Exploration and Grading of Possible Genes from 183 Bacterial Strains by a Common Protocol to Identification of New Genes: Gene Trek in Prokaryote Space (GTPS)

Takehide Kosuge,1 Takashi Abe,1 Toshihisa Okido,1 Naoto Tanaka,1,2,3 Masaki Hirahata,1,2 Yutaka Maruyama,1,2 Jun Mashima,1 Aki Tomiki,1 Motoyoshi Kurokawa,4 Ryutaro Himeno,4 Satoshi Fukuchi,1 Satoru Miyazaki,1,3 Takashi Gojobori,1 Yoshio Tateno,1 and Hideaki Sugawara1,*

Center for Information Biology, DNA Data Bank of Japan, National Institute of Genetics, The Graduate University for Advanced Studies (SOKENDAI), Mishima, Shizuoka 411-8540, Japan;1, Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Corporation (JST), Chiyoda-ku, Tokyo 102-8666, Japan;2, Faculty of Pharmaceutical Science, Tokyo University of Science, Noda, Chiba 278-8510, Japan;3 and Advanced Center for Computing and Communication, RIKEN, Wako, Saitama 351-0198, Japan4

(Received 29 June 2006; revised 9 November 2006; published online 13 December 2006)

Abstract

A large number of complete microorganism genomes has been sequenced and submitted to the public database and then incorporated into our complete genome database, Genome Information Broker (GIB, http://gib.genes.nig.ac.jp/). However, when comparative genomics is carried out, researchers must be aware that there are protein-coding genes not confirmed by homology or motif search and that reliable protein-coding genes are missing. Therefore, we developed a protocol (Gene Trek in Prokaryote Space, GTPS) for finding possible protein-coding genes in bacterial genomes. GTPS assigns a degree of reliability to predicted protein-coding genes. We first systematically applied the protocol to the complete genomes of all 123 bacterial species and strains that were publicly available as of July 2003, and then to those of 183 species and strains available as of September 2004. We found a number of incorrect genes and several new ones in the genome data in question. We also found a way to estimate the total number of orthologous genes in the bacterial world.

Key words: comparative genome; gene prediction; annotation; gene grading; bacterial genome

1. Introduction

It is imperative to study and understand bacteria because they are integrated with, and mutually dependent on, all other living systems and the ecosystem. Although there are many approaches to scientific understanding of bacteria; it may be natural to start with genes. After all, it is the bacterial genome that makes a microorganism our friend, foe or otherwise.

Since complete sequencing of the Haemophilus influenzae genome,1 the whole genome sequencing method2 has been applied extensively to bacterial species, and complete genome sequences have been determined one after another. Currently, whole genome sequences are available for >300 bacterial strains from the Genome Information Broker (GIB, http://gib.genes.nig.ac.jp/)3 website of the DNA Data Bank of Japan (DDBJ). The tens of thousands of genes stored in GIB reflect the fact that bacteria are remarkably diverse, as well as, flexible so that they may survive in various environmental conditions. Furthermore, innumerable bacterial species remain unidentified.4–6 Nevertheless, even though
bacterial species are countless, there should be a limit to the number of their genes. We thus ponder the question: how many genes are there in the bacterial world?

As the reports of complete genome sequences have accumulated, another question has become more serious and urgent: how real are the genes that reportedly exist in the genome of a bacterial species? The most thoroughly studied species is *Escherichia coli* K-12, which is reported to possess 4302 protein-coding genes in its genome. However, no one has yet confirmed that every one of these purported genes functions and produces the corresponding RNA and protein. This extends to other species as well. There may be several reasons for this frustrating reality, one being that researchers find it difficult to examine all genes experimentally and thus resort to computerized homology searches. Homology searching is an unquestionably powerful method for exploring gene functions, but it is confined to the *in silico* domain where the parameters particular to a software tool determine the outcome. It is generally observed that different sets of parameters lead to different results for the same data input.

