16–20 Nucleic Acids Research, 1998, Vol. 26, No. 1

DNA Data Bank of Japan at work on genome sequence data

Yoshio Tateno*, Kaoru Fukami-Kobayashi, Satoru Miyazaki, Hideaki Sugawara and Takashi Gojobori

Center for Information Biology, National Institute of Genetics, Yata, Mishima 411, Japan

Received September 2, 1997; Revised and Accepted October 15, 1997

ABSTRACT

We at the DNA Data Bank of Japan (DDBJ) (http://www.ddbj.nig.ac.jp) have recently begun receiving, processing and releasing EST and genome sequence data submitted by various Japanese genome projects. The data include those for human, *Arabidopsis thaliana*, rice, nematode, *Synechocystis sp.* and *Escherichia coli*. Since the quantity of data is very large, we organized teams to conduct preliminary discussions with project teams about data submission and handling for release to the public. We also developed a mass submission tool to cope with a large quantity of data. In addition, to provide genome data on WWW, we developed a genome information system using Java. This system (http://mol.genes.nig.ac.jp/ecoli/) can in theory be used for any genome sequence data. These activities will facilitate processing of large quantities of EST and genome data.

INTRODUCTION

Since the publication of the papers outlining the problems of sequencing entire genomes (1,2), great progress has been made. Two problems have been faced in this area, however; one is the necessary advancement in sequencing technology, and the other concerned the handling of massive amounts of sequence data produced. The first problem has been less burdensome, thanks to various aspects of technological breakthroughs (e.g., 3,4).

The International Nucleotide Sequence Databases (INSD) have been actively involved in dealing with the second problem. INSD is a tripartite international collaboration between the EMBL Nucleotide Sequence Database, GenBank and the DNA Data Bank of Japan (DDBJ). In fact, large amounts of EST (5) and STS (6) data for human, mouse, nematode, rice and other organisms have been submitted, processed and published by INSD. The statistics of the most recent data release (DDBJ release 30) indicate that >65% of the total amount are EST and STS data for various organisms. Note that the three data banks daily exchange data they collect and process, so that each bank provides virtually the same quality and amount of data. In addition, INSD has received, processed and released data on entire genome sequences of bacterial and yeast species (7–14).

We believe that EST and genome sequence data have unmeasurable value not only in the field of biology but also in medicine and agriculture. In particular, the recent achievement of sequencing the whole genome of *Helicobacter pylori* (11) is remarkable because, with this information, we will be able to study the etiology of stomach ulcers at the molecular level, and develop an effective and efficient medicine for preventing and treating the disease. A recent report (15) also revealed that an international consortium for sequencing the genome of *Plasmodium falciparum* will be launched soon. The consortium will contribute to elucidating the molecular etiology of malaria and developing a cure for it.

As we ourselves have been carrying out molecular evolutionary studies, we are particularly interested in the origin and evolution of the genome as a whole. It is now clear that present genomes have not originated from a unique common ancestral form but have been organized with parts of eubacterial, archaebacterial and eukaryotic chromosomes. For example, the initial enzyme of chorismate biosynthesis of *H.pylori* is much closer to a homologue of *Arabidopsis thaliana* than to that of *Escherichia coli*, while the tRNA synthetases of *H.pylori* are more similar to eubacterial homologues than eukaryotic ones (11). It is necessary to keep this in mind when we construct a phylogenetic tree for a homologous group for a particular gene. We may obtain a totally different tree for a different homologous group depending on the origin of the gene. It has also been shown that the arrangement of genes in a genome has changed dynamically in the course of evolution (16). Along this line, it is interesting to refer to the finding that the genome of *Saccharomyces cerevisiae* resulted from duplication of the entire genome of its ancestor ~100 million years ago (17).

We need to mine genome sequence data for the riches they contain in view of molecular evolution and information biology. In this paper we report our activities at DDBJ focusing mainly on the collection, processing and release of genome sequence data. We also briefly refer to a strategy for coping with incoming mass EST and genome sequence data.

