GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Evaluating the annotation of protein-coding genes in bacterial genomes: *Chloroflexus aurantiacus* strain J-10-fl and *Natrinema* sp J7-2 as case studies

H.X. Zhang, S.J. Li and H.Q. Zhou

School of Life Science and Technology,
Key Laboratory for Neuroinformation of Ministry of Education,
University of Electronic Science and Technology of China, Chengdu, China

Corresponding author: H.X. Zhang
E-mail: johnzhangchina2007@gmail.com

Genet. Mol. Res. 13 (4): 10891-10897 (2014)
Received January 21, 2014
Accepted April 11, 2014
Published December 19, 2014
DOI http://dx.doi.org/10.4238/2014.December.19.10

ABSTRACT. Gene annotation plays a key role in subsequent biochemical and molecular biological studies of various organisms. There are some errors in the original annotation of sequenced genomes because of the lack of sufficient data, and these errors may propagate into other genomes. Therefore, genome annotation must be checked from time to time to evaluate newly accumulated data. In this study, we evaluated the gene density of 2606 bacteria or archaea, and identified 2 with extreme values, the minimum value (*Chloroflexus aurantiacus* strain J-10-fl) and maximum value (*Natrinema* sp J7-2), to conduct genome re-annotation. In the genome of *C. aurantiacus* strain J-10-fl, we identified 17 new genes with definite functions and eliminated 34 non-coding open-reading frames; in the genome of *Natrinema* sp J7-2, we eliminated 118 non-coding open reading frames. Our re-annotation procedure may provide a reference for improving the annotation of other bacterial genomes.

Key words: Gene density; Genome re-annotation; Missed genes; Over-annotated genes

INTRODUCTION

In recent years, thousands of complete genome sequences from various species have been published, and more and more sequences will continue to be published in the coming years. The wealth of sequence data affords the opportunity to mine useful information or knowledge from the data. In addition, the need for accurately annotated genomes, which consist of 2 main parts, identification of the protein-coding portions and an understanding of the biological meaning of this information, increases. Because genome annotation is directly relevant to the research results on which it is based, the quality of these data is critical. Inaccurate genome annotation may influence subsequent studies (Yu et al., 2012). Once a genome-sequencing project is complete and the information is submitted to a public database, researchers must manually evaluate the original annotation and include updates. Various prokaryotic genomes have been re-annotated and the 2 main types of re-annotations are frequently performed.

First, over-annotated protein-coding regions, which are falsely predicted to be positive samples, should be eliminated from the original annotation. For example, Bocs et al. (2002) found that 34% of the original annotations for 26 complete prokaryotic genomes were non-coding open reading frames (ORFs). Gundogdu et al. (2007) reduced the total number of originally annotated proteins from 1654 to 1643 by eliminating coding sequences (CDS) or merging adjacent CDS in *Campylobacter jejuni* NCTC11168. Based on the Z-curve method, Guo and Yu (2007) found that more than 30 of 294 originally annotated genes did not encode proteins in the genome of *Amsacta moorei entomopoxvirus*. Using another graphical method, Yu and Sun (2010) confirmed this hypothesis. In addition, a large number of falsely predicted genes have been eliminated from *Aeropyrum pernix* K1 based on the combination of the Z-curve method and K-means clustering by Guo et al. (2004) and Guo and Lin (2009).

Second, some real genes may be missed for various reasons, which could be identified through other methods, such as *ab initio* gene detection. For example, Camus et al. (2002) identified 75 new genes by re-examining the genome using manual or automatic analysis methods and another 7 based on experimental evidence in the genome of *Mycobacterium tuberculosis* strain H37Rv. This is not an isolated case, as 8 potential genes have been added to the archaeon *Pyrobaculum aerophilum* genome by Du et al. (2011). In addition, Okamoto and Yamada (2011) added 9 novel ORFs in the *Streptococcus pyogenes* SF370 genome based on shotgun proteomic analysis. Zhou et al. (2011) found 278 new CDSs based on similarity searching and another 147 CDSs through the transcriptions of detectable mRNA in *Xanthomonas campestris*.

In this study, we conducted 2 types of re-annotation for the prokaryotic complete genomes of *Chloroflexus aurantiacus* strain J-10-fl (Tang et al., 2011) and *Natrinema* sp J7-2 (Feng et al., 2012). *C. aurantiacus* strain J-10-fl, as an aerobic facultative bacterium, can conduct photosynthesis under anaerobic conditions through processes that are a mix of both purple bacteria and green sulfur bacterium photosynthesis. Updated annotations of this bacterial strain may contribute to studies examining the origin of photosynthesis and the evolution of photosynthesis diversity. Similarly, re-annotating *Natrinema* sp J7-2, an extreme halophilic archaeon, may help researchers understand evolutionary relationships between haloarchaea.

