Journal of Bioinformatics and Computational Biology

THE PURITY MEASURE FOR GENOMIC REGIONS LEADS TO HORIZONTALLY TRANSFERRED GENES

YUTA TANIGUCHI
Department of Informatics, Kyushu University,
Fukuoka, Japan,
yuta.taniguchi@inf.kyushu-u.ac.jp

YASUHIRO YAMADA
Interdisciplinary Graduate School of Science and Engineering, Shimane University,
Matsue, Japan,
yamada@cis.shimane-u.ac.jp

OSAMU MARUYAMA
Institute of Mathematics for Industry, Kyushu University,
Fukuoka, Japan,
om@imi.kyushu-u.ac.jp

SATORU KUHARA
Department of Bioscience & Biotechnology, Faculty of Agriculture, Kyushu University,
Fukuoka, Japan,
kuhara@grt.kyushu-u.ac.jp

DAISUKE IKEDA
Department of Informatics, Kyushu University,
Fukuoka, Japan,
daisuke@inf.kyushu-u.ac.jp

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Sequence analysis is important to understand a genome, and a number of approaches such as sequence alignments and hidden Markov models have been employed. In the field of text mining, the *purity measure* is developed to detect unusual regions of a string without any domain knowledge. It is reported in that work that only RNAs and transposons are shown to have high purity values. In this work, the purity values of regions of various bacterial genome sequences are computed, and those regions are analyzed extensively. It is found that mobile elements and phages as well as RNAs and transposons have high purity values. It is interesting that they are all classified into a group of horizontally transferred genes. This means that the purity measure is useful to predict horizontally transferred genes.

Keywords: Horizontally Transferred Genes; Composition; Purity Measure

1. Introduction

Sequence analysis is an established approach to elucidate the structure of genome sequences and their functional regions. Sequence alignments and hidden Markov models have been heavily used to analyze genome sequences[5]. These traditional tools are very successful to identify functional regions.

In the field of text mining, the *purity* of a string is introduced by Yamada et al.[25] to find unusual regions of an input string. In that work, at first, the purity measure is applied to Japanese blog texts, and it is shown to successfully detect artificially copied and inserted parts of the texts. Secondly, the measure is also applied to only two genome sequences of *Escherichia coli* and *Bacillus subtilis*, and it is reported that the top 100 regions with high purity values are likely to correspond to RNAs and transposons.

It should be noted here that these genes can be considered to be horizontally transferred genes[15], which are closely related to the evolution of microorganism genomes. Horizontal gene transfer is considered as one of the primary reasons for bacterial genetic diversity[6] because transferring genes involves larger changes than mutation. However, recall that in that work[25] only the top 100 regions with high purity values are evaluated. As a result, their implications are very limited. It is interesting to examine what kind of functional regions are identified by the purity measure.

In this paper, we have extensively investigated the effectiveness of the purity measure to bacterial genome sequences. Although in the previous work[25], only RNAs and transposons are characterized by the purity measure, our result shows that mobile elements and phages as well as RNAs and transposons have high purity values. Interestingly, those found genes can be considered to be horizontally transferred genes[11,26]. Therefore, this fact means that the purity measure is useful to predict horizontally transferred genes.

2. Related Work

Roughly speaking, to characterize genomic regions computationally, there are two approaches: the one that is based on biological knowledge and the one that utilize the compositional characteristics of a sequence. From the viewpoint, the purity measure is considered to be a kind of a *compositional measure* and to belong to the latter approach.

In the former approach, many methods are proposed based on different domain knowledge such as homology of sequences[8,16,19] and structures common to particular genes[17,20]. Though those methods are very effective to detect regions similar to well-known regions, the power of those methods are limited because such domain knowledge are biased toward previously identified regions. Hence, it is hard to discover novel functional regions compared to the latter approach.

On the other hand, in the latter approach, functional regions are characterized by the variations of compositions of the regions. Previously, numerous kinds of com-

positional features have been proposed nucleotide-level composition[1], di-nucleotide abundance[21], probability[13] and complexity[14,24]. The variations of such compositional features are usually measured compared to a *standard* model such as an average composition[1] and a probabilistic model of a background sequence[13].

This composition-based approach is also widely used for characterization of horizontally transferred genes[4,9,18,22]. Basically, in those methods, only the compositions of low-order nucleotide, i.e. mono-, di- and tri-nucleotide, are considered. Although, relatively longer oligonucleotides have been employed[18], only a restricted set of nucleotide sequences are usually considered.

In contrast to them, the purity measure, which is also a compositional measure, takes account of every part of a sequence regardless of the length of the part. Furthermore, the measure doesn't assume any models to measure the variation, and it depends on a universally applicable assumption instead. Hence, it can be considered that the purity measure characterizes horizontally transferred genes better and more robustly.

3. Methods

3.1. *Purity Measure*

The purity measure quantifies unusualness of substrings of a given string under the reasonable assumption such that shorter substrings occur more frequently than longer ones. With the assumption, given a string T and a substring x of T, we can consider x to be unusual in T if most substrings of x appear the same number of times as x. It is obvious because most substrings of x are much shorter than x, and such substrings are considered to appear more frequently than x in T under the assumption. To quantify how much x deviates from the assumption, in that work[25], three measures are proposed based on statistics called *probability*, *entropy* and *difference* respectively. In this study, we use only the measure using probability statistic because it was used in their experiments on genome sequences and transposons were captured well. We call the measure just the purity measure in the rest of this paper.

