Common Repeat Sequences in Bacterial Genomes

Yo-Cheng Chang Chuan-Hsiung Chang*

Institute of Bioinformatics, National Yang-Ming University, Taipei, Taiwan, 112, ROC

Received 10 May 2003; Accepted 10 June 2003

Abstract

Repeat sequences are widely appearing in bacterial genomes. However, detailed functions and mechanisms of most repeats are still unknown. We explored four types of repeats in 58 fully sequenced bacterial genomes, including 60 chromosomal and 22 plasmid sequences. For each bacterial genome sequence, the repeat density in number and in length were both calculated and compared, and potential common repeats of four repeat types were also recognized. The results showed most chromosomal sequences contain repeat sequences and a large number of plasmid sequences do not have repeat sequences. The density of repeats is not correlated with either the current taxonomy or the genome size of bacteria. Only the reverse complement repeats are common in most bacterial genomes. The four types of repeats might be different in terms of functions, evolutions and mechanisms. For further investigations, the development of a new methodology with combined statistical analysis and detailed biological knowledge of each bacterial organism will be needed.

***Keywords*:** Bacterial genome, Repeat sequence

Introduction

DNA repeats can be defined as sequences sharing extensive similarity with other sequences of the same genome. It is usually assumed that repeats arise by successive duplications and with several causal mechanisms. Usually, repeats in bacteria are divided into two subclasses: low-complexity short repeats and longer repeats. The first category consists of short nucleotides, typically ranging from 3-5 nucleotides in length, repeated many times in a contiguous region. These low-complexity repeats are very abundant in certain eukaryotic genomes, in which they have been widely studied [1]. Although less abundant in archaeal and bacterial genomes [2], the mechanisms of their origin [3], their function [4], the consequences for genome dynamics [5], and the structural constraints imposed on the chromosome [6] have all being studied.

Longer repeats include transposable elements, minisatellites (mostly in eukaryotic genomes), large tandem repeats and spaced repeats. They are widely distributed in archaeal and bacterial genomes. These longer repeats can be divided into four types: forward (direct) repeats, reverse (inverted) repeats, complement repeats, and reverse complement (palindromic) repeats [7, 8]. Forward repeats are two or more repeating DNA fragments appeared in the same orientation and on the same strand. For example, a sequence fragment ATGGC and another sequence fragment ATGGC are forward repeats. Reverse repeats are two or more repeating DNA fragments appeared in inverted orientation and on the same strand. For example, a sequence fragment ATGGC and another sequence fragment CGGTA are reverse repeats. Complement repeats are two or more repeating DNA fragments appeared in the same orientation but on the complement strand, i.e. in complement orientation on the same strand. For example, a sequence fragment ATGGC and another sequence fragment TACCG are reverse repeats. Reverse complement repeats are two or more repeating DNA fragments appeared in inverted orientation but on the complement strand, i.e. in reverse complement orientation on the same strand. For example, a sequence fragment ATGGC and another sequence fragment GCCAT are reverse complement repeats.

More recently, a number of studies have supported the notion that repeats are likely to be a highly significant source of informative markers for bacterial genomes [9-13]. This probably reflects the important contribution of repeats to the adaptation of bacteria. Repeats appear to contribute to phenotypic variation in bacteria in at least two ways. Repeats located within the regulatory region of a gene can constitute an on/off switch of gene expression at the transcriptional level [14, 15]. Similarly, repeats located within coding regions can induce a reversible premature ending of translation when a mutation changes the number of repeats [16-18]. However, detailed mechanisms and functions of most repeats are still unknown.

The main question we tackled in this work concerns the common repeat sequences of bacterial genomes. We recognized four types of repeats in fully sequenced bacterial

*Corresponding author: Chuan-Hsiung Chang
Tel: +886-2-28267316; Fax: +886-2-28206754
E-mail: cchang@ym.edu.tw

Table 1. Bacterial genomes analyzed. (a) list of 60 chromosomal sequences (b) list of 22 plasmid sequences.