MATERIAL AND METHODS

Data source

The complete genomes of *C. aurantiacus* strain J-10-fl (RefSeq accession No. NC_010175) and *Natrinema* sp J7-2 (RefSeq accession No. NC_018224) were used in this study. The current annotation of *C. aurantiacus* strain J-10-fl was obtained on November 11, 2011 and contained 3853 potential genes with 56.7% G + C content. The latter, which has 64.3% G + C content, was originally annotated with 4239 potential genes and was obtained on July 27, 2012. Both sequences are available from the Refseq ftp site (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/).

Elimination of over-annotated genes

To eliminate over-annotated genes, 2 types of DNA sequences were generated as training sets. The first set included known-function genes with defined names and served as the positive set for the training model. The randomly shuffled sequences of the positive samples become the negative samples. Next, non-coding proteins from the hypothetical protein collection were eliminated based on the training results of positive and negative samples. A web server known as Zfisher, which is designed to evaluate the coding potentials of questionable genes in the sequenced bacterial or archaeal genomes based on the Z-curve method, can be used for this process (Guo et al., 2013). Zfisher is freely available at http://147.8.74.24/Zfisher/.

Identifying missed genes

In this study, 2 programs, Prodigal (Hyatt et al., 2010) and ZCURVE (Guo et al., 2003; Guo and Zhang, 2006), were used to identify new candidate genes (Guo et al., 2013). Both programs have been successfully used to identify genes in various microbes and are freely available at http://compbio.ornl.gov/prodigal/ and http://tubic.tju.edu.cn/Zcurve_B/. First, we chose the intersection of both candidate genes with different 5'-termini for comparison with the original annotation. In addition, the overlap rate of this intersection against the annotated genes was required to be low (<20%). Next, the new candidate genes were submitted to the BLAST program to compare the query sequence with the NCBI nr database to identify all homologous genes. If 1 candidate met the following conditions, E-value <1e-20, Query cover >60%, Ident >50%, and a low length difference with the counterpart (diff <20%), it was regarded as a genuine gene.

RESULTS AND DISCUSSION

Linear correlation between genome size and gene number

There is a strong positive correlation between genome size and gene number in sequenced bacterial or archaeal genomes, as shown in Figure 1. Thus, the gene number increases (or decreases) proportionally with genome size. In Figure 1, the number of originally annotated genes in *Natrinema* sp J7-2 (red square in Figure 1) was much higher than the expected number. In contrast, the gene number of the original annotation in the *C. aurantiacus* strain J-10-fl (green triangle in Figure 1) was much lower than the expected number. Therefore, these 2 species were re-annotated.

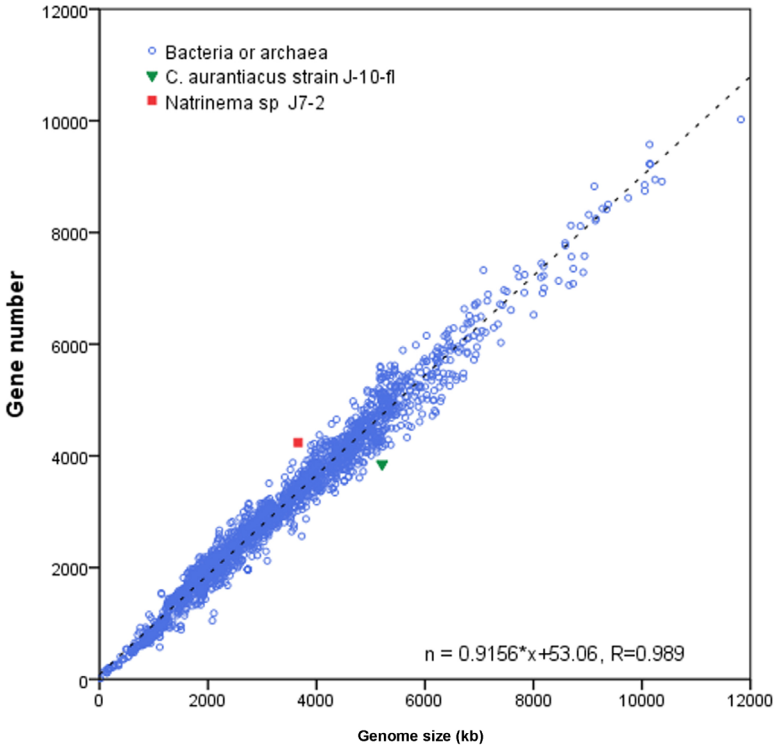

Figure 1. Linear correlation between genome size and gene number among 2606 bacteria or archaea published before September, 2013. Values show significant linear dependence (R = 0.989). The red square above the regression line represents the *Natrinema* sp J7-2 and the *Chloroflexus aurantiacus* strain J-10-fl below the regression line is marked as a green triangle.