Formally the purity measure is defined as follows. Let $\mathbb{N}$ 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 $x = a_1 a_2 \dots a_n \in \Sigma^*$ 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 $a_i \dots a_j$ 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:

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

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

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

For a string $T, x \in \Sigma^*$, $freq_T(x)$ is defined as $freq_T(x) = \left|pos_T(x)\right|$. 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 calculated as follows:

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

This definition of the purity value formulates the unusualness of x as the fraction of the substrings of x that only appear within x in T.

We can compute the frequency $freq_T(x)$ of a substring x of T efficiently with memory space and run time proportion to the length of T with the aid of an appropriate data structure, such as suffix trees or suffix arrays[10].

3.2. *Target Substrings to Evaluate with the Purity Measure*

Although we can simply apply the purity measure to all the substrings of the input string to find substrings with highest purity values, it is practically impossible to do so since there are $\mathcal{O}(n^2)$ substrings for an input string of length n. Hence, we focus only on repeated substrings as many methods for mobile elements do[2,23].

Additionally, an equivalence relation is employed to reduce redundancy further by grouping substrings and considering only representatives of resulting equivalence classes. Here we consider an equivalence relation proposed by Blumer et al.[3]

Let x and $y = \alpha x \beta$ be substrings of an input string T, where $\alpha, \beta \in \Sigma^*$. If $freq_T(x) \neq freq_T(y)$, we can think there are distinct usages of x in T other than being used as a part of y, and hence it would be better to distinguish x and y. On the contrary, if $freq_T(x) = freq_T(y)$, x is considered to be redundant because x always occurs as a part of y with α and β as a prefix and suffix respectively. In this example, x and y are grouped together by the equivalence relation.

There are at most $\mathcal{O}(n)$ equivalence classes for an input string T of length n. The longest element of each equivalence class is chosen as a representative. In the following sections, we call the representatives of the equivalence classes *Blumer strings*.

4. Results and Discussion

In this section, experiments on eleven bacterial genome sequences are presented, which were conducted to show what kind of functional regions are found by the purity measure and how strongly such functional regions and purity values are associated. Note that our experiment is not intended to precisely identify functional regions of genome sequences. Instead, we aim to visually and quantitatively show the power of the purity measure on various genome sequence without any complex procedure other than the purity measure.

Table 1. Bacterial genome sequences that are used in our experiments. For each sequence, an accession number of GenBank, sequence length, G+C content and organism name are shown. Popular sequences with various lengths and G+C contents were chosen.

Accession	Length	G+C (%)	Organism
NC_000911.1	3,573,470	47.7	Synechocystis sp. PCC 6803
NC_000913.2	4,639,675	50.8	Escherichia coli str. K-12 substr. MG1655
NC_000964.3	4,215,606	43.5	Bacillus subtilis subsp. subtilis str. 168
NC_002695.1	5,498,450	50.5	Escherichia coli O157:H7 str. Sakai
NC_002946.2	2,153,922	52.7	Neisseria gonorrhoeae FA 1090
NC_003228.3	5,205,140	43.2	Bacteroides fragilis NCTC 9343
NC_007517.1	3,997,420	59.5	Geobacter metallireducens GS-15
NC_008261.1	3,256,683	28.4	Clostridium perfringens ATCC 13124
NC_010572.1	8,545,929	72.2	Streptomyces griseus subsp. griseus NBRC 13350
NC_012973.1	1,576,758	39.2	Helicobacter pylori B38
NC_015431.1	1,153,998	23.8	Mycoplasma mycoides subsp. capri LC str. 95010

Therefore, our experimental procedure is simply as follows: (1) enumerate all Blumer strings of a sequence, (2) compute the purity value for each string of them, (3) output the strings with purity values, and (4) analyze these strings together with annotated regions of the input sequence. Table 1 describes the eleven bacterial genomes briefly and also includes the accession numbers of GenBank files from which we retrieve annotation information. As shown in the table, from popular sequences, we chose the sequences so that sequences with various lengths and G+C contents are included. In the following experiments, we only used a single strand of a genome sequence which is included in a GenBank file of the sequence.

Figure 1 shows the number of substrings which have larger purity values than a given threshold value for each genomes. The horizontal axes represent the purity values of substrings. Note that the vertical axes represent the logarithmic numbers of substrings. It is shown in the figure that there are around 10^6 Blumer strings. However, as threshold value increases, the numbers of substrings decrease rapidly, and only the small number of substrings have large purity values.

By reducing the redundancy of substrings dramatically, the equivalence relation makes it possible for us to analyze those substrings comprehensively. Although the restriction to Blumer strings might seem to be excessive, as one of the following figures shows below, these strings still appear ubiquitously in sequences, and we think it does not matter.

We conducted all the experiments on a single Linux machine composed of Intel Core i7 3.4 GHz and 16 GB of memory. We implemented the procedure described above in C++. Source code were all compiled with GCC 4.8.1.

