

DDBJ with new system and face

H. Sugawara, O. Ogasawara, K. Okubo, T. Gojobori and Y. 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 13, 2007; Revised October 1, 2007; Accepted October 2, 2007

ABSTRACT

DDBJ (<http://www.ddbj.nig.ac.jp>) collected and released 1 880 115 entries or 1 134 086 245 bases in the period from July 2006 to June 2007. The released data contains the high-throughput cDNAs of cricket and high-quality draft genome of medaka among others. Our computer system has been upgraded since March 2007. Another new aspect is an efficient data retrieval tool that has recently been equipped and served at DDBJ. It is called All-round Retrieval for Sequence and Annotation, which enables the user to search for keywords also in the Feature/Qualifier of the International Nucleotide Sequence Database Collaboration (<http://www.insdc.org/>). We will also replace our home page with a more efficient one by the end of 2007.

INTRODUCTION

Through our service we have witnessed dramatic advancements in biology and the related areas in the past 20 or more years. For example, using genome sequence data for eubacteria, archaeobacteria and eukaryotes, some authors constructed a tree of life, which is the phylogenetic tree of the three super-kingdoms (1,2). Others reported a way to predict the number of genes at least in the bacterial world (3). The dramatic advancements prove our simple idea that the more data we collect and serve the more people make use of it for various purposes.

On the other hand, the recent development of sequencing machines such as 454 (by 454 Life Sciences), Solexa (by Illumina, Inc.) and SOLiD (by Applied Biosystems) makes us worrisome as well. According to some estimate, 5–10 tera bases will be sequenced by Solexa at one sequencing facility in a month in the near future. With the further development of the sequencing technology the whole genome of a person may repeatedly be submitted in the near future, as few examples warn (4). To cope with the expected situation of sequencing genes and genomes, we have recently upgraded our computer system and installed an efficient keyword search tool. We think that

the new computer and tool serve our data submitters and users better and make our job more effective and efficient.

In this article we will report on the data submissions to DDBJ in the past year, replacement of our computer system with an upgraded one, a new data retrieval tool and a new home page.

DATA SUBMISSIONS TO DDBJ IN THE LAST YEAR

In the period from July 2006 to June 2007, DDBJ collected and released the original data of 1 880 115 entries or 1 134 086 245 bases that were classified into the 19 International Nucleotide Sequence Database Collaboration (INSDC) divisions (5). More than 90% of the submissions came from Japanese researchers, and the rest were mainly from Chinese and Korean researchers.

The released data includes the high-throughput cDNAs (HTC) of cricket, *Gryllus bimaculatus* submitted from Tokushima University (6). The data amount is 32 010 entries that can be obtained through anonymous FTP with the file name, *Gryllus bimaculatus_HTC_070726_1.seq.gz*. Also included is 700 Mb of the high-quality draft genome data of medaka, *Oryzias latipes*, which was submitted from University of Tokyo and National Institute of Genetics (7). The data was carefully assembled and upgraded from the WGS data that was reported in our previous paper (8). The given accession numbers are BAAF03000000 (Hd-rR, version 0.9), BAAF04000000 (Hd-rR, version 1.0) BAAE01000000 (HNI) and ACAA00000001-ACAAA0356693 (5' SAGE tags). Although draft genome sequences of two fugu (blowfish) species are available, the high-quality draft genome of medaka will be quite useful particularly for the study of vertebrate evolution. The submitters of the genome data discussed, for example, that the medaka genome preserved its ancestral karyotype for more than 300 million years (7).

It is also noted that the current number of bacterial species/strains in the complete bacteria genome data repository, the Genome Information Broker, (GIB, <http://gib.genes.nig.ac.jp/>) (9), at DDBJ is 569 and keeps on growing rapidly. The species added in the past year include *Methanococcus maripaludis* (by Joint Genome Institute), *Saccharopolyspora erythraea* (by University of Cambridge), *Francisella tularensis* subsp. *tularensis*

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

(by UT Southwestern Medical Center), *Desulfotomaculum reducens* (by Joint Genome Institute), *Burkholderia vietnamiensis* (by Joint Genome Institute), *Herminiimonas arsenicoxydans* (by Genoscope), *Geobacillus thermodenitrificans* (by Nankai University), *Corynebacterium glutamicum* (by RITE) and many others. We also serve a complete virus genome data repository, GIB for Viruses (GIB-V, <http://gib-v.genes.nig.ac.jp/>) that now contains 31 486 virus genomes and genomic segments.