1. *Aeropyrum pernix* K1 complete genome	31. *Mycobacterium leprae* strain TN complete genome
2. *Agrobacterium tumefaciens* strain C58 circular chromosome; complete	32. *Mycobacterium tuberculosis* complete genome
3. *Aquifex aeolicus* complete genome	33. *Mycoplasma genitalium* G37 complete genome
4. *Archaeoglobus fulgidus* complete genome	34. *Mycoplasma pneumoniae* M129 complete genome
5. *Bacillus halodurans* C-125; complete genome	35. *Mycoplasma pulmonis* (strain UAB CTIP) complete genome
6. *Bacillus subtilis* complete genome	36. *Neisseria meningitidis* serogroup B strain MC58 complete genome
7. *Buchnera sp* APS complete genome	37. *Pasteurella multocida* PM70 complete genome
8. *Campylobacter jejuni* complete genome	38. *Pseudomonas aeruginosa* PA01; complete genome
9. *Caulobacter crescentus* complete genome	39. *Pyrococcus abyssi* complete genome
10. *Chlamydia muridarum*; complete genome	40. *Pyrococcus horikoshii* OT3 complete genome
11. *Chlamydia pneumoniae* complete genome	41. *Rickettsia conorii* Malish 7; complete genome
12. *Chlamydia trachomatis* complete genome	42. *Rickettsia prowazekii* strain Madrid E; complete genome
13. *Chlamydophila pneumoniae* AR39; complete genome	43. *Sinorhizobium meliloti* 1021 complete genomes
14. *Chlamydophila pneumoniae* J138; complete genome	44. *Staphylococcus aureus* strain Mu50; complete genome
15. *Clostridium acetobutylicum* ATCC824 complete genome	45. *Staphylococcus aureus* strain N315; complete genome
16. *Deinococcus radiodurans* R1 complete chromosome 1	46. *Streptococcus pneumoniae* complete genome
17. *Deinococcus radiodurans* R1 complete chromosome 2	47. *Streptococcus pneumoniae* R6 complete genome
18. *Escherichia coli* K-12 MG1655 complete genome	48. *Streptococcus pyogenes* strain SF370 serotype M1; complete genome
19. *Escherichia coli* O157:H7 EDL933; complete genome	49. *Sulfolobus solfataricus* complete genome
20. *Escherichia coli* O157:H7; complete genome	50. *Sulfolobus tokodaii* complete genome
21. Genomic sequence of a Lyme disease spirochete; *Borrelia burgdorferi*	51. *Synechocystis* PCC6803 complete genome
22. *Haemophilus influenzae* Rd complete genome	52. *Thermoplasma acidophilum*; complete genome
23. *Halobacterium sp* NRC-1 complete genome	53. *Thermoplasma volcanium*; complete genome
24. *Helicobacter pylori* 26695 complete genome	54. *Thermotoga maritima* complete genome
25. *Helicobacter pylori*; strain J99 complete genome	55. *Treponema pallidum* complete genome
26. *Lactococcus lactis* subsp lactis IL1403 complete genome	56. *Ureaplasma urealyticum* complete genome
27. linear chromosome; complete sequence	57. *Vibrio cholerae* chromosome I; complete chromosome
28. *Mesorhizobium loti* complete genome; complete sequence	58. *Vibrio cholerae* chromosome II; complete chromosome
29. *Methanobacterium thermoautotrophicum* delta H complete genome	59. *Xylella fastidiosa*; complete genome
30. *Methanococcus jannaschii* complete genome	60. *Yersinia pestis* strain CO92; complete genome

(a)

