IPSJ Transactions on Bioinformatics Vol. 3 95–107 (Dec. 2010)

Original Paper

Computational Identification of Discriminating Features of Pathogenic and Symbiotic Type III Secreted Effector Proteins

Koji Yahara,[†1,†2,†4] Ying Jiang[†1]
and Takashi Yanagawa[†3]

Type III secretion systems (T3SS) deliver bacterial proteins, or "effectors", into eukaryotic host cells, inducing physiological responses in the hosts. Effector proteins have been considered virulence factors of pathogenic bacteria, but T3SSs have now been found in symbiotic bacteria as well. Whether any physicochemical difference exists between the two types of effectors remains unknown. In this work, we combined computational statistical and machine learning methods to identify features that could be responsible for the difference. For computational statistical method we used generalized Bayesian information criterion and kernel logistic regression, and for machine learning method we used support vector machine. It was clearly shown that differences in amino acid composition exist between pathogenic and symbiotic effector proteins. All identified discriminating features were those of amino acid composition and average residue weight, and their classification performance could be nearly identical to that using all physicochemical features, with sensitivity and specificity of over 80%. Further analysis on the seven discriminating features by graphical modeling revealed three dominant features among them. Moreover, amino acid regions that were distinctive for the seven features were explored by sliding window analysis. This study provides a methodological basis and important insights into the functional differences between pathogenic and symbiotic T3SS effectors.

1. Introduction

Type III secretion systems (T3SS) are complex secretion machines that deliver bacterial proteins called effectors into eukaryotic host cells through an injectisome during infection [1),2)]. T3SS-secreted effector proteins induce physiological responses in their hosts, such as cytoskeletal rearrangement to promote bacterial attachment and invasion, interference with cellular trafficking processes, cytotoxicity [2)], induction of apoptosis of macrophages [3)], disruption of tight junctions [4)], and microtubule destabilization [5)]. These effector protein functions are considered causes of virulence in pathogenic bacteria such as *Yersinia* species (spp.), *Chlamydia* spp., *Salmonella* spp., *Shigella* spp., and enteropathogenic *Escherchia coli*. However, T3SSs are also found in symbiotic bacteria [6),7)], and a genome analysis of a Chlamydia-related symbiont of free-living amoebae suggests that the origins of T3SSs may be unrelated to virulence [8)].

Common features of T3SS effector proteins in pathogenic and symbiotic bacteria can be identified by computational methods [9),10)]. While T3SS effector proteins were originally not thought to share any common features [11)], recent studies using machine learning approaches have identified commonalities in the N-terminus of effectors, mainly in amino acid composition. One study [9)] analyzed both pathogenic and symbiotic T3SS effector proteins, and found a signature in the N-terminus that is taxonomically universal and conserved.

The symbiotic T3SS effector proteins, however, have different functions than the pathogenic effectors. Symbiotic effectors of rhizobia, for example, modulate host-plant reactions, that lead to the formation of functional nodules [12),13)]. Putative effector proteins of the tsetse fly endosymbiont, *Sodalis glossinidius*, specifically facilitate the host cell cytoskeletal rearrangements necessary for bacterial entry, although the number of genes encoding effector proteins is smaller in the symbiotic regions than in the homologous islands in pathogenic bacteria [14)]. Homologs of the symbiotic regions are also found in endosymbionts of grain weevils, *Sitophilus oryzae* and *S. zeamais*, in which T3SS genes are suggested to function during a specific stage of weevil development [14)]. Even if the signature amino acid sequence in the N-terminus is conserved among pathogenic and symbiotic T3SS effector proteins, these functional differences exist. We were interested in finding the physicochemical differences between pathogenic and symbiotic T3SS effector proteins that might be responsible for these functional differences.

In this work, we combined computational statistical and machine learning approaches to address this issue. From a dataset of physicochemical features pre-

†1 Division of Biostatistics, Kurume University School of Medicine
†2 Division of Infectious Diseases, Kurume University School of Medicine
†3 Biostatistics Center, Kurume University
†4 Life Science Systems Department, Fujitsu Kyushu Systems Ltd.

pared from pathogenic and symbiotic T3SS effector proteins, discriminating features of amino acid composition were determined using generalized Bayesian information criterion, kernel logistic regression and support vector machine (SVM). Further analysis on seven discriminating features by graphical modeling revealed three dominant features among them. Moreover, amino acid regions that were distinctive for the seven features were explored by sliding window analysis.

2. Materials and Methods

2.1 Dataset

We collected the 57 currently available amino acid sequences of symbiotic T3SS effector proteins from the literature [9),15)], and the same number of amino acid sequences for pathogenic T3SS effector proteins [9)]. The accession number (Uniprot ID), protein name and organism name for the sequences are shown in **Tables S1** and **S2** in **Appendix A.2**.

For each effector protein amino acid sequence, we calculated the physicochemical features, 41 in total, such as charge, isoelectric point, number of proteolytic enzyme or reagent cleavage sites, mole percentage of each amino acid and amino acid groups defined in EMBOSS [16)], and signal peptide probability. The list of 41 physicochemical features used in this study is in **Table 1**. Among them, following features have been examined in studies of T3SS effector proteins: amino acid composition and secondary structure [9)], charge [17)], cleavage sites [18)], and signal peptide probability [15)]. Others (No.1, 4, and 7–10 in Table 1) are general physicochemical features of proteins and added because there was no prior knowledge about differences between pathogenic and symbiotic T3SS effector proteins. Signal peptide probability was calculated by SignalP 3.0 [19)], and others features were calculated by EMBOSS [16)]. These were used as attributes in our classification analysis.

2.2 Feature Selection

We first used the Lepage test for the location-dispersion difference between the two groups [20)]. The top 10 discriminating features were chosen by the order of their p-values in the test statistics. The p-values of all of these candidate features were less than 0.001.

For these candidate features, we examined all combinations, $2^{10} - 1$, as explanatory variables in the kernel logistic regression (KLR), which is one of the kernel-learning methods suitable for binary-pattern recognition problems [21),22)]. Let y_i be a binary observed variable and $p(\mathbf{x}_i)$ be its conditional distribution given $\mathbf{x}_i$, i.e., $p(y_i = 1|\mathbf{x}_i)$, then the likelihood function is given by

$$L = \prod_{i=1}^{n} p(\mathbf{x}_i)^{y_i}(1 - p(\mathbf{x}_i))^{(1-y_i)} \quad (1)$$

and log-likelihood function become

Table 1 Biochemical features used as attributes of effector proteins.