In the case of software tools for finding genes, the situation may be even worse because we have to deal not only with software parameters but also with algorithms. We can thus argue that a significant percentage of the computationally deduced genes of the bacterial species displayed by GIB and other databases will not be reproduced exactly by using one gene finder tool with a certain set of parameters. This includes doubly or triply inferred genes from the literature. If we allow this practice to continue, we will be forced into dealing with many invalid genes in the near future. We are presented with two choices. The first is to begin experimentally proving or disproving every gene in question, as Roberts proposed. The second, more feasible choice is to apply a closely examined tool to all genome sequence data available from GIB and elsewhere and to deduce genes all over again. With this approach, we can expect to reduce inconsistency and invalidity and to reproduce more reliable data.

2. Materials and Methods

2.1. Genome sequences used in this study

We have provided for GIB to continuously receive and install complete bacterial genome sequences released by the International Nucleotide Sequence Database Collaboration (INSDC, i.e. DDBJ/EMBL/GenBank, http://www.insdc.org/). We examined complete bacterial genome data stored in GIB for 123 species and strains released by July 2003 (ver. 2003). Thereafter, we added and updated data for the 183 species and strains that were released by September 2004 (ver. 2004). Analysis of the 303 species and strains (ver. 2005) is now in progress. An open reading frame (ORF) was defined in the protocol as a possible protein-coding region.

2.2. Steps in the exploration and evaluation of ORFs

Analysis involves three main steps (Fig. 1):

(i) Search for and masking of RNA coding genes in the genome sequence.

(ii) Prediction of boundaries for ORFs.

(iii) Evaluation of the predicted ORFs based on the result of homology and motif searches.

**Figure 1.** The workflow of GTPS. Steps in the workflow are described in the Section Materials and Methods.
Evaluation of the predicted ORFs based on the results of homology and motif searches.

These steps are described in detail in the following sections.

2.2.1. Search for and masking of RNA coding genes in the genome sequence RNA coding genes (ribosomal RNA [rRNA], transfer RNA [tRNA] and other non-coding RNA [ncRNA]) were first searched for and masked. For rRNA, the regions annotated as rRNA in the microbial genome as released by GIB were searched for and masked. For tRNA, we used the tRNAscan-SE (ver. 1.23) for prediction of the tRNA region. This program has been commonly used in previously reported studies. 

2.2.2. Prediction of boundaries for ORFs From various ORF prediction programs applied to microbial genomes (Fig. 2), we chose GLIMMER 2.0, the latest version as of September 2004. Because GLIMMER 2.0 can produce a learning dataset for ORF prediction by using the input genomic sequence in question only, it makes a highly accurate prediction of ORFs. The important parameter in using GLIMMER 2.0 is the minimum length of the ORF region. In predicting ORFs for the microbial genomes that have been reported to date, each annotation team sets the minimum length differently (Supplementary Table 1). Consequently, the predicted results are greatly affected by the differences in the minimum length that is set. This explains in part why there are inconsistencies in the annotation information. To determine the optimum setting for the minimum ORF length, GLIMMER 2.0 was applied to 183 strains by using four minimum ORF lengths, 15AA, 30AA, 60AA and 90AA (Supplementary Table 2). In comparison to the ORFs in the INSDC entries, the predicted numbers were too large at the 15AA and 30AA settings and too small at the 90AA setting. The highest rate of concordance with the existing annotations is obtained by setting the minimum length at 60AA. (Supplementary Tables 1 and 2 are available at www.dnares.oxfordjournals.org).

However, we noticed that the minimum length of the protein sequences in the UniProtKB/Swiss-Prot database was 17AA (O52748, Heterocyst inhibition signaling peptide). Therefore, to improve accuracy of the ORF prediction, we set the minimum length at 60AA and at 15AA. When two ORFs predicted with each of the two minimum lengths were matched at the 3’ ends, the longer ORF was chosen for the following steps.

To confirm the 5’ end of each ORF in a circular genome, we debugged and significantly modified the source code of RBSfinder. GLIMMER 3 has been released after ver. 2004. We now evaluate the parameter setting of GLIMMER 3 for analysis of the entire genomes of 303 bacterial species and strains published as of February 2006.

