Discrimination of symbiotic/parasitic bacterial type III secretion system effector protein using principal component analysis

Yuuichi Nakano[1] Mitsuo Iwadate[2,a)] Hideaki Umeyama[2,b)] Y-h. Taguchi[1,c)]

Abstract: Type III secretion system (T3SS) effector protein is a part of bacterial secretion systems. T3SS exists in the pathogenic and symbiotic bacteria. How the T3SS effector proteins in these two classes differ from each other should be interesting. In this paper, we proposed the usage of principal component analysis based linear discriminant analysis that discriminates T3SS effector proteins between plant pathogenic, animal pathogenic and plant symbiotic bacteria by the accuracy of 0.77. We also hypothesized that the feature vector proposed by Yahara et al represents protein structure, possibly protein folds defined in Structural Classification of Proteins (SCOP) database.

Keywords: Type III secretion system, principal component analysis, linear discriminant analysis

1. Introduction

Type III secretion system (T3SS) is employed by a number of Gram-negative bacterial pathogens to inject toxins into eukaryotic cells [1]. Termination of proper T3SS functionality by understanding T3SS functionality may allow us to suppress infection of these bacteria. Thus, understanding the characteristic feature of T3SS effector protein is important. Actually, numerous researches tried to discriminate T3SS effector proteins from other proteins [2], [3], [4], [5], [6].

In contrast to these efforts, trials that discriminate symbiotic T3SS effector proteins from pathogenic T3SS effector proteins were seldom. Since symbiotic bacteria never tried to inject toxins into eukayotic cells, the distinction between symbiotic and pathogenic proteins is important. Yahara et al [7] recently proposed numerical methods that discriminate symbiotic T3SS effector proteins from pathogenic T3SS effector proteins. After selecting features using generalized Bayesian information criterion (GBIC) of kernel logistic regression (KLR), they discriminate between symbiotic and pathogenic proteins by support vector machine (SVM). Also, the seven features were reported to discriminate symbiotic and pathogenic T3SS effector proteins.

In this paper, we proposed the alternative and simpler method that discriminate symbiotic T3SS effector proteins from pathogenic T3SS effector proteins. We also proposed the method that estimate optimal number of principal component for the discrimination without evaluating the performance of cross validation.

[1] Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112–8551, Japan
[2] Department of Biological Science, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112–8551, Japan
a) iwadate@bio.chuo-u.ac.jp
b) umeyama@bio.chuo-u.ac.jp
c) tag@granular.com

Table 1 Proteins replaced because of uniprot modification.

Yahara et al	this study
O84118	B0B9M4
O84119	B0B9M5
Q56027	E1WAC6
Q8X2D5	P0DJ88
Q9RPH0	D0ZPH9
Q9RPQ1	B0B9M3

In addition to this, since Yahara et al's original dataset turned out to include errors, we introduced an alternative dataset, by which animal pathogenic, plant pathogenic and plant symbiotic bacteria were shown to be discriminated well with each other. Yahara's feature vector was also suggested to represent protein folds defined in Structural Classification of Proteins (SCOP) database.

2. Materials and Methods

2.1 T3SS effector proteins

T3SS effector protein sequences were obtained from uniprot[*1], based on protein IDs listed in Tabels S1 and S2 provided by Yahara et al [7]. Protein sequences obtained were formatted as fasta format. Since two to three years passed since Yahara et al's work, some modifications took place in uniprot. Then, the six proteins listed in Table 1 were replaced with alternative ones.

2.2 Feature extraction

EMBOSS[*2] and SignalIP[*3]provide features used for the discrimination, although "Number of cleavage sites between signal sequence and mature exported protein" was not used.

43 features used for discrimination were listed in Table 2.

[*1] http://www.uniprot.org/
[*2] http://emboss.sourceforge.net/
[*3] http://www.cbs.dtu.dk/services/SignalP/