NEW COMPUTER SYSTEM

In July 2007, we celebrated the 20th anniversary of the public release of the DNA data. Our first release in July 1987 contained only 66 entries or 108 970 bases that were typed in from published papers. These numbers may be impressive in the comparison with the corresponding ones as of June 2007, 13 371 690 entries or 8 988 178 758 bases. This tremendous increase in the numbers perhaps reflects the remarkable advancement of research in biology and the related areas in Japan in the past 20 years. The ever-increasing amount of the data also makes us worry about our hardware and software facilities.

In March 2007, we completely replaced our computer system with an upgraded one. Major upgraded aspects are as follows. (i) The increase in the number of entries in making the flat files from 300 000 to 1 000 000 entries/day, (ii) the decrease in processing time in making a huge flat file; in case of four rice chromosomes, from 110 to 13 min, (iii) the decrease in processing time from 120 to 13 min for updating the live-list that lists the accession numbers and dates of the public release of the released entries; it is weekly updated to exchange the information about the currently released data with the EMBL Bank and GenBank, (iv) the increase in the number of ESTs in data processing from 40 000 to 800 000 entries/h and (v) the increase in the number of queries accepted at once by 1.5 times. Therefore, we will be able to cope with the increase in the number of data submissions for the next several years.

NEW KEYWORD SEARCH TOOL

Recently, we have installed a high-speed keyword search tool, All-round Retrieval for Sequence and Annotation (ARSA, <http://arsa.ddbj.nig.ac.jp/top-e.html>). The search logic behind ARSA is called SIGMA, which was invented by Arikawa and his colleagues (10,11). For a given query SIGMA makes it possible to retrieve all the right entries by checking the contents of a database just once, no matter how the query is complicated. The one time checking makes keyword search fast. SIGMA does not need an index file, which means that search can be made against the currently available data. SIGMA is implemented on the Shunsaku search engine developed by Fujitsu. The search engine operates in parallel for divided data, which makes the search even faster. ARSA also has a large scalability with an increasing amount of data. In theory, one search can be completed within 10 s irrespective of the data size and the query formula.

If the data increases more than 10 times larger than the current amount, however, we may have to increase the number units in the Shunsaku accordingly to keep the present search speed.

ARSA covers 23 databases including DDBJ, UniProt, PFAM, PDB and LENZYME. A special feature of ARSA is that it can also incorporate the terms defined by the Feature/Qualifier of INSDC. While this feature is very helpful for us to annotate the submitted data, it enables our user to perform data retrieval by using terms in the Feature/Qualifier. For example, you can search for CDSs (protein coding sequences) located on human Y chromosome, as shown in Figure 1. In the figure, the query formula is given on the top, and a part of the hit entries is given below with the accession numbers. By clicking one of the numbers you can see its contents. HUM in the last column stands for the human division. You can download the search result in Flat File, FASTA or XML, and also choose the items in the search results to be displayed on the computer screen and directly download them in tab-limited format. We also provide you with WebAPI (<http://xml.nig.ac.jp/> > <http://xml.nig.ac.jp/>) (12) so that you can customize ARSA by writing a program in Perl or JAVA. We will soon include KEGG (<http://www.genome.ad.jp/kegg/>) in ARSA and make the 24 databases simultaneously retrievable for common keywords.

NEW FACE OF DDBJ

We updated our home page (HP) in 2005 (13). We are again in the process of updating it rather drastically this time. Since the present HP holds many contents that have been added in an irregular sequence without much consideration for consistency, it is not really convenient now for our data submitters and users. The main point of the updating thus is to reach the almost every content with three clicks or less, which is now a common practice in making use of a HP. In the new HP when you click one of our main services, data submission, data retrieval, ftp/SOAP, statistics and inquiry, you can get the whole view of all contents for each service at once, and easily go to the one of them that you wish. The new HP will replace the present one by the end of 2007. We hope the new HP on the new computer system and tool will be more attractive to our data submitters and users worldwide.

