

Tandem Repeats in 44 Prokaryotic Genomes and Genome Evolution

Satoshi Mizuta¹

slmizu@cc.hirosaki-u.ac.jp

Hikaru Munakata¹

gs02416@si.hirosaki-u.ac.jp

Abulimiti Aimaiti¹

gs02408@si.hirosaki-u.ac.jp

Kenji Oosawa²

kenji@nms.gunma-u.ac.jp

Toshio Shimizu¹

slsimi@si.hirosaki-u.ac.jp

¹ Faculty of Science and Technology, Hirosaki University, Hirosaki 036-8561, Japan

² Department of Nano-Material Systems, Graduate School of Engineering, Gunma University, Kiryu 376-8515, Japan

Keywords: tandem duplication, genome evolution, genome rearrangement, whole genome duplication

1 Introduction

Genomes are full of various types of repetitive sequences, such as microsatellite, minisatellite, interspersed repeats, tandem repeats, and so on. These repetitive sequences are replete with information that would bring about a better understanding of not only genome evolution but also gene evolution. A tandem repeat (TR) is a DNA sequence piece containing two or more consecutive homologous elements that were arisen as a result of tandem duplication. It is of particular interest and importance with its relation to the gene arrangement or evolution. A TR found across a few coding regions or within a coding region should contain some traces of the gene rearrangement such as gene duplication, gene recombination, or internal gene duplication, which are considered to be the principle mechanisms of protein repertoire expansion[1, 2, 3]. In this study, we tried to investigate genome evolutions by analyzing equivalent TR-pairs detected in 44 prokaryotic genomes.

2 Method and Results

Table 1 shows the 44 prokaryotic genomes analyzed in this article. They were selected as representatives of the individual lineages in the phylogenetic tree of prokaryotic species. Their sequence data were downloaded from GenBank[5] and TRs were detected on them by three inspectors through the CC method[4]. The numbers of the detected TRs for each genome are indicated at the last column of Table 1.

All the detected TRs were compared pairwise by SSEARCH[6] with the default parameter values, match(5), mismatch(-4), gap creation penalty(-16), and gap extension penalty(-4), and the option of E-value ≤ 0.001 . Among the results, those pairs in which the proportion of the overlap to the longer sequence is larger than 0.5 are identified as equivalent TR-pairs.

The equivalent TR-pairs found in each genomes were arranged on a circle which symbolically expressed the genome, and the central angles of the TR-pairs were calculated. If a genome had experienced a whole genome duplication once (1R), there may be TR-pairs which have central angles around 180° in the genome; if twice (2R), there may be those with the central angles around 90° and 270° as well as 180° , and so on.

Fig. 1 shows the central angle distribution for *S. pneumoniae*, where the evidence of 2R whole genome duplication is explicitly recognized.

3 Discussions

We analyzed the equivalent TR-pairs and found the evidences of 2R whole genome duplication on 16 genomes. For further investigation in such as protein repertoire expansion, however, the relationship between TRs and coding regions must be studied.

References

- [1] Shimizu, T., Mitsuke, H., Noto, K., and Arai, M., Internal Gene Duplication in the Evolution of Prokaryotic Transmembrane Proteins, *J. Mol. Biol.*, 339:1–15, 2004.
- [2] Teichmann, S.A., Rison, S.C.G, Thornton, J.M., Riley, M., Gough, J., and Chothia, C., The Evolution and Structural Anatomy of the Small Molecule Metabolic Pathways in *Escherichia coli*, *J. Mol. Biol.*, 311:693–708, 2001.
- [3] Vogel, C., Berzuini, C., Bashton, M., Gough, J., and Teichmann, S.A., Supra-domains: Evolutionary Units Larger than Single Protein Domains, *J. Mol. Biol.*, 336:809–823, 2004.
- [4] Yoshida, T., Obata, N., and Oosawa, K., Color-coding reveals tandem repeats in the *Escherichia coli* genome, *J. Mol. Biol.*, 298:343–349, 2000.
- [5] <http://www.ncbi.nlm.nih.gov/>
- [6] <http://fasta.bioch.virginia.edu/>

Table 1: List of genomes analyzed. The third column gives the numbers of the detected TRs for the individual genomes.