Table 2 43 features employed for protein discrimination

No.	Features	Softwares	Programs
1	The number of antigenic sites in proteins	EMBOSS	antigenic
2	The number of fragment generated by Trypsin digestion	EMBOSS	digest
3	The number of sites that belong to predicted helix	EMBOSS	garnier
4	The number of sites that belong to predicted sheet	EMBOSS	garnier
5	The number of sites that belong to predicted turn	EMBOSS	garnier
6	The number of sites that belong to predicted coil	EMBOSS	garnier
7	Hydrophobic moment	EMBOSS	hmoment
8	Average Residue Weight	EMBOSS	pepstats
9	Charge	EMBOSS	pepstats
10	Isoelectric Point	EMBOSS	pepstats
11	A280 Molar Extinction Coefficient	EMBOSS	pepstats
12	A280 Extinction Coefficient	EMBOSS	pepstats
13	Improbability of expression in inclusion bodies	EMBOSS	pepstats
14	Amino Acid Ratio:Ala(Alanine)	EMBOSS	pepstats
15	Amino Acid Ratio:Cys(Cysteine)	EMBOSS	pepstats
16	Amino Acid Ratio:Asp(Aspartic acid)	EMBOSS	pepstats
17	Amino Acid Ratio:Glu(Glutamic acid)	EMBOSS	pepstats
18	Amino Acid Ratio:Phe(Phenylalanine)	EMBOSS	pepstats
19	Amino Acid Ratio:Gly(Glycine)	EMBOSS	pepstats
20	Amino Acid Ratio:His(Histidine)	EMBOSS	pepstats
21	Amino Acid Ratio:Ile(Isoleucine)	EMBOSS	pepstats
22	Amino Acid Ratio:Lys(Lysine)	EMBOSS	pepstats
23	Amino Acid Ratio:Leu(Leucine)	EMBOSS	pepstats
24	Amino Acid Ratio:Met(Methionine)	EMBOSS	pepstats
25	Amino Acid Ratio:Asn(Asparagine)	EMBOSS	pepstats
26	Amino Acid Ratio:Pro(Proline)	EMBOSS	pepstats
27	Amino Acid Ratio:Gln(Glutamine)	EMBOSS	pepstats
28	Amino Acid Ratio:Arg(Arginine)	EMBOSS	pepstats
29	Amino Acid Ratio:Ser(Serine)	EMBOSS	pepstats
30	Amino Acid Ratio:Thr(Threonine)	EMBOSS	pepstats
31	Amino Acid Ratio:Val(Valine)	EMBOSS	pepstats
32	Amino Acid Ratio:Trp(Tryptophan)	EMBOSS	pepstats
33	Amino Acid Ratio:Tyr(Tyrosine)	EMBOSS	pepstats
34	Amino Acid Ratio:Tiny(A+C+G+S+T)	EMBOSS	pepstats
35	Amino Acid Ratio:Small(A+B+C+D+G+N+P+S+T+V)	EMBOSS	pepstats
36	Amino Acid Ratio:Aliphatic(A+I+L+V)	EMBOSS	pepstats
37	Amino Acid Ratio:Aromatic(F+H+W+Y)	EMBOSS	pepstats
38	Amino Acid Ratio:Non-polar(A+C+F+G+I+L+M+P+V+W+Y)	EMBOSS	pepstats
39	Amino Acid Ratio:Polar(D+E+H+K+N+Q+R+S+T+Z)	EMBOSS	pepstats
40	Amino Acid Ratio:Charged(B+D+E+H+K+R+Z)	EMBOSS	pepstats
41	Amino Acid Ratio:Basic(H+K+R)	EMBOSS	pepstats
42	Amino Acid Ratio:Acidic(B+D+E+Z)	EMBOSS	pepstats
43	Ratio of signal peptide cleavage sites	SignalP	

2.3 Linear discriminant analysis based upon principal component analysis

Principal component analysis (PCA) was applied to the normalized (mean zero, variance one) feature vectors. Then, symbiotic and pathogenic T3SS effector proteins are discriminated using linear discriminant analysis (LDA) with optimal number of principal components (PCs) estimated based on leave one out cross validation. Semi-supervised[8] discriminations was also employed; PCA was applied to all of samples without classification information of test samples, while LDA was applied to samples excluding samples in test set as usual.

2.4 Estimation of useful features for discrimination

Suppose that x_i is normalized ith feature. Then jth PC, PC_j, is

$$PC_j = \sum_{i=1}^{43} a_i^j x_i$$

Then discriminant function

$$z = \sum_j c_j PC_j$$

can be expressed as

$$z = \sum_j c_j \sum_{i=1}^{43} a_i^j x_i = \sum_{i=1}^{43} \left[\sum_j c_j a_i^j \right] x_i$$

Here

$$C_i \equiv \sum_j c_j a_i^j$$

is the amount of contribution of the ith feature to discriminant function.

2.5 Estimation of optimal number of PCs without evaluation of cross validation

LDA was performed in two ways; with and without cross validation. For the cross validation, we exclude one symbiotic T3SS effector protein or one pathogenic T3SS effector protein from the training set. Then, excluded the proteins was classified based on LDA trained. Since there are N symbiotic and pathogenic proteins, in total $2N$ trials. The prediction for excluded proteins with cross validation was not compared to the true classification, but to the prediction without cross validation. Suppose the prediction with cross validation of ith symbiotic or pathogenic protein when ith symbiotic protein or pathogenic protein are excluded is A_i, where A_i takes 1(0) when ith protein is predicted to be

Table 3 Performance of semi-supervised PCA based LDA.

		True	
		Symbiotic	Pathogenic
Predicted	Symbiotic	49	8
	Pathogenic	8	49

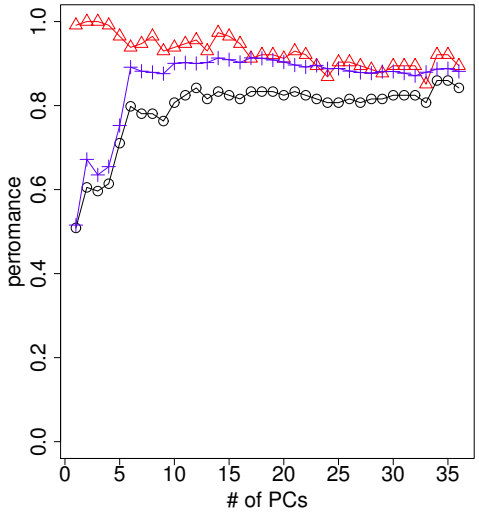

Fig. 1 PC number dependency of Accuracy (black circle), AUC (blue cross) and over fitting measure A (red triangle) for semi-supervised PCA based LDA

symbiotic(pathogenic). A'_i is the prediction of ith symbiotic or pathogenic protein without cross validation. Then, averaged accuracy A is defined as

$$A \equiv 1 - \frac{\sum_{i=1}^{2N} | A_i - A'_i |}{2N}$$

Then denote A when PCs up to kth PC are used for discrimination as $A(k)$. The standard errors up to $A(k)$ is defined as

$$\delta A(k)^2 \equiv \frac{\langle (A(k) - \langle A(k) \rangle_k)^2 \rangle_k}{k}$$

where $\langle Q_k \rangle_k \equiv \sum_{k'=1}^{k} Q_{k'}/k$. Optimal k is decided such that $\delta A(k)$ takes minimum.

