Tandem Repeats in Prokaryotic Genomes

Jiewen Liu[1]
srjiewen@si.hirosaki-u.ac.jp

Hikaru Munakata[1]
s498056@si.hirosaki-u.ac.jp

Abulimiti Aimaiti[1]
srabuli@si.hirosaki-u.ac.jp

Akito Taneda[1,2]
taneda@si.hirosaki-u.ac.jp

Kenji Oosawa[3,4]
kenji@math.nagoya-u.ac.jp

Toshio Shimizu[1,2]
slsimi@si.hirosaki-u.ac.jp

[1] Department of Information Science, Graduate School of Science, Hirosaki University, 3 Bunkyo-cho, Hirosaki 036-8561, Japan
[2] Department of Electronic and Information System Engineering, Faculty of Science and Technology, Hirosaki University, 3 Bunkyo-cho, Hirosaki 036-8561, Japan
[3] Graduate School of Mathematics, Nagoya University, Chikusa-ku, Nagoya 464-8602, Japan
[4] Protonic NanoMachine Project, ERATO, JST, 1- 7 Hikaridai, Seika, Souraku, Kyoto 619-0237, Japan

Keywords: color-coding method, individual difference, long tandem repeat, Java applet

1 Introduction

Genomes contain many perfect and approximate tandem repeats. In general, tandem repeats are generated by duplication which changes genome structures, so that it is one of the important processes of genome evolution. So far, many detection tools for tandem repeats have been proposed. In the present study, by using color-coding method proposed by Yoshida *et al.* [4], we have detected tandem repeats which are relatively long and have low sequence similarity. Color-coding method is a visualization technique in which nucleotides are substituted by squares with user-defined colors corresponding to a nucleotide type, and the color-coded sequence is aligned to two dimensions with a column width. When a tandem repeat exists in the sequence region, the tandem repeat can be recognized as some recognizable pattern. In literature, we can find papers in which prokaryotic tandem and interspersed repeats are studied [1, 3]. In the papers, on the basis of the analysis on repeat length, sequence identity, and the position in the genome it has been proposed that 1) tandem repeats are the first step of interspersed repeats (not including the sequences with different origin such as transposons), 2) tandem repeats control the expression and /or the structure of immune-related proteins by varying the number of repeats, and so on. Thus, systematic survey of tandem repeats in prokaryotic genomes is important for studying evolution and functions of prokaryotic genomes.

2 Repeats Mining and Results

For the detail of color-coding method, see the original paper [4]. In the present study, tandem repeats mining is performed by visual-inspection of well-trained three operators. However, it turned out that the results obtained by the operators show differences due to individual difference. This is because in a color-coded sequence the boundaries between tandem repeats and flanking non-repetitive regions of the sequence can be ambiguous, particularly this is the case when the sequence identity is low and/or the length of the repeat unit is long. To clarify the individual difference in color-coding, we compare the tandem repeats of three species (*H. influenzae*, *D. radiodurans*, and *Y. pestis*) detected by the operators and determine criterions (e.g., for the tandem repeats with a unit length longer than 100 bp and a sequence identity higher than 60 % individual differences are small). In addition, we compare the results of the color-coding and tandem repeats finder [2] and it is found that color-coding

can detect more tandem repeats for those with long unit length (longer than 30 bp). Table 1 shows summary of the tandem repeats we have found in thirty prokaryotic genomes. In the present study, a Java applet for color-coding method has also been developed. An example screen shot of a beta version of the Java applet is shown in Fig. 1. The tandem repeats data and the Java color-coding viewer will be available from `http://bioinfo.si.hirosaki-u.ac.jp/~repeat/`.