4.1. *Sequence Maps*

We rendered regions that correspond to Blumer strings on sequence maps for each of all examined genome sequences to visually and qualitatively understand the association between purity values and regions for biological functions. Figure 2 illustrate

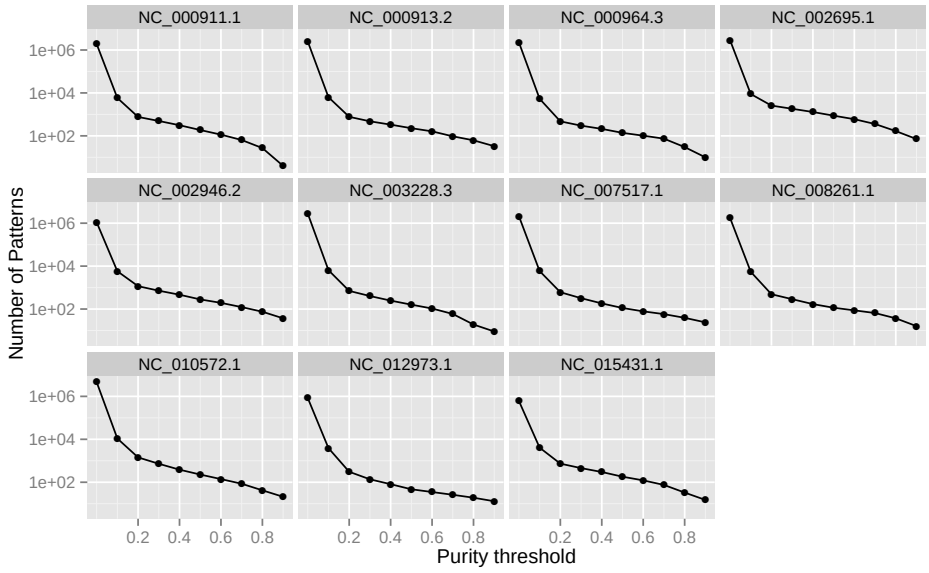

Fig. 1. The graphs show the number of substrings with purity values larger than a given threshold as the threshold value increases. The horizontal axes represents the threshold values of purity, and the vertical axes represents the logarithmic number of substrings which have larger purity values than the threshold values.

how to see the maps, and Fig. 3, 4, 5 and 6 shows some sequence maps out of them.

A map shows a sequence, and each row consists of three tracks. The first track provides annotation information written in a GenBank file. The all annotated regions were classified into seven functional types by simply checking whether some specific terms are included in feature qualifier values or not. For example, regions whose annotations contain either qualifiers "function", "product" or "note" with qualifier values including a word "phage" are classified as *phage*. The detailed description of the complete rules are omitted here. Then, the annotated regions were colored according to their types on a map. The seven types and their corresponding colors are: mobile element (red), rearrangement hot spot element (*Rhs*[7,12]; orange), phage (yellow), transposon and transposase (green), tRNA (light blue), rRNA (blue) and others (transparent, i.e. not shown in maps). Those six types of functional regions are considered to be horizontally transferred genes or related to insertion of such genes.

The second track shows local G+C contents for each non-overlapping l-mers, i.e. total ratios of "g" and "c" in the sequence of each l-mer, where $l = 1000$. Every l-mer is colored according to their G+C contents: l-mer with highest G+C content is colored in red, and l-mer with lowest G+C content is colored in green. This coloring is also shown in the legend located at the bottom of a map. This track is intended to provide a simple compositional information, which are often used to

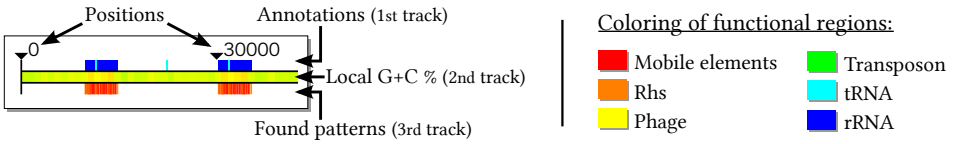

Fig. 2. How to see the sequence maps below. The diagram on the left hand side shows the part of a sequence map and describes the structure. In the first track of a sequence map, some functional regions are drawn with the colors as shown in right hand side of this figure.

characterize horizontally transferred genes[9].

Finally, the third track shows Blumer strings. For each string, all corresponding regions were located and colored according to the purity value of the string. Since the purity values also ranges between from zero to one as with G+C content, they are colored in the same way: a string with high purity value is colored in red, and a string with low purity value is colored in green. Note that only Fig. 3 shows all Blumer strings while other sequence maps doesn't show Blumer strings with smaller purity values than 0.5.

In the figures, it is shown that the purity measure captures RNAs, transposons and Rhss very well. Though those substrings are usually not wide enough to entirely cover every functional regions, short substrings covers those regions together. Also phages seems to be captured by the purity measure although some part of the regions are not covered.

It can also seen that purity and local G+C content are not closely related. While, in any figures, the purity measure mostly gives higher values only to horizontally transferred genes and the variation of the purity is very specific to those regions, the G+C contents vary regardless of whether the regions are horizontally transferred genes or not. Hence, it can be said that the local G+C content is not enough to characterize horizontally transferred genes.