3. Results and Discussions

3.1 Original data set provided by Yahara et al

3.1.1 Performance

In Table 3, we have shown the performance of semi-supervised PCA based LDA. Accuracy is 0.86 for the optimal number of PCs of 34. This value of accuracy is as good as Yahara et al[7]'s fine tuned non-linear methods, while ours was a simple and classical linear method.

In Fig. 1, PC number dependency of the Area Under the Curve (AUC) value of the Receiver Operating Statistic Curve (ROC) was shown. AUC for optimal number of PC (=17) was 0.91. This means that the performance was good enough compared to the simpleness of the method.

3.1.2 Useful features for discrimination

Fig. 2 shows C_i (see Materials and Methods) for optimal number of PCs (=17) for AUC. Upward (downward) features correspond to pathogenic (symbiotic) expressive features.

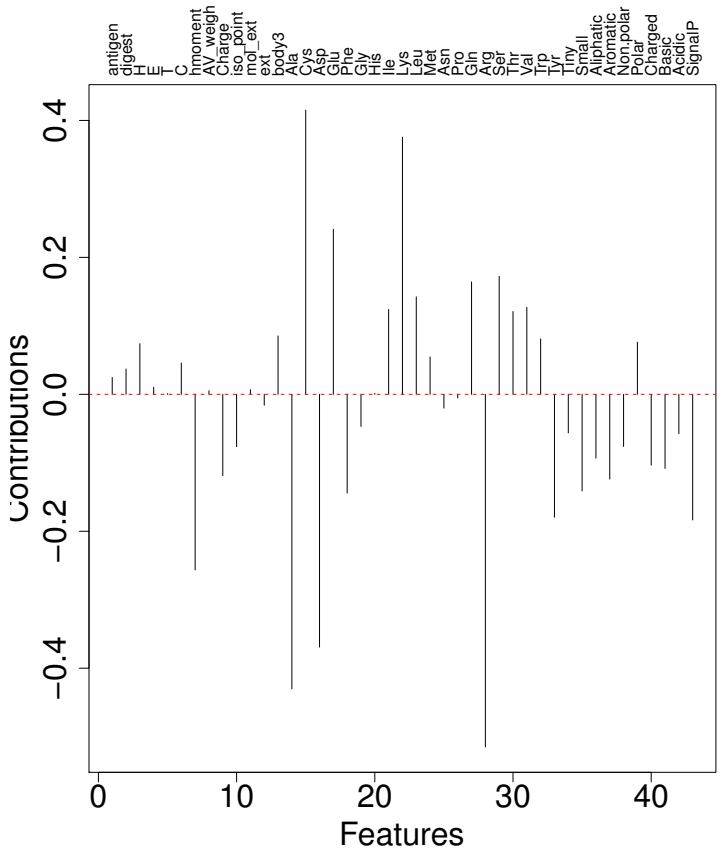

Fig. 2 The amount of contribution of ith feature to discrimination function, C_i. Positive (negative) values contribute to pathogenesis (symbiosis). The number (=17) of PCs used was the optimal number for AUC.

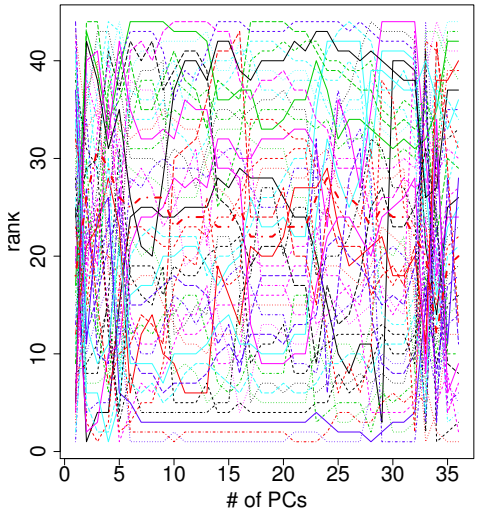

Fig. 3 The rank of amount of contribution of ith feature to discrimination function, C_i as a function of PCs. Bold-red-broken-doted curve indicates ranks of zero. Features ranked above (below) the rank of zero contribute to pathogenesis (symbiosis)

Apparently, although there are several features specific to either symbiosis or pathogenesis, it is not a case. Although Fig. 3 shows the rank of each feature's contribution as a function of the number of PCs; upper (lower) features are pathogenic (symbiotic), clearly it heavily fluctuates. Since AUC and accuracy does not depend upon number of PCs if the number of PCs exceeds 10

(Fig. 1), the fluctuation indicates that there are numerous combination of features in order to achieve good performance. Only exceptions are highly symbiosis specific features (i.e., at the bottom lines in Fig. 3). Yahara et al also reported that there were limited number of symbiosis-specific-features while very little number of pathogenesis-specific-features existed. These symbiotic-specific features in our study are Arginine, Alanine, Aspartic acid and hydrophobic moment in this order. Yahara et al also reported that both Alanine and Aspartic acid are highly symbiosis-specific, but reported neither Arginie nor hydrophobic moment. Fig. 4 shows the boxplot of the ranks of each features contribution when the number of PCs is between 10 and 30 where both AUC and accuracy constantly take higher values (Fig. 1). Only limited number of features that represent each amino acid ratio are definitely symbiosis or pathogenesis specific; e.g., Ala, Asp, and Arg for symbiosis and Lys for pathogenesis. Although hydrophobic moment and polar are symbiosis and pathogenesis specific respectively, their specificity is relatively weak compared with that of individual amino acid ratios. This means, in spite of the good performance of discriminations, it is not easy to understand what discriminates between pathogenesis and symbiosis, unfortunately.

