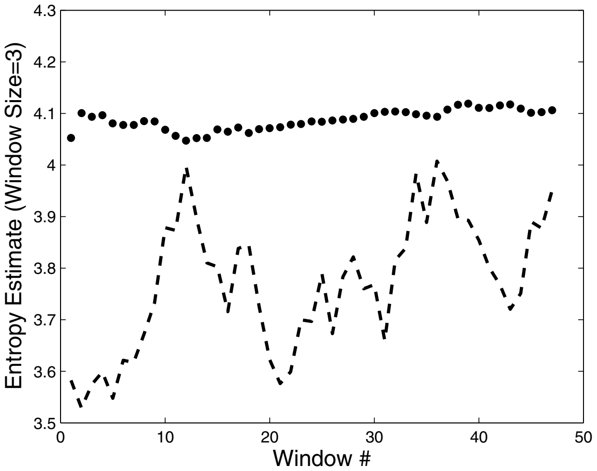

Figure 1: *Entropy estimate using a sliding window size of 3 in positions 2-47. Dotted line represents the background set, dashed line represents the substrate training set.*

3 Information Divergence

For this work, we apply a symmetric version of the Kullback-Liebler distance [3]. Given two discrete probability distributions P and Q with N elements, the symmetric Kullback-Liebler distance is defined as

$$D_s(P\|Q) \equiv D(P\|Q) + D(Q\|P) \qquad (2)$$

where

$$D(P\|Q) = \sum_{i=1}^{N} p_i \log_2 \frac{p_i}{q_i}$$

is generally referred to as the Kullback-Liebler distance or the relative entropy. To characterize an unknown protein sequence as being close or far from the substrate distribution, D_s is evaluated over a series of sliding windows. For a given observation and window size W, we estimate the symbol probability as $\frac{n_i}{W}$ where n_i is the number of times the i^{th} amino acid appears in the window. A hard decision about the membership of a sequence is then made at each position according to the following algorithm. We consider the first $L = 50$ amino acids of an unknown sequence. For a window size of W, there will be $K = L - W + 1$ positions to consider. At the l^{th} position:

1. Construct Q by measuring $q_i = \frac{n_i}{W}$ ($i = 1, \cdots, 20$) for the unknown sequence.
2. Calculate $D_s(P^k|Q)$ for $k = 1, 2$ where P^1 is the background distribution and P^2 is the substrate distribution.
3. Perform a hard decision according to the rule: Choose category 2 if $D_s(P^1|Q) > D_s(P^2|Q)$; otherwise, choose category 1.

Finally, to decide upon the membership of a given sequence, examine the decision for all $l = 1, \cdots, K$ instances and choose the majority. In other words, over K instances there will be k_1 instances in favor of the background and k_2 instances in favor of the substrate distribution. A score S is created by taking the difference $S = k_1 - k_2$. For the purposes of robustness, we run our algorithm three times with window sizes $W = 1, 2, 3$. For each sequence being tested, we take the minimum score for each of the three tests.

4 Results

Using the above algorithm, results of data mining genomes from organisms known to possess a TTSS are presented. In particular, we examine Pseudomonas syringae, Yersinia enterocolitica and Salmonella typhimurium. In addition to identifying known substrates within these organisms, candidate substrates are also indicated. Finally, experimental results indicating that this technique has isolated a Type III secretion signal are presented.

5 Acknowledgements

Lisa Schecter and Magdalen Lindeberg are to be thanked for helpful discussions and providing the substrate data sets.

References

[1] A. Collmer, M. Lindeburg, T. Petnicki-Ocwieja, D. Schneider, and J. Alfano. Genomic mining type III secretion system effectors in Pseudomonas syringae yields new picks for all TTSS prospectors. *Trends in Microbiology*, 10(10):462–469, 2002.

[2] G. Cornelis and F. Van Gijsegem. Assembly and function of Type III secretory systems. *Annual Review of Microbiology*, 54:735–774, 2000.

[3] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley, New York, NY, 1991.

[4] T. Petnicki-Ocwieja, D. Schneider, V.C.Tam, S. Chancey, L. Shan, Y. Jamir, L. Schechter, M. Janes, C. Buell, X. Tang, A. Collmer, and J. Alfano. Genomewide identification of protein ssecreted by the Hrp type III protein secretion system of Pseudomonas syringae pv. tomato DC3000. *PNAS*, 99(11):7652–7657, 2002.