No.	Description
1	Number of potentially antigenic regions of a protein sequence$^{\star 1}$
2	Number of proteolytic enzyme or reagent cleavage sites$^{\star 1}$
3	Number of secondary structure$^{\star 1}$
4	Hydrophobic moment$^{\star 1}$
5	Average residue weight$^{\star 1}$
6	Charge$^{\star 1}$
7	Isoelectric point$^{\star 1}$
8	Molar extinction coefficient$^{\star 1}$
9	Extinction coefficient at 1 mg/ml$^{\star 1}$
10	Probability of protein expression in E. coli inclusion bodies$^{\star 1}$
11–30	Mole percentage of each amino acid$^{\star 1}$ 11:Ala, 12:Cys, 13:Asp, 14:Glu, 15:Phe, 16:Gly, 17:His, 18:Ile, 19:Lys, 20:Leu, 21:Met, 22:Asn, 23:Pro, 24:Gln, 25:Arg, 26:Ser, 27:Thr, 28:Val, 29:Trp, 30:Tyr
31	Mole percentage of tiny amino acids$^{\star 1}$ (A+C+G+S+T)
32	Mole percentage of small amino acids$^{\star 1}$ (A+B+C+D+G+N+P+S+T+V)
33	Mole percentage of aliphatic amino acids$^{\star 1}$ (A+I+L+V)
34	Mole percentage of aromatic amino acids$^{\star 1}$ (F+H+W+Y)
35	Mole percentage of non-polar amino acids$^{\star 1}$ (A+C+F+G+I+L+M+P+V+W+Y)
36	Mole percentage of polar amino acids$^{\star 1}$ (D+E+H+K+N+Q+R+S+T+Z)
37	Mole percentage of charged amino acids$^{\star 1}$ (B+D+E+H+K+R+Z)
38	Mole percentage of basic amino acids$^{\star 1}$ (H+K+R)
39	Mole percentage of acidic amino acids$^{\star 1}$ (B+D+E+Z)
40	Number of cleavage sites between signal sequence and mature exported protein$^{\star 1}$
41	Signal peptide probability$^{\star 2}$

$\star 1$ Calculated by EMBOSS [16)].
$\star 2$ Calculated by SignalP [19)].

$$\log L = \sum_{i=1}^{n} y_i \log \frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)} + \log(1-p(\mathbf{x}_i)) \tag{2}$$

in which the unknown quantity $p(\mathbf{x}_i)$ is modeled using the radial basis kernel function $K(\mathbf{x}_j, \mathbf{x}_i)$ as

$$f(\mathbf{x}_i) = \log \frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)} = \sum_{j=0}^{n} \alpha_j K(\mathbf{x}_j, \mathbf{x}_i) \tag{3}$$

where

$$\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\sigma \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2) \tag{4}$$

and σ is the kernel parameter, n is the number of observations, and j is the index of observation which forms a pair with i-th observation in the radial basis kernel function. Let $\hat{\boldsymbol{\alpha}}$ be the solution of the vector of the regression coefficients in Eq. (3). It is calculated using the following penalized log-likelihood function

$$\frac{1}{n}\left\{\sum_{i=1}^{n} y_i f(\mathbf{x}_i) - \log[1+\exp(f(\mathbf{x}_i))]\right\} - \frac{\lambda}{2}\boldsymbol{\alpha}^T R \boldsymbol{\alpha} \tag{5}$$

where

$$R = \begin{pmatrix} 1 & \mathbf{1}_n' \\ \mathbf{1}_n & \mathbf{K} \end{pmatrix}$$

by Fisher's scoring methods.

To select the best combination of the 10 candidate features, we used a generalized Bayesian information criterion (GBIC) [23]. Using the likelihood function $L(\boldsymbol{\alpha})$ in Eq. (1) and the multivariate normal prior density $\pi(\boldsymbol{\alpha}|\lambda)$ for the parameter vector $\boldsymbol{\alpha}$ defined by

$$\pi(\boldsymbol{\alpha}|\lambda) = (2\pi)^{-r/2}(n\lambda)^{r/2}|R|_+^{1/2} \exp\left(-\frac{n\lambda}{2}\boldsymbol{\alpha}' R \boldsymbol{\alpha}\right). \tag{6}$$

GBIC is defined as

$$\mathrm{GBIC} = -2\log \int L(\boldsymbol{\alpha})\pi(\boldsymbol{\alpha}|\lambda) d\boldsymbol{\alpha} \tag{7}$$

and R is the same as that of Eq. (5), r is the rank of R, and $|R|_+$ is the product of r nonzero eigenvalues of R. Once $\hat{\boldsymbol{\alpha}}$ is obtained, GBIC is calculated through the Laplace approximation

$$\begin{aligned} &-2\log \int \exp(n l_\lambda(\boldsymbol{\alpha})) d\boldsymbol{\alpha} \\ &= -2\log\left\{\frac{(2\pi/n)^{(n+1)/2}}{|J_\lambda(\hat{\boldsymbol{\alpha}})^{1/2}|}\exp(n l_\lambda(\hat{\boldsymbol{\alpha}}))\right\}\{1+O(n^{-1})\} \end{aligned} \tag{8}$$

where

$$\begin{aligned} l_\lambda(\boldsymbol{\alpha}) &= \frac{1}{n}\log L(\boldsymbol{\alpha}) + \frac{1}{n}\log \pi(\boldsymbol{\alpha}|\lambda) \\ J_\lambda(\hat{\boldsymbol{\alpha}}) &= -\frac{\partial^2 l_\lambda(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} \end{aligned}$$

GBIC was computed for each combination of 10 features, and the combination with the minimum GBIC was determined as explanatory variables of KLR. During the feature selection, values of kernel parameter σ and hyperparameter λ were given in the range of 10^{-3} to 10^3 (σ) or to 10^4 (λ) for each set of explanatory features.

2.3 Classification Performance

Classification using discriminating features identified by GBIC of KLR was conducted by SVM based on the approximate relationship between KLR and the SVM [21]. To determine the advantage of the identified discriminating features, a misclassification rate was calculated by leave-one-out cross-validation for each combination of k-features that attained the minimum GBIC in ${}_{10}C_k$ combinations ($k = 1, \cdots, 10$). The results are summarized in a figure which illustrates the misclassification rates, with the number of features on the horizontal axis. We used 'svm' function of e1071 package (E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel) in R.

2.4 Graphical Modeling

To obtain a deeper understanding of discriminating features identified by GBIC of KLR, it was useful to graphically represent correlated relationships among it. We used graphical modeling developed by Imoto, et al. [24),25)] which combines nonlinear nonparametric regression with radial basis and Bayesian network, and was originally developed for estimating genetic networks and functional relationships between genes. Non-linear nonparametric regression enabled us to capture directed dependencies among the features without advance knowledge about their

relationships. Bayesian network is a powerful, graph-theoretic approach for expressing correlated relationships among variables as networks. Details are explained in the **Appendix A.1**.

Calculations were conducted by MATLAB R2008b on the basis of NETLAB toolbox [26], Bayes Net Toolbox [27], and BNT structure learning package [28].

2.5 Sliding Window Analysis

While discriminating features identified by GBIC of KLR were calculated from full-length amino acid sequences, we considered that their differences between pathogenic and symbiotic T3SS effectors proteins might be evident in some amino acid region. In order to explore such regions that were distinctive for identified features, sliding windows analysis was conducted. N-terminal regions from the 1^{st} to 97^{th} residue were analyzed, with the window size varying from 8 to 50, and the starting position varying from 1 to 50. Using discriminating features which attained the smallest minimum GBIC of KLR and the lowest misclassification rate, a dataset was created for each window, and classification performance was evaluated by the leave-one-out cross-validation using SVM.

3. Results

3.1 Identification of Discriminating Features

A plot of minimum GBIC for ${}_{10}C_k$ combination of features used in KLR was given in **Fig. 1** taking the number of features, k, on the horizontal axis. Sets of discriminating features with a minimum GBIC in ${}_{10}C_k$ combinations are in **Table 2**. Clearly, all identified discriminating features were those of amino acid composition and average residue weight.

The figure shows that the minimum GBIC tends to decrease as the number of features increase, take the smallest value when the number of features is seven, and increase at greater than seven features.

