Probabilistic Model for Purity Values of Bacterial Genome Sequences

Yuta Taniguchi[1*], Ryuji Masui[2], Toshihiro Aoyama[2], Daisuke Ikeda[3]

[1] Shikoku Innovative and Collaborative Organization for Industry, Academia and Government, Tokushima University, Tokushima 770-8506, Japan.
[2] Department of Electronic and Information Engineering, Suzuka National College of Technology, Mie 510-0294, Japan.
[3] Department of Informatics, Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan.

* Corresponding author. Tel.: +81-88-656-9774; email: yuta.taniguchi@tokushima-u.ac.jp
Manuscript submitted January 10, 2015; accepted April 12, 2015.
doi: 10.17706/ijbbb.2015.5.5.288-295

Abstract: In sequence analysis, quantifying characteristics of a genomic region is helpful to identify functions of the region, and many computational measures have been proposed. Purity measure is a computational measure, and its potential to characterize horizontally transferred genes has been shown in the literatures. However, in the previous studies, only repeating regions were studied, and also the statistical property of purity values, evaluation values of the measure, has not been uncovered. In this paper, we propose a generative model for *k*-mers and evaluate the accuracy of distributions of purity values predicted by the probabilistic model. We train the model for each of 14 bacterial genomes, and our analysis shows that our model predicts the distributions successfully for six genome sequences. It is also shown that our model makes better predictions for shorter *k*-mers.

Key words: Sequence analysis, purity measure, generative model, *k*-mers, prediction, .

1. Introduction

Sequence analysis is a widely used approach for understanding genomes. Traditional methods such as sequence alignments and hidden Markov Models have been successfully applied to this end [1]. For identifying functions of regions of a genome sequence, existing studies have also employed the characteristics of regions and proposed many computational measures for quantifying the characteristics.

One of the approaches to quantify characteristics of genomic regions is to use biological knowledge, and there exist many methods based on various domain knowledge [2]-[4]. Another approach focuses on the compositional characteristics of a sequence and does not need any domain knowledge. This approach includes nucleotide-level composition [5], di-nucleotide abundance [6], probability [7] and complexity [8], [9].

Purity measure [10] is another compositional measure, which was proposed in the field of text mining for finding unusual regions of an input string. It has been shown [11] that this measure certainly characterize genes such as mobile elements, phages, RNAs and transposons which can be considered as horizontally transferred genes [12]-[14] as well as existing measures [15]-[17]. Since horizontal gene transfer is considered as one of the primary reasons of bacterial genetic diversity [18], finding such genes could lead to evidences of the

hypothesis.

However, in the previous studies of purity measure, only repeating regions, which appear at least twice in an entire sequence, were examined with the measure. Moreover, the statistical property of the evaluation values of the measure, that we call purity values, has not been uncovered. Thus, currently, there is no definitive way to discriminate regions of horizontal transferred genes from others using purity values. Revealing statistical properties of purity values would enable us, for example, to decide a threshold of purity values for the discrimination in an objective fashion.

In this paper, we aim to find out statistical properties of purity values. To this end, we propose a generative model for *k*-mers and evaluate the accuracy of distributions of purity values predicted by the model. Making histograms of the purity values, we have found that they form a bell curve-*like* distribution. Then we propose a probabilistic model for *k*-mers based on a binomial model. We trained our model for each of 14 bacterial genome sequences that we chose, and evaluate the accuracies of predicted distributions of purity values comparing them to the observed distributions.

2. Method

First we explain the definition of purity measure before proposing a probabilistic model for its evaluation values.

2.1. Purity Measure

Given a string T and a substring x of T, purity measure quantifies how many substrings of x appear the same number of times as x. In other words, how many substrings of x appear only as substrings of x in T. Obviously, most substrings of x are much shorter than x, and such substrings are considered to appear more frequently than x in T. Hence, we could consider x is unusual if most substrings of x appear only as parts of x. Yamada *et al*. [10] proposed three different definitions of purity measure called *probability*, *entropy* and *difference*. In this study, we only use the *probability* definition as with the previous study [11]. We call the measure based on the definition just purity measure in the rest of this paper.

Formally the purity measure is defined as follows. Let N be the set of non-negative integers. Let Σ be a finite set of *characters*. We call Σ an *alphabet*. We denote a set of finite sequences of zero or more characters by Σ^* and call its element a *string*. The length of a string $x \in \Sigma^*$ is denoted by $|x|$. For a string $x = a_1 a_2 \ldots a_n \in \Sigma^*$ of length n, the i-th character a_i of x is denoted by $x[i]$ for a positive integer i, and a contiguous part $a_i \ldots a_j$ of x is denoted by $x[i:j]$ for positive integers i and j such that $i \leq j$ and called a *substring* of x.