2.2.3. Evaluation of the predicted ORFs based on results of homology and motif searches In this study, protein-coding regions, such as for a leader peptide and intron-containing genes, were taken from INSDC and put into predicted ORFs. When GLIMMER 2.0 could not predict ORFs in the corresponding location in the INSDC data where a protein-coding sequence existed, the coding sequence was regarded one of our predicted
ORFs. After predicting all ORFs for all species, we performed a blastp 2.2.6 search for them against the genes of BCT division in the DDBJ Amino Acid Database (DAD), release 24, ver. 2003 and DDBJ, release 28, for ver. 2004. The DAD is a database that contains all the translated amino acid sequences from the corresponding DNA sequences in INSDC. In the blastp search, we used the following three conditions as the threshold for the blast hit: I (E-value \( \leq 10^{-40} \); sequence identity of aligned residues \( \geq 70\% \); mutual coverage \( \geq 70\% \) between subject and query), II (E-value \( \leq 10^{-40} \), sequence identity of aligned residues \( \geq 60\% \); mutual coverage \( \geq 60\% \) between subject and query), and III (E-value ignored; sequence identity of aligned residues \( \geq 50\% \); mutual coverage \( \geq 50\% \) between subject and query). For the blast analysis, we used a 16-node PC cluster system (CPU: Dual Xeon 3.2 GHz). The computation time for the analysis was 8 h for the 123 genomes in ver. 2003 and 12.5 h for the 183 genomes in ver. 2004. Next, we performed a motif search using InterProScan 3.2 (InterPro 6.2) for ver. 2003 and 3.3 (InterPro 7.2) for ver. 2004. A motif hit was assigned when the region predicted to have InterPro motifs was \( \geq 30\% \) of the ORF length. In this study we used 64-CPU RIKEN Super Combined Cluster System. The computation time for InterProScan was 420 h for the 123 genomes in ver. 2003 and 624 h for the 183 genomes in ver. 2004.

Flags were assigned to all of the ORFs to indicate the minimum length in GLIMMER 2.0 prediction, existence of Shine-Dalgarne sequence and results of the homology and motif searches. Details of the flag information are provided as Supplementary Table 4. (Supplementary Table 4 is available at www.dnares.oxfordjournals.org) The ORFs were compared with the corresponding genes in INSDC annotation. When the corresponding gene had a ‘/pseudo’ qualifier and the ORF was included in the gene or overlapped with the gene in-frame, a pseudo flag was assigned to the ORF. In addition, the ORFs were graded on an A–E or X scale to classify them according to the blastp and InterProScan results (Fig. 3). Grade A included highly reliable ORFs that are well supported by blastp and motif scans. Grades B and C included ORFs with poor homology and motif results and are thus less reliable than Grade A. Grades A and B were each divided into four sub-grades (AAAA, AAA, AA and A; BBBB, BBB, BB and B) with respect to the InterProScan results. Grade D included hypothetical proteins with no motifs. Grades A, B and C were used to indicate reliable genes for the computer simulation of gene divergence in the bacterial world.

The predicted ORFs were also graded with respect to the degree of matching to INSDC ORFs as follows: 1 (identical), 2 (3’-matched), 3 (newly predicted in this study) and 4 (not predicted by GLIMMER 2.0). Finally whole ORFs were graded for reliability and the degree of matching to INSDC in combination (e.g. AAAA1, BB2, C3 or D4). Potential genes were defined as ORFs of Grades AAAA1–D3 (Fig. 3). All false positives obtained by GLIMMER 2.0 and taken from the INSDC data, which were not predicted by GLIMMER 2.0, were graded as E or X.

In the final step, a protein product name was automatically assigned to each ORF according to the method described in Supplementary Table 5. (Supplementary Table 5 is available at www.dnares.oxfordjournals.org) It took about 3 months to complete the entire job though all resources were dedicated to this project.

2.3. Development of the GTPS viewer

We developed a system for browsing the results of our GTPS analysis. This browser is available on the Web (GTPS Viewer, http://gtps.ddbj.nig.ac.jp/). Users can view the results of our ver. 2003 and ver. 2004 analyses, and the database will be annually updated and expanded. GTPS can show gene information for each genome, homology analysis results, motif search results, reliability grades and link to Genomes-TO-Protein structures and functions (GTOP). GTPS can also show candidate orthologous gene sets based on homology searching. The results can be downloaded for each genome in DDBJ flat file and tab-delimited formats.