3.2 Classification Performance of Discriminating Features

Misclassification rates using the identified discriminating features (shown in Table 2) are plotted in **Fig. 2**, taking the number of features on horizontal axis. The plot of minimum GBICs (Fig. 1) and misclassification rates showed parallel tendencies. While classification performance was about 80 to 85 percent for the three to ten features, the best classification performance was obtained using a

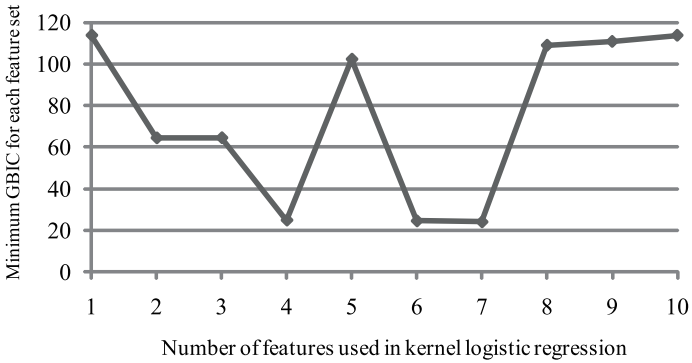

Fig. 1 Plot of minimum GBIC against number of features used in kernel logistic regression. Minimum GBIC from ${}_{10}C_k$ combinations of features ($k = 1, 2, \cdots, 10$) used in kernel logistic regression.

combination of the seven features that attained the smallest minimum GBIC (Fig. 2 A). The best performance with seven features was nearly identical to the result obtained when all 41 features were used. The seven discriminating features had a specificity of 85.5% and a sensitivity of 83.1% (Fig. 2 B).

3.3 Graph Structure Showing Correlated Relationships and Dominant Features among the Seven Discriminating Features

While the best classification performance was obtained using a combination of the seven features, the difference among the three to ten features was not substantial. In addition, the seven discriminating features identified by GBIC of KLR were correlated: about 28.6% among ${}_7C_2 - 7$ non-diagonal correlation matrix elements had Pearson's linear correlation coefficients larger than 0.5. To obtain a deeper understanding of them, we illustrated the relationship among the discriminating features by the directed graphical modeling technique (**Fig. 3**). The figure indicates that among the seven features, mole percentages of alanine and isoleucine ('Ala' and 'Ile' in Fig. 3) are dominant features that are directly contributed to the discrimination. Other features such as mole percentages of small, acidic, tiny amino acids, and aspartic acid contribute to the discrimination through those of alanine and isoleucine. Mole percentages of tiny, acidic amino acids and aspartic acid are related to those of alanine and isoleucine through av-

Table 2 Discriminating features identified by GBIC and kernel logistic regression.

Number of features	No. in Table 1	Description of each feature
1	18	Ile (mole percentage)
2	5	Average residue weight
	11	Ala (mole percentage)
3	25	Arg (mole percentage)
	31	Tiny amino acids (mole percentage)
	32	Small amino acids (mole percentage)
4	11	Ala (mole percentage)
	13	Asp (mole percentage)
	17	His (mole percentage)
	25	Arg (mole percentage)
5	11	Ala (mole percentage)
	13	Asp (mole percentage)
	17	His (mole percentage)
	25	Arg (mole percentage)
	31	Tiny (mole percentage)
6	5	Average residue weight
	11	Ala (mole percentage)
	25	Arg (mole percentage)
	31	Tiny amino acids (mole percentage)
	32	Small amino acids (mole percentage)
	38	Basic amino acids (mole percentage)
7	5	Average residual weight
	11	Ala (mole percentage)
	13	Asp (mole percentage)
	18	Ile (mole percentage)
	31	Tiny amino acids (mole percentage)
	32	Small amino acids (mole percentage)
	39	Acidic amino acids (mole percentage)
8	5	Average residual weight
	11	Ala (mole percentage)
	13	Asp (mole percentage)
	17	His (mole percentage)
	25	Arg (mole percentage)
	31	Tiny amino acids (mole percentage)
	32	Small amino acids (mole percentage)
	39	Acidic amino acids (mole percentage)
9	5	Average residual weight
	11	Ala (mole percentage)
	13	Asp (mole percentage)
	17	His (mole percentage)
	25	Arg (mole percentage)
	31	Tiny amino acids (mole percentage)
	32	Small amino acids (mole percentage)
	38	Basic amino acids (mole percentage)
	39	Acidic amino acids (mole percentage)
10	5	Average residue weight
	11	Ala (mole percentage)
	13	Asp (mole percentage)
	17	His (mole percentage)
	18	Ile (mole percentage)
	25	Arg (mole percentage)
	31	Tiny amino acids (mole percentage)
	32	Small amino acids (mole percentage)
	38	Basic amino acids (mole percentage)
	39	Acidic amino acids (mole percentage)

erage residue weight. Indeed, mole percentage of isoleucine has been identified by GBIC of KLR when the number of feature is one, and the combination of average residue weight and mole percentage of alanine has also been identified when the number of features is two (Table 2). Furthermore, Fig. 2 shows that classification accuracy is as much as about 70% for the mole percentage of isoleucine, and nearly 80% for a combination of alanine and average residue weight.

3.4 Amino Acid Regions That Were Distinctive for the Seven Discriminating Features

Results of sliding window analysis with variable window sizes and starting points are shown in **Table 3**. The region that gave the best classification by the seven discriminating features was 48–95 residues from the N-terminus (N48–95), which gave a classification accuracy of 83.3% (**Fig. 4**). Notably, almost all regions with the second and third highest classification accuracy overlapped with this region (Table 3), supporting that this is a distinctive region for the seven discriminating features.

3.5 Summary of Differences of the Seven Discriminating Features

The differences of the seven features between pathogenic and symbiotic T3SS effector proteins are summarized in **Table 4**, with "+" meaning "more common in symbiotic proteins". Results are given for full-length amino acid sequences and the region that was distinctive for the seven discriminating features, N48–95. The

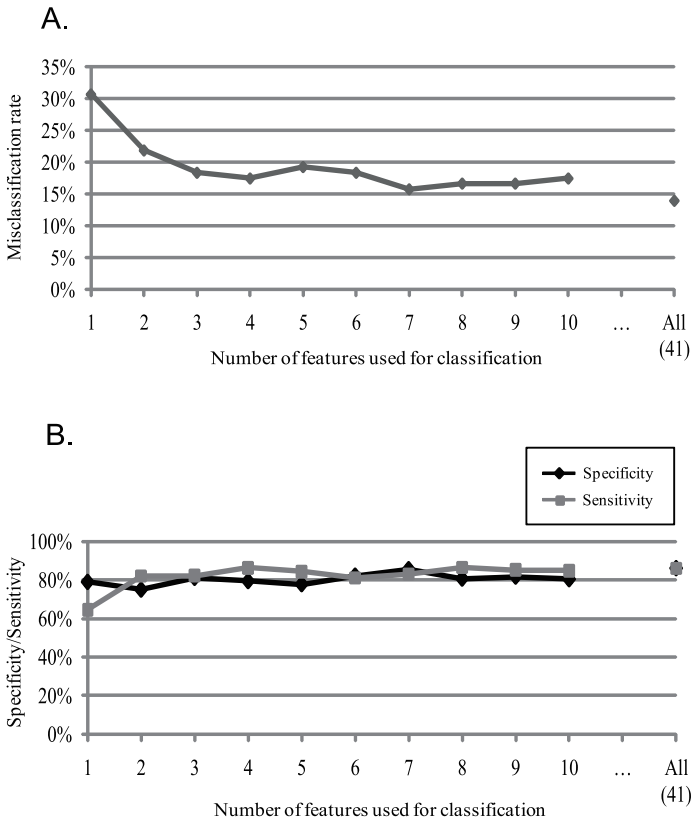

Fig. 2 Classification performance using the discriminating features identified by GBIC of KLR. Misclassification rate for each combination of k-features that attained the minimum GBIC in ${}_{10}C_k$ combinations ($k = 1, \cdots, 10$). Classification using all 41 features was also conducted, and the misclassification rate is at "All (41)" of the x-axis. (A) Misclassification rate. (B) Specificity and sensitivity.