For a string $x \in \Sigma^*$, $sub(x)$ is defined as follows:

$$sub(x) = \{\langle i, j\rangle \in \mathrm{N}^2 \mid 1 \leq i \leq j \leq |x|\}.$$

For a string T and a string $x \in \Sigma^*$, we define $pos_T(x)$ as follows:

$$pos_T(x) = \{\langle i, j\rangle \in sub(T) \mid T[i{:}j] = x\}.$$

For a string T and a string $x \in \Sigma^*$, $freq_T(x)$ is defined as $freq_T(x) = |pos_T(x)|$. Intuitively, $sub(x)$ represents a set of the all substrings of x, $pos_T(x)$ is a set of occurrences of x in T, and $freq_T(x)$ is the frequency of x in T.

Definition 1. Given an input string T and a substring $x = T[i{:}j]$ of T, the purity value of x on T is defined as follows:

$$purity_T(x) = \frac{|\{\langle k, l\rangle \in sub(x) \mid freq_T(x[k:l]) = freq_T(x)\}|}{|sub(x)|}$$

This definition of the purity measure quantifies a characteristic of *x* in *T* as the fraction of the substrings of *x* that only appear as parts of *x* in *T*.

The definition requires us to compute the frequencies of a target string x and its substrings. Employing a naïve way, that is to search a genome sequence for these substrings and to count their occurrences, can easily become impractical when we try to evaluate many regions of a sequence. Special data structures such as suffix trees and suffix arrays provide efficient algorithms that can solve this problem [19], and we can construct a practical algorithm to compute purity values.

2.2. Probabilistic Model

In this section, we discuss the distribution of purity values. Since purity values highly depend on the length of a target substring, we focus only on fixed-length substrings of length *k*, which are called *k*-mers, in this paper. We have found that distributions of the purity values of *k*-mers form bell curve-*like* distributions which are slightly different from Gaussian distributions. Hence, we propose a probabilistic model of purity values on top of *binomial models* rather than Gaussian models based on the observations.

Suppose we are given an input string *T* and an integer *k*, where $0 \leq k \leq |T|$. Let *x* be a substring of *T* such that $1 \leq |x| \leq k$, and let *y* be a substring of *T* such that $|y| = k$. We call *x* a *specific substring* of *y* if and only if *x* appears only as a part of *y* in *T*. Now, the purity value of the substring *y* can be described as the ratio of specific substrings of *y* to all the substrings of *y*. Since a *k*-mer always has $n = k(k+1)/2$ substrings, a distribution of purity values of *k*-mers depends only on the number of specific substrings of every *k*-mer.

Our basic idea is to regard a *k*-mer as a set of *n* independent substrings. Let's suppose we randomly choose every element of such a set from either specific substrings or non-specific ones. Assuming choice probability p for a specific substring and $(1-p)$ for a non-specific substring, a final number X of specific substrings in an assembled set follows a binomial distribution $B(n, p)$, and the probability that the number becomes l is written as $\Pr[X = l] = \binom{n}{l} p^l (1-p)^{n-l}$. Let Z a *k*-mer chosen from a sequence randomly. The probability that Z's purity value $purity_T(Z)$ is equal to q is written as follows:

$$\Pr[purity_T(Z) = q] = \Pr[X = nq] = \binom{n}{nq} p^{nq} (1-p)^{n-nq}.$$

Then, given a set of purity values $\{q_i\}$ of *k*-mers, we can determine the parameter p by maximum likelihood estimation as follows:

$$\hat{p} = \operatorname{argmax}_p \prod_i \Pr[purity_T(Z) = q_i] = \frac{\sum_i q_i}{\sum_i 1}.$$

Furthermore, we replace the parameter p with another variable m to make the model parameter more intuitive. As a longer substring has more possibility to be a specific substring of some other substrings, we assume that every substring of any *k*-mers whose length is longer than *m* is a specific substring of the *k*-mer. Therefore, we can denote $p = (k-m)(k-m+1) \,/\, k(k+1)$. Solving the equation for m, we finally obtain an optimal parameter value $\hat{m}$ as

$$\hat{m} = k + \frac{1}{2} - \frac{1}{2}\sqrt{1 + 8n\frac{\sum_i q_i}{\sum_i 1}}.$$

Finally, our probabilistic model only has a single parameter m. We can train our model for a genome sequence by giving a set of purity values of *k*-mers included by the sequence.

