Published online 15 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D13–D15***
doi:10.1093/nar/gkl908

DDBJ working on evaluation and classification of bacterial genes in INSDC

Hideaki Sugawara, Takashi Abe, Takashi Gojobori and Yoshio Tateno*

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima 411-8540, Japan

Received September 4, 2006; Revised October 10, 2006; Accepted October 13, 2006

ABSTRACT

DNA Data Bank of Japan (DDBJ) (http://www.ddbj.nig.ac.jp) newly collected and released 12 927 184 entries or 13 787 688 598 bases in the period from July 2005 to June 2006. The released data contain honeybee expressed sequence tags (ESTs), re-examined and re-annotated complete genome data of *Escherichia coli* K-12 W3110, medaka WGS and human MGA. We also systematically evaluated and classified the genes in the complete bacterial genomes submitted to the International Nucleotide Sequence Database Collaboration (INSDC, http://insdc.org) that is composed of DDBJ, EMBL Bank and GenBank. The examination and classification selected 557 000 genes as reliable ones among all the bacterial genes predicted by us.

INTRODUCTION

In the period from July 2005 to June 2006, DNA Data Bank of Japan (DDBJ) collected and released the original data of 12 927 184 entries or 13 787 688 598 bases. Among them ~90% were submitted by Japanese researchers and the rest were mainly submitted by Chinese and Korean researchers.

The ever-increasing DNA data submissions to International Nucleotide Sequence Database Collaboration (INSDC) have made a profound impact and contribution not only to the research community of life sciences but also to those of medicine, pharmacology, agriculture and others. However, a problem inconspicuous in the past has been more and more discernible and serious with the increasing number of the data submissions. The problem is that the majority of the submissions contain genes that were not examined or confirmed *in vivo*/*vitro*, but inferred by homology search or the like *in silico*. This *in silico* inference has been repeated by data submitters by using homology search against those inferred *in silico*, making the problem deepen and intractable. Consequently, a large number of submitted genes have been described as being 'hypothetical' or 'homologous to a particular gene'. Sometimes, 'a gene homologous to a particular gene' was found in fact to be an entirely different gene by *in vivo*/*vitro* experiments (1–3). Another aspect of this problem is that the *in silico* inferences were made by using various computer tools with various parameters. Therefore, strictly speaking, we cannot freely compare one hypothetical gene with another. This aspect makes the problem even worse.

To resolve the problem we have to examine every doubtful gene in INSDC *in vivo*/*vitro*, as proposed by Roberts (1). However, this approach is currently not quite feasible, because the number of such genes in INSDC is intractably large and increasing. The second and more feasible choice perhaps is to evaluate the genes in INSDC *in silico* by the same tool and parameters, and classify them into the degree of reliability. We have carried out the second approach for the bacterial genes in our GIB database (http://gib.genes.nig.ac.jp) (4). In this report we will also summarize our *in silico* approach and results.

NEW DATA SUBMISSIONS TO DDBJ IN THE LAST YEAR

The submitted and released data in the period mentioned above include 53 359 entries of honeybee (*Apis mellifera*) expressed sequence tags (ESTs) submitted by RIKEN (http://genome.gsc.riken.jp). They also include the complete genome sequence of *Escherichia coli* K-12 W3110 with accession number AP009048. The genome sequence was newly annotated by the collaboration among Marine Biological Laboratory in USA, Nara Advanced Institute of Science and Technology, Institute of Basic Biology and DDBJ in Japan (5). The *E.coli* genome data are supposed to be the most comprehensive and accurate data in all bacterial genome data submitted so far to INSDC. The genome data can be retrieved at the GIB database of DDBJ (visit also http://ecoli.aist-nara.ac.jp).