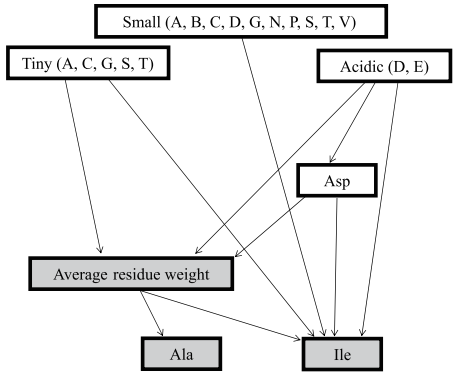

Fig. 3 Graph structure showing correlated relationships and dominant features among the seven discriminating features. Directed dependencies detected by nonparametric regression are depicted by arrows whose heads indicate response variables and tails indicate explanatory variables. Colours are the dominant discriminating features identified by GBIC, when the number of features is one or two (see Table 2).

Table 3 Results of sliding window analysis using the seven discriminating features.

Region	Misclassification rate	Starting point	Window size
N48–95	0.167	48	48
N49–95	0.175	49	47
N48–93	0.184	48	46
N48–96	0.184	48	49
N49–89	0.184	49	41
N49–90	0.184	49	42
N49–96	0.184	49	48
N9–36	0.184	9	28
N40–89	0.193	40	50
N47–93	0.193	47	47
N47–96	0.193	47	50
N48–92	0.193	48	45
N48–94	0.193	48	47
N49–93	0.193	49	45
N50–96	0.193	50	47
N65–97	0.193	65	33

patterns of differences were almost equivalent between full-length amino acid sequences and the distinctive region, revealing that the discriminating signatures of the seven features were evident in this region. Directions of the differences were as follows: as for mole percentage of amino acids, isoleucine decreased in symbiotic proteins, while the other amino acids (alanine, aspartic acids, acidic amino acids, tiny amino acids, small amino acids) increased in symbiotic proteins. The tendency was found both in full-length amino acid sequences and the distinctive region (N48–95).

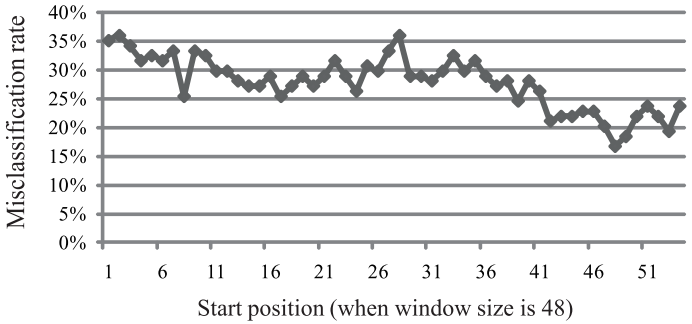

Fig. 4 Plot of misclassification rate by sliding window analysis with window size 48. As shown in Table 3, misclassification rate is lowest when the analysis start position is 48 and the window size is 48 (i.e., for region N48–95), which gives the best classification performance by the seven discriminating features.

Table 4 Summary of differences of the seven discriminating features.

	pathogen (full-length)		symbiont (full-length)		direction⋆1	
Feature	Mean	SD	Mean	SD	Mean	SD
Ile (Molar %)	5.74	2.25	3.98	1.52	-	-
Average residue weight	109.30	4.12	108.98	2.80	-	-
Ala (Molar %)	8.28	3.07	10.99	2.80	+	-
Asp (Molar %)	4.49	1.91	16.01	1.60	+	-
Acidic (Molar %)	10.79	3.91	11.70	2.21	+	-
Tiny (Molar %)	31.58	6.98	32.94	3.90	+	-
Small (Molar %)	51.97	6.95	54.53	4.41	+	-
	pathogen (N48–95)		symbiont (N48–95)		direction⋆1	
Feature	Mean	SD	Mean	SD	Mean	SD
Ile (Molar %)	5.30	4.09	3.84	2.81	-	-
Average residue weight	109.26	6.41	109.79	4.34	+	-
Ala (Molar %)	9.06	4.55	10.78	5.25	+	+
Asp (Molar %)	3.07	2.27	5.88	3.46	+	+
Acidic (Molar %)	8.92	5.69	11.15	4.91	+	-
Tiny (Molar %)	32.35	10.64	33.08	8.05	+	-
Small (Molar %)	51.68	10.95	53.91	6.69	+	-

4. Discussion

In this work, we identified the discriminating features between pathogenic and symbiotic T3SS effector proteins, using a large combination of physicochemical features, analyzed by GBIC of KLR. Their classification performance could be nearly identical to that using all physicochemical features, with sensitivity and specificity of over 80%.

The seven discriminating features which attained the smallest minimum GBIC of KLR and the lowest misclassification rate were those of amino acid composition and average residue weight. Moreover, all of the identified discriminating features were those of amino acid composition and average residue weight, which are correlated primary properties of proteins. No other higher-order property was selected by GBIC of KLR on every number of features (from one to ten) used in KLR (Table 2).

Interestingly, recently reported common features of T3SS effectors were also found to be amino acid composition or shared sequence motif [9),10)]. Especially, the common features of T3SS effectors were reported to be enrichment or depletion of several amino acids (alanine, aspartic acid, threonine, glutamic acid, proline, leucine, serine, threonine) in the N-terminus [9)]. It was reported that, when these differences in individual amino acid composition and other features are combined together, discrimination between T3SS effectors and other control proteins was made possible with sensitivity of ∼71% and selectivity of ∼85%. Similarly, it is conceivable that discrimination between pathogenic and symbiotic T3SS effector proteins in our analysis is made possible by combining differences in individual amino acid composition.

The distinctive region for the seven discriminating features was 48–95 residues from the N-terminus. The classic signal peptide secretion signal is 15–40 residues from the N-terminus [29)]. Common features of T3SS effectors proteins were recently found to be embedded in 30 [10)] or up to 50 residues [9)] at the N-terminus. These findings are complementary with ours because the differences between pathogenic and symbiotic effector proteins are thought to have arisen after the common features in the N-terminus. Although common features are conserved, differences in amino acid composition occur, presumably because of different relationships of pathogens, and symbionts with their hosts.

While classification performance was about 80 to 85 percent for the three to ten features identified by GBIC of KLR, Fig. 1 shows that the minimum GBIC

⋆1 From pathogenic to symbiotic ("+" means "more in symbiotic proteins").

became larger when the number of features used in KLR increased from four to five, or became more than seven. In both cases, the misclassification rate increased (Fig. 2). Table 2 shows that the difference between the four and five features identified by KLR was mole percentage of tiny amino acids. A combination of four features, specifically mole percentage of Arg, Ala, Asp, His, and mole percentage of tiny amino acids are considered to reduce model validity and classification performance. Considering more than seven features also had an adverse effect, presumably because unnecessary information was added to the model.