3. Results and Discussion

3.1. ORFs predicted by GTPS analysis

Table 1 shows the classification of ORFs according to our grading system, A (best) to X (worst). The number of the potential genes (AAAA1–D3) registered in ver. 2003 was 370 876 and in ver. 2004 was 551 246. The number of ORFs registered in ver. 2003 INSDC was 362 828 and in ver. 2004 INSDC was 537 312. Of the ORFs registered in INSDC (6690 and 9657 genes for ver. 2003 and ver. 2004, respectively) ~18% were not classified into possible protein-coding genes in this study. Comparison of ORFs between GTPS and INSDC is shown in Table 2. Of the possible protein-coding genes predicted by GTPS, ~70% were identical to INSDC genes and ~4% were newly found by GTPS. Researchers who submitted data to INSDC did not find these new ORFs. We also used the UniProtKB/Swiss-Prot database to confirm our new ORFs. The new ORFs overlapped those found by using UniProtKB/Swiss-Prot. For example, a newly found ORF between 8457 and 8678 bp of the Aquifex aeolicus VF5 genome was the same as one coding for 50S ribosomal protein L29 with accession no. P56613 in UniProtKB/Swiss-Prot. In addition to this, many new ORFs showed significant identities to those coding for the proteins registered in UniProtKB/Swiss-Prot (Fig. 4).
The number of bacterial genes is expected to correlate with the size of the bacterial genome, because a bacterial genome is packed with genes without large intergenic regions and introns. We plotted the correlation between them using ver. 2004 of GTPS and that of INSDC (Fig. 5). It is clear that the number of genes correlates with the genome size. We thus consider that bacterial genomes have evolved in coordination between genetic complexity and genome size. For example, bacterial parasites such as Buchnera sp. do not possess many

![Figure 3. ORFs grades. ORFs were graded as follows. Grades A–C: functional information available; Grade D: hypothetical protein; Grade E: hypothetical protein supported by poor coverage in blastp and Grade X: no information available. An unknown protein is a hypothetical protein for which the product and notes are given in Supplementary Table 7. A putative membrane protein is a protein for which the product is given in Supplementary Table 8. An unknown motif is a motif that includes ‘unknown’ in its InterPro description. A significant motif is a known motif. Potential genes were defined as ORFs of Grades AAAA1–D3. (Supplementary Tables 7 and 8 are available at www.dnares.oxfordjournals.org)](image)

| Grades | Ver. 2003 | Ver. 2004 |
|--------|-----------|-----------|
| AAAA–A | 284,127   | 431,672   |
| BBBB–B | 7208      | 10,250    |
| C      | 4680      | 7,511     |
| D      | 79,779    | 107,382   |
| E      | 6788      | 10,225    |
| X      | 466,681   | 687,110   |
| Total (AAAA–X) | 848,383 | 1,254,150 |
| Potential genes (AAAA1–D3) | 370,876 | 551,246 |
| INSDC  | 362,828   | 537,312   |

The number of bacterial genes is expected to correlate with the size of the bacterial genome, because a bacterial genome is packed with genes without large intergenic regions and introns. We plotted the correlation between them using ver. 2004 of GTPS and that of INSDC (Fig. 5). It is clear that the number of genes correlates with the genome size. We thus consider that bacterial genomes have evolved in coordination between genetic complexity and genome size. For example, bacterial parasites such as Buchnera sp. do not possess many...
genes, and the genome is small. Therefore, using the linear relationship in Fig. 5, one can estimate the number of genes in a genome of a given size, or vice versa. In the GTPS study, the *Aeropyrum pernix* K1 genome decreased by 1006 genes, whereas the *Mycobacterium leprae* TN genome increased by 549 genes. On the whole linearity holds better for GTPS than for INSDC, suggesting that the different researchers who submitted data used different tools and parameters for finding ORFs. Examples of less efficient gene predictions for the three organisms mentioned above are shown in Fig. 6. The GTPS moved ORFs without showing homology and motif data to lower grades. In the case of *M. leprae* TN, most of the 549 genes were to code for hypothetical proteins that were predicted on the flanking regions of pseudogenes. One such example is shown in Fig. 4.