3. Result and Discussion

We train our probabilistic model for various bacterial genomes with four different lengths of *k*-mers. Table 1 shows 14 bacterial genome sequences for which our model is trained. We chose the genomes from popular ones so that both gram-positive and gram-negative genomes, and sequences with various lengths and G+C contents are included. We only use a single strand of a genome sequence which is included in a GenBank file retrieved from NCBI's RefSeq database [20]. We tried $k = 30, 50, 70, 90$ for the lengths of *k*-mers.

Table 1. Bacterial Genome Sequences Used in Our Analysis

Accession No.	Length	Organism
NC_000117.1	1,042,519	Chlamydia trachomatis D/UW-3/CX
NC_000911.1	3,573,470	Synechocystis sp. PCC 6803
NC_000913.2	4,639,675	Escherichia coli str. K-12 substr. MG1655
NC_000962.2	4,411,532	Mycobacterium tuberculosis H37Rv
NC_000964.3	4,215,606	Bacillus subtilis subsp. subtilis str. 168
NC_002695.1	5,498,450	Escherichia coli O157:H7 str. Sakai
NC_002946.2	2,153,922	Neisseria gonorrhoeae FA 1090
NC_003228.3	5,205,140	Bacteroides fragilis NCTC 9343
NC_007517.1	3,997,420	Geobacter metallireducens GS-15
NC_008261.1	3,256,683	Clostridium perfringens ATCC 13124
NC_009882.1	1,257,710	Rickettsia rickettsii str. 'Sheila Smith'
NC_010572.1	8,545,929	Streptomyces griseus subsp. griseus NBRC 13350
NC_012973.1	1,576,758	Helicobacter pylori B38
NC_015431.1	1,153,998	Mycoplasma mycoides subsp. capri LC str. 95010

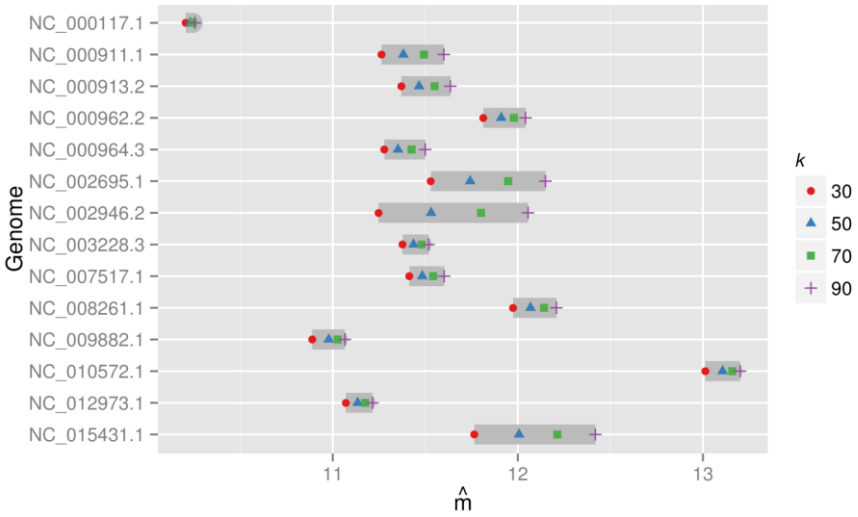


Fig. 1. Relationships between $\hat{m}$ and genome.

For Each Sequence, an accession number in NCBI's RefSeq database, sequence length and organism name are shown.

Fig. 1 shows the parameter values of $\hat{m}$ obtained by training for every combination of genome and value of *k*. In the figure, we cannot see clear relationship between $\hat{m}$ and *k*. To see the relationship from a different perspective, we made another plot shown in Fig. 2. This figure shows the same for every combination of sequence length and value of *k*. However, there is still no clear relationship in the figure. For

the most genome sequences, it seems that $\hat{m}$ depends on genomes more than on the lengths of *k*-mers, and thus we could use the same model for prediction of the purity distribution regardless of *k* for some cases.

In Fig. 3, for every genome sequence, a plot of observed distribution and theoretical one of purity values of 30-mers is shown. It is shown in the figures that our model successfully predicts the distributions of six genomes, NC_000117.1, NC_000911.1, NC_000913.2, NC_000964.3, NC_002695.1, and NC_003228.3, out of 14 genomes. For the rest of sequences, although mean values of purity values are predicted well, variance of the distributions are very different from those of theoretical ones.