The sequence dataset of T3SS effector proteins poses a challenge. Because sequence identities among these proteins are low, and their length varies, obtaining fully aligned datasets is difficult or sometimes impossible. We could not analyze T3SS proteins by sequence-alignment methods, so we prepared a feature dataset by calculating the physicochemical properties of each effector protein. The features chosen were easy to calculate compared to other features, such as genomic context, evolutionary based features, and regulatory networks [30]. We also examined the differences between pathogenic and symbiotic effectors as a feature-selection problem, using GBIC of KLR, and classification using SVM. This approach provided a methodological basis for future research examining characteristic features of T3SS effector proteins.

The two previous studies that examined common features of T3SS effector proteins also used feature-selection methods. One conducted feature selection by a greedy hill-climbing search in combination with correlated feature selection [31] based on the WEKA (Waikato Environment for Knowledge Analysis) machine learning toolbox [32]. The selected number of discriminating features was 10, more than half of which were amino acid composition. Another study used a recursive feature elimination approach of SVM [10]. A minimal set of 88 features was found to retain the ability to classify secreted effectors. Although the datasets in these studies were different from ours, our method was comparably effective in capturing essential differences with lower features.

In this study, we used GBIC of KLR for computational statistical method, and used SVM and leave-one-out cross-validation for machine learning method. Although both of information criterion and cross-validation can be used to feature selection, these results often disagree. Since GBIC of KLR is more sophisticated method than simple cross-validation, enabling selection of features that maximize posterior probability given observed data, we selected it primarily. We anticipate the combination would lead to better classification.

The identified discriminating features were used for classification, and for elucidating their correlated relationships and dominant features among them using graphical modeling that combined non-linear nonparametric regression and Bayesian network. Although these techniques are usually used for estimating gene networks from microarray expression data, the combination of them, with feature selection, was a powerful method for a deeper understanding of the meaning of the discriminating features.

This is the first study to identify discriminating features between pathogenic and symbiotic T3SS effector proteins, using a combination of computational statistical and machine learning approaches. The discriminating features of amino acid composition and average residue weight, the dominant features among the seven discriminating features, and the amino acid regions that were distinctive for them were revealed. This study will provide a methodological basis for future research, and provides important insight about the functional differences between pathogenic and symbiotic T3SS effectors.

Acknowledgments The computational calculations were carried out at the Human Genome Center at the Institute of Medical Science, the University of Tokyo. This work was supported by a grant from the Science and Technology Foundation of Japan to Koji Yahara.

References

1) Cornelis, G.R.: The type III secretion injectisome, *Nat Rev Microbiol*, Vol.4, No.11, pp.811–825 (2006).
2) Coburn, B., Sekirov, I., Finlay, B.B.: Type III secretion systems and disease, *Clin Microbiol Rev*, Vol.20, No.4, pp.535–549 (2007).
3) Hernandez, L.D., Pypaert, M., Flavell, R.A. and Galan, J.E.: A Salmonella protein causes macrophage cell death by inducing autophagy, *J Cell Biol*, Vol.163, No.5, pp.1123–1131 (2003).
4) Boyle, E.C., Brown, N.F. and Finlay, B.B.: Salmonella enterica serovar Typhimurium effectors SopB, SopE, SopE2 and SipA disrupt tight junction structure and function, *Cell Microbiol*, Vol.8, No.12, pp.1946–1957 (2006).

5) Yoshida, S., Katayama, E., Kuwae, A., Mimuro, H., Suzuki, T., et al.: Shigella deliver an effector protein to trigger host microtubule destabilization, which promotes Rac1 activity and efficient bacterial internalization, *Embo J*, Vol.21, No.12, pp.2923–2935 (2002).
6) Beeckman, D.S. and Vanrompay, D.C.: Bacterial Secretion Systems with an Emphasis on the Chlamydial Type III Secretion System, *Curr Issues Mol Biol*, Vol.12, No.1, pp.17–42 (2009).
7) Coombes, B.K.: Type III secretion systems in symbiotic adaptation of pathogenic and non-pathogenic bacteria, *Trends Microbiol*, Vol.17, No.3, pp.89–94 (2009).
8) Horn, M., Collingro, A., Schmitz-Esser, S., Beier, C.L., Purkhold, U., et al.: Illuminating the evolutionary history of chlamydiae, *Science*, Vol.304, No.5671, pp.728–730 (2004).
9) Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., et al.: Sequence-based prediction of type III secreted proteins, *PLoS Pathog*, Vol.5, No.4, e1000376 (2009).
10) Samudrala, R., Heffron, F. and McDermott, J.E.: Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems, *PLoS Pathog*, Vol.5, No.4, e1000375 (2009).
11) Grynberg, M. and Godzik, A.: The signal for signaling, found, *PLoS Pathog*, Vol.5, No.4, e1000398 (2009).
12) Kambara, K., Ardissone, S., Kobayashi, H., Saad, M.M., Schumpp, O., et al.: Rhizobia utilize pathogen-like effector proteins during symbiosis, *Mol Microbiol*, Vol.71, No.1, pp.92–106 (2009).
13) Masson-Boivin, C., Giraud, E., Perret, X. and Batut, J.: Establishing nitrogen-fixing symbiosis with legumes: How many rhizobium recipes?, *Trends Microbiol*, Vol.17, No.10, pp.458–466 (2009).
14) Dale, C. and Moran, N.A.: Molecular Interactions between Bacterial Symbionts and Their Hosts, *Cell*, Vol.126, No.3, pp.453–465 (2006).
15) Lower, M. and Schneider, G.: Prediction of type III secretion signals in genomes of gram-negative bacteria, *PLoS One*, Vol.4, No.6, e5917 (2009).
16) Rice, P., Longden, I. and Bleasby, A.: EMBOSS: The European Molecular Biology Open Software Suite, *Trends Genet*, Vol.16, No.6, pp.276–277 (2000).
17) Yao, Q., Cui, J., Zhu, Y., Wang, G., Hu, L., et al.: A bacterial type III effector family uses the papain-like hydrolytic activity to arrest the host cell cycle, *Proc. Natl Acad Sci USA*, Vol.106, No.10, pp.3716–3721 (2009).
18) Chisholm, S.T., Dahlbeck, D., Krishnamurthy, N., Day, B., Sjolander, K., et al.: Molecular characterization of proteolytic cleavage sites of the Pseudomonas syringae effector AvrRpt2, *Proc. Natl Acad Sci USA*, Vol.102, No.6, pp.2087–2092 (2005).
19) Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S.: Improved prediction of signal peptides: SignalP 3.0, *J Mol Biol*, Vol.340, No.4, pp.783–795 (2004).
20) Lepage, Y.: A combination of Wilcoxon's and Ansari-Bradley's statistics, *Biometrika*, Vol.58, No.1, pp.213–217 (1971).
21) Zhu, J. and Hastie, T.: Kernel Logistic Regression and the Import Vector Machine, *Journal of Computational and Graphical Statistics*, Vol.14, No.1, pp.1081–1088 (2001).
22) Cawley, G.C. and Talbot, N.L.: Efficient approximate leave-one-out cross-validation for kernel logistic regression, *Machine Learning*, Vol.71, No.2-3, pp.243–264 (2008).
23) Konishi, S., Ando, T. and Imoto, S.: Bayesian information criteria and smoothing parameter selection in radial basis function networks, *Biometrika*, Vol.91, No.1, pp.27–43 (2004).
24) Imoto, S., Goto, T. and Miyano, S.: Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression, *Pac Symp Biocomput*, Vol.7, pp.175–186 (2002).
25) Imoto, S., Sunyong, K., Goto, T., Aburatani, S., Tashiro, K., et al.: Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network, *Proc. IEEE Comput Soc Bioinform Conf*, Vol.1, pp.219–227 (2002).
26) Nabney, I.: *NETLAB: Algorithms for pattern recognition*, Springer, London (2001).
27) Murphy, K.P.: The Bayes net toolbox for matlab, *Computing Science and Statistics*, Vol.33, pp.331–350 (2001).
28) Leray, P. and Francois, O.: BNT structure learning package: Documentation and experiments. Laboratoire PSI, Universite et INSA de Rouen (2006).
29) von Heijne, G.: Signal sequences: The limits of variation, *J Mol Biol*, Vol.184, No.1, pp.99–105 (1985).
30) Burstein, D., Zusman, T., Degtyar, E., Viner, R., Segal, G., et al.: Genome-scale identification of Legionella pneumophila effectors using a machine learning approach, *PLoS Pathog*, Vol.5, No.7, e1000508 (2009).
31) Hall, M.: Correlation-based feature selection for machine learning, University of Waikato (1998).
32) Witten, I.H. and Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann (2005).