Species	Size(Mb)	# TRs
Proteobacteria		
<i>Brucella melitensis</i> 16M	2.12	224
<i>Buchnera</i> sp. APS	0.64	206
<i>Campylobacter jejuni</i> NCTC11168	1.64	332
<i>Escherichia coli</i> K-12 MG1655	4.64	299
<i>Haemophilus influenzae</i> Rd	1.83	59
<i>Helicobacter pylori</i> J99	1.64	155
<i>Helicobacter pylori</i> 26695	1.67	89
<i>Neisseria meningitidis</i> MC58	2.27	567
<i>Rickettsia conorii</i> Malish 7	1.27	183
<i>Rickettsia prowazekii</i> Madrid E	1.11	113
<i>Vibrio cholerae</i> El Tor N16961	2.96	134
<i>Yersinia pestis</i> CO92	4.65	901
Firmicutes		
<i>Bacillus halodurans</i> C-125	4.20	266
<i>Bacillus subtilis</i> 168	4.21	230
<i>Clostridium perfringens</i> 13	3.03	1072
<i>Lactococcus lactis</i> IL1403	2.37	263
<i>Mycoplasma genitalium</i> G-37	0.58	126
<i>Mycobacterium leprae</i> TN	3.27	282
<i>Mycoplasma pneumoniae</i> M129	0.82	129
<i>Mycoplasma pulmonis</i>	0.96	37
<i>Staphylococcus aureus</i> N315	2.81	581
<i>Streptococcus pneumoniae</i> R6	2.04	436
<i>Streptococcus pyogenes</i> SF370	1.85	323
<i>Ureaplasma urealyticum</i>	0.75	154
Other Bacteria		
<i>Aquifex aeolicus</i> VF5	1.55	180
<i>Borrelia burgdorferi</i> B31	0.91	116
<i>Chlamydia muridarum</i>	1.07	130
<i>Chlamydia pneumoniae</i> CWL029	1.23	152
<i>Chlorobium tepidum</i> TLS	2.15	463
<i>Chlamydia trachomatis</i>	1.04	128
<i>Deinococcus radiodurans</i> R1	2.65	1003
<i>Synechocystis</i> sp. PCC6803	3.57	366
<i>Thermosynechococcus elongatus</i> BP-1	2.59	235
<i>Thermotoga maritima</i> MSB8	1.86	216
<i>Treponema pallidum</i> Nichols	1.14	174
Archaea		
<i>Archaeoglobus fulgidus</i> DSM4304	2.18	322
<i>Aeropyrum pernix</i> K1	1.67	209
<i>Methanococcus jannaschii</i> DSM2661	1.66	589
<i>Methanopyrus kandleri</i> AV19	1.69	662
<i>Methanobacterium thermoautotrophicum</i> deltaH	1.75	315
<i>Pyrobaculum aerophilum</i> IM2	2.22	308
<i>Pyrococcus horikoshii</i> OT3	1.74	230
<i>Thermoplasma acidophilum</i>	1.56	275
<i>Thermoplasma volcanium</i> GSS1	1.58	308

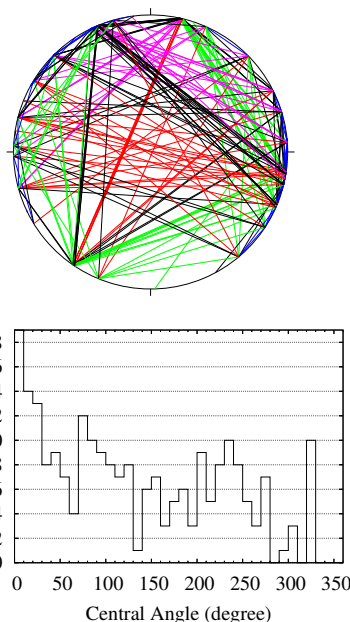


Figure 1: Central angle distribution of the equivalent TR-pairs for the *S. pneumoniae* genome. In the upper diagram, the circle symbolically expresses the genome, and blue, green, red, magenta, and black lines correspond to the central angles of less than 30° , between 60° and 120° , between 150° and 210° , between 240° and 300° , and others, respectively.