Nucleic Acids Research, 2004, Vol. 32, Database issue ***D31–D34***
DOI: 10.1093/nar/gkh127

DDBJ in the stream of various biological data

S. Miyazaki, H. Sugawara, K. Ikeo, T. Gojobori and Y. Tateno*

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

Received September 16, 2003; Revised October 3, 2003; Accepted October 23, 2003

ABSTRACT

In the past year we at DDBJ (http://www.ddbj.nig.ac.jp) have made a steady increase in the number of data submissions with a 50.6% increment in the number of bases or 46.5% increment in the number of entries. Among them the genome data of man, ascidian and rice hold the top three. Our activity has extended to providing a tool that enables sequence retrieval using regular expressions, and to launching our SOAP server and web services to facilitate the acquisition of proper data and tools from a huge number of biological data resources on websites worldwide. We have also opened our public gene expression database, CIBEX.

INTRODUCTION

Recent advances in biology and bioinformatics are perhaps epitomized by the emergence of new fields such as transcriptomics, proteomics and phenomics. Protein–protein interaction and pathway analysis will also be promising subjects in these disciplines. However, the emergence of these fields and subjects does not undermine the importance of DNA sequences—actually, quite the opposite to that.

In July 2003 the 19th International Congress of Genetics (ICG) was held in Melbourne. We believe that ICG is one of the most comprehensive international conferences in the field of genetics. The main theme of the congress this time was 'Genomics—The Linkage to Life', as symbolizing the cutting edge of genetics today. The first speaker of the congress was Francis Collins, who emphasized that we were not yet in the post-genome era but right in the genome era. What he partly implied in his lecture, we think, is that there are still many things to do in biology and bioinformatics directly using DNA sequence data. In fact, we do not yet know the exact total number of genes in the human genome not to mention most of their functions. It is also noted that the number of BLAST searches (1–3) against the DDBJ database has continued to grow. Occasionally, we face the delightful and difficult situation in which the number of BLAST searches alone exceeds the capacity of our computer system, which is composed of several high-speed servers. The *in silico* aspect of biology has widened its niche.

The number of data submissions to DDBJ has steadily increased in the past year. Since >95% of the submissions to DDBJ were made by Japanese researchers, the data submitted somewhat reflect the trend in biological research in Japan, as will be mentioned below. We will report our recent activities in this paper.

DATA COLLECTION AT DDBJ

Since the collection of authentic and original DNA sequence data is one of our core activities, we have continuously refined our data submission tools, Sakura (4) and MST (5), for the convenience of data submitters. We believe that this has helped data submitters to perform their data submissions more easily and efficiently. In fact, we collected 1 033 046 154 bases or 1 594 276 entries in the past year alone, indicating a 50.6% increment in the number of bases or 46.5% in the number of entries since last year. These numbers are similar to those of the previous year (6) securing a steady increase in the number of data submissions to DDBJ.

In Table 1 we list the top 30 species for which DNA sequence data were submitted to DDBJ in the past year. As mentioned above, Japanese researchers contributed most of them. In the table, *Homo sapiens* ranked the first, as expected from the fact that the Japanese human genome sequencing team took third place, after the United States and United Kingdom teams, in making a contribution to sequencing the whole human genome. In particular the Japanese team made a significant contribution to sequencing chromosomes 21 (7), 22 (8), 11 (9), 18 (9) and 6 (10). Note that the number of bases in Table 1 is much less than the total number that the Japanese human genome team has sequenced and submitted, because it is the number for the last year only. Now one can get access to the human genome sequence of 2.9 Gb in length with an accuracy of 99.99% and coverage of 98% at DDBJ. This means that one can conduct data retrieval against the entire euchromatic regions of all the human chromosomes. In addition, the quality (Phrap) value for the genome sequence is accessible.