Appendix

A.1 Details of Graphical Modeling Applied in This Study

We denote the most discriminating features identified by the GBIC as $\mathbf{x} = \{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ where $\mathbf{x}_j$ $(j = 1, \cdots, p)$ is n-dimensional column vector. Correlated relationships among them are explored through nonparametric regression by taking a response variable, for example $\mathbf{b}_j$, as $\mathbf{x}_j$ $(j = 1, \cdots, p)$, and a subset of remaining variables, for example $\mathbf{a}_k$ $(k = 1, \cdots, q,\ 1 \leq q \leq p-1)$, as explanatory variables. A nonparametic regression model using radial basis func-

tion network [23] is given by

$$b_i = \gamma_0 + \sum_{m=1}^{M} \gamma_m \psi_m(\mathbf{a}_i) + \epsilon_i \ (i = 1, \cdots, n) \tag{9}$$

where $\epsilon_i \sim N(0,\ \sigma^2)$, and

$$\psi_m(\mathbf{a}_i) = \psi_m(||\mathbf{a}_i - \mathbf{c}_m||^2) = \exp\left(-\frac{||\mathbf{a}_i - \mathbf{c}_m||^2}{2vh_m^2}\right),\ m = 1, \cdots, M \tag{10}$$

where M and v are given constants; M is the number of basis function and v is a hyperparameter that controls the widths of the basis functions. We abbreviate the subscript j ($j = 1, \cdots, p$) and k ($k = 1, \cdots, q$) for convenience. To estimate parameters in the model, we fix v to be 1, and M to $p-1$ for convenience. A K-means clustering based procedure is applied to estimate center $\mathbf{c}_m$ and width h_m in advance for each m. The observations $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are divided into M clusters C_m ($m = 1, \cdots, M$) by K-means algorithm, and estimates of $\mathbf{c}_m$ and h_m are given by

$$\mathbf{c}_m = \frac{1}{n_m} \sum_{i \in C_m} \mathbf{a}_i,\ \ h_m = \frac{1}{n_m q} \sum_{i \in C_m} ||\mathbf{a}_i - \mathbf{c}_m||^2 \tag{11}$$

where n_m is the number of the observations which belong to the m-th cluster C_m. The unknown parameters $\gamma_0, \cdots, \gamma_m$ ($m = 1, \cdots, M$) are estimated by the maximum-likelihood method.

To determine directed dependencies among $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ we apply a Bayesian network method [24),25)] based on the nonparametric regression. Considering a graph G with $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ as nodes, the directed edges of the graph are given, for example, from $\mathbf{a}_1, \cdots, \mathbf{a}_q$ to $\mathbf{b}_j$ when these variables satisfy Eq. (9). The key issue is to select $\mathbf{a}_1, \cdots, \mathbf{a}_q$ to $\mathbf{b}_j$ from $\{\mathbf{x}_1, \cdots, \mathbf{x}_p\}$ for each $\mathbf{b}_j$, which is called the selection of the best graph structure G. The selection is conducted by maximizing the posterior probabilities, represented by

$$P_{post}(G|\mathbf{x}) \propto P(G) \int P(\boldsymbol{\theta}, \mathbf{x}|G) d\boldsymbol{\theta} \tag{12}$$

where $\boldsymbol{\theta}$ is a parameter vector used in the graph G (in this case $\gamma_0, \cdots, \gamma_m$ and σ). We assume $P(G)$ is uniform, introduce hyperparameter λ, and denote the joint distribution of $\boldsymbol{\theta}$ and G by $P(\boldsymbol{\theta}, G, \lambda)$. Since

$$P(\mathbf{x}|\boldsymbol{\theta}, G) = \prod_{i=1}^{n} f(\mathbf{x}^{(i)}; \boldsymbol{\theta}, G) = \prod_{j=1}^{p} \prod_{i=1}^{n} f_j(b_{ij}; \{a_{i1}, \cdots, a_{iq}\}, \boldsymbol{\theta}_j, G)$$

where f_j is the density function of $\mathbf{b}_j$, it follows that

$$\begin{aligned}
&\int P(\boldsymbol{\theta}, \mathbf{x}, G) d\boldsymbol{\theta} \\
&= \int P(\mathbf{x}|\boldsymbol{\theta}, G) P(\boldsymbol{\theta}, G, \lambda) d\theta \\
&\propto \int P(\mathbf{x}|\boldsymbol{\theta}, G) P_{prior}(\boldsymbol{\theta}|G, \lambda) d\boldsymbol{\theta} \\
&= \prod_{j=1}^{p} \int \prod_{i=1}^{n} f_j(b_{ij}; \{a_{i1}, \cdots, a_{iq}\}, \boldsymbol{\theta}_j, G) P_{prior,j}(\boldsymbol{\theta}_j|G, \lambda_j) d\boldsymbol{\theta}_j
\end{aligned}$$

Putting

$$\begin{aligned}
&BNRC(G) \\
&= -2 \log \left\{ \prod_{j=1}^{p} \int \prod_{i=1}^{n} f(b_{ij}; \{a_{i1}, \cdots, a_{iq}\}, \boldsymbol{\theta}_j, G) P_{prior}(\boldsymbol{\theta}_j|G, \lambda_j) d\boldsymbol{\theta}_j \right\} \\
&= \sum_{j=1}^{p} BNRC(G_j)
\end{aligned} \tag{13}$$

we select the best graph structure G that minimize $BNRC(G)$. For structure learning, we use the greedy hill-climbing algorithm [25)].

In the following, we abbreviate the subscript j for convenience. To calculate $BNRC(G)$, we calculate the integration using the Laplace approximation

$$\begin{aligned}
&-2 \log \int \exp\{n l_\lambda(\boldsymbol{\theta}|\mathbf{b}, \mathbf{a})\} d\boldsymbol{\theta} \\
&= \frac{(2\pi/n)^{r/2}}{|J_\lambda(\hat{\boldsymbol{\theta}})|^{1/2}} \exp\{n l_\lambda(\hat{\boldsymbol{\theta}}|\mathbf{b}, \mathbf{a})\} \{1 + O(n^{-1})\}
\end{aligned} \tag{14}$$

where r is the dimension of $\boldsymbol{\theta}$, and

Table S1 Amino acid sequences of symbiotic effector proteins used in this study.