Since we combined Blumer strings and the purity measure, it may be considered that the successful results are obtained mainly by Blumer strings instead of the purity measure. Note that Blumer strings are interspersed all over the sequence in Fig. 3 although most of them have lowest purity values, i.e. colored green. In the figure, it is shown that Blumer strings are not so biased. Hence, Blumer strings are not specific to regions of horizontally transferred genes, and we can conclude that horizontally transferred genes are captured thanks to the purity measure.

4.2. *Function-wise Analysis*

We further examined the purity values to quantitatively show the association between functions of regions and them. First, for every region annotated in GenBank files, the most optimal string that best covers only the single region was chosen from *all* the Blumer strings of the genome sequence. Then, the selected strings were split into groups according to their corresponding regions' function types, and statistics

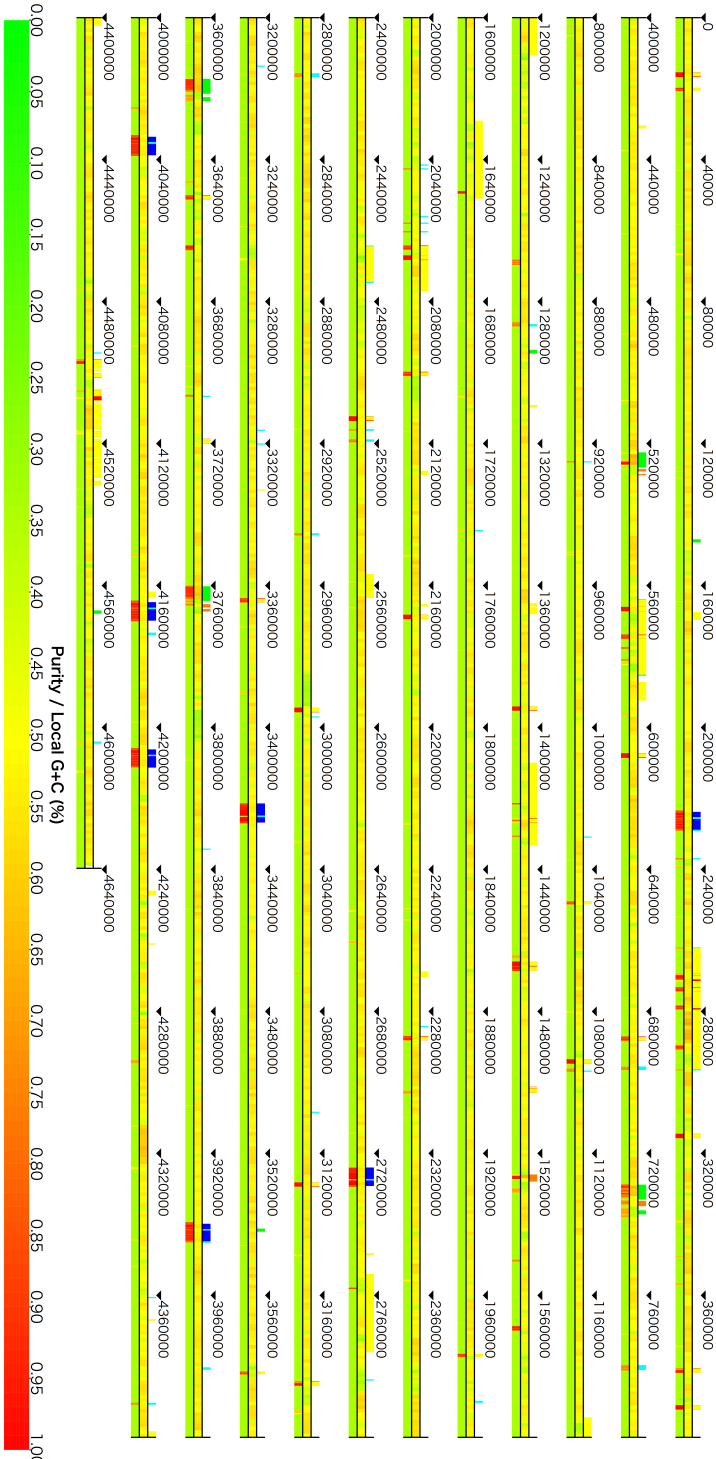

Fig. 3. All Blumer strings rendered according to their purity values on the sequence of *Escherichia coli str. K-12* [GenBank:NC_000913].