1. *Aquifex aeolicus* plasmid ece1; complete plasmid sequence.	12. *Methanococcus jannaschii* small extra-chromosomal element; complete
2. *Buchnera sp.* APS plasmid pLeu DNA; complete sequence.	13. plasmid Ti; complete sequence.
3. *Buchnera sp.* APS plasmid pTrp DNA; complete sequence.	14. *Sinorhizobium meliloti* plasmid pSymA complete plasmid sequence.
4. *Chlamydia muridarum* plasmid pMoPn; complete sequence.	15. *Sinorhizobium meliloti* plasmid pSymB; complete sequence.
5. *Deinococcus radiodurans* R1 megaplasmid MP1; complete plasmid	16. *Staphylococcus aureus* plasmid pN315B DNA; complete sequence.
6. *Deinococcus radiodurans* R1 plasmid CP1; complete plasmid sequence.	17. *Staphylococcus aureus* plasmid VRSAp; complete sequence.
7. *Halobacterium sp.* NRC-1 plasmid pNRC100; complete plasmid sequence.	18. *Xylella fastidiosa* plasmid pXF1.3; complete sequence.
8. *Halobacterium sp.* NRC-1 plasmid pNRC200 complete genome.	19. *Xylella fastidiosa* plasmid pXF51; complete sequence.
9. *Mesorhizobium loti* plasmid pMLa DNA; complete genome; complete	20. *Yersinia pestis* plasmid pCD1.
10. *Mesorhizobium loti* plasmid pMLb DNA; complete genome; complete	21. *Yersinia pestis* plasmid pPCP1.
11. *Methanococcus jannaschii* large extra-chromosomal element; complete	22. *Yersinia pestis* plasmid pPMT1.

(b)

Table 2. (a) Repeat densities of 60 chromosomal sequences, (b) Repeat densities of 22 plasmid sequences.