Uniprot ID	Protein name	Organism
Q9ANH8	Bll1810 protein	*Bradyrhizobium japonicum*
Q9ANI9	Bll1798 protein	*Bradyrhizobium japonicum*
Q2LDQ5	HsvB type III effector	*Pantoea agglomerans* pv. *gypsophilae gypsophilae*
A2I8A2	HsvB type III effector	*Pantoea agglomerans* pv. *gypsophilae*
Q9F0I3	Y4yA	*Rhizobium fredii*
Q9F0I4	Y4yB	*Rhizobium fredii*
Q9EUG5	Y4yA	*Rhizobium fredii*
Q9EUG6	Y4yB	*Rhizobium fredii*
Q84H14	Putative type III secreted effector LopT	*Photorhabdus luminescens*
Q6MCQ9	Putative CPAF (Chlamydia protease-like activity factor)	*Protochlamydia amoebophila* (strain UWE25)
A1WKP8	Type III effector Hrp-dependent outers	*Verminephrobacter eiseniae* (strain EF01-2)
Q9ANJ0	Bll1797 protein	*Bradyrhizobium japonicum*
P13835	Avirulence protein B	*Pseudomonas syringae* pv. *glycinea*
Q887D0	Effector protein hopM1	*Pseudomonas syringae* pv. *tomato*
Q888Y7	Type III effector HopQ1-1	*Pseudomonas syringae* pv. *tomato*
Q52537	AvrPmaA1 protein	*Pseudomonas syringae*
Q886L1	Type III effector HopAF1	*Pseudomonas syringae* pv. *tomato*
Q88BF6	Type III effector HopY1	*Pseudomonas syringae* pv. *tomato*
Q889A9	Type III effector HopAJ1	*Pseudomonas syringae* pv. *tomato*
Q87V79	Type III effector HopAN1	*Pseudomonas syringae* pv. *tomato*
Q882F0	Type III helper protein HopP1	*Pseudomonas syringae* pv. *tomato*
Q8RP03	Type III effector HopPtoA1Pma	*Pseudomonas syringae* pv. *maculicola*
Q9ANJ7	Blr1789 protein	*Bradyrhizobium japonicum*
Q888Y1	Type III effector HopR1	*Pseudomonas syringae* pv. *tomato*
Q87W07	Type III effector HopI1	*Pseudomonas syringae* pv. *tomato*
Q08370	Protein hrmA	*Pseudomonas syringae* pv. *syringae*
Q87WF7	Type III effector HopT1-2	*Pseudomonas syringae* pv. *tomato*
Q87X57	HopPtoE	*Pseudomonas syringae* pv. *tomato*
Q87W42	HopPtoG	*Pseudomonas syringae* pv. *tomato*
Q88A09	Type III effector HopH1	*Pseudomonas syringae* pv. *tomato*
Q881L7	Type III effector HopL1	*Pseudomonas syringae* pv. *tomato*
Q9K2L5	ORF2	*Pseudomonas syringae* pv. *phaseolicola*
Q87W46	Type III effector HopV1	*Pseudomonas syringae* pv. *tomato*
Q9AND2	Bll1858 protein	*Bradyrhizobium japonicum*
Q88AB8	Type III effector HopAS1	*Pseudomonas syringae* pv. *tomato*
Q7PC62	Effector protein hopAE1	*Pseudomonas syringae* pv. *syringae* (strain B728a)
Q7PC42	Putative type III effector HolPtoACPsy	*Pseudomonas syringae* pv. *syringae* (strain B728a)
Q52530	Avirulence gene D (Fragment)	*Pseudomonas syringae* pv. *phaseolicola*
Q9L6W4	Putative uncharacterized protein	*Pseudomonas syringae* pv. *tomato*
Q9F3T4	Probable cysteine protease avirulence protein avrPpiC2	*Pseudomonas syringae* pv. *pisi*
Q52394	AvrPphE protein	*Pseudomonas syringae* pv. *phaseolicola*
Q9RBW3	Effector protein hopAB1	*Pseudomonas syringae* pv. *phaseolicola*
Q888W0	Type III effector HopAI1	*Pseudomonas syringae* pv. *tomato*
Q7PC45	Type III effector HopAG1	*Pseudomonas syringae* pv. *syringae* (strain B728a)
Q9AMW4	Putative cysteine protease yopT-like blr2058	*Bradyrhizobium japonicum*
P11437	Avirulence protein A	*Pseudomonas syringae* pv. *glycinea*
Q52432	Avirulence protein (Fragment)	*Pseudomonas syringae*
Q9F3T6	Effector protein AvrPphD	*Pseudomonas syringae* pv. *phaseolicola*
Q52389	Putative uncharacterized protein	*Pseudomonas syringae*
Q9JP32	Type III effector HopN1	*Pseudomonas syringae* pv. *tomato*
Q87W65	Effector protein hopAD1	*Pseudomonas syringae* pv. *tomato*
Q87XS5	Type III effector HopAK1	*Pseudomonas syringae* pv. *tomato*
Q9L6W3	HrpK	*Pseudomonas syringae* pv. *tomato*
Q89TL5	Blr1904 protein	*Bradyrhizobium japonicum*
Q9ANJ9	Blr1787 protein	*Bradyrhizobium japonicum*
Q9ANM7	ID84	*Bradyrhizobium japonicum*
Q89TQ6	Blr1788 protein	*Bradyrhizobium japonicum*

$$l_\lambda(\boldsymbol{\theta}|\mathbf{b},\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n}\log f(b_i;\{a_{i1},\cdots,a_{iq}\},\boldsymbol{\theta},G) + \frac{1}{n}\log P_{prior}(\boldsymbol{\theta}|G,\lambda)$$

$$J_\lambda(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 l_\lambda(\boldsymbol{\theta}|\mathbf{b},\mathbf{a})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$$

$$\hat{\boldsymbol{\theta}} = \arg\max\{l_\lambda(\boldsymbol{\theta}|\mathbf{b},\mathbf{a})\}$$

We use a multivariate normal distribution with mean vector zero and diagonal covariance matrix whose i-th element is λ for $\log P_{prior}(\boldsymbol{\theta}|G,\lambda)$. Thus,

$$\log P_{prior}(\boldsymbol{\theta}|G,\lambda) = -\frac{r}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\boldsymbol{\gamma}^T\Sigma^{-1}\boldsymbol{\gamma} \quad (15)$$

Furthermore,

Table S2 Amino acid sequences of pathogenic effector proteins used in this study.