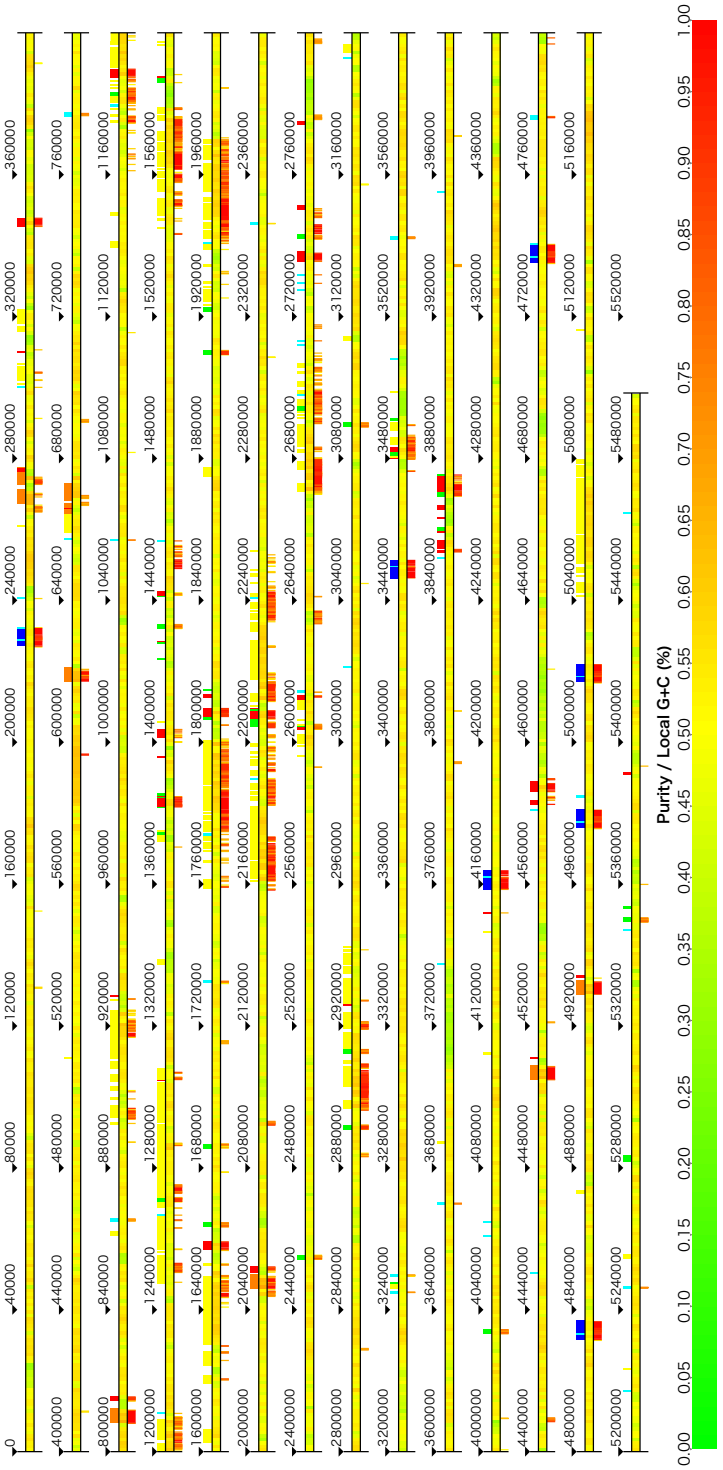

Fig. 4. Blumer strings whose purity values are greater than 0.5 on the sequence of *Escherichia coli O157:H7 str. Sakai* [GenBank:NC_002695].

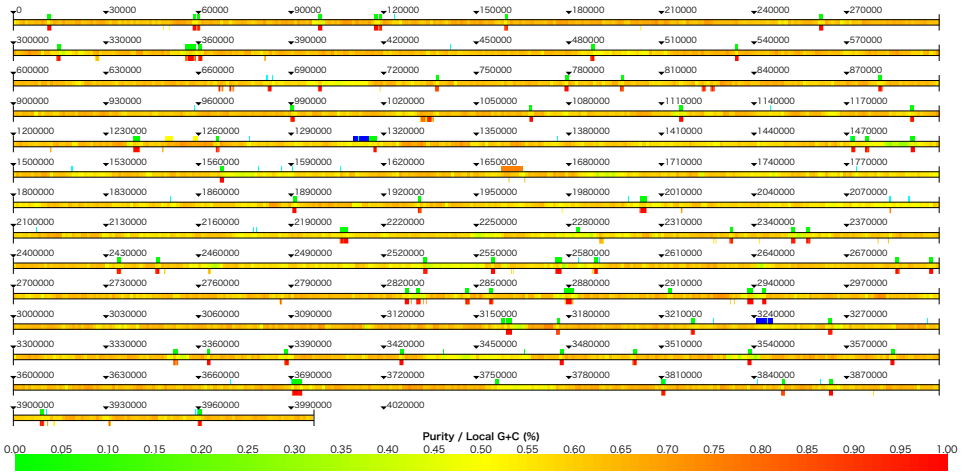

Fig. 5. Blumer strings whose purity values are greater than 0.5 on the sequence of *Geobacter metallireducens GS-15* [GenBank:NC_007517].

Fig. 6. Blumer strings whose purity values are greater than 0.5 on the sequence of *Mycoplasma mycoides subsp. capri LC str. 95010* [GenBank:NC_015431].

of the associations were analyzed. Here, the same seven types of functions described above were used.