We made a histogram of the corresponding number of amino acids in the new ORFs predicted by applying GTPS to the data of ver. 2004 to know if the difference in the number of ORFs between the GTPS and INSDC predictions is due to gene size. At first, the newly predicted ORFs were selected from Grades AAAA3, AAA3, A3, BBBB3, BBB3, BB3, B3 and C3. ORFs with corresponding locations in the INSDC sequence data for which any information such as gene feature, repeat region feature, or other characteristics were then removed. The remaining 9088 ORFs with no information were used for making the histogram. As seen in Fig. 7, 6739 new ORFs from the GTPS prediction correspond to short proteins with 20–100 amino acids. This indicates that the minimum number of nucleotides in ORF prediction was set at >100 amino acids in some genome projects. Supplementary Table 1 also supports
In the GTPS prediction for the ver. 2004 data, 581,490 ORFs coding for 20–100 amino acids were predicted, but most of them fell into the worst or second-worst grades (5125 ORFs in Grade E and 500,433 ORFs in Grade X). There were 39,538 ORFs classified into Grade D that were hypothetical proteins. There were 36,394 ORFs classified into Grades AAAA–C according to results of homology or motif searches, 6739 of these were revealed to be new genes. As shown in Fig. 4, some of the new ORFs coding for 20–100 amino acids show high identity to proteins in UniProtKB/Swiss-Prot. GTPS with no size limitation was able to predict ORF coding for 2159 amino acids in the *Bordetella bronchiseptica* RB50 and *Bordetella parapertussis* 12822 strains.

In this study, ORFs of complete genomes deposited in INSDC and stored in GIB were re-evaluated. In the GTPS analysis, all complete bacterial genomes are simultaneously re-evaluated by blastp and InterProScan analyses for the same version of a database. Note that we cannot compare ORF prediction results one with another if they were obtained by using different versions of reference databases (e.g. InterPro) and different versions of gene prediction tools. Neither can we assign consistent grades of reliability to predicted ORFs. GTPS leads to more consistent, comprehensive and reliable results in ORF prediction than those obtained by any extant method.

### 3.2. *E. coli* genome annotation

*E. coli* is a model organism, and it had been studied extensively before the whole genome was sequenced. In this section, we compare annotation of the *E. coli* genome in ver. 2003, ver. 2004 and INSDC. It is noted that by updating *E. coli* genome data, the number of bases in the genome increased from 4,639,221 in ver. 2003 to 4,639,675 in ver. 2004. However, the total number of the predicted ORFs decreased from 9845 (ver 2003) to 9768 (ver 2004). The number of ORFs in Grades AAAA–D3 also decreased from 4695 (ver 2003) to 4688 (ver 2004). The ratio of the number of ORFs in Grades AAAA–D3 to the total number of ORFs was 48% in both ver. 2003 and ver. 2004. In ver. 2003 125 ORFs turned out to code for a different amino acid sequence than those in INSDC, as did 111 ORFs in ver. 2004. However, it was not feasible for GTPS to clarify whether the
changes were due to sequence errors or other causes. As we continue the annual GTPS analysis, it will be possible for us to detect hidden sequence errors.

Comparison of the *E. coli* ORFs in the two versions and those in INSDC is summarized in Table 3. Whereas the number of the new ORFs decreased in ver. 2004, the number of 3'-matched ORFs increased. In addition, 193 new ORFs in ver. 2003 matched the 3' terminals of INSDC ORFs, indicating that GTPS predicted the 193 ORFs in advance of the INSDC ORFs update. An ORF from the present comparison was adopted to the ORF of ECK4368 in an international community annotation.8

### 3.3. Simulation of the wide range of gene divergence in the bacterial world

An interesting question is whether or not the number of reliable ORFs will increase according to the increasing number of new bacterial genomes to be sequenced and submitted in the future. To answer this question, we conducted computer simulations focusing on the diversity of gene functions in bacterial genomes. Homologous ORFs with (90% identity were defined as having the same function. The percentage of homology was obtained by using CD-HIT.32,33 When the ORFs are clustered into groups by homology, the number of groups indicates the degree of divergence in gene function. If the number of groups becomes saturated despite an increase in the number of genomes, we will be able to regard the saturated number as the number of unique genes in the bacterial world. With this in mind, we conducted the following simulation.