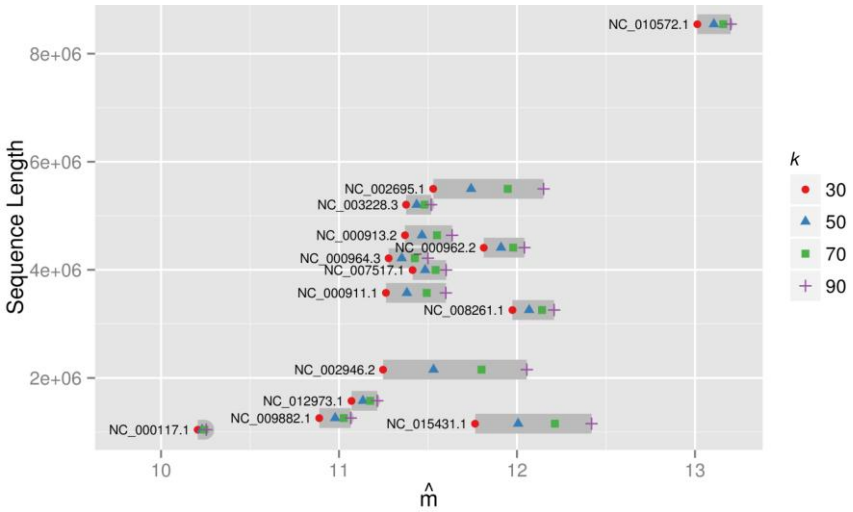


Fig. 2. Relationships between $\hat{m}$ and sequence length.

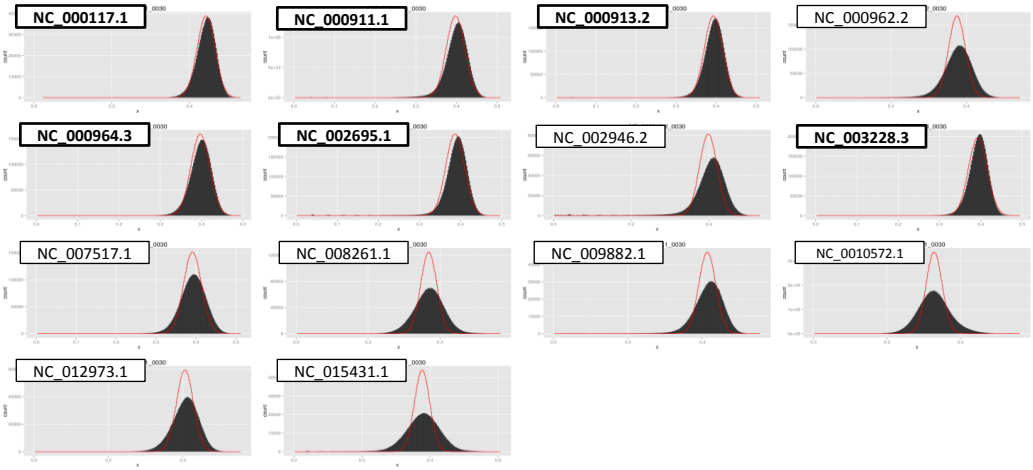


Fig. 3. Predicted distributions (red lines) and observed distributions (black areas) of purity values for 30-mers of every genome sequence. We can see that relatively good predictions are made on genomes of NC_000117.1, NC_000911.1, NC_000913.2, NC_000964.3, NC_002695.1, and NC_003228.3. These accession numbers are emphasized in the above figure.

We also emphasized the names of genomes which look better than the others in the previous figure.

We quantitatively evaluate how well our model predicted the distributions. We computed three quantities to measure accuracy of the predictions: *overlap ratio*, *mode gap* and *variance gap*. An overlap ratio is the fraction of common area on a plot. A mode gap and a variance gap are the absolute difference of modes and variances of theoretical distribution and observed one respectively. Fig. 4 shows those three quantities for every genome and *k* values. We can see the accuracies of the predictions get worse as *k* gets bigger. It is also shown that the accuracy depends on genome very much.