SN	Direct_D_L	Direct_D_N	Inverted_D_L	Inverted_D_N	Complement_D_L	Complement_D_N	Palindrome_D_L	Palindrome_D_N
1	1.516	1228	0.958	877.4	0.152	214.4	1.184	1097
2	0	0	0	0	0	0	0	0
3	1.223	1047	0.76	701.3	0.17	239.1	0.857	799.3
4	1.558	1207	0.906	834.6	0.117	163.9	0.758	705.1
5	1.768	1356	1.296	1172	0.181	240.6	1.081	976.6
6	0	0	1.477	1317	0.039	57.42	1.172	1082
7	0	0	0	0	0	0	0	0
8	1.76	1485	1.218	1069	0.074	109.7	2.159	1971
9	2.383	1977	1.608	1349	0.08	123.5	2.082	1915
10	0.985	733.1	0.859	783.6	0.056	76.68	1.294	1191
11	0.855	734	0.913	841.3	0.01	15.44	1.258	1148
12	0.803	679.1	0.823	760.7	0.028	40.29	1.241	1144
13	0.827	724.5	0.915	842.4	0.009	13.82	1.26	1149
14	0.805	725.6	0.901	829.1	0.009	13.86	1.263	1152
15	2.744	2021	1.564	1344	0.311	407.8	1.909	1756
16	2.238	1901	1.291	1094	0.064	106.8	0.65	597.7
17	0.975	812.4	0.558	497.2	0.252	419.5	0.739	693.6
18	2.069	1509	1.412	1273	0.302	377.4	0.908	836.8
19	2.246	1720	1.489	1341	0.249	318.9	0.927	853.4
20	2.235	1731	1.505	1355	0.256	328.6	0.925	851
21	0	0	0	0	0	0	0	0
22	2.27	1596	1.119	1006	0.241	355.2	1.055	959.5
23	1.995	1689	1.262	1105	0.061	115.7	2.168	1974
24	2.965	1861	0.944	853.2	0.462	679.9	1.248	1155
25	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0
27	1.271	1121	1.098	988.1	0.018	29.88	1.366	1260
28	3.249	2421	1.804	1542	0.287	379.5	1.92	1764
29	2.163	1421	0.783	723.4	0.167	246.7	1.183	1089
30	3.272	1862	1.118	984.4	0.459	720.7	1.553	1428
31	1.198	1066	1.325	1126	0.08	133.7	1.431	1296
32	5.185	2642	1.593	1391	0.712	1002	2.072	1900
33	1.484	1179	0.923	846.4	0.215	320.6	1.739	1591
34	2.627	1661	0.768	716.6	0.409	591.6	1.329	1237
35	4.679	2057	1.614	1415	1.13	1575	2.663	2399
36	0	0	0	0	0	0	0	0
37	1.917	1380	1.159	1045	0.354	467.3	1.046	958.1
38	3.014	2343	1.646	1402	0.187	269.8	1.982	1822
39	0	0	0.819	758.6	0.108	150.1	1.049	971
40	1.497	1098	0.852	786.3	0.194	270.3	1.137	1050
41	2.167	1023	1.286	1073	0.29	365.7	2.041	1879
42	0	0	1.071	959	0.108	140.3	2.485	2252
43	1.832	1462	1.512	1307	0.13	177.1	1.952	1791
44	2.91	2148	1.306	1155	0.606	835.3	1.714	1582
45	3.001	2173	1.291	1142	0.66	903.8	1.728	1595
46	2.835	1709	1.075	968.6	0.879	1211	0.999	929.7
47	3	1546	1.03	923.7	0.512	654.4	0.998	925.1
48	0	0	0	0	0	0	0	0
49	2.511	1611	1.268	1089	0.234	312.8	1.453	1344
50	2.284	1665	1.392	1216	0.12	178.5	1.598	1475
51	2.164	1399	1.06	957.6	0.1	139.4	1.7	1600
52	1.009	871.6	0.836	784.7	0.04	57.51	1.2	1117
53	0.908	789.4	0.902	840.5	0.089	118	1.321	1224
54	1.371	1076	0.914	842.1	0.216	289.7	0.967	892.1
55	1.338	782.1	0.705	659.9	0.189	254	1.356	1251
56	0	0	1.255	1090	0.115	188.9	3.246	2933
57	1.751	1270	1.241	1132	0.28	347.2	0.994	911.5
58	1.03	793.6	0.853	785.2	0.067	109.1	0.925	846.8
59	5.056	2616	1.045	940.9	0.546	849.1	0.951	874.9
60	2.562	1890	1.385	1255	0.439	614.1	0.983	903.6

(a)

SN	Direct_D_L	Direct_D_N	Inverted_D_L	Inverted_D_N	Complement_D_L	Complement_D_N	Palindrome_D_L	Palindrome_D_N
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0
8	0.591	528.2	1.055	925	0.12	180.6	1.485	1283
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0
13	0.3	205.3	0.306	279.9	0.019	51.32	1.437	1316
14	0	0	0	0	0	0	0	0
15	1.156	1024	1.099	985.5	0.02	31.49	1.887	1733
16	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0
20	0.164	113.8	0.249	213.4	0.878	1138	1.351	1237
21	0	0	0.458	416.1	0	0	0.77	728.3
22	0.209	187.1	0.378	343	0	0	0.813	758.8

(b)

genomes, and tried to find particular repeat sequences that appeared in most bacterial genomes, i.e. common repeat sequences in bacterial genomes. These common repeat sequences probably play a novel role in these bacterial genomes. Knowing this function will help us to understand more about the blueprints of bacterial genomes in terms of repeat structures and mechanisms.

Materials and Methods

Detecting repeats

There are numerous computational methods for detecting repeats from genomic DNA sequences. These include algorithms that locate repeated substrings [19-21], such as the Tandem Repeats Finder [21], as well as programs for identifying known repeats, such as the RepeatMasker [22]. Most of these tools have some restrictions on the maximum length of the input sequence. Recently, however, new systems based on suffix trees, such as REPuter [7, 8], have overcomed this size limitation, at least within biologically realistic input sizes. REPuter is a highly efficient computational tool that can find all the exact or inexact repeats in sequences and output a summary report of the repetitive structure of a sequence. The REPuter package also includes a visualization tool to generate repeat graphs, which are useful for identifying the positions of repeats. Therefore, we established an automated system based on REPuter to recognize all exact and inexact repeat sequences in 58 fully sequenced bacterial genomes (downloaded from GenBank database), including 60 chromosomal sequences and 22 plasmid sequences (Table 1). All the results from REPuter were parsed into MySQL database by a custom Perl script for further analysis.