GENOME PROJECTS IN JAPAN

The statistics for data submissions to DDBJ show that EST data were first submitted in 1992, and genome data in 1996. Since then the proportion of EST submissions has steeply increased, and it is now >60% of the total submissions to DDBJ. Major species for

*To whom correspondence should be addressed. Tel: +81 559 81 6857; Fax: +81 559 81 6858; Email: ytateno@genes.nig.ac.jp

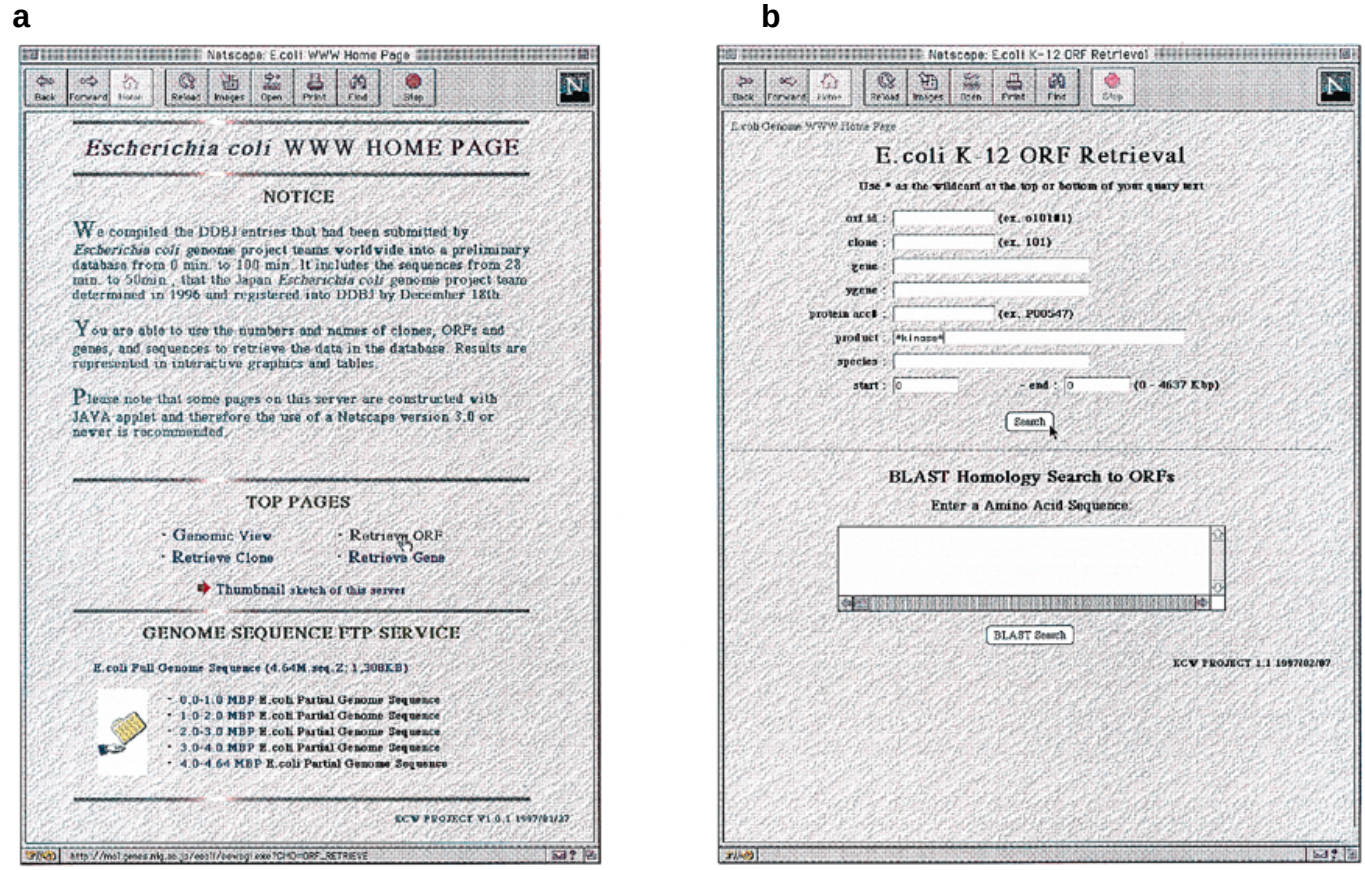


Figure 1. (**a**) Top page of the *E.coli* genome information service. It shows four functions: Genomic View, Retrieve Clone, Retrieve ORF and Retreive Gene. The size of the genome in mega base pairs is also given. (**b**) Page for ORF retrieval. The retrieval can be made for a given ORF id, clone number, gene name, product name and others. The figure shows the result obtained by retrieval for a keyword, kinase.