The fluctuation seen in Figs. 3 and 4 is very distinct from Yahara et al's results that specified well defined and stable seven features. Possibly, although there could be more combinations of features that can discriminate pathogenic and symbiotic proteins as good as seven selected proteins, they were overlooked because of their employment of GBIC, which resulted in apparent stability of feature selections.

It is also interesting that amino acid ratios that were distinct between T3SS effector protein and other proteins, i.e., Leu, Glu, Asp and Ala [6] were also distinct between pathogenic and symbiotic T3SS effector proteins (Fig. 4). These were depleted proteins in T3SS effector proteins. This may suggest that these proteins whose molecular percentages were distinct between symbiotic and pathogenic play critical roles in T3SS systems.

3.1.3 Optimal number of PC without cross validation evaluation

In PCA-based LDA, the parameter to be optimized was the number of PCs used for LDA. Usually, the number of PCs used for LDA was optimized such that cross validation performance is maximized. However, this criterion always gives us "optimal" number of PCs, even if it is not an optimal number, because there is always the maximum performance of cross validation and the number of PCs attributed to the maximum performance. It does not guarantee us that it is truly optimal. If not, the number of PCs used is highly sample-dependent, thus it may not be an optimal number for other independent samples. In order to clarify this point, we checked how the amount of over fitting grows when the number of PCs increases (see Material and methods).

Fig. 5 shows $\delta A(k)$ as a function of k, which is the measure of over fitting. In contrast to the expectation, $\delta A(k)$ does not have any minimum values but remains almost constant for $k > 15$. This suggests that, even if we added the features, the quality of discrimination did not decrease. Thus, the optimal numbers of PCs, i.e., any numbers more than 10, is trustable and will be valid for the independent samples. Unfortunately, not many symbiotic T3SS effector proteins were found, we could not test our method for independent samples. However, in the future when more symbiotic T3SS effector proteins are found, we can test this conjecture; optimal number of PCs that discriminate symbiotic T3SS effector proteins from pathogenic T3SS effector proteins is robust and take any values between 10 and 30.

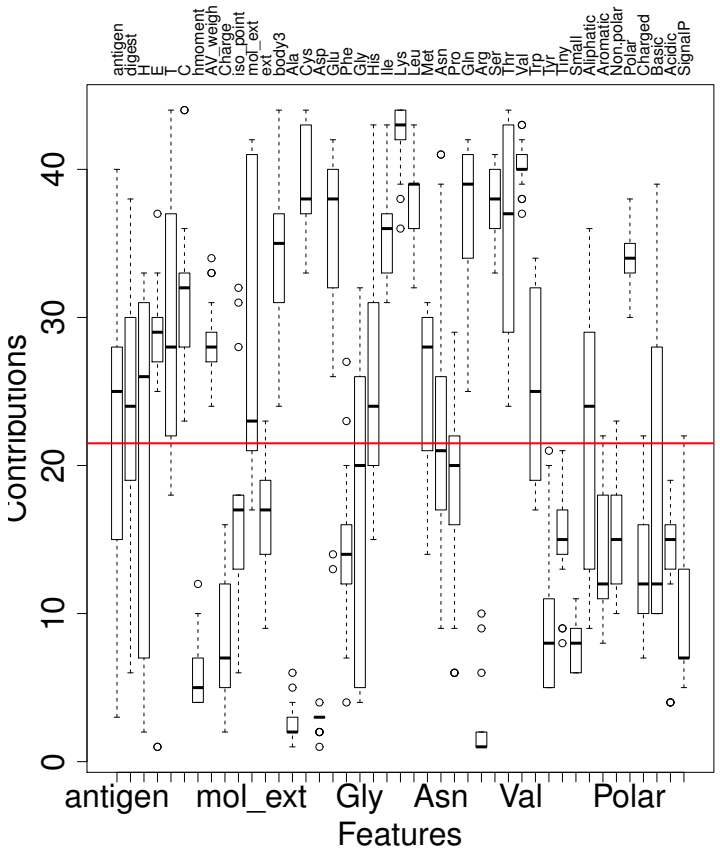

Fig. 4 Boxplot of the rank of amount of contribution of ith feature to discrimination function, C_i when the number of PCs is between 10 and 30. Bold red horizontal line indicates mean rank. Features ranked upper (lower) than this line contribute to pathogenesis (symbiosis)

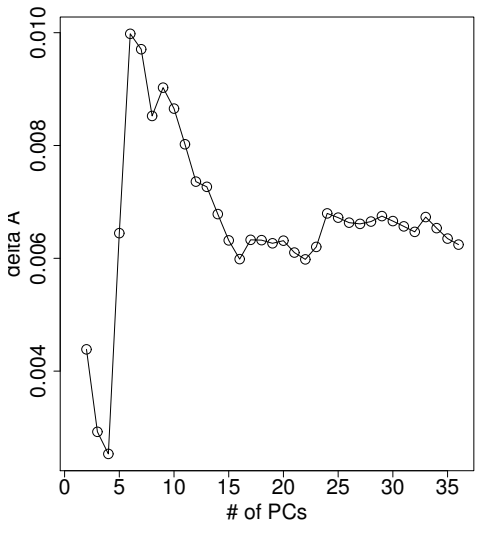

Fig. 5 $\delta A(k)$ as a function of k, the number of PCS used for discrimination.