Density of repeats

In order to characterize the repeats, we used two measures of density, the density in number D_N and the density in length D_L, for the quantitative analysis. They are defined as:

D_N = number of repeat copies/size of genome (Mb)

$$D_L = 100 \times \frac{\text{size of repeat sequence (bp)}}{\text{size of genome (bp)}}$$

The biological interpretation of these measures may be quite different: D_N can be assimilated as the rate of amplification (a balance between duplication and deletion processes) and D_L is more connected to the history of the genomes, i.e. a measure of the redundancy tolerated by a genome sequence.

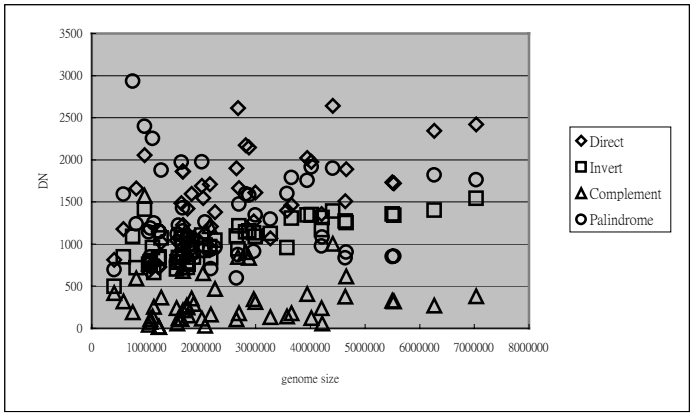
Finding common repeat sequences of bacterial genomes

All repeat sequences are directly parsed from the output files of REPuter. This is no problem for exact repeat sequences. For inexact repeat sequences, a pattern of the repeat sequences was saved as for all of its possibilities. For example, the pattern of repeat sequence GAGC[TG]C was saved as two records, GAGCTC and GAGCGC. For each repeat pattern, we used SQL query to a value N_x. N_x denotes the number of genome sequences that this repeat pattern appeared in. The code x consists of four members to denote four types of repeat sequences, i.e., forward (direct) repeats, reverse (inverted) repeats, complement repeats, and reverse complement (palindromic) repeats.

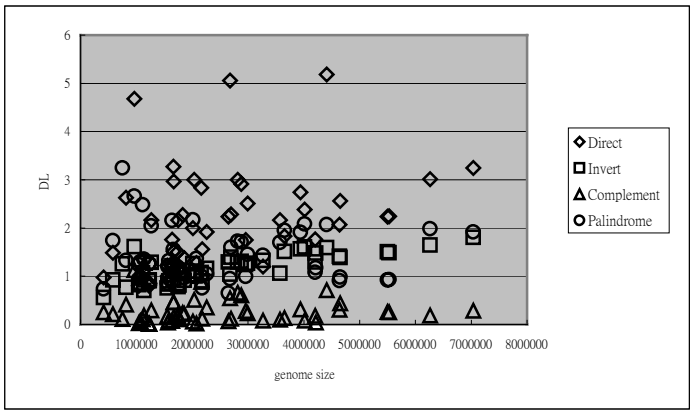
Results and Discussion

Density of repeats

Table 2 shows the densities of four types of repeats (direct,



(a)



(b)