which submissions have been made include rice, nematode and human. In addition, three noteworthy genome projects have been undertaken in Japan; on *Synechocystis* (12), *E.coli* (e.g. 14) and yeast (18). The results of the three projects were submitted and processed at DDBJ and are now retrievable through INSD. The project on *Synechocystis* at the Kazusa DNA Research Institute succeeded in sequencing the complete genome sequence of a species for the first time in Japan. The Japanese *E.coli* genome team had long made efforts in sequencing the whole genome and finally finished a stretch covering 0–68.8 minutes. The sequence data were annotated in collaboration between the genome team and DDBJ (19). The US team, however, went a bit farther in completing the whole genome (13). At any rate, the availability of the complete genome sequence of *E.coli* will have a great impact on many areas of biology, because this organism has been extensively studied for many years. The *E.coli* genome could be regarded as the standard in bacterial genomes to be sequenced. The genome sequence data will be quite useful as a 'DNA language dictionary', when elucidating functions of a stretch of DNA for which only a base pair arrangement is known.

Currently, several genome projects are underway in Japan. The human genome project has been sequencing mainly chromosome 21 (20–22). One of the outcomes of this project could be the elucidation of the molecular etiological mechanism of Down's syndrome. Actually, a stretch of 9 kb in the chromosome has been considered to be responsible for causing the syndrome (23), though the actual mechanism in which this stretch plays a role remains to be elucidated. The Kazusa DNA Research Institute has begun a project for sequencing the *A.thaliana* genome (24), and subsequently submitted data to DDBJ. At present, the data on 20 clones or 1 622 296 bases from this institute are available through INSD. The institute has participated in an international consortium of the *Arabidopsis* genome project, and is supposed to be responsible for sequencing a certain proportion of the total genome. Thus, the institute will keep on sequencing it and submitting data.

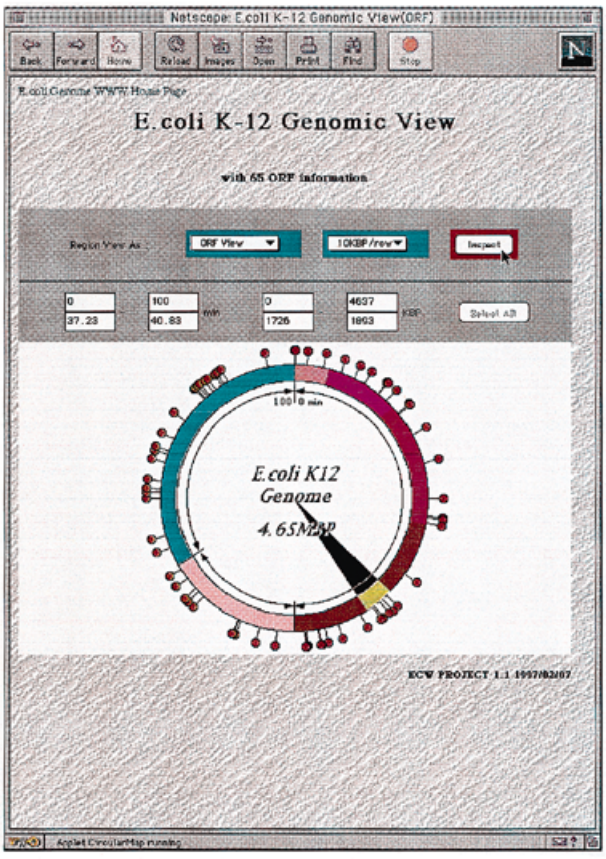
Also noteworthy is a *Caenorhabditis elegans* genome project led by Y. Kohara at our institute. The project has submitted ~42 000 EST sequence data to DDBJ, and will continue submissions. This is one of the most extensive *C.elegans* genome projects in the world. Another typical project in Japan is the rice genome project at the National Institute of Agricultural Research. They have actively sequenced rice EST sequences, and submitted them to DDBJ. The project is undoubtedly the most productive rice genome project in the world.