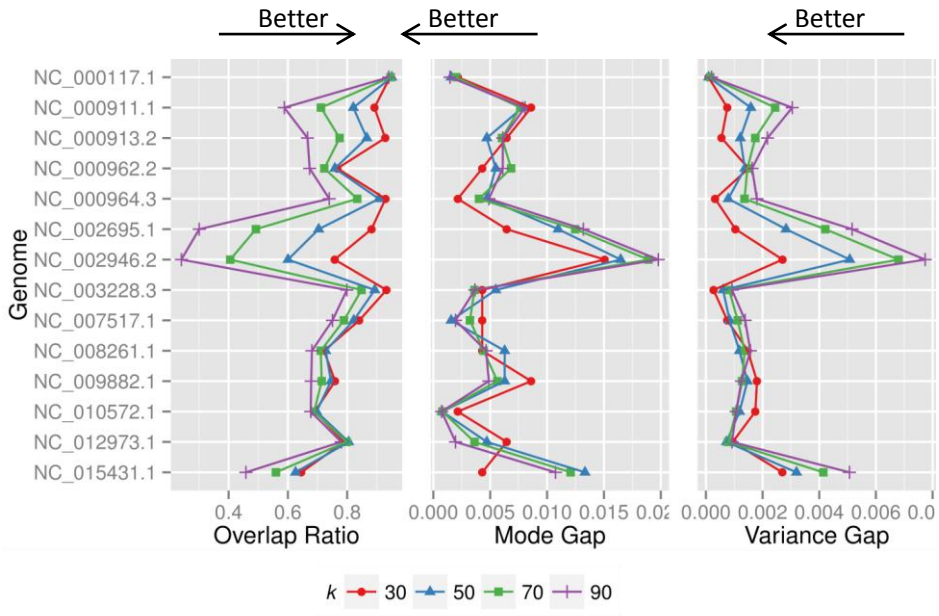


Fig. 4. Plots of three quantities for measuring accuracy of predictions made by our model.

From the above analysis, we conclude that our model captures the statistical properties of purity values very well for about half of genome sequences we tested though there is a room for improvement in cases of long *k*-mers and the rest of genome sequences.

4. Conclusion

We proposed a generative model for *k*-mers of genome sequences and analyzed the accuracy of distributions of purity values predicted by the model. Our model is based on a binomial distribution and has only a single parameter which can be easily determined from a set of purity values. We train our model for 14 bacterial genome sequences with different lengths of *k*-mers, and the accuracies of the predictions made by the model are investigated with three quantitative measures. Our analysis shows that our model can predict the distribution of purity values very well for six genome sequences and it makes better prediction for shorter *k*-mers. We conclude that we have successfully uncovered the statistical properties of purity values at least for a part of bacterial genome sequences. We believe our result will become the foundation of future development of the discriminator of horizontal transferred genes.

Acknowledgment

This work was supported by JSPS Grant-in-Aid for Scientific Research (B) Grant Number 24300059.

References

[1] Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

[2] Fouts, D. E. (2006). Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research*, *34(20)*, 5839-5851.

[3] Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. (2008). Prophinder: A computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*, *24(6)*, 863-865.

[4] Rho, M., Choi, J. H., Kim, S., Lynch, M. & Tang, H. (2007). De novo identification of LTR retrotransposons in eukaryotic genomes. *BMC Genomics*, *8(1)*, 90.

[5] Akhter, S., Aziz, R. K., & Edwards, R. A. (2012). PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Research*, *40(16)*, e126.

[6] Srividhya, K. V., Alaguraj, V., Poornima, G., Kumar, D., Singh, G. P., Raghavenderan, L., Katta, A. V. S. K. M., Mehta, P., & Krishnaswamy, S. (2007). Identification of prophages in bacterial genomes by dinucleotide relative abundance difference. *PLoS ONE*, *2(11)*, e1193.

[7] Ikeda, D. & Suzuki, E. (2009). Mining peculiar compositions of frequent substrings from sparse text data using background texts. W. Buntine, M. Grobelnik, D. Mladeni & J. Shawe-Taylor (Eds.), *Machine Learning and Knowledge Discovery in Databases*, *Volume 5781 of Lecture Notes in Computer Science* (pp. 596-611). Springer Berlin Heidelberg.

[8] Kargar, M. & An, A. (2010). Evaluation of different complexity measures for signal detection in genome sequences. *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology* (pp. 422-425). ACM: New York, NY, USA.

[9] Wan, H. & Wootton, J. C. (2000). A global compositional complexity measure for biological sequences: AT-rich and GC-rich genomes encode less complex proteins. *Computers & Chemistry*, *24(1)*, 71–94.

[10] Yamada, Y., Nakatoh, T., Baba, K. & Ikeda, D. (2012). Mining pure patterns in texts. *Proceedings of IIAI International Conference on Advanced Applied Informatics* (pp. 285-290).