Elimination of 118 over-annotated genes in *Natrinema* sp J7-2

For most species, a number of over-annotated genes were identified as potential genes when the *ab initio* gene finder was used for gene prediction. Three groups of genes were marked off from the original annotation. As previously discussed, the known-function genes with defined names constituted the first group. The third group contained genes encoding hypothetical proteins. Thus, the remaining genes were classified as the second group. The first group was able to serve as the genuine genes. However, uncertainty exists regarding whether the third group encoded proteins. Thus, we primarily eliminated non-coding ORFs from the third group in the study.

Using the Zfisher server, we identified 118 hypothetical genes as non-coding ORFs in *Natrinema* sp J7-2 (Table 1). The accuracy of 5-fold validation was 100%. The scatterplot of the nucleotide distribution of 972 known-function genes and 118 predicted non-coding ORFs is shown in Figure 2. Most dots of the predicted non-coding ORFs were detached from the known-function genes. In addition, nearly all of the G + C content of known-function genes at the 3rd codon positions were higher than those at the 2nd codon position, so these genes were located far above the diagonal. However, most non-coding ORFs were located uniformly around the diagonal, suggesting that the G + C content at the 2nd position was approximately

equal to that at the 3rd position. There was an apparent distinction between the known-function genes and non-coding ORFs for the nucleotide composition. Zhang and Chou (1994) suggested that the function of encoded proteins would impose restrictions on the nucleotide composition of genes, and several studies have verified this argument in high G + C content prokaryotic genomes (Chen and Zhang, 2003; Guo, 2007). Similarly, we predicted 34 non-coding ORFs in *C. aurantiacus* strain J-10-fl with an accuracy of 5-fold validation of 99.9%. Names of the 34 ORFs are listed in Table 2. We removed these non-coding ORFs from the gene list.

Table 1. Synonymous codes of the 118 ORFs identified as non-coding in *Natrinema* sp J7-2.

NJ7G_0065	NJ7G_0081	NJ7G_0095	NJ7G_0108	NJ7G_0157	NJ7G_0170	NJ7G_0175	NJ7G_0200
NJ7G_0234	NJ7G_0275	NJ7G_0291	NJ7G_0306	NJ7G_0323	NJ7G_0441	NJ7G_0475	NJ7G_0525
NJ7G_0546	NJ7G_0564	NJ7G_0569	NJ7G_0585	NJ7G_0670	NJ7G_0676	NJ7G_0703	NJ7G_0777
NJ7G_0857	NJ7G_0867	NJ7G_0871	NJ7G_0911	NJ7G_0934	NJ7G_0970	NJ7G_0972	NJ7G_0982
NJ7G_0988	NJ7G_0996	NJ7G_1001	NJ7G_1016	NJ7G_1059	NJ7G_1062	NJ7G_1067	NJ7G_1094
NJ7G_1152	NJ7G_1215	NJ7G_1220	NJ7G_1245	NJ7G_1263	NJ7G_1372	NJ7G_1391	NJ7G_1399
NJ7G_1419	NJ7G_1437	NJ7G_1571	NJ7G_1630	NJ7G_1645	NJ7G_1662	NJ7G_1679	NJ7G_1728
NJ7G_1807	NJ7G_1849	NJ7G_1852	NJ7G_1884	NJ7G_1945	NJ7G_2021	NJ7G_2036	NJ7G_2095
NJ7G_2106	NJ7G_2121	NJ7G_2139	NJ7G_2250	NJ7G_2374	NJ7G_2384	NJ7G_2480	NJ7G_2592
NJ7G_2601	NJ7G_2609	NJ7G_2625	NJ7G_2652	NJ7G_2656	NJ7G_2688	NJ7G_2717	NJ7G_2721
NJ7G_2742	NJ7G_2754	NJ7G_2943	NJ7G_2999	NJ7G_3042	NJ7G_3175	NJ7G_3183	NJ7G_3273
NJ7G_3317	NJ7G_3318	NJ7G_3356	NJ7G_3404	NJ7G_3406	NJ7G_3441	NJ7G_3452	NJ7G_3453
NJ7G_3469	NJ7G_3537	NJ7G_3625	NJ7G_3628	NJ7G_3630	NJ7G_3640	NJ7G_3699	NJ7G_3705
NJ7G_3724	NJ7G_3727	NJ7G_3812	NJ7G_3906	NJ7G_3945	NJ7G_3999	NJ7G_4002	NJ7G_4061
NJ7G_4063	NJ7G_4114	NJ7G_4146	NJ7G_4192	NJ7G_4229	NJ7G_4281		