Figure 1. Densities of four types of repeats versus the genome size. (a) D_N (b) D_L

inverted, complement, and palindrome) in 60 chromosomal sequences and 22 plasmid sequences, including both D_N and D_L. There are 7 chromosomal sequences that do not have any types of repeats. They are *Agrobacterium tumefaciens strain* C58, *Buchnera sp.* APS, *Borrelia burgdorferi*, *Helicobacter pylori* J99, *Lactococcus lactis* IL1403, *Neisseria meningitidis* MC58, and *Streptococcus pyogenes* SF370 M1. Moreover, there are 4 chromosomal sequences do not have direct repeats. They are *Bacillus subtilis*, *Pyrococcus abyssi*, *Rickettsia prowazekii* Madrid E, and *Ureaplasma urealyticum*. We mapped these 11 genomes to the current taxonomy of bacteria used in GenBank. However, we did not find any associations between them. Some genomes, such as *Streptococcus pyogenes* and *Lactococcus lactis*, are in the same subdivision. But most of these 11 genomes are discretely distributed in the taxonomy tree with no clear correlations. In each subdivision, one organism may have high repeat density, but another one may have low or zero repeat density.

Most plasmid sequences do not have any types of repeats. However, 5 of the 22 plasmid sequences still contain repeats. They are *Halobacterium sp.* NRC-1 plasmid pNRC200, plasmid Ti, *Sinorhizobium meliloti* plasmid pSymB, *Yersinia pestis* plasmid pCD1, and *Yersinia pestis* plasmid pPMT1. With the exception of *Sinorhizobium meliloti* plasmid pSymB, the other 4 plasmid sequences have low density of direct repeats and moderate densities of the other three types of repeats.

Figure 1 shows the densities of four types of repeats versus the genome size. According to the results, inverted and complement repeats seem to be more stable and in relatively low densities, but direct and palindromic repeats are more diverse and in relatively high densities. However, the density of repeats does not seem to correlate well with the genome size. Small chromosome may have high density of repeats and large chromosome may have low density of repeats, and vice versa.

Common repeat sequences of bacterial genomes

The main question we tackled in this work concerns the common repeat sequences of bacterial genomes. We defined a number N_x to denote the number of genome sequences that a repeat pattern appeared in, where x denotes the four types of repeats. Figure 2 shows the logarithm of number of repeat patterns versus each N_x. We found only palindromic repeats have the potential to be common in most genomes because its maximum N_x is 50, near the total number of analyzed genomes (60). The low maximum N_x of direct repeats (14), inverted repeats (7), and complement repeats (28) implies these three types of repeats are not commonly present. Therefore, we decided to focuse on the common palindromic repeats for further detailed analysis.

Moreover, in the region of N_x = 10~45 in Figure 2, each N_x have 10^1~10^2 common palindromic repeat patterns. It implies the existence of subgroups of common palindromic

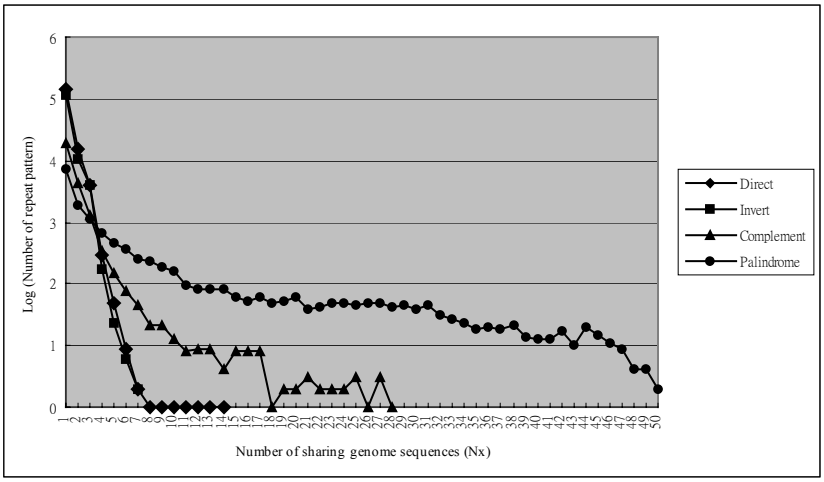


Figure 2. Log(Number of repeat patterns) for each number of sharing genome sequences.