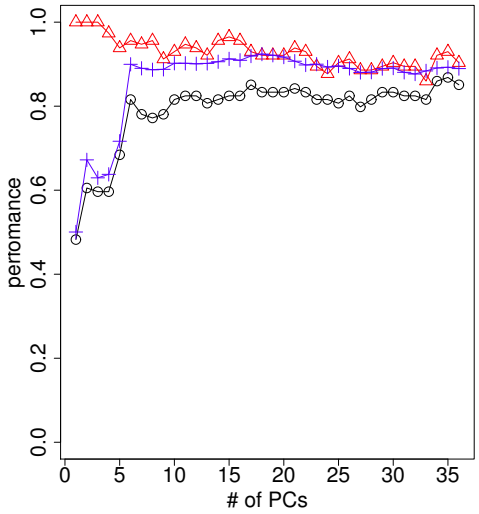

Fig. 6 PC number dependency of Accuracy (black circle), AUC (blue cross) and over fitting measure *A* (red triangle) for not-semi-supervised PCA based LDA.

3.1.4 Not semi-supervised discrimination

The reason why we employed semi-supervised discrimination was because semi-supervised discrimination has more tendencies that discriminate test samples including outliers, since discriminant function can be constructed so as to deal with outliers well later when trained function was applied. However, as can be seen in the previous sections, the present samples and methods have less possibility to be overfitted. In this case, semi-supervised method may not have to be required. Employment of unnecessary semi-supervised method might twist the results. In order to check this point, we have repeated the same procedure without semi-supervised learing; i.e., features of test set were not included into training set even without classification information.

Fig. 6 shows the results. They are quite similar to Fig. 1. Thus, the employment of semi-supervised discrimination did not twist the outcomes. Thus, it is better to continue to use semi-supervised learning in order to be prepared for the cases where we face the treatment of outliers someday.

3.2 An alternative data set

3.2.1 Data preparation

Due to the results presented in the above, the methodological details of discrimination do not seem to matter, but feature vector selected by Yahara et al seems to be critical. Hereafter we call this as Yahara Feature Vector (YFV) even if a few features lack or are replaced with others. Thus, the point is, what YFV represents. During the process of this investigation, Yahara's data set turned out to be erroneous. More than half of proteins identified as symbiotic in their analysis were taken from *Pseudomonas syringae*, which is the famous plant pathogen. So we newly identified these proteins taken from *P. syringae* as plant pathogenic proteins, while the former pathogenic proteins are identified as animal pathogenic proteins since they were mostly taken from animal pathogen genus, e.g. *Yersinia, Salmonella, Chlamydia* and from *E. Coli* O157. Since some almost identical proteins had distinct uniprot ID, only one of them remains within the newly compiled animal pathogenic protein list, in order to reduce redundancy. Symbiotic proteins were represented by an newly added list of plant symbiotic proteins taken from *Sinorhizobium fredii* and *Bradyrhizobium japonicum* [9], which are famous plant symbiotic bacteria. Table 4 shows the full list of proteins. Then, now the problem is the three classes discrimination between animal pathogenic, plant pathogenic and plant symbiotic proteins.

3.2.2 Performance

Then we have applied PCA-based LDA to discriminate three classes. Table 5 shows the best performance obtained by PCA-based LDA with optimal number (29) of PCs. Accuracy is 0.77,

Table 5 Performance of semi-supervised PCA based LDA.

			True		
			Plant		animal
			Symbiotic	Pathogenic	
Predicted	Plant	Symbiotic	46	6	0
		Pathogenic	11	28	10
	Animal		2	3	36

which can be regarded to be well as a three classes discrimination. The almost half of predicted plant pathogenic proteins clearly consist of misclassified plant symbiotic or animal pathogenic proteins. It is natural since plant pathogen shares the host with plant symbiotic while plant pathogen shares pathogenesis with animal pathogen. When we recompute accuracy by merging plant pathogenic and symbiotic proteins into one group, accuracy as a two classes discrimination increases up to 0.89, which is as large as accuracy obtained for the discrimination of Yahara's original data set. This is possibly the reason why Yahara et al apparently successfully discriminated between symbiotic and pathogenic proteins in spite that their symbiotic proteins include many pathogenic proteins. They did not discriminate between symbiotic and pathogenic proteins, but discriminated between proteins that belong to bacteria with animal hosts and proteins that belong to bacteria with plant hosts. This warns us that incorrect labelling of data set can derive wrong conclusion.

Fig. 7 shows the two dimensional discrimination by PCA-based LDA. Plant pathogenic proteins were clearly located between plant symbiotic and animal pathogenic proteins. It is reasonable as mentioned in the above.

3.2.3 What does YFV represent?

Although Yahara et al's results included some mistakes, the proposal of YFV was definitely excellent. Why does it work so well? One of the reasons is possibly because YFV does not include anything specific to symbiosis or pathogenesis. This possibly enabled them to discriminate animal and plant bacteria, successfully but unintentionally. Since YFV works very well, this should represent implicitly something critical to protein properties.

Here we hypothesizes that YFV represents something related to protein structure. Previously, one of authors demonstrated that only amino acid composition can discriminate protein fold well[10]. Since most part of YFV consists of amino acid ratio, it is not surprising if YFV can discriminate protein fold well. And if each class has more class specific protein folds, it is not strange even if YFV can discrminate between animal pathogenic, plant

Table 4 Full list of T3SS proteins in the alternative data set consists of 59 plant symbiotic proteins, 37 plant pathogenic proteins and 46 animal pathogenic proteins. Folds in SCOP indicated beside UNIPROT IDs were estimated by 3D-BLAST.