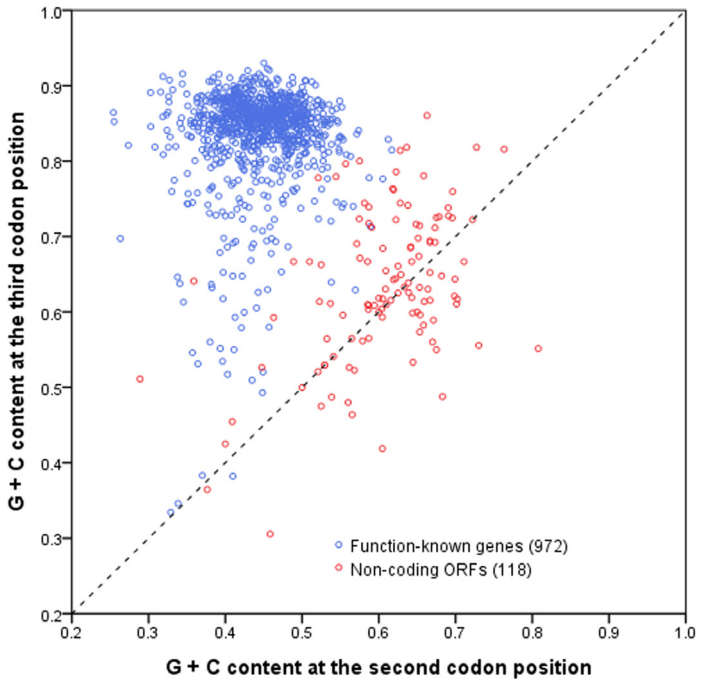

Figure 2. Distribution of GC2 *vs* GC3 for the known-function genes and non-coding ORFs in *Natrinema* sp J7-2. The x-axis and y-axis represent the values of GC2 and GC3, respectively. The 972 blue dots indicate known-function genes and the 118 red dots indicate non-coding ORFs.

Table 2. Synonymous codes of the 34 ORFs identified as non-coding in *Chloroflexus aurantiacus* strain J-10-fl.

Caur_0111	Caur_0176	Caur_0177	Caur_0296	Caur_0366	Caur_0398	Caur_0430	Caur_0454
Caur_0653	Caur_0669	Caur_1000	Caur_1406	Caur_1506	Caur_1755	Caur_1846	Caur_2077
Caur_2159	Caur_2291	Caur_2294	Caur_2298	Caur_2757	Caur_2894	Caur_2928	Caur_2999
Caur_3082	Caur_3167	Caur_3181	Caur_3235	Caur_3259	Caur_3361	Caur_3455	Caur_3623
Caur_3657	Caur_3800						

Identifying 34 missed genes in *C. aurantiacus* strain J-10-fl

Some genuine genes may not be identified during the process of genome annotation. Several main steps for detecting these missed genes include *ab initio* gene identification, BLAST comparison, and filtering based on specific rules. Applying this method to *C. aurantiacus* strain J-10-fl, we found 34 new genes, including 17 genes with definite function (Table 3). Among these 17 new genes, only 3 genes have some overlapping bases with the originally annotated genes. However, we found no new genes satisfying our criteria in *Natrinema* sp J7-2. Here, we only retained these new genes with very high similarities, which may result in missing some genuine genes without significant similarities to known genes.

Table 3. Details of the 17 new genes found in the genome of *Chloroflexus aurantiacus* strain J-10-fl.

Location	Cover (%)	Ident. (%)	E-value	Overlap (%)	Potential function
361359-361886	99	100	2e-119	0	Molecular chaperone-like protein
361849-362346	92	100	4e-100	0	Molecular chaperone-like protein
452398-452844	84	99	4e-66	1	Major facilitator superfamily protein
1110604-1111272	87	67	8e-86	0	Allergen V5/Tpx-1 family protein
1127668-1128426	97	52	8e-63	0	Iron-sulfur binding protein
1137206-1137769	99	89	1e-106	0	Transposase IS4 family protein
1262170-1262742	99	99	3e-98	0	Type II secretion system protein E
2315614-2316093	99	61	1e-67	0	Type II site-specific deoxyribonuclease
2565985-2566377	90	61	5e-49	1	Deoxyribodipyrimidine photo-lyase
2575190-2575699	92	67	1e-73	0	Modification methylase
2778475-2778876	98	53	2e-29	0	Mobile element protein
3325450-3325752	92	56	5e-24	8	Molybdenum ABC transporter ATP-binding protein
3859638-3860336	99	100	2e-148	0	ATP-dependent protease La
4240373-4240747	94	75	2e-60	0	DNA methylase N-4/N-6 domain-containing protein
4312197-4312628	99	52	5e-45	0	Transposase, IS4 family protein
4339812-4340069	88	93	4e-40	0	Transposase IS5 family protein-like protein
4644781-4645554	74	51	4e-42	0	Peptidase S8/S53 subtilisin kexin sedolisin

