Vol.2011-BIO-25 No.37
2011/6/24

Feature extraction for discriminance of symbiotic/parasitic bacterial type III effector protein using principal component analysis

Yuichi Nakano[†1] and Y-h. Taguchi[†1]

Type three secretion system (T3SS) is a protein machinery found in several-bacteria. Using this machinery, bacteria infect the eukaryotic cell. The problem is that similar machinery can be found also in symbiotic bacteria. Thus, it is important if there are any difference between pathogenic and symbaiotic T3SS. In this paper, we have investigated this difference using simple method, principal component analysis together with linear discriminant analysis. These two types proteins have clear difference between them.

1. Introduction

Type three secretion system (T3SS)[1] is common machinery for bacteria to invade eucaryotic cells. During process, several kinds of effector proteins are injected into cells. These usually affect physiology slightly. Since there are two kinds of bacteria, pathogenic and symbiotic bacteria. Although former causes deceases in host cells, the later does not cause anything critical to host cells. Thus, it is important if there are fundamental difference of effector proteins between pathogenic and symbiotic bacteria. If there are no differences, most of poisonnic processes are caused by other proteins than effector proteins. However, if there are differences, there are some possibilities that pathogenic effector protein itself can cause deceases.

Recently, Yahara et al[2] has investigate this difference with machine learning method and found seven most-discriminating features. In this paper, we have purposed same goal with simpler methods, i.e., principal component analysis and linear discriminant analysis. The results that we obtained are similar to theirs.

Table 1 Accession numbers of proteins used in this study.

Symbiotic
AAG45728.1 AAG45731.2 AAG60780.1 AAG60840.1 NP_444148.1
NP_444155.1 NP_768429.1 NP_768450.1 NP_768544.1 NP_768698.1
AAG45730.2 AAG60738.1 AAG60781.1 NP_443964.1 NP_768427.1
NP_768428.1 YP_025399.1 YP_052973.1 YP_162947.1 YP_293633.1
AAG45729.1 P55724.1 P55730.1 Q0FF52 Q2LDQ5
Q2RKJ5 Q2RPK8 Q84H14 Q139Z8 Q221V6
Pathogenic
A6M3R1 A9K514 A9R9K1 B4TH61 B0HZP7
Q1MQX3 Q1MQX5 Q254G9 Q252Q1 Q663I2
A6M3R2 A6M3T7 A9K4S2 A9R9K6 B0HZP0
B0HZP3 B0HZP4 B0HZP6 B0HZP9 B0HZQ3
AAZ37972.1 AAZ38042.1 BAD20871.1 CAC05870.1 CAF25383.1
CAF25398.1 CAF25400.1 CAF25401.1 CAL13559.1 NP_224395.1

2. Materials and Methods

2.1 Protein sequences

The protein sequences for pathogenic and symbiotic bacteria are taken from two references[3],[4]. Table 1 shows the list of 30 symbiotic and 30 pathogenic proteins used in this study.

2.2 Principal component analysis and linear discriminant analysis

First, we have computed 44 features employed by Yahara et al[2] by EMBOSS[5] and SignalP[6]. Then, we have scaled all feature such that they have zero mean and unity variance. After that, we have applied principal component analysis (PCA). Linear discriminant analysis (LDA) with leave-one-out cross-validation is performed with changing number of used principal components. Optimal number of components is decided. Further analysis is done based upon this number of principal components.

3. Results

Dependent upon number of proteins used, we have achieved accuracy from 0.85 to 0.90 (See Table 2). It is clearly comparative to or even better than Yahara et al's results (see Fig.2,[2], sensitivity and specificity are about 0.85) where they have used more complicated machine learning method. Thus, we

[†1] Department of Physics, Chuo University, Tokyo, Japan

Table 2 Performance to discriminate symbiotic/pathogenic T3SS effector proteins. # of PCs means number of principal components used in LDA.

	Predcited					
True	S	P	S	P	S	P
symbiotic(S)	9	1	20	0	27	3
pathogenic(P)	1	9	1	19	2	28
# of PCs	2		5		8	

Table 3 List of significant features to discriminate symbiotic/pathogenic proteins. P(S) stands for pathogenic(symbiotic). Amino acid rich in P(S) is underlined(in bold face). The last column indicates seven siginificant features[2].

Features	rich in	
Number of potentially antigenic regions of a protein sequence	S	
Number of proteolytic enzyme or reagent cleavege sites	S	
Number of alpha helix	S	
Number of beta sheet	S	
Hydrophobic moment	S	
Average residue weight	S	S
Ala	P	P
Asp	P	P
Gly	P	
Ile	S	S
Lys	S	
Leu	S	
Asn	S	
Pro	P	
Arg	P	
Mole percentage of tiny amino acids (Ala,Cys,Gly,Ser,Thr)	P	P
Mole percentage of small amino acids (tiny amino acids ,Asp,**Asn**,Thr,Val)	P	P
Mole percentage of aliphatic amino acids (Ala,**Ile**,**Leu**,Val)	S	
Mole percentage of aromatic amino acids (Phe,His,Trp,Tyr)	P	
Mole percentage of charged amino acids (Asp,**Leu**,His,**Leu**,Arg)	P	
Mole percentage of acidic amino acids (Asp,Glu)	P	P
Signal peptide probability	S	

conclude that our simpler method can have ability to investigate the difference of effector proteins between pathogenic and symbiotic bacteria. Examples of features which discriminate two categories independent of number of proteins used are listed in Table 3. These are consistent to those by Yahara et al. There are some discrepancies between features. For example, Asn is rich in symbiotic protein, but included into small amino acid which is rich in pathogenic proteins. However, generally, pathogenic proteins consist of smaller amino acids. Symbiotic proteins consist of larger amino acid, thus have more complex structure. Thus, the later have more secondary structures. Biological reason of this should be understood in the feature. Since our method is much simpler than theirs, ours are more suitable than Yahara et al's method.

4. Conclusion

In this paper, we have investigated difference of T3SS effector proteins between pathogenic and symbiotic proteins. Consistent features to discriminate these two from each other includes Yahara et al's seven significant features[2].

5. Acknowledgement

This work was supported by KAKENHI 23300357.

References

1) Dean, P.: Functional domains and motifs of bacterial type III effector proteins and their roles in infection. *FEMS Microbiol Rev.* doi: 10.1111/j.1574-6976.2011.00271.x. (2011)

2) Yahara, K., Jiang, Y., and Yanagawa, T,: Computational identification of discriminating features of pathogenic and symbiotic type III secreted effector proteins, *IPSJ SIG Technical Report*, 2010-BIO-21, No.16, pp.1-8, (2010).

3) Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., Niinikoski, A., Mewes, H. W., Horn, M. and Rattei, T.:Sequence-based prediction of type III secreted proteins, *PLoS Pathog.* vol.5, e1000376 (2009).

4) Lower, M. and Schneider, G.: Prediction of Type III Secretion Signals in Genomes of Gram-Negative Bacteria, *PLoS One*, vo.4, e5917 (2009).

5) Rice, P., Longden, I. and Bleasby, A.: EMBOSS: the European Molecular Biology Open Software Suite, *Trends Genet.*, vol.16, pp.276-7, (2000)

6) Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H.: Locating proteins in the cell using TargetP, SignalP, and related tools, *Nature Protocols*, vol.2, pp.953–971 (2007).