In addition, RIKEN submitted 1.89 million human MGA entries. The whole data can be obtained by ftp (ftp://ftp.ddbj.nig.ac.jp/database/mga). As for WGS ~220 000 entries of medaka (*Oryza latipes*) strain Hd-rR were submitted

*To whom correspondence should be addressed. Tel: +81 55 981 6857; Fax: +81 55 981 6858; Email: ytateno@genes.nig.ac.jp

The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors

from the University of Tokyo. The medaka entries were later assembled to be 6790 CON entries that are available at DDBJ with accession nos DF076466–DF083206 and DG000001–DG000024 and the University of Tokyo (http://medaka.utgenome.org). Note that MGA and WGS are not included in the ordinary INSDC divisions (6), and thus the entries in those categories are not counted in the number of collected and released entries or bases mentioned above.

BACTRIAL GENE EVLUATION AND CLASSIFICATION

To pursue the bacterial gene evaluation and classification, we first decided two things, the database to be targeted and the tools to be used. As to the database we used GIB that included all complete bacterial genome sequences submitted to INSDC. To be more specific we used GIB ver. 2003 and ver. 2004, in which the former included 123 bacterial species/strains installed by July 2003 and the latter included 183 bacterial species/strains installed by September 2004. As for the tools we employed tRNAscan-SE (ver. 1.23) (7), RBSfinder (8), Glimmer 2.0 (9,10) and InterPro (vers. 6.2 and 7.2) (11). Although the submitters of those bacterial genome data specified or inferred genes in the genomes with or without annotations, we independently searched for genes by using the databases and tools mentioned.

We set out on the gene search first to mask various RNA-coding genes including non-coding RNA on the genome in question by using tRNAscan-SE (*tRNA* genes), GIB (*rRNA* genes) and Rfam (12) (non-coding genes). We then predicted pairs of boundaries of ORFs in the unmasked genomic region by Glimmer 2.0 with two minimum lengths, 15 and 60 amino acids. For each predicted ORF RBSfinder was applied to ascertain the region in the vicinity of the gene start to which the ribosome bound. Those passed the two procedures were then subject to Blastp search against the proteins of the bacterial division (BAC) in the DAD databases (releases 24 corresponding to GIB ver. 2003 and release 28 to GIB ver. 2004) at DDBJ. DAD includes all the translated amino acid sequences from the nucleotide sequences of INSDC. In the Blastp search the mutual coverage and mutual homology between the predicted and translated ORF and the subject (a translated sequence in BAC) were computed. Finally, the predicted and translated ORF was examined if it contained a known or an unknown motif by InterPro.

The predicted ORFs were classified into six grades, A (highest) to X (lowest), with respect to the Blastp and InterPro results. The A and B grades were further divided into four sub-grades each accordingly to the InterPro results. The difference in the grade between the A and the corresponding B grades was due to the Blastp results that the former referred to the 'mutual' values and the latter to 'one-way' values. The highest grade (AAAA) ORF satisfied that the values of mutual coverage and homology were ⩾70% and contained at least one known motif in it. The C grades ORF did not have any homologue in BAC, but contained at least one known motif. The D grades ORF met the same Blastp requirement as the highest grades, but did not contain a known or an unknown motif. It was also described as a hypothetical gene by the submitter. The remaining ORFs in the E and X grades did not hit any extant sequence or motif. The present classification is essentially operational and not scientifically decisive. Nevertheless, the classification will be meaningful and useful, because it was made consistently by the methods and parameters that are all known to the public. In this respect, it is possible that the ORFs in the AAA and lower grades will rank up with further data submissions and development of the related methods in the future.

As a result, 848 383 and 1 254 150 ORFs were predicted in total for GIB ver. 2003 and 2004, respectively. Among them

Grade	Blastp hit: Coverage	InterProScan hit: Subject	Number of predicted ORFs: Ver. 2003	Ver. 2004
AAAA / AAA / AA / A; BBBB / BBB / BB / B	alignment/subject ≧70% & or alignment/ORF ≧70%	known motif	AAAA - A: 283,247	431,672
		unknown motif		
		no hit	BBBB - B: 7,208	10,250
		mishit		
C	no hit	known motif	4,680	7,511
D (hypothetical)	≧70% &	no hit	79,779	107,382
		Total	374,914	556,815
		INSDC	362,828	537,312

Mishit means to hit a putative membrane protein or an unknown protein

Figure 1. Classification of bacterial genes submitted to INSDC Four of the total six grades are presented. While '&' means that the mutual coverage and homology between a predicted ORF and a gene in INSDC are both ⩾70% in Blastp alignment, 'or' means that one-way coverage and homology between them are so.