ACKNOWLEDGEMENTS

We thank all DDBJ staff for the collection, annotation and distribution of the original data. We are also grateful to H. Yamada and T. Hosokawa for providing us with valuable information on ARSA. DDBJ is funded by Ministry of Education, Culture, Sports, Science and Technology (MEXT) with Management experience grants for national university cooperation. Funding to pay the Open Access publication charges for this article was provided by Japan Society for Promotion of Science Grant no. 16255006.

Conflict of interest statement. None declared.

Query: (/ENTRY/DDBJ/feature-table/feature{/f_key == 'source' AND /f_qual/qualifier{/q_name == 'chromosome' AND /q_value == 'Y' } }) AND /ENTRY/DDBJ/feature-table/feature/f_key == 'CDS'

1,033 results found in DDBJ Download all the contents displayed in Tab Separated Value format

Download in TSV

Download the listing results.

Primary Accession Number	Definition	Division
AF000979	Homo sapiens testis-specific Basic Protein Y 1 (BPY1) mRNA, complete cds	HUM
AF000980	Homo sapiens testis-specific Basic Protein Y 2 (BPY2) mRNA, complete cds	HUM
AF000981	Homo sapiens testis-specific ChromoDomain Y isoform 1 (CDY1) mRNA, alternatively spliced, complete cds	HUM
AC006335	Homo sapiens BAC clone RP11-492C2 from Y, complete sequence	HUM
AF000984	Homo sapiens dead box, Y isoform (DBY) mRNA, alternative transcript 2, complete cds	HUM
AF000985	Homo sapiens dead box, Y isoform (DBY) mRNA, alternative transcript 1, complete cds	HUM
AF000986	Homo sapiens chromosome Y ubiquitin specific protease 9 (USP9Y) mRNA, complete cds	HUM
AF000987	Homo sapiens eIF-1A, Y isoform (EIF1AY) mRNA, complete cds	HUM

Figure 1. Example of keyword search by ARSA. Keywords used are 'source' (Feature), 'chromosome' (Qualifier belonging to 'source') and CDS (Feature). 'Chromosome' has a value attribute to which 'Y' is given for specifying chromosome Y.

REFERENCES

- Wolf, Y.I., Rozogin, I.B., Grishin, N.V. and Koonin, E.V. (2002) Genome trees and the tree of life. *Trends Genet.*, **18**, 472–479.
- Fukami-Kobayashi, K., Minezaki, Y., Tateno, Y. and Nishikawa, K. (2007) A tree of life based on protein domain organizations. *Mol. Biol. Evol.*, **24**, 1181–1189.
- Kosuge, T., Abe, T., Okido, T., Tanaka, N., Hirahata, M., Maruyama, Y., Mashima, J., Tomiki, A., Kurokawa, M. *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.*, **13**, 245–254.
- Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
- 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.
- Shinmyo, Y., Mito, T., Uda, T., Nakamura, T., Miyawaki, K., Ohuchi, H. and Noji, S. (2006) Brachyenteron is necessary for morphogenesis of the posterior gut but not for anteroposterior axial elongation from the posterior growth zone in the intermediate-germband cricket *Gryllus bimaculatus*. *Development*, **133**, 4539–4547.
- Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K. *et al.* (2007) The medaka draft genome and insights into vertebrate genome evolution. *Nature*, **447**, 714–719.
- Sugawara, H., Abe, T., Gojobori, T. and Tateno, Y. (2007) DDBJ working on evaluation and classification of bacterial genes in INSDC. *Nucleic Acids Res.*, **35**, D13–D15.
- Fumoto, M., Miyazaki, S. and Sugawara, H. (2002) Genome Information Broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more. *Nucleic Acids Res.*, **30**, 66–68.
- Arikawa, S. (1981) One-way sequential search systems and their powers. *Bull. Math. Stat.*, **19**, 69–85.
- Arikawa, S., Shinohara, T. and Takeya, S. (1989) SIGMA: A text database management system. *Berliners Informatik Tag*, 72–81.
- Sugawara, H. and Miyazaki, S. (2003) Biological SOAP servers and web services provided by the public sequence data bank. *Nucleic Acids Res.*, **31**, 3836–3839.
- Tateno, Y., Saitou, N., Okubo, K., Sugawara, H. and Gojobori, T. (2005) DDBJ in collaboration with mass-sequencing teams on annotation. *Nucleic Acids Res.*, **33**, D25–D28.