Table 3. Top 30 common palindromic repeat patterns shared in most genomes.

Repeat pattern	Number of sharing genome
aaaaattttt	50
aaaatatttt	50
agcagctgct	49
aggaattcct	49
gaagatcttc	49
ttttgcaaaa	49
aaaacgtttt	48
gaaatatttc	48
tcctgcagga	48
ttttcgaaaa	48
aaaagctttt	47
aaatcgattt	47
agaatattct	47
atttgcaaat	47
ggaatattcc	47

Repeat pattern	Number of sharing genome
tcaatattga	47
tgctgcagca	47
ttcagctgaa	47
tttcatgaaa	47
aaagtacttt	46
aaatatattt	46
aaattaattt	46
aacagctgtt	46
agaaatttct	46
agcaattgct	46
atcagctgat	46
gaacatgttc	46
tggaattcca	46
ttctgcagaa	46
tttttaaaaa	46

repeat patterns. This phenomenon is unique to palindromic repeats. With the conclusions drawn from Table 2 and Figure 2, one may find the four types of repeats seem have their own specificity. Direct repeats are diverse and in relatively high densities, and are not common to all genomes. Inverted repeats are stable and in relatively low densities, and are not common to all genomes. Complement repeats are stable and in relatively low densities, and are to some extent common to all genomes. Palindromic repeats are diverse and in relatively high densities, and are common to all genomes with some potential subgroups. From their unique specificity, we may infer that these four types of repeats may be totally different in terms of functions, evolutions and mechanisms.

Table 3 shows the top 30 common palindromic repeat patterns shared in most genomes ($N_x > 45$). We found most of them are A-T rich. In a general view of a genome, G-C rich elements are usually involved in coding jobs, such as ORFs, and A-T rich elements usually perform regulatory functions, such as promoters and terminators. Therefore, we supposed the functions of these common palindromic repeats are related to gene regulations. This hypothesis can also explain why the density of palindromic repeats is so diverse and relatively high. At the biological pathway level, each organism has its own unique regulatory system. Therefore, the required number of palindromic repeats is also unique and variable. Moreover, if the palindromic repeats really perform in an important and novel regulation mechanism, and this mechanism is needed in most organisms, then it must be present in relatively high density. Nevertheless, this hypothesis still needs to be verified.

Conclusions

Our results support the following conclusions:

1. The occurrences of repeat sequences are not related to current taxonomy of bacterial organisms. In the same subdivision, one organism may have high repeat density, but another one may have low or zero repeat density.
2. Most plasmid sequences do not have repeat sequences, and most chromosomal sequences have repeat sequences. However, both of them have some exceptions.
3. The occurrences of repeat sequences are not correlated well with chromosome size. We did not find any strong correlation between repeat density and chromosome size. Large chromosome may have more repeats, but the repeat density may not increase concurrently.
4. Only reverse complement (palindromic) repeats are common to most bacterial genomes. This result is not only significant, but also interesting. We supposed that the mechanisms of evolution and function are different between palindromic repeats and other types of repeats. There are many palindromic repeats shared by most bacterial genomes, but there are also many palindromic repeats unique to only one bacterial genome.

With the conclusions above, we supposed the functions and evolutions of repeat sequences might be involved in more complicated mechanisms. They can not be clearly realized by simple statistical analysis. Conventionally we only use simple statistical methods to analyze repeat sequences. Now, according to our results and conclusions, further development for a new methodology combined with statistical analysis and detailed biology of each bacterial organism will be required and suggested.

Acknowledgenents

The financial support for this research work was provided partly by a National Research Program for Genomic Medicine (NRPGM) grant (NSC 91-3112-B-010-015) to C-H Chang from the National Science Council of R.O.C.