For each of those projects, we at DDBJ have formed a team to discuss with the project people the submission, processing and release of data, before submissions are made. We also developed a mass submission tool to deal with a large quantity of data. These

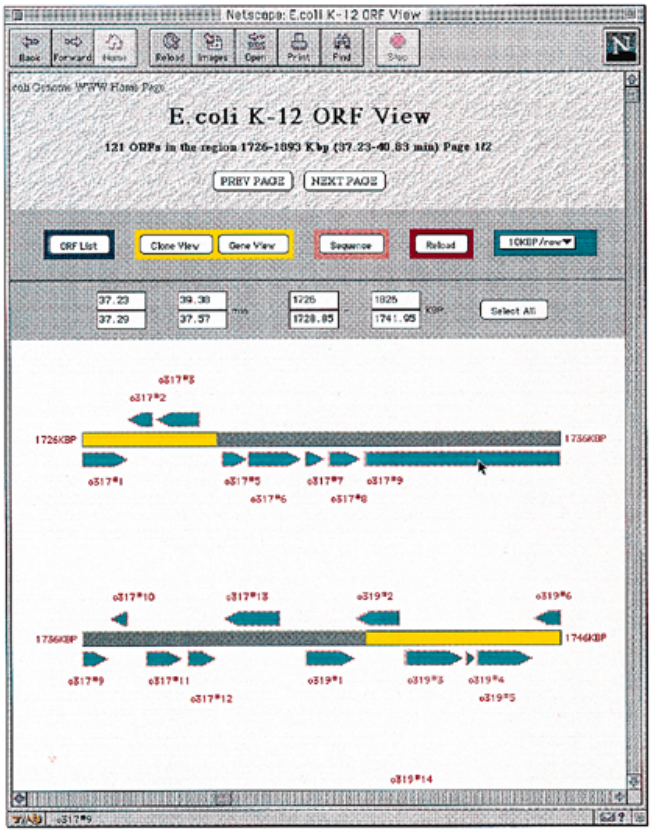
a


Netscape: ORF SEARCH RESULT
E.coli Genome WWW Home Page
ORF LIST: Found 65 matches containing orf id [], clone [], gene [], ygene [], protein acc# [], product [*kinase*], species [], start-end: [0-0] (PAGE 1/4)
PRODUCT
SCORE
PREV PAGE
NEXT PAGE
GENOMIC VIEW
ECW PROJECT 1.1 1997/02/07

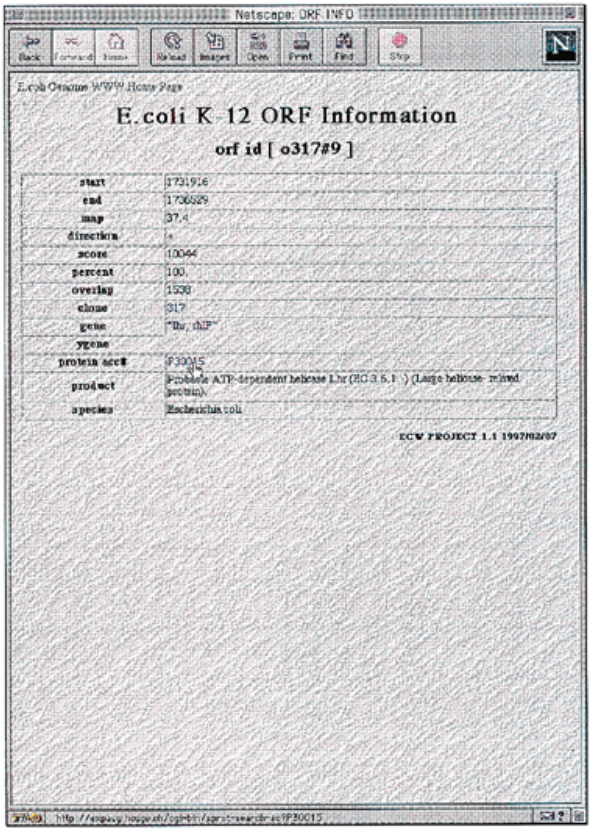
b


Netscape: E.coli K-12 Genomic View(ORF)
E.coli Genome WWW Home Page
E.coli K-12 Genomic View
with 65 ORF information
Region View As :
ORF View
1CKBP/row
Inquest
E.coli K12
Genome
4.65MBP
ECW PROJECT 1.1 1997/02/07