Plant Symbiotic (59 proteins): Q89E69_BRAJA, Q89E68_BRAJA, Q89E67_BRAJA, Q89E66_BRAJA, Q89TH9_BRAJA, Y4ME_RHISN (d.144.1.7), Y4YB_RHISN (d.268.1.2), H7C6G4_BRAJA (c.10.2.6), Q89Y78_BRAJA (c.56.2.1), Q89F81_BRAJA (c.69.1.1), Q89F80_BRAJA (f.38.1.1), PANB1_BRAJA (c.1.12.8), H7C7V0_BRAJA, Q89GY6_BRAJA (c.94.1.1), H7C7T1_BRAJA, Q89TK8_BRAJA, Q89TP0_BRAJA, H7C6G2_BRAJA, H7C6I2_BRAJA (d.118.8.2), Q89TN8_BRAJA, Q89TQ4_BRAJA, H7C6P6_BRAJA, Q79UN8_BRAJA, Q89EF4_BRAJA (d.128.1.4), Q89GY8_BRAJA (b.70.1.1), H7C827_BRAJA (b.80.1.5), H7C6S1_BRAJA, Q89TN2_BRAJA, Q89TW7_BRAJA (b.29.1.11), H7C813_BRAJA, Y4YJ_RHISN, H7C6M8_BRAJA (b.26.1.2), Y4YQ_RHISN (b.26.1.2), C4PL66_BRAEL (d.14.1.5), NOPX_RHISN, H7C823_BRAJA (a.114.1.1), NOLV_RHISN, Q9ANH4_BRAJP, NOLU_RHISN, NOLT_RHISN (d.226.1.1), Q89TR9_BRAJA, NOPP_RHISN, Q89TU9_BRAJA (c.10.2.6), Y4FR_RHISN (c.10.2.7), NOPL_RHISN, Y4LO_RHISN, Q89TX4_BRAJA, H7C6R2_BRAJA, C4PL71_BRAEL (c.109.1.4), C4ALD1_RHISN, C4PL56_BRAEL, Y2140_BRAJA (d.3.1.10), Y4ZC_RHISN (d.3.1.10), H0HX55_9RHIZ (d.3.1.10), YP_346620.1, YP_004473413.1, A1WKP8_VEREI (c.146.1.1), Q9EUG5_RHIFR (c.1.6.1), Q9EUG6_RHIFR (d.268.1.2)
Plant Pathogenic (37 proteins): AVRB_PSESG (e.45.1.1), HOPM1_PSESM, Q888Y7_PSESM (c.70.1.1), Q52537_PSESX, Q886L1_PSESM, Q88BF6_PSESM, Q889A9_PSESM (d.2.1.6), Q87V79_PSESM (c.146.1.1), Q882F0_PSESM (a.118.5.1), Q8RP03_PSEYM (c.2.1.2), Q888Y1_PSESM (a.238.1.1), Q87W07_PSESM (a.2.3.1), HRMA_PSESY, Q87WF7_PSESM, Q87X57_PSESM, Q87W42_PSESM, Q88A09_PSESM (d.92.1.7), Q881L7_PSESM, Q9K2L5_PSESH (d.166.1.4), Q88AB8_PSESM, HOAE1_PSEU2, Q7PC42_PSEU2 (c.150.1.2), Q52530_PSESH, Q9L6W4_PSEUB (c.150.1.2), AVRP2_PSESJ (d.3.1.10), Q52394_PSESH, HPAB1_PSESH, Q888W0_PSESM (c.43.1.2), Q7PC45_PSEU2 (d.113.1.4), AVRA_PSESG, Q52432_PSESX (a.15.1.1), AVRD1_PSESH, Q52389_PSESX, Q9JP32_PSEUB, HOPAD_PSESM (c.3.1.5), Q87XS5_PSESM (b.80.1.2), Q9L6W3_PSEUB (c.3.1.4)
Animal Pathogenic (46 proteins): A6M3N5_YERPE (a.24.11.1), O30783_CHLCI, O34020_CHLCI (f.22.1.1), SOPE_SALTM (a.168.1.1), INCE_CHLT2, INCF_CHLT2, O84235_CHLTR, O84236_CHLTR, TARP_CHLTR, A6M3U5_YERPE (c.10.2.6), SPAN_SALTY (b.1.18.2), SOPD_SALTY (d.68.2.1), SPTP_SALTY (c.45.1.2), Q3KMQ0_CHLTA (a.238.1.3), Q3KMQ1_CHLTA, Q46210_CHLCI (f.37.1.1), SIPA_SALTS (a.257.1.1), SIFA_SALTY (a.118.8.1), A9R9K8_YERPG (c.45.1.2), Q56935_YERPU, Q57QR2_SALCH (a.22.1.3), Q663L9_YERPS (c.10.2.1), Q7BS06_YEREN, SOPE2_SALTY (a.168.1.1), A9ZER0_YERPE (a.243.1.2), Q7DB81_ECO57, Q7DB85_ECO57, Q824H6_CHLCV (a.238.1.3), ESFU2_ECO57, Q8XC86_ECO57, SOPA_SALTY (b.80.8.1), Q93KU8_YEREN (c.10.2.1), SSPH2_SALT1 (c.10.2.6), A9ZFE7_YERPE (d.144.1.7), INCD_CHLT2, Q9Z7W9_CHLPN (a.238.1.4), Y572_CHLPN, Q9Z8L4_CHLPN (a.243.1.3), Q9Z8P6_CHLPN, Q9Z8P7_CHLPN, Q9Z8Z8_CHLPN (a.25.1.2), Q9Z9F5_CHLPN (a.25.1.1), B0A3S3_YERPE (d.3.1.10), B0A3S4_YERPE, B0HNN9_YERPE, B2NN32_ECO57

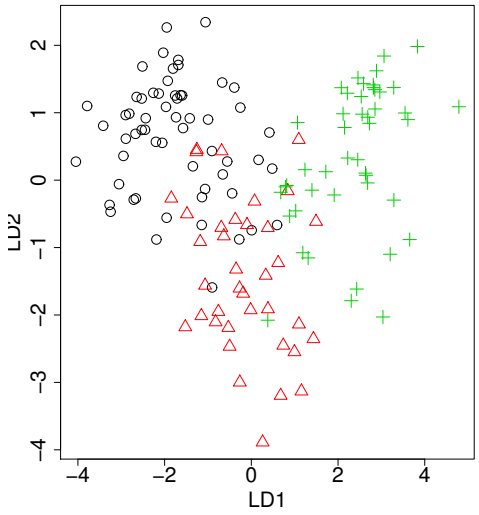

Fig. 7 Two dimensional discrimination between plant symbiotic (black circle), plant pathogen (red triangle) and animal pathogen (green cross). The horizontal (vertical) axis represents the first (second) discriminant function.

pathogenic and plant symbiotic proteins.