To evaluate the coverage of a string for an annotated region, the *F-measure* was employed, which is a well-known measure and widely used in the field of Information Retrieval. Given a genome sequence $T \in \Sigma^*$, a candidate substring x of T and a functional region $r = \langle i, j \rangle \in sub(T)$, we define the F-measure $F(x, r)$ for x and r

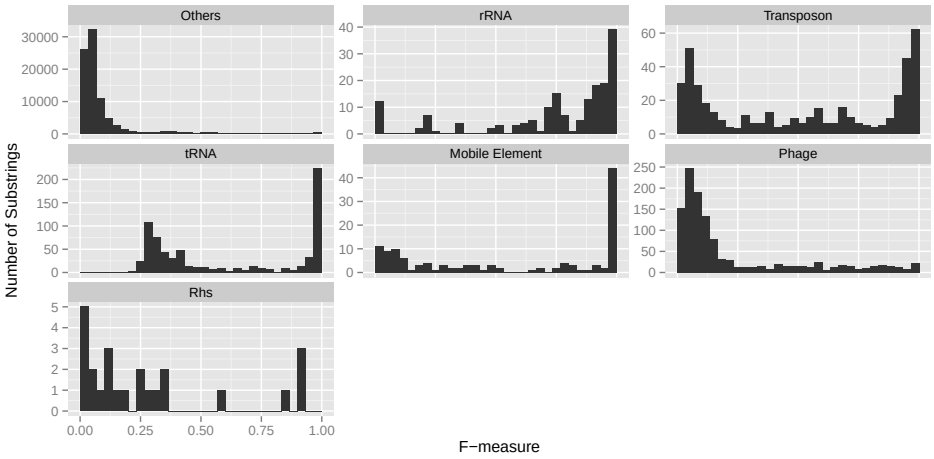

Fig. 7. For each functional type, here are the histograms of the F-measure values of *all* the pairs of a region and its optimal Blumer's string from all the genome sequences. The horizontal axes represent F-measure values, and the vertical axes represent the total frequencies of F-measure values within each "bin" of F-measure value.

as follows:

$$F(x,r) = \max\left\{ H\left(\frac{overlap(\langle i,j\rangle,\langle k,l\rangle)}{j-i}, \frac{overlap(\langle i,j\rangle,\langle k,l\rangle)}{l-k} \right) \;\middle|\; \forall\langle k,l\rangle \in pos_T(x) \right\},$$

where

$$H(p,q) = \frac{2pq}{p+q},$$

$$overlap(\langle a,b\rangle,\langle c,d\rangle) = \begin{cases} \min(b,d) - c + 1 & (a \le c \le b) \\ \min(b,d) - a + 1 & (c \le a \le d). \end{cases}$$

Briefly speaking, the value of F-measure will be one if and only if a string x covers the entire region r, and the value will be zero when x and r are not overlapped.

The distribution of the F-measure values for all the annotated regions is shown in Fig. 7. It shows histograms of the *F-measure values* of *all* the pairs of a region and its optimally corresponding Blumer's string from all the genome sequences. In a histogram, the horizontal axis express F-measure value, and the vertical axes express the total frequency of F-measure values within each "bin" of F-measure value. From the figure, it can be seen that there are still a lot of strings that poorly cover annotated regions even though they are the optimally covering strings for the regions. This means that most of the annotated regions, except for rRNAs, have very different boundaries from those of Blumer strings. In contrast, interestingly, half of rRNAs have corresponding Blumer strings with quite high F-measure values greater than 0.87, and 84% of rRNAs have Blumer strings with F-measure values higher than 0.5.

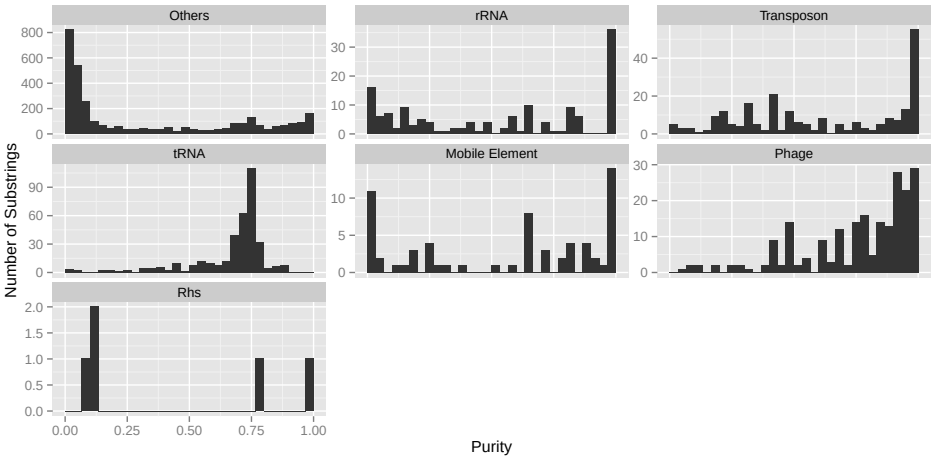

Fig. 8. Purity value histograms for annotated regions with high F-measure values (≥ 0.5). A distribution of purity values across all the genome sequences is shown for each function type. The horizontal axes represents purity values, and the vertical axes represents the number of regions.