References

[1] Katti MV, Ranjekar PK, Gupta VS "Differential distribution of simple sequence repeats in eukaryotic genome sequences." *Mol Biol Evol,* 18:1161-1167, 2001.

[2] Le Fleche P, Hauck Y, Onteniente L, Prieur A, Denoeud F, Ramisse V, Sylvestre P, Benson G, Ramisse F, Vergnaud G "A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*." *BMC Microbiol,* 1:2, 2001.

[3] Levinson G, Gutman GA "High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12." *Nucleic Acids Res,* 15:5323-5338, 1987.

[4] van Belkum A, van Leeuwen W, Scherer S, Verbrugh H "Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes." *Res Microbiol,* 150:617-626, 1999.

[5] Levinson G, Gutman GA "Slipped-strand mispairing: a major mechanism for DNA sequence evolution." *Mol Biol Evol,* 4:203-221, 1987.

[6] Yeramian E and Buc H "Tandem repeats in complete bacterial genome sequences: sequence and structural analyses for comparative studies." *Res Microbiol,* 150: 745-754, 1999.

[7] Kurtz S, Schleiermacher C "REPuter: Fast Computation of Maximal Repeats in Complete Genomes." *Bioinformatics,* 15(5): 426-427, 1999.

[8] Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R "REPuter: The Manifold Applications of Repeat Analysis on a Genomic Scale." *Nucleic Acids Res,* 29(22): 4633-4642, 2001.

[9] van Belkum A, Scherer S, van Leeuwen W, Willemse D, van Alphen L, Verbrugh H "Variable number of tandem repeats in clinical strains of Haemophilus influenzae." *Infect Immun*, 65: 5017-5027, 1997.

[10] Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME "Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within Bacillus anthracis." *J Bacteriol,* 182: 2928-2936, 2000.

[11] Frothingham R, Meeker-O'Connell WA "Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats." *Microbiology,* 144: 1189-1196, 1998.

[12] Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C "Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome." *Mol Microbiol,* 36: 762-771, 2000.

[13] Adair DM, Worsham PL, Hill KK, Klevytska AM, Jackson PJ, Friedlander AM, Keim P "Diversity in a variable-number tandem repeat from Yersinia pestis." *J Clin Microbiol,* 38: 1516-1519, 2000.

[14] van Ham SM, van Alphen L, Mooi FR, van Putten JP "Phase variation of H. influenzae fimbriae: transcriptional control of two divergent genes through a variable combined promoter region." *Cell,* 73: 1187-1196, 1993.

[15] Weiser JN, Love JM, Moxon ER "The molecular mechanism of phase variation of H. influenzae lipopolysaccharide." *Cell,* 59: 657-665, 1989.

[16] Bayliss CD, Field D, Moxon ER "The simple sequence contingency loci of Haemophilus influenzae and Neisseria meningitidis." *J Clin Invest,* 107: 657-666, 2001.

[17] Henderson IR, Owen P, Nataro JP "Molecular switches - the ON and OFF of bacterial phase variation." *Mol Microbiol,* 33: 919-32, 1999.

[18] Wang G, Ge Z, Rasko DA, Taylor DE "Lewis antigens in Helicobacter pylori: biosynthesis and phase variation." *Mol Microbiol,* 36:1187-1196, 2000.

[19] Leung M-Y, Blaisdell BE, Burge C, Karlin S "An efficient algorithm for identifying matches with errors in multiple long molecular sequences." *J Mol Biol,* 221: 1367-1378, 1991.

[20] Kannan SK, Myers EW "An algorithm for locating nonoverlapping regions of maximal alignment score." *SIAM J Comput,* 25: 648-662, 1996.

[21] Benson G "Tandem repeats finder: a program to analyze DNA sequences." *Nucleic Acids Res* 27: 573-580, 1999.

[22] RepeatMasker: http://ftp.genome.washington.edu/RM/RepeatMasker.html