In order to confirm this conjecture, we need to know protein structure, but unfortunately, only limited number of proteins were reported to have known protein structures. In order to infer protein structure, we used FAMS[11]. Then FAMS provided us model structure for 27 out of 46 animal pathogenic, 19 out of 37 plant pathogenic and 27 out of 59 plant symbiotic proteins. In order to see if other program can provide more model proteins, we employed phyre2[12][*4] and found that the performance of providing model protein structure was as good as FAMS. Finally, we checked that if I-TASSER[13][*5] can provide feasible model proteins for proteins for which neither FAMS nor phyre2 could provide model protein. I-TASSER was known to achieve the best performances in the recent Critical Assessment of protein Structure Prediction (CASP)[*6] that is a community-wide, worldwide experiment for protein structure prediction taking place every two years since 1994. At the moment, we did not check all of proteins without model protein structures that FAMS or phyre2 predicted, proteins that have already been checked were not accompanied with the feasible model proteins provided by I-TASSER. Although I-TASSER provides some model protein structures, feasibility for these model proteins was poor since no positive C-scores are associated with them. In their analysis[13], feasible model protein structure should be accompanied with more than 0.5 TM-score. Considering error-bars, model proteins with positive C-scores were known to have more than 0.5 TM-score with high possibilities. Thus, I-TASSER does not seem to provide us feasible model proteins for proteins to which neither FAMS nor phyre2 could provide good model proteins.

Based upon obtained model protein structure, we predicted protein fold within SCOP[14][*7] (Table 4). For this purpose, 3D-BLAST[15][*8] was employed. Then, amino acid sequences of 27 out of 46 animal pathogenic, 19 out of 37 plant pathogenic and 27 out of 59 plant symbiotic proteins were replaced with randomly selected proteins with same fold from non-redundant set of proteins, SCOP 1.75 (40%)[*9] that is representative proteins taken from SCOP 1.75B with removing proteins that share more than 40 % sequence identity. YFVs were computed for these set and discrimination was applied. The obtained accuracy averaged over 100 random sampling was 0.45 (see a typical example of two di-

[*4] http://www.sbg.bio.ic.ac.uk/phyre2/
[*5] http://zhanglab.ccmb.med.umich.edu/I-TASSER/
[*6] http://www.predictioncenter.org/casp10/
[*7] http://scop.mrc-lmb.cam.ac.uk/scop/
[*8] http://3d-blast.life.nctu.edu.tw/
[*9] http://scop.berkeley.edu/downloads/scopseq-1.75B/astral-scopdom-seqres-gd-sel-gs-bib-40-1.75B.fa

mensional discrimination in Fig. 8). Although it was much less than accuracy obtained for original (true) set, 0.77, it was still significantly better than accuracy expected for simply random selection, 0.33. In actual, random permutation of proteins between three classes reduced the accuracy down to 0.35 that was averaged over 10 independent random permutations. *P*-values that accuracy 0.45 can be obtained accidentally from randomly permutation was only 3×10^{-4}, which is small enough to be significant. This means, three classes proteins have surely class specific protein folds defined in SCOP (Table 6) and YFV possibly reflected these protein folds more or less.

Table 6 Protein folds distribution for plant symbiotic (PS), plant pathogenic (PP), and animal pathogenic (AP) proteins, estimated by 3D-blast. Subtotals indicate numbers of proteins in each class (a, b, c, d, e, or f). Percentages (%) are equal to the numbers of proteins in each class divided by total number of proteins in each category (PS, PP or AP). Since this information is especially useful for drug discovery, correspondence between protein ID and protein fold is listed in Table 4.

fold	PS	PP	AP
c.1	2	0	0
c.10	3	0	4
c.109	1	0	0
c.146	1	1	0
c.150	0	2	0
c.2	0	1	0
c.3	0	2	0
c.43	0	1	0
c.45	0	0	2
c.56	1	0	0
c.69	1	0	0
c.70	0	1	0
c.94	1	0	0
subtotal	10	6	6
%	**37**	**32**	**22**
d.113	0	1	0
d.118	1	0	0
d.128	1	0	0
d.14	1	0	0
d.144	1	0	1
d.166	0	1	0
d.2	0	1	0
d.226	1	0	0
d.268	2	0	0
d.3	3	1	1
d.68	0	0	1
d.92	0	1	0
subtotal	10	7	3
%	**37**	**37**	**11**

fold	PS	PP	AP
a.114	1	0	0
a.118	0	1	1
a.168	0	0	2
a.2	0	1	0
a.22	0	0	1
a.238	0	1	4
a.24	0	0	1
a.243	0	0	2
a.25	0	1	2
a.257	0	0	1
subtotal	1	4	14
%	**4**	**21**	**52**
b.1	0	0	1
b.26	2	0	0
b.29	1	0	0
b.70	1	0	0
b.80	1	1	1
subtotal	5	1	2
%	**18**	**5**	**7**
e.45	0	1	0
subtotal	0	1	0
%	**0**	**5**	**0**
f.22	0	0	1
f.37	0	0	1
f.22	1	0	0
subtotal	1	0	2
%	**4**	**0**	**7**
total	27	19	27