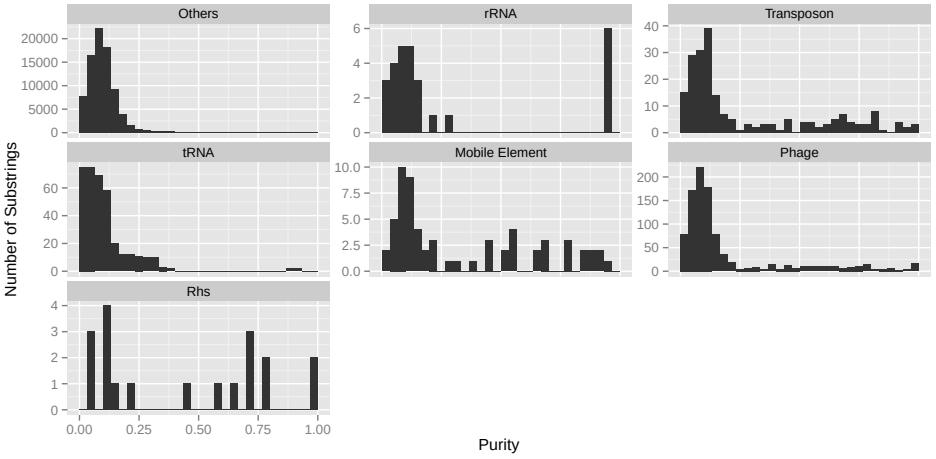

Fig. 9. Purity value histograms for annotated regions with low F-measure values (< 0.5). A distribution of purity values across all the genome sequences is shown for each function type. The horizontal axes represents purity values, and the vertical axes represents the number of regions.

It is considered that only strings with high F-measure values are subject to the functions of their corresponding regions. Hence, we divided those optimal strings into two groups of high F-measure values (≥ 0.5) and low F-measure values (< 0.5).

Figure 8 and 9 show the histograms of *purity values* for both the groups: the former is for the group of high F-measure values and the latter is for the other group

Table 2. The numbers of Blumer strings categorized according to purity values and their corresponding functional types. Here only strings with high F-measure values (≥ 0.5) are considered.

	HTGs	Others
Purity ≥ 0.5	734	1035
Purity < 0.5	286	2209

of low F-measure values. In the histograms, the horizontal axes express purity value, and the vertical axes express the total frequency of purity values. In Fig. 8, it is shown that most of the strings that matches horizontally transferred genes have high purity values, and the other strings have low purity values. On the other hand, in Fig. 9, the purity measure gives low evaluations to almost all substrings even if the substring matches a horizontally transferred genes a little. Hence, it is likely that the purity measure gives high purity values to substrings which covers horizontally transferred genes well and low purity values to the other substrings.

This observation was tested statistically by being stated as follows: "F-measure values and the purity values are positively-correlated in the case of horizontally transferred genes and not correlated in the case of other genes". We statistically tested this hypothesis on these functional types except for "others". Spearman's rank correlation of the F-measure values and the purity values were computed for the both cases, and tested a null hypothesis "correlation is less than or equal to 0" respectively. For horizontally transferred genes, a correlation coefficient of 0.58 and a p-value less than 0.01 are obtained. Thus we can say there are correlation between the F-measure values and the purity values under the choice of significance level of 1%. On the other hand, for the other genes, a correlation coefficient of -0.18 and p-value of 1 were obtained. Hence, we can not say F-measure values and the purity values of regions other than horizontally transferred genes are positively-correlated.

It is also calculated that how many of the horizontally transferred genes have purity values no less than 0.5, how many of the other genes have purity values less than 0.5 and so on. Table 2 shows these numbers for the four cases. Note that only the Blumer strings with F-measure values no less than 0.5 are considered here. From the table, it can be seen that Blumer strings with high purity value are not necessarily associated with horizontally transferred genes and the false discovery rate of 0.59 is fairly high. On the other hand, it can also be said that 72% of Blumer strings which correspond to horizontally transferred genes have high purity values and 68% of them corresponding to the other genes have low purity values.

From these results, we conclude that the purity measure certainly characterizes horizontally transferred genes collectively as unusual regions of genome sequences while it is not sufficient to distinguish horizontally transferred genes from the others.

5. Conclusions

We extensively investigated the usefulness of the purity measure for genome sequences. Interestingly, our sequence maps indicated visually the considerable association between the purity values of Blumer strings and several horizontally transferred genes. On the other hand, it also showed that G+C contents are not so suggestive of these genes. Our quantitative examination on Blumer strings showed that particularly rRNAs are likely to be covered well by these strings and there is significant correlation between the functionality of horizontally transferred genes and purity values of corresponding Blumer strings. Moreover, around 70% of annotated regions with high F-measure values can be correctly classified into horizontally transferred genes or other genes according to purity values of their corresponding Blumer strings while the false discovery rate of 0.59 is not so low.

Our analysis suggests that the purity measure characterizes horizontally transferred genes collectively. Therefore, this fact means that the purity measure has potential to be a useful tool to predict horizontally transferred genes.