Table 1: The total number and density of tandem repeats for thirty prokaryotic genomes. Density is defined as (total number of repeats) / (genome size) and its unit is per Mbp. This table is slightly redundant because of individual differences of operators.

	species	total	density
Bacteria	*A. aeolicus*	208	134
	B. halodurans	317	75
	B. subtilis	238	56
	B. burgdorferi	131	143
	B sp. APS	206	322
	C. jejuni	365	222
	C. muridarum	153	143
	C. trachomatis	155	149
	C. pneumoniae CWL029	163	132
	D. radiodurans	1082	409

	species	total	density
Bacteria	*E. coli K12*	51	11
	H. influenzae Rd	72	39
	H. pylori 26695	97	58
	H. pylori J99	180	110
	L. lactis subsp. Lactis	294	124
	M. genitalium	189	325
	M. pneumoniae	154	189
	M. pulmonis	41	43
	R. conorii	212	167
	R. prowazekii	119	107

	species	total	density
Bacteria	*S. sp. PCC 6803*	383	107
	T. pallidum	174	153
	U. urealyticum	175	232
	V. cholerae chr.1	147	50
	Y. pestis	983	211
Archaea	*A. pernix*	240	144
	A. fulgidus	324	149
	M. jannaschii	663	398
	T. acidophilum	275	176
	T. volcanium	309	195

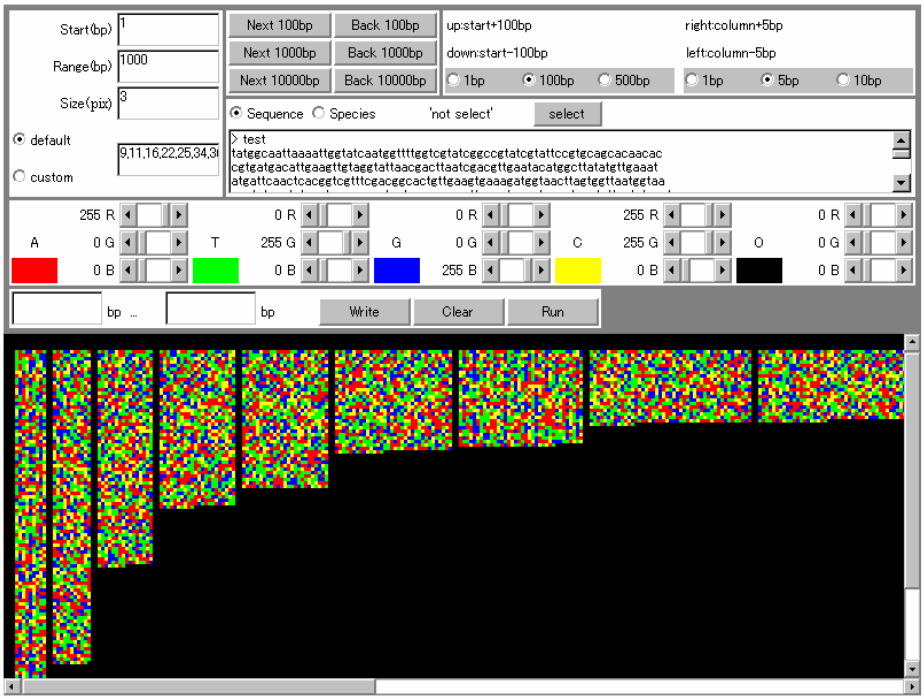

Figure 1: A screen shot of Java color-coding viewer. Users can change color, width and other parameters.

References

[1] Achaz, G., Rocha, E.P.C., Netter, P., and Coissac, E., Origin and fate of repeats in bacteria, *Nucleic Acids Res.*, 30:2987–2994, 2002.

[2] Benson, G., Tandem repeats finder: A program to analyze DNA sequences, *Nucleic Acids Res.*, 27:573–580, 1999.

[3] Rocha, E.P.C. and Blanchard, A., Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution, *Nucleic Acids Res.*, 30:2031–2042, 2002.

[4] Yoshida, T., Obata, N., and Oosawa, K., Color-coding reveals tandem repeats in the *Esherichia coli* genome, *J. Mol. Biol.*, 298:343–349, 2000.