[11] Taniguchi, Y., Yamada, Y., Maruyama, O., Kuhara, S. & Ikeda, D. (2013). The purity measure for genomic regions leads to horizontally transferred genes. *Journal of Bioinformatics and Computational Biology*, *11(06)*, 1343002, PMID: 24372031.

[12] Kidwell, M. (1992). Horizontal transfer of P elements and other short inverted repeat transposons. *Genetica*, *86(1-3)*, 275-286.

[13] Heinemann, J. & Kurenbach, B. (2009). Horizontal transfer of genes between microorganisms. *Encyclopedia of Microbiology (Third Edition)*, Academic Press, Oxford, 597.

[14] Yap, W. H., Zhang, Z., & Wang, Y. (1999). Distinct types of rRNA operons exist in the genome of the actinomycete thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon. *Journal of Bacteriology*, *181(17)*.

[15] Dufraigne, C., Fertil, B., Lespinats, S., Giron, A., & Deschavanne, P. (2005). Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Research*, *33(1)*, e6.

[16] Garcia-Vallv, S., Romeu, A., & Palau, J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes, *Genome Research*, 10(11).

[17] Tsirigos, A., & Rigoutsos, I. (2005). A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. *Nucleic Acids Research*, *33(12)*, 3699–3707.

[18] Dutta, C., & Pan, A. (2002). Horizontal gene transfer and bacterial diversity. *Journal of Biosciences*, *27(1)*.

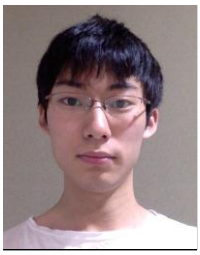
[19] Gusfield, D. (1997). *Algorithms on Strings, Trees and Sequence.* Cambridge University Press, New York.

[20] Tatusova, T., Ciufo, S., Federhen, S., Fedorov, B., McVeigh, R., O'Neill, K., Tolstoy, I., & Zaslavsky, L. (2015). Update on RefSeq microbial genomes resources. *Nucleic Acids Research*, *43(D1)*, D599–D605.

Yuta Taniguchi was born in Japan in 1984. He received his bachelor degree and master degree in science, and doctor degree in information science from Kyushu University in 2009, 2011 and 2014, respectively.

He is currently a project researcher in Shikoku Innovative and Collaborative Organization for Industry, Academia and Government, Tokushima University. The latest publications include articles in the field of data mining and bioinformatics. His research

interests include data mining and machine learning.



Ryuji Masui was born in Mie, Japan on August 10, 1993. In 2014, he obtained the associate degree in Electronic and Information Engineering of National Institute of Technology, Suzuka College, Japan.

In 2014, he started his bachelor of computer science at Kyoto University, Japan. He is interested in theory of computation, discrete optimization, machine learning.

Toshihiro Aoyama was born on January 19, 1974 in Aichi. He received the M.Sc. degree in information engineering from Toyohashi University of Technology in 1998, and the PhD degree in electronics and information engineering from Toyohashi University of Technology in 2002.

He has been working at National Institute of Technology, Suzuka College since 2004. He is an associate professor in the Department of Electronic and Information Engineering. He worked at BSI RIKEN as a researcher 2002-2004. He has 14 peer reviewed journal papers.

Dr. Aoyama has participated in the Information Processing Society of Japan, Japanese Neural Network Society and Japanese Society of Bioinformatics. He awarded by Japan Society of Information and Knowledge in 2014 for his work on the system of digital repository.

Daisuke Ikeda was born in Japan in 1971. He received his bachelor degree, master degree, and doctor degree in science from Kyushu University in 1994, 1996, and 2004, respectively.

He is currently an associate professor in the Department of Informatics, Kyushu University. He is also a visiting associate professor of National Institute of Informatics, Japan. He is a member of General Analysis Section, International Center for Space Weather Science and Education Center, Kyushu University, and a member of Research and Development Department, Kyushu University Library. He became an associate professor in Kyushu University in 2004. The latest publications include articles in the field of data mining, e-science, bioinformatics. His research interests include data analysis, such as data mining and machine learning, and data infrastructure, such as database and information retrieval.

Dr. Ikeda is serving and has served as a PC member at conferences in data mining, such as International Conference on Advanced Data Mining and Applications (ADMA), Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), and Discovery Science (DS). He is a member of Association for Computing Machinery and Information Processing Society of Japan.