374 914 and 556 815 ORFs were, respectively, classified into the AAAA to D grades (Figure 1). The numbers are slightly larger than those in BAC, 362 828 and 537 312, respectively. The major reason for the differences is considered to be due to the minimum length required in Glimmer 2.0. Glimmer with a shorter minimum length tends to predict more ORFs than that with a longer one. It is possible that submitters generally tend to set the minimum length at a larger value than a suitable one. In fact, however, many bona fide genes are small, 60–300 nt in length. Therefore, of the total INSDC bacterial genes, 556 815 genes may be currently reliable ones. Among them 78% belong to the AAAA to A grades. The whole results are available at our GTPS (Gene Trek in Prokaryote Space) viewer (http://gtps.ddbj.nig.ac.jp) (4). On this viewer we have also listed the ORFs that were newly found by the GTPS analysis (4). The new ORFs were all classified into the AAAA to C grades. We are now in the process of extending the examination and classification using GIB ver. 2005 that includes the released data from DDBJ by February 2006. In this case we use Glimmer 3.0 that has currently been available. The extension may produce more reliable genes than the current ones. When the difference in the number between two consecutive GTPS analyses is small enough, we will be able to state that the number is close to the total number of genes in the bacterial world.

CONCLUDING REMARKS

As proposed by Roberts (1), it will be profoundly meaningful that every suspicious gene in INSDC is examined *in vivo*/*vitro*. The examination will undoubtedly lead not only to the correction of many wrongly annotated genes but also the finding of new genes. We hope that the proposal will be implemented in the future. For this implementation our evaluation and classification will also be useful. We may have to seriously think about what has been done and what has not in information biology, because we are now in flood of data and sometimes drowned in it.

ACKNOWLEDGEMENTS

We thank all staff of DDBJ for the data collection, annotation, release, software development, and gene evaluation and classification. DDBJ is funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) with the management expenses grant for national university cooperation. Funding to pay the open access publication charges for this article was provided by the Japan Society for the Promotion of Science and Grant no. 16255006.

Conflict of interest statement. None declared.

REFERENCES

1. Roberts,R.J. (2004) Identifying protein function—a call for community action. *PLoS Biol*., **2**, E42.
2. Heurgue-Hamard,V., Champ,S., Engstrom,A., Ehrenberg,M. and Buckingham,R.H. (2002) The *hemK* gene in *Escherichia coli* encodes the N5-glutamine methyltransferase that modifies peptide release factors. *EMBO J*., **21**, 769–778.
3. Nakahigashi,K., Kubo,N., Narita,S., Shimaoka,T. and Goto,S. (2002) *HemK*, a class of protein methyl transferase with similarity to DNA methyl transferases, methylates polypeptide chain release factors, and *hemK* knockout induces defects in translational termination. *Proc. Natl Acad. Sci. USA*, **99**, 1473–1478.
4. Kosuge,T., Abe,T., Okido,T., Tanaka,N., Hirahata,M., Maruyama,Y., Mashima,J., Tomiki,A., Kurokawa,M., Himeno,R. *et al.* (2006) Exploration and grading of possible genes in 183 bacterial strains by a common fine protocol lead to new genes: Gene Trek in Prokaryote Space (GTPS). *DNA Res.,* in press.
5. Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R., Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T. *et al.* (2006) *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. *Nucleic Acids Res*., **34**, 1–9.
6. Okubo,K., Sugawara,H., Gojobori,T. and Tateno,Y. (2006) DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res*., **34**, D6–D9.
7. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*., **25**, 955–964.
8. Suzek,B.E., Ermolaeva,M.D., Schreiber,M. and Salzberg,S.L. (2001) A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, **17**, 1123–1130.
9. Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*., **26**, 544–548.
10. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*., **27**, 4636–4641.
11. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al*. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*., **29**, 37–40.
12. Griffiths,J.S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res*., **31**, 439–441.