The second place in Table 1 is *Ciona intestinalis* (ascidian) whose genome was sequenced by N. Satoh of Kyoto University and his colleagues (11). This species is now considered as a serious candidate for the closest species to the origin of chordata and is paid strong attention. Therefore, the sequence data of this species will contribute to elucidating the evolutionary origin of chordata, if the data are widely used and

*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, all authors should be regarded as joint First Authors

Table 1. Top 30 species for which DNA sequence data were submitted to DDBJ in the past year

	Organism	Bases	Entries
1	*Homo sapiens*	205 960 983	93 326
2	*Ciona intestinalis*	202 387 657	323 232
3	*Oryza sativa*	200 280 615	46 371
4	*Lotus corniculatus*	41 244 748	70 522
5	*Mus musculus*	24 198 182	37 625
6	*Lotus japonicus*	13 725 627	1261
7	*Bombyx mori*	9 158 630	13 871
8	*Synthetic construct*	8 824 439	55 294
9	*Bos taurus*	6 756 852	13 665
10	*Nicotiana tabacum*	5 047 374	9593
11	*Porphyra yezoensis*	5 039 707	10 638
12	*Zinnia elegans*	4 898 133	9721
13	*Caenorhabditis elegans*	4 439 328	9731
14	*Hordeum vulgare*	3 534 217	6605
15	*Dugesia japonica*	3 164 718	6461
16	*Cryptomeria japonica*	2 441 154	4736
17	*Arabidopsis thaliana*	1 925 925	622
18	*Hordeum vulgare*	1 563 184	2889
19	*Thermus thermophilus*	1 205 571	1162
20	*Pan troglodytes verus*	1 107 378	2043
21	*Dictyostelium discoideum*	1 036 793	2337
22	*Hemicentrotus pulcherrimus*	1 029 245	2473
23	*Sus scrofa*	967 882	1397
24	*Ctenocephalides felis*	928 443	1882
25	*Rattus norvegicus*	832 791	716
26	*Macaca fascicularis*	823 098	442
27	*Mus* sp.	680 855	877
28	*Drosophila melanogaster*	657 842	1154
29	Uncultured bacterium	645 368	1107
30	*Pinus radiata*	614 941	985

analyzed. It is genome sequence data that could clearly and quantitatively unite the most primitive chordata such as *C.intestinalis* and the furthest chordata such as *H.sapiens*. This makes evolutionary analysis possible even for such a distantly related pair.

Oryza sativa is ranked third in Table 1. There are two main studies (12,13) in which Japanese researchers took the initiative in sequencing and analyzing the genome of this species. The complete sequence of chromosome 1 of 43.3 Mb in length has been shown to contain 6756 protein coding genes of which 3161 (46.8%) share homology with those of *Arabidopsis thaliana* (12). This finished sequence shows that the quality of a finished sequence is important for further possible research not only by the submitters themselves but also by others (12). In the other study (13) the authors collected and sequenced 28 469 full-length cDNA clones of *O.sativa*, and have indicated that ~20 000 transcription units exist in the whole genome of this species. It is noteworthy that there are three different terms: coding gene, full-length cDNA and transcription unit, to denote the traditional gene defined in the early 20th century. We may have to redefine what a gene is now in the genome era. Otherwise, there will be some confusion over, for example, the determination of the exact total number of genes in the human genome.

There are other interesting species among the top 30 species such as *Pan troglodytes* and *Macaca fascicularis.* The Japanese chimpanzee genome sequencing team has been in collaboration with the Asian and European teams, and the genome sequence data of chromosome 22 was released from DDBJ/EMBL/GenBank in October 2003. The Japanese monkey is also an interesting species, because it is very close to man and could be used for experiments more freely than chimpanzees.

OTHER ACTIVITIES AT DDBJ

Currently, there are a huge number of biological data resources including databases and data retrieval and analysis tools available online worldwide. However, it is often laborious or almost impossible for ordinary biologists to write computer programs that would enable them to find useful resources, to prepare and send a proper query and then to process the outcome. To help such researchers, we have implemented a Simple Object Access Protocol (SOAP) server and web services at http://www.xml.nig.ac.jp (14).

Sequence homology searches using BLAST or FASTA (15) are one of the most popular practices against DDBJ and perhaps the EMBL Bank and GenBank. However, they sometimes face the difficulty of not finding sequences homologous or similar to their queries in the database. This difficulty may occur more often as more new sequences accumulate at DDBJ. For such researchers it may be enough to find a similar part of sequences to their queries in a specified form. A functionally or structurally important region in a gene or a protein is usually made up of a small number of bases or amino acids in a specified arrangement. To help them, we have provided a tool, SQmatch, that enables data retrieval against the DDBJ database by giving a query in the regular expression (Fig. 1). There are 10 different regular expressions that can be used in SQmatch (see the website, http://sqmatch.ddbj.nig.ac.jp/sqmatch_en.html for details).