ACKNOWLEDGMENTS

We thank Prof. Feng-Biao Guo for providing invaluable support and inspiring discussion. Research supported by the Fundamental Research Funds for the Central Universities of China (Grants #ZYGX2013J100 and #ZYGX2013J101). The Project was also sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

REFERENCES

Bocs S, Danchin A and Médigue C (2002). Re-annotation of genome microbial coding-sequences: finding new genes and inaccurately annotated genes. *BMC Bioinformatics* 3: 5.

Camus JC, Pryor MJ, Médigue C and Cole ST (2002). Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* 148: 2967-2973.

Chen LL and Zhang CT (2003). Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. *Biochem. Biophys Res. Commun.* 306: 310-317.

Du MZ, Guo FB and Chen YY (2011). Gene re-annotation in genome of the extremophile *Pyrobaculum aerophilum* by using bioinformatics methods. *J. Biomol. Struct. Dyn.* 29: 391-401.

Feng J, Liu B, Zhang Z, Ren Y, et al. (2012). The complete genome sequence of *Natrinema* sp. J7-2, a haloarchaeon capable of growth on synthetic media without amino acid supplements. *PLoS One* 7: e41621.

Gundogdu O, Bentley SD, Holden MT, Parkhill J, et al. (2007). Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. *BMC Genomics* 8: 162.

Guo FB (2007). The distribution patterns of bases of protein-coding genes, non-coding ORFs, and intergenic sequences in *Pseudomonas aeruginosa* PA01 genome and its implications. *J. Biomol. Struct. Dyn.* 25: 127-133.

Guo FB and Zhang CT (2006). ZCURVE_V: a new self-training system for recognizing protein-coding genes in viral and phage genomes. *BMC Bioinformatics* 7: 9.

Guo FB and Yu XJ (2007). Re-prediction of protein-coding genes in the genome of *Amsacta moorei entomopoxvirus*. *J. Virol. Methods* 146: 389-392.

Guo FB and Lin Y (2009). Identify protein-coding genes in the genomes of *Aeropyrum pernix* K1 and *Chlorobium tepidum* TLS. *J. Biomol. Struct. Dyn.* 26: 413-420.

Guo FB, Ou HY and Zhang CT (2003). ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.* 31: 1780-1789.

Guo FB, Wang J and Zhang CT (2004). Gene recognition based on nucleotide distribution of ORFs in a hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res* 11: 361-370.

Guo FB, Xiong L, Teng JL, Yuen KY, et al. (2013). Re-annotation of protein-coding genes in 10 complete genomes of *Neisseriaceae* family by combining similarity-based and composition-based methods. *DNA Res.* 20: 273-286.

Hyatt D, Chen GL, Locascio PF, Land ML, et al. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119.

Okamoto A and Yamada K (2011). Proteome driven re-evaluation and functional annotation of the *Streptococcus pyogenes* SF370 genome. *BMC Microbiol.* 11: 249.

Tang KH, Barry K, Chertkov O, Dalin E, et al. (2011). Complete genome sequence of the filamentous anoxygenic phototrophic bacterium *Chloroflexus aurantiacus*. *BMC Genomics* 12: 334.

Yu JF and Sun X (2010). Reannotation of protein-coding genes based on an improved graphical representation of DNA sequence. *J. Comput. Chem.* 31: 2126-2135.

Yu JF, Jiang DK, Xiao K and Jin Y (2012). Discriminate the falsely predicted protein-coding genes in *Aeropyrum pernix* K1 genome based on graphical representation. *Math. Comput. Chem.* 67: 845-866.

Zhang CT and Chou KC (1994). A graphic approach to analyzing codon usage in 1562 *Escherichia coli* protein coding sequences. *J. Mol. Biol.* 238: 1-8.

Zhou L, Vorholter FJ, He YQ, Jiang BL, et al. (2011). Gene discovery by genome-wide CDS re-prediction and microarray-based transcriptional analysis in phytopathogen *Xanthomonas campestris*. *BMC Genomics* 12: 359.