c


Netscape: E.coli K-12 ORF View
E.coli K-12 ORF View
121 ORFs in the region 1726-1893 Kbp (37.23-40.83 min) Page 1/2
PREV PAGE
NEXT PAGE
ORF List
Clone View
Gene View
Sequence
Reload
1CKBP/row
Select All
1726KBP
1756KBP
1746KBP

d


Netscape: ORF INFO
E.coli K-12 ORF Information
orf id [o317#9]
start 1731916
end 1736929
map 37.4
direction +
score 10044
percent 100.
overlap 1538
clone 317
gene "lhr, rhlF"
ygene
protein acc# P30015
product Probable ATP-dependent helicase Lhr (EC 3.6.1 -) (Large helicase-related protein).
species Escherichia coli
ECW PROJECT 1.1 1997/02/07

have facilitated the process from submission to release of genome sequence data at DDBJ.

GENOME INFORMATION SERVICE AT DDBJ

As mentioned above, we collected, processed and released *E.coli* genome data in collaboration with the Japanese *E.coli* genome team. During the collaboration we realized that we would serve researchers better if we published the genome data not only through our ordinary database but also on the World Wide Web (WWW). To implement the latter service, we developed a genome information broker operating on WWW. Since the details of this software will be published elsewhere, we introduce it briefly here.

The genome information broker was developed using a Java applet in order to facilitate graphic, dynamic and interactive processing of genome sequence data and to display them on the computer screen. The Java applet is employed for processing genome information that is retrieved from the DDBJ database by a CGI program also developed by us. The broker is divided mainly into three parts, for browsing information on genome regions, retrieving clone information, and retrieving information on ORFs and genes.

We applied the broker first to the *E.coli* genome sequence data mentioned above. As a result, we are now providing genome data on the WWW. The home page for this information service is illustrated in Figure 1a. The page shows that the service is four-fold, the Genomic View, Retrieve Clone, Retrieve ORF and Retrieve Gene. The first function is driven by the browsing information part in the broker. The second one is carried out by the retrieving clone information part, and the last two are realized by retrieving information on the ORF and gene part in the broker.

When you click the Retrieve ORF function, for example, you are led to the page shown in Figure 1b. On this page we can retrieve data with respect to a given ORF, clone, gene, or to a gene product. You can also carry out a BLAST homology search (25,26) for a given nucleotide or amino acid sequence. If you retrieve for a keyword, kinase, for example, you obtain the result given in Figure 2a. The figure lists part of the retrieved ORFs with the identification number, starting and ending positions of the sequence, location on the plus or minus strand, clone number, possible gene name, protein product, and others.

If you click the Genome View button on the same page, you move to the page given in Figure 2b. On this page a circular chromosome of *E.coli* is illustrated with many protruding spikes with a small circle on the top. Each spike corresponds to an ORF/gene in Figure 2a, so that you will be able to locate the chromosomal position of an ORF/gene in base number or minute. If you want to know more detailed chromosomal locations for ORF/genes, you first indicate the relevant region in the chromosome, select the ORF View function, and click the Inspect button as shown in the figure. Then the broker will guide you to the page shown in Figure 2c, in which the ORFs/gene is designated as a bar with a sharp end along a region of the genome. The sharp end refers to the 3′ terminal, and the blunt end to the 5′ terminal. An ORF/gene designated as a bar with the sharp end on the right is located on the plus strand, and the one with the sharp end on the left is located on the minus strand. The two figures tell you the physical relationships between an ORF/gene and other genes in question on the chromosome. They also enable us to compare chromosomal locations of particular homologous genes between different species, and perhaps leads us to making inferences about the evolution of gene arrangements in genomes. In Figure 2d detailed information on an ORF/gene can be given. You can get access to the *E.coli* genome information system on http://mol.genes.nig.ac.jp/ecoli/ . This information system has apparently attracted many researchers worldwide; it is currently accessed >8000 times a month.