4. Conclusion

In this paper, we have applied semi-supervised PCA based LDA for the discrimination between symbiotic and pathogenic T3SS effector proteins, using 43 physico-chemical features that can be calculated solely from amino acid sequences. In spite of the good performance of discriminations, there were no stable and small number of features that can discriminate symbiotic and pathogenic proteins, because of many combinations of features that successfully discriminate symbiotic and pathogenic features. All of these combinations were suggested to be robust, since little over fittings turned out to take place. In addition to this, using newly proposed and more feasible protein sets, Yahara's feature vector (YFV) can successfully discriminate three classes, i.e., plant symbiotic, plant pathogenic and animal pathogenic proteins,

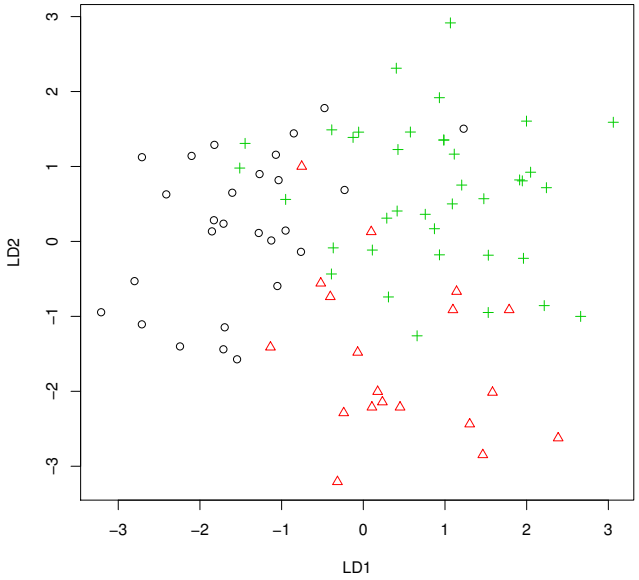

Fig. 8 A typical two dimensional discrimination when T3SS proteins are randomly replaced with proteins selected within same protein fold. Other notations are the same as Fig. 7

with PCA-based LDA. This definitely demonstrated the powerful and general ability of YFV for the discrimination of proteins. YFV was further suggested to reflect protein folds defined in SCOP data base.

Acknowledgments This study was supported by KAKENHI, 23300357 and Chuo University Joint Research Grant.

References

[1] Izore, T., Job, V. and Dessen, A.: Biogenesis, regulation, and targeting of the type III secretion system, *Structure*, Vol. 19, No. 5, pp. 603–612 (2011).
[2] Yang, Y., Zhao, J., Morgan, R. L., Ma, W. and Jiang, T.: Computational prediction of type III secreted proteins from gram-negative bacteria, *BMC Bioinformatics*, Vol. 11 Suppl 1, p. S47 (2010).
[3] Yang, Y.: Identification of novel type III effectors using latent Dirichlet allocation, *Comput Math Methods Med*, Vol. 2012, p. 696190 (2012).
[4] Wang, Y., Zhang, Q., Sun, M. A. and Guo, D.: High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles, *Bioinformatics*, Vol. 27, No. 6, pp. 777–784 (2011).
[5] Schechter, L. M., Valenta, J. C., Schneider, D. J., Collmer, A. and Sakk, E.: Functional and computational analysis of amino acid patterns predictive of type III secretion system substrates in Pseudomonas syringae, *PLoS ONE*, Vol. 7, No. 4, p. e36038 (2012).
[6] 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, No. 4, p. e1000376 (2009).
[7] Yahara, K., Jiang, Y. and Yanagawa, T.: Computational Identification of Discriminating Features of Pathogenic and Symbiotic Type III Secreted Effector Proteins, *Information and Media Technologies*, Vol. 6, No. 1, pp. 39–51 (2011).
[8] Chapelle, O., Schölkopf, B., Zien, A. et al.: *Semi-supervised learning*, Vol. 2, MIT press Cambridge (2006).
[9] Kimbrel, J. A., Thomas, W. J., Jiang, Y., Creason, A. L., Thireault, C. A., Sachs, J. L. and Chang, J. H.: Mutualistic co-evolution of type III effector genes in Sinorhizobium fredii and Bradyrhizobium japonicum, *PLoS Pathog.*, Vol. 9, No. 2, p. e1003204 (2013).
[10] Taguchi, Y. H. and Gromiha, M. M.: Application of amino acid occurrence for discriminating different folding types of globular proteins, *BMC Bioinformatics*, Vol. 8, p. 404 (2007).
[11] Umeyama, H. and Iwadate, M.: *FAMS and FAMSBASE for Protein Structure*, (online), DOI: 10.1002/0471250953.bi0502s04, John Wiley & Sons, Inc. (2002).
[12] Kelley, L. A. and Sternberg, M. J.: Protein structure prediction on the Web: a case study using the Phyre server, *Nat Protoc*, Vol. 4, No. 3, pp. 363–371 (2009).

[13] Zhang, Y.: I-TASSER server for protein 3D structure prediction, *BMC Bioinformatics*, Vol. 9, p. 40 (2008).
[14] Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, Vol. 247, No. 4, pp. 536–540 (1995).
[15] Tung, C. H., Huang, J. W. and Yang, J. M.: Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database, *Genome Biol.*, Vol. 8, No. 3, p. R31 (2007).