In May 2003 the representatives of the EMBL Bank, GenBank and DDBJ held the 16th International DNA Data Banks Collaborative Meeting at the NCBI in the United States. The purpose of this annual meeting is to discuss and solve practical problems raised in daily database work. Among many items discussed at this meeting we particularly spent time on the relaxation of the length limit of a submitted sequence, third party annotation (TPA) and a common XML data exchange format among the three data banks. As a result, we have decided to relax the length limit (350 kb) taking into account well-used software tools that will be affected by the relaxation. The three data banks have recently started accepting TPA submissions. TPA data are defined as submitted data that are composed of one or more related DNA sequences, of which the submitters themselves might sequence some of these, conducted experiments on the sequences and published a paper about them. The publication of a pertinent paper is required, because we at the three data banks cannot evaluate every TPA submission. We have so far collected more than 30 TPA submissions at DDBJ that will be made public when the pertinent papers are published. The three data banks have also begun working together to define the common XML format in order to realize data exchange in XML among them in the near future.

CONCLUDING REMARKS

Although we have placed emphasis on the importance of DNA sequence data in biology and bioinformatics, we have never regarded other biological data as less important in our

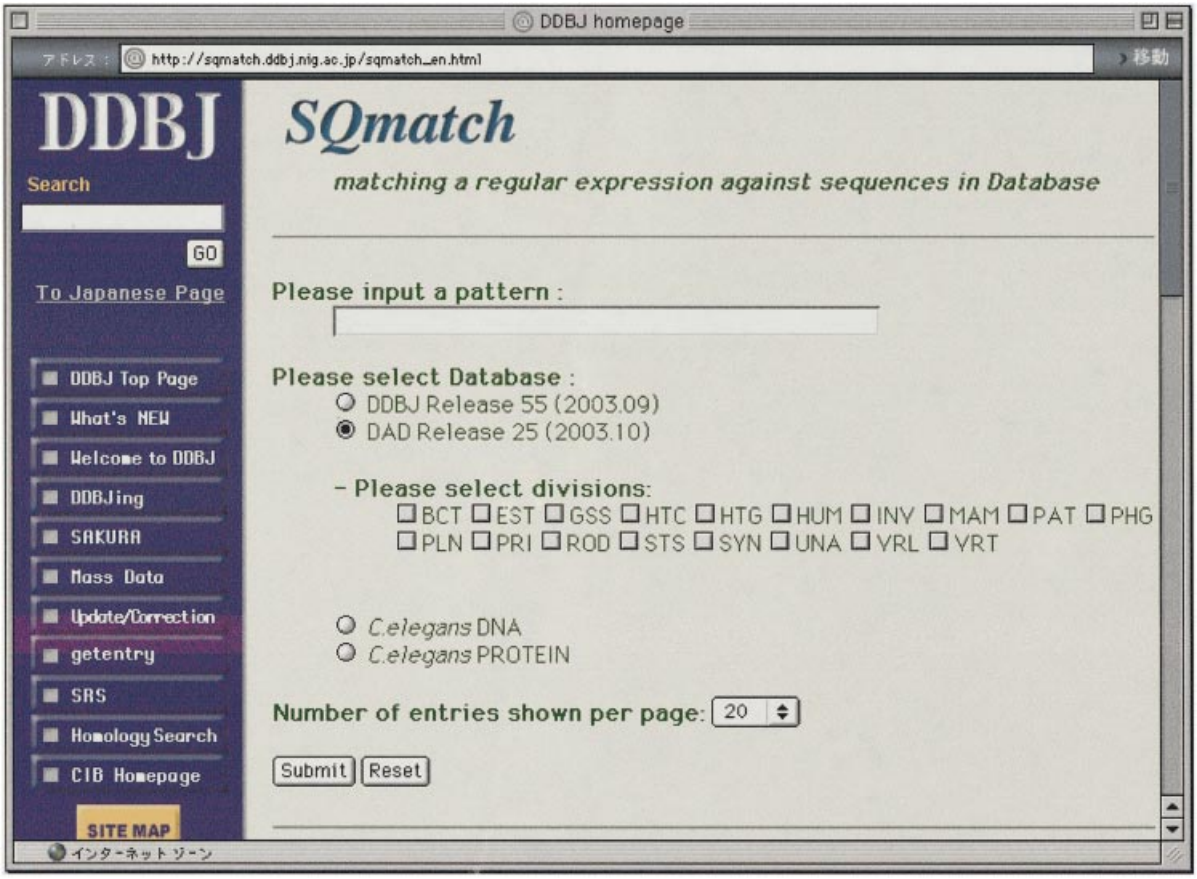

Figure 1. First page of the website of SQmatch at DDBJ.