References

1. Akhter S, Aziz RK, Edwards RA, PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies, *Nucleic Acids Research* **40**(16):e126, 2012. doi:10.1093/nar/gks406.
2. Bergman CM, Quesneville H, Discovering and detecting transposable elements in genome sequences, *Briefings in Bioinformatics* **8**(6):382–392, 2007. doi:10.1093/bib/bbm048.
3. Blumer A, Blumer J, Haussler D, Ehrenfeucht A, Chen MT, Seiferas J, The smallest automation recognizing the subwords of a text, *Theoretical Computer Science* **40**:31–55, 1985.
4. Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P, Detection and characterization of horizontal transfers in prokaryotes using genomic signature, *Nucleic Acids Research* **33**(1):e6, 2005. doi:10.1093/nar/gni004.
5. Durbin R, Eddy S, Krogh A, Mitchison G, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998. ISBN 9780521629713.
6. Dutta C, Pan A, Horizontal gene transfer and bacterial diversity, *Journal of Biosciences* **27**(1):27–33, 2002. doi:10.1007/BF02703681.
7. Feulner G, Gray JA, Kirschman JA, Lehner AF, Sadosky AB, Vlazny DA, Zhang J, Zhao S, Hill CW, Structure of the rhsA locus from Escherichia coli K-12 and comparison of rhsA with other members of the rhs multigene family, *Journal of Bacteriology* **172**(1):446–456, 1990.
8. Fouts DE, Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences, *Nucleic Acids Research* **34**(20):5839–5851, 2006. doi:10.1093/nar/gkl732.
9. Garcia-Vallv S, Romeu A, Palau J, Horizontal gene transfer in bacterial and archaeal complete genomes, *Genome Research* **10**(11):1719–1725, 2000. doi:10.1101/gr.130000.
10. Gusfield D, *Algorithms on Strings, Trees and Sequence*, Cambridge University Press, New York, 1997. ISBN 9780521585194.
11. Heinemann J, Kurenbach B, Horizontal transfer of genes between microorganisms,

in *Encyclopedia of Microbiology (Third Edition)*, ed., in Chief: Moselio Schaechter E, third edition ed., Academic Press, Oxford, pp. 597–606, 2009. doi:10.1016/B978-012373944-5.00009-2.

12. Hill CW, Sandt CH, Vlazny DA, Rhs elements of Escherichia coli: a family of genetic composites each encoding a large mosaic protein, *Molecular Microbiology* **12**(6):865–871, 1994.
13. Ikeda D, Suzuki E, Mining peculiar compositions of frequent substrings from sparse text data using background texts, in *Machine Learning and Knowledge Discovery in Databases*, eds., Buntine W, Grobelnik M, Mladeni D, Shawe-Taylor J, Springer Berlin Heidelberg, pp. 596–611, 2009. doi:10.1007/978-3-642-04180-8_56.
14. Kargar M, An A, Evaluation of different complexity measures for signal detection in genome sequences, *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, ACM, New York, NY, USA, pp. 422–425, 2010. ISBN 978-1-4503-0438-2. doi:10.1145/1854776.1854845.
15. Kidwell M, Horizontal transfer of P elements and other short inverted repeat transposons, *Genetica* **86**(1-3):275–286, 1992. doi:10.1007/BF00133726.
16. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R, Prophinder: a computational tool for prophage prediction in prokaryotic genomes, *Bioinformatics* **24**(6):863–865, 2008. doi:10.1093/bioinformatics/btn043.
17. McCarthy EM, McDonald JF, LTR_STRUC: a novel search and identification program for LTR retrotransposons, *Bioinformatics* **19**(3):362–367, 2003. doi:10.1093/bioinformatics/btf878.
18. Pride DT, Blaser MJ, Identification of horizontally acquired genetic elements in helicobacter pylori and other prokaryotes using oligonucleotide difference analysis, *Genome Letters* **1**(1):2–15, 2002.
19. Quesneville H, Nouaud D, Anxolabhre D, Detection of new transposable element families in drosophila melanogaster and anopheles gambiae genomes, *Journal of Molecular Evolution* **57**(1):S50–S59, 2003. doi:10.1007/s00239-003-0007-2.
20. Rho M, Choi JH, Kim S, Lynch M, Tang H, De novo identification of LTR retrotransposons in eukaryotic genomes, *BMC Genomics* **8**(1):90, 2007. doi:10.1186/1471-2164-8-90.
21. Srividhya KV, Alaguraj V, Poornima G, Kumar D, Singh GP, Raghavenderan L, Katta AVSKM, Mehta P, Krishnaswamy S, Identification of prophages in bacterial genomes by dinucleotide relative abundance difference, *PLoS ONE* **2**(11):e1193, 2007. doi:10.1371/journal.pone.0001193.
22. Tsirigos A, Rigoutsos I, 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, 2005. doi:10.1093/nar/gki660.
23. Volfovsky N, Haas B, Salzberg S, A clustering method for repeat analysis in DNA sequences, *Genome Biology* **2**(8):research0027.1–research0027.11, 2001. doi:10.1186/gb-2001-2-8-research0027.
24. Wan H, Wootton JC, A global compositional complexity measure for biological sequences: AT-rich and GC-rich genomes encode less complex proteins, *Computers & Chemistry* **24**(1):71–94, 2000. doi:10.1016/S0097-8485(00)80008-X.
25. Yamada Y, Nakatoh T, Baba K, Ikeda D, Mining pure patterns in texts, *IIAI International Conference on Advanced Applied Informatics*, pp. 285 –290, 2012. doi:10.1109/IIAI-AAI.2012.75.
26. Yap WH, Zhang Z, Wang Y, 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):5201–5209, 1999.