Uniprot ID	Protein name	Organism
A6M3N5	Translocated host-GTPase-activating protein	*Yersinia pestis* CA88-4125
O30783	Inclusion membrane protein C	*Chlamydophila caviae*
O34020	CopN protein	*Chlamydophila caviae*
O52623	Guanine nucleotide exchange factor sopE	*Salmonella typhimurium*
O84118	Inclusion membrane protein E	*Chlamydia trachomatis*
O84119	Inclusion membrane protein F	*Chlamydia trachomatis*
O84235	Inclusion Membrane Protein B	*Chlamydia trachomatis*
O84236	Inclusion Membrane Protein C	*Chlamydia trachomatis*
O84462	Translocated actin-recruiting phosphoprotein	*Chlamydia trachomatis*
O85239	Protein kinase YopO	*Yersinia enterocolitica*
P0C2N1	Cysteine protease yopT1	*Yersinia enterocolitica*
A6M3U5	Leucine-rich 15-repeat translocated effector protein	*Yersinia pestis* CA88-4125
P27475	Cysteine protease yopT	*Yersinia enterocolitica*
P40613	Surface presentation of antigens protein spaN	*Salmonella typhimurium*
P40722	Sop effector protein sopD	*Salmonella typhimurium*
P74873	Effector protein sptP	*Salmonella typhimurium*
Q05608	Protein kinase ypkA	*Yersinia pseudotuberculosis*
Q3KMQ0	Inclusion membrane protein A	*Chlamydia trachomatis*
Q3KMQ1	Inclusion membrane protein G	*Chlamydia trachomatis*
Q46210	Inclusion membrane localized protein	*Chlamydophila caviae*
Q56027	Cell invasion protein sipA	*Salmonella typhimurium*
Q56061	Protein sifA	*Salmonella typhimurium*
A9R9K8	Protein-tyrosine-phosphatase YopH	*Yersinia pestis* bv. Antiqua (strain Angola)
Q56921	Protein kinase A	*Yersinia enterocolitica*
Q56935	Yop targeting protein yopK, yopQ	*Yersinia pseudotuberculosis*
Q57QR2	Outer protein	*Salmonella choleraesuis*
Q663I2	Yop proteins translocation protein H	*Yersinia pseudotuberculosis*
Q663L9	YopM; putative targeted effector protein	*Yersinia pseudotuberculosis*
Q7BRY7	Yop effector YopE	*Yersinia enterocolitica*
Q7BRY8	Yop effector YopH	*Yersinia enterocolitica*
Q7BRZ4	Secreted protein YopR	*Yersinia enterocolitica*
Q7BS06	YopQ	*Yersinia enterocolitica*
Q7CQD4	Guanine nucleotide exchange factor sopE2	*Salmonella typhimurium*
A9ZER0	Type III secretion protein YopR	*Yersinia pestis* biovar Orientalis str. IP275
Q7DB81	EspD	*Escherichia coli* O157:H7
Q7DB85	EspF	*Escherichia coli* O157:H7
Q824H6	Putative uncharacterized protein	*Chlamydophila caviae*
Q8X2D5	EspF-like protein	*Escherichia coli* O157:H7
Q8XC86	EspB	*Escherichia coli* O157:H7
Q8ZNR3	Secreted effector protein of Salmonella	*Salmonella typhimurium*
Q93KQ5	Yop effector YopP	*Yersinia enterocolitica*
Q93KU8	Yop effector YopM	*Yersinia enterocolitica*
Q93RN4	Cysteine protease yopT	*Yersinia pseudotuberculosis*
Q9RPH0	Leucine-rich repeat protein	*Salmonella typhimurium*
A9ZFE7	Protein kinase YopO	*Yersinia pestis* biovar Orientalis str. IP275
Q9RPQ1	Inclusion membrane protein D	*Chlamydia trachomatis*
Q9Z7W9	CPj0585 protein	*Chlamydia pneumoniae*
Q9Z7Y1	Uncharacterized protein CPn_0572/CP_0177/CPj0572/CpB0594	*Chlamydia pneumoniae*
Q9Z8L4	CopN	*Chlamydia pneumoniae*
Q9Z8P6	Inclusion Membrane Protein C	*Chlamydia pneumoniae*
Q9Z8P7	Inclusion Membrane Protein B	*Chlamydia pneumoniae*
Q9Z8Z8	Inclusion membrane protein A	*Chlamydia pneumoniae*
Q9Z9F5	Putative uncharacterized protein	*Chlamydia pneumoniae*
B0A3S3	Cysteine protease YopT	*Yersinia pestis* biovar Orientalis str. F1991016
B0A3S4	YopK protein	*Yersinia pestis* biovar Orientalis str. F1991016
B0HNN9	Effector protein YopJ	*Yersinia pestis* biovar Antiqua str. B42003004
B2NN32	NleB	*Escherichia coli* O157:H7 str. EC4196

$$\log f(b_i;\{a_{i1},\ldots,a_{iq}\},\boldsymbol{\theta},G)$$
$$=-\frac{1}{2}(\log 2\pi-2\log\sigma)-\frac{b_i-\sum_{m=0}^{M}\gamma_m\psi_m(\{a_{i1},\cdots,a_{iq}\})}{2\sigma^2} \quad (16)$$

To obtain parameter estimates $\hat{\boldsymbol{\theta}}$ which maximize $l_\lambda(\boldsymbol{\theta}|\mathbf{b},\mathbf{a})$ in Eq. (14), an iterative procedure is applied. First, initially fixed λ, σ^2, $\boldsymbol{\gamma}=(\gamma_0,\cdots,\gamma_M)^T$ is estimated by solving regularized least-square function based on $l_\lambda(\boldsymbol{\theta}|\mathbf{b},\mathbf{a})$

$$E=\frac{1}{2}\sum_{i=1}^{n}\left(b_i-\sum_{m=0}^{M}\gamma_m\psi_m(\{a_{i1},\cdots,a_{iq}\})\right)^2+\frac{\lambda}{2}\boldsymbol{\gamma}^T\boldsymbol{\gamma} \quad (17)$$

as

$$\hat{\boldsymbol{\gamma}} = \left(\boldsymbol{\psi}^T(\mathbf{a})\boldsymbol{\psi}(\mathbf{a}) + \lambda I\right)^{-1}\boldsymbol{\psi}(\mathbf{a})^T\mathbf{b} \quad (18)$$

where $\boldsymbol{\psi}(\mathbf{a}) = (\psi_0(\{a_{i1}, \cdots, a_{iq}\}), \cdots, \psi_M(\{a_{i1}, \cdots, a_{iq}\}))$. Second, σ^2 is estimated by

$$\hat{\sigma}^2 = ||\mathbf{b} - \boldsymbol{\psi}(\mathbf{a})\hat{\boldsymbol{\gamma}}||^2/n \quad (19)$$

Third, λ is estimated by maximizing $l_\lambda(\boldsymbol{\theta}|\mathbf{b}, \mathbf{a})$. The procedure is repeated until $\boldsymbol{\gamma}$, σ^2, λ converge. $BNRC(G)$ is computed using the parameter estimates. The best graph structure G that attain the minimum $BNRC(G)$ is selected.

A.2 Supplementary Data

Supplementary tables with amino acid sequences of effector proteins used in this study are in **Tables S1** and **S2**.

(Received May 7, 2010)
(Accepted September 9, 2010)
(Released December 9, 2010)

(Communicated by *Takenao Ookawa*)

Koji Yahara was born in 1981. He received his M.S. degree in life science from the University of Tokyo in 2006. He has been working in Life Science Systems Department, Fujitsu Kyushu Systems Ltd. since 2006 and been a Ph.D. student in Divisions of Biostatistics and Infectious Diseases, Kurume University School of Medicine since 2008. He has been engaging in research areas of bioinformatics, biostatistics, and genomics. He is a member of IPSJ. He was awarded from the Science and Technology Foundation of Japan and the Science and Technology Foundation of Japan.

Ying Jiang was born in 1982. She received her B.S. degree from Department of Mathematics, Shimane University in 2006, and conferred her M.S. degree from Department of Biostatistics, Kurume University in 2008. She is a Ph.D. student in Department of Biostatistics, Kurume University School of Medicine now. She is a member of International Biometric Conference.

Takashi Yanagawa was born in 1940. He received his Ph.D. from Kyushu University in 1970. He devoted his 33 years of life in research and education in Kyushu University. He was awarded a title of Professor Emeritus by Kyushu University in May 2004. He has been a professor of the Biostatistics Center, Kurume University, since 2004. He has been engaging in research areas of statistical science and biostatistics. He was awarded from the Biometric Society of Japan and Japan Statistical Society.