We also applied the broker to the genomes of *Haemophilus influenzae* (7), *Mycoplasma genitalium* (8), *Methanococcus jannaschii* (9), *Synechocystis* PCC6803 (12) and *Mycoplasma pneumoniae* (27). Though the genome information systems for these species are not yet as complete as the *E.coli* genome information system, they are available on http://ddbjs4d.genes.nig.ac.jp:8880/ . The broker can be applied in theory to any chromosome irrespective of whether its shape is circular or linear. Actually, we are now extending its functions to be used also for linear chromosomes. We will continue to apply it to genome sequence data, whenever available and suitable, like the data on the newly released genome of *H.pylori* and the genome of *P.falciparum* to be completed in the near future.

By using the genome information system, you can check if the *E.coli* genome contains homologues to a stretch of DNA sequence or an amino acid counterpart of interest. In this way, even if you know nothing about your sequence except for the nucleotide arrangement, you could perhaps obtain a very good clue for inferring its function. Since this system also provides you with information on the flanking regions of a particular ORF/gene, you will be able to carry out PCR for finding a homologue in other species, using the information for designing the primers. In molecular phylogeny this approach is particularly useful, when you lack data on a homologue of a species for which you want to clarify its phylogenetic position among related species.

CONCLUDING REMARKS

As comprehensive data banks of nucleotide sequences, we face two problems now; one is to how to deal with a huge number of fragmented sequences like EST, and the other is to cope with a very long stretch of a genome sequence. At the annual INSD Collaborator's Meeting held at the National Center for Biotechnology Information in USA in 1997 we discussed how to deal with ever-increasing EST and genome sequence data. One of the outcomes of this meeting was to create a new division (called the

Figure 2. (**a**) List of ORFs retrieved. They are given with the identification number, chromosomal position, clone number, gene name and possible protein product. (**b**) Circular presentation of the genome with the ORFs retrieved. The ORFs are shown by protruding spikes. You can select any region of the genome as indicated in light green in the genome. (**c**) ORFs/genes represented along the chromosomal regions. An ORF/gene is illustrated as a bar with a sharp point. The blunt end refers to 5′ terminal and the sharp end to 3′ terminal of an ORF/gene. An ORF/gene with the sharp point on the right is located in the plus strand and the one with the left sharp point is located in the minus strand. (**d**) Detailed information about an ORF/gene. For a particular ORF/gene such pieces of information as the starting and ending positions in the genome, map position, possible gene name, possible products are obtained.

constructed or CON division), which stores for retrieval constructed sequences that are made up of ordered fragmented sequences on the same chromosome. We know, however, that this is not enough.

Until several years ago the data amount in INSD was doubled every five years or so. It is now doubled almost every year, because of advancements of sequencing technology and the start of genome projects worldwide. Perhaps, we can say that we have come to the second leap. The first one was of course brought about by the invention of the sequencing methods 20 years ago (28,29). One of the obvious problems in this steep growth concerns information storage. With many on-going genome projects and more to come in mind, we can expect that the growth rate will continue to increase for the next ten years at least. One report (30) notes that the human genome projects will continue until at least the year 2005 before the entire genome is sequenced. Thus, sooner or later we will have to alleviate the storage burden. As the EST data greatly surpasses the others in total amount in the INSD databases, we may have to start with them.

Since the EST data are known to be largely redundant, we have to try to reduce redundancy and keep unique sequences only. If EST sequences on the same chromosome are combined together, the storage problem would further be alleviated. This will also help save time in retrieval from the INSD databases, unless the combined sequence is very long. We think that EST data are important, mainly because they are useful for gene tagging, tissue typing and expression profiling. We do not believe, however, that EST data are permanently important and thus should be kept forever. If, for example, the human genome is completely sequenced and provided with proper annotations at INSD, we will no longer have to keep human EST data. One of the most important goals of INSD, we believe, is to provide complete genome sequence data of high quality for as many species as possible.