activities at DDBJ. In fact, we collaborated with the Japan Biological Information Research Center to host an international human cDNA annotation jamboree (H-invitational I) in September 2002. At this jamboree about 120 experts from many countries gathered together and annotated human cDNA data from several different angles including protein 3D structure, gene expression and pathways. We have also participated in the activity of the Microarray Gene Expression Data Society (MGED, http://www.mged.org), which has worked on the standardization of microarray data (16), data sharing and exchange and ontology. Along this line, we have opened our public gene expression database, CIBEX (17).

ACKNOWLEDGEMENTS

We thank all the DDBJ members for making it possible to run DDBJ well in collaboration with the EMBL Bank and GenBank. In particular we thank Sadahiko Misu for the preparation of materials for this paper. We are also grateful to the Ministry of Education, Science, Culture, Sports and Technology for their enduring financial support.

REFERENCES

1. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
2. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
3. Zhang,J. and Madden,T.L. (1997) PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res.*, **7**, 649–656.
4. Yamamoto,H., Tamura,T., Isono,K., Gojobori,T., Sugawara,H., Nishikawa,K., Saitou,N., Imanishi,T., Fukami-Kobayashi,K., Ikeo,K. *et al.* (1996) SAKURA: A new data submission system of DDBJ to meet users' needs in the age of mass production of DNA sequences. In Akutsu,T., Asai,K., Hagiya,M., Kuhara,S., Miyano,S. and Nakai,K. (eds), *The Proceedings of the Seventh Workshop on Genome Informatics*. Universal Academy Press, Tokyo, pp. 204–205.
5. Tateno,Y., Fukami-Kobayashi,K., Miyazaki,S., Sugawara,H. and Gojobori,T. (1998) DNA Data Bank of Japan at work on genome sequence data. *Nucleic Acids Res.*, **26**, 16–20.
6. Miyazaki,S., Sugawara,H., Gojobori,T. and Tateno,Y. (2003) DNA Data Bank of Japan (DDBJ) in XML. *Nucleic Acids Res.*, **31**, 13–16.
7. Hattori,M., Fujiyama,A., Taylor,T.D., Watanabe,H., Yada,T., Park,H.-S., Toyoda,A., Ishii,K., Totoki,Y., Chol,D.-K. *et al.* (2000) The DNA sequence of human chromosome 21. *Nature*, **405**, 311–319.
8. Dunham,I., Shimizu,N., Roe,B.A., Chissoe,S. Hunt,A.R., Collins,J.E., Bruskiewich,R., Beare,D.M., Clamp,M., Smink,L.J. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.
9. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* the International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
10. The MHC Sequencing Consortium (1999) Complete sequencing and gene map of a human major histocompatibility complex. *Nature*, **401**, 921–923.

11. Satoh,N., Satou,Y., Davidson,B. and Levine,M. (2003) *Ciona intestinalis*: an emerging model for whole-genome analyses. *Trends Genet.*, **19**, 776–781.
12. Sasaki,T., Matsumoto,T., Yamamoto,K., Sakata,K., Baba,T., Katayose,Y., Wu,J., Niimura,Y., Cheng,Z., Nagamura,Y. *et al.* (2002) The genome sequence and structure of rice chromosome 1. *Nature*, **420**, 312–316.
13. Kikuchi,S., Satoh,K., Nagata,T., Kawagashira,N., Doi,K., Kishimoto,N., Yazaki,J., Ishikawa,M., Yamada,H., Ooka,H. *et al.* Rice Full-Length cDNA Consortium; National Institute of Agrobiological Sciences Rice Full-Length cDNA Project Team; Foundation of Advancement of International Science Genome Sequencing & Analysis Group; RIKEN (2003) Collection, mapping and annotation of over 28 000 cDNA clones from japonica rice. *Science*, **301**, 376–379.
14. 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.
15. Lipman,D.J. and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.
16. Editorial (2002) Microarray standards at last. *Nature*, **419**, 323.
17. Ikeo,K., Ishi-i,J., Tamura,T., Gojobori,T. and Tateno,Y. (2004) CIBEX: Center for Information Biology Gene Expression Database. *Biologies*, in press.