As to dealing with complete genome sequence data, such a database management system as the genome broker introduced above is helpful, though its functions must be extended and refined. If we serve the complete genome data for a species separately from that for another species by the broker, we would perhaps be able to cope with continuously incoming genome data. It is of course unnecessary to mention that the computer and network technology for both hardware and software should be advanced further.

REFERENCES

1 Dulbecco, R. (1986) *Science*, **231**, 1055–1056.
2 Cantor, C.R. (1990) *Science*, **248**, 49–51.
3 Venter, J.C., Smith, H.O. and Hood, L. (1996) *Nature*, **381**, 364–366.
4 Hawkins, T.L., McKernan, K.J., Jacotot, L.B., MacKenzie, J.B., Richardson, P.M. and Lander, E.S. (1997) *Science*, **276**, 1887–1889.
5 Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F. *et al*. (1991) *Science*, **252**, 1651–1656.
6 Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989) *Science*, **245**, 1434–1435.
7 Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage., A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M. *et al*. (1995) *Science*, **269**, 496–512.
8 Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelly, J.M. *et al*. (1995) *Science*, **270**, 397–403.
9 Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D. *et al*. (1996) *Science*, **273**, 1058–1073.
10 Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H. Galibert, F., Hoheisel, J.D., Jacq, C. Johnston, M. *et al*. (1996) *Science*, **274**, 546–567.
11 Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A. *et al*. (1997) *Nature*, **388**, 539–547.
12 Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S. *et al*. (1996) *DNA Res*., **3**, 109–136.
13 Blattner, F.R., Plunkett, G.,III, Bloch, C.G., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glaaner, J.D., Rode, C.K., Mayhew, G.F. *et al*. (1997) *Science*, **277**, 1453–1462.
14 Aiba, H., Baba, T., Hayashi, K., Inada, T., Isono, K., Itoh, T., Kasai, H., Kashimoto, K.,Kimura, S., Kitakawa, M. *et al*. (1996) *DNA Res*., **3**, 363–377.
15 Butler, D. (1997) *Nature*, **388**, 701.
16 Watanabe, H., Mori, H., Itoh T. and Gojobori, T. (1997) *J. Mol. Evol*., **44**, S57–S64.
17 Wolfe, K.H. and Shields, D.C. (1997) *Nature*, **387**, 708–713.
18 Murakami, Y., Naitou, M., Hagiwara, H., Shibata, T., Ozawa, M., Sasanuma, S., Sasanuma, M., Tsuchiya, Y., Soeda, E., Yokoyama, K. *et al*. (1995) *Nature Genet.*, **10**, 261–268.
19 O'Brien C. (1997) *Nature*, **385**, 472.
20 Ohira, M., Ichikawa, H., Suzuki, E., Iwaki, M., Suzuki, K., Saito-Ohara, F., Ikeuchi, T., Chumakov, I., Tanahashi, H., Tashiro, K. *et al*. (1996) *Genomics*, **33**, 65–74.
21 Ohira, M., Ootsuyama, A., Suzuki, E., Ichikawa, H., Seki, N., Nagase, T., Nomura, N. and Ohki, M. (1996) *DNA Res.*, **3**, 9–16.
22 Shimizu, N., Antonarakis, S.E., Van Brockhoven, C., Patterson, D., Gardiner, K., Nizetic, D., Cresu, N., Delabar, J.-M., Korenberg, J., Reeves, R. *et al*. (1995) *Cytogenet. Cell Genet.*, **70**, 147–182.
23 Eki, T., Abe, M., Naitou, M., Sasanuma, S., Nohata, J., Kawashima, K., Ahmad, I., Hanaoka, F. and Murakami, Y. (1997) *DNA Seq*., **7**, 153–164.
24 Sato, S., Kotani, H., Nakamura, Y., Kaneko, T., Asamizu, E., Fukami, M., Miyajima, N. and Tabata, S. (1997) *DNA Res*. in press.
25 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol*., **215**, 403–410.
26 Gish, W. and States, D.J. (1993) *Nature Genet*., **3**, 266–272.
27 Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C. and Herrmann, R. (1996) *Nucleic Acids Res*., **24**, 4420–4449.
28 Maxam, A.M. and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 560–564.
29 Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
30 Report (1997) *Science*, **276**, 1505.