First, we randomly sampled 10 genomes from the 183 genomes and clustered their predicted ORFs by using CD-HIT. We then computed the ratio of the clustered ORFs to all ORFs among the sampled genomes was obtained. This procedure was repeated 10 times to calculate the average ratio over the repeats. Next, the whole procedure was repeated, but the number of sampled genomes was increased by 10 at a time until the number reached 180. These 180 comprised the simulation set. We conducted two simulations for (90 and 100% homology in clusterings. The results are shown in Fig. 8.

As shown in Fig. 8, the ratio increased with the increase in the number of genomes. The ratio is not yet saturated at 180 genomes (29.5% of the genes among 180 genomes were clustered and presumed to have the same function.). Therefore, we could not determine the total number of bacterial genes by using 183 genomes. We need to have genome sequences of strains from every branch of the phylogenetic tree for bacteria. We will then be able to infer the total number of gene functions from the bacterial world.

Essentially, we estimated the total number of ‘orthologous’ genes in the prokaryote world. An orthologous gene of a species is defined as one that has diverged from the gene of its common ancestor by speciation. Therefore, an orthologous gene, like cytochrome c gene, functions equally in all species that carry it. Most orthologous genes submitted to INSDC were predicted to be ORFs by software tools for gene prediction or homology searching. It is thus natural to treat such an orthologous gene in different species by the same

---

**Table 3. Comparison of *E. coli* ORFs in INSDC with those in ver. 2003 and ver. 2004**

|                | Ver. 2003 | ver. 2004 |
|----------------|-----------|-----------|
| Common with INSDC | 3281      | 3128      |
| 3'-matched     | 721       | 1054      |
| New ORFs (only predicted in this study) | 674       | 476       |

---

**Figure 7.** Distribution of amino acid sequence lengths of new genes predicted by GTPS. The newly predicted ORFs, which were not annotated in INSDC, were selected from Grades AAAA3, AAA3, AA3, A3, BBBB3, BBB3, BB3, B3 and C3 of ver. 2004 and used for drawing a histogram. The number of newly predicted ORFs and their amino acid sequence lengths are shown.
standard or criteria when one examines and evaluates its authenticity. In our GTPS analysis, we used only bacterial genome sequence data, and we disregarded the attached biological information or annotation. What we did is also called annotation, but it may differ from what is usually used. We believe that our criteria and the results of our examination and evaluation of as many bacterial genes as available at INSDC based on the same criteria will be quite useful and helpful, particularly for those who carry out homology searches against INSDC in their studies including genome sequencing projects and for those who try to experimentally examine the authenticity of a hypothetical gene.

3.4. Future aspects of GTPS

Users of our GTPS database can see the details of protocols and look into lines of evidence that support our prediction and grading of ORFs on the basis of blastp and InterProScan analyses (visit http://gtps.ddbj.nig.ac.jp/). For example, while some researchers want to use only Grade A ORFs, others may include Grade D ORFs. Therefore, researchers can have confidence in selecting and using candidate genes from the GTPS database without visiting multiple sites on the Internet and repeating large-scale computation. Even genome sequencing projects may use the GTPS database for annotation of newly sequenced genomes. It is noted that GTPS will reflect results of proteome analysis through quality-controlled secondary databases including InterPro, because we continue GTPS analysis by using the latest tools and databases, and we update our GTPS database at least once a year. We will increase the frequency of updates by improving each step of the GTPS analysis and using more powerful computer resources.

Venter et al.\textsuperscript{34} reported and submitted massive amounts of data on 800 000 sequences and 1 500 000 genes into the whole genome shotgun (WGS) section of INSDC. The amount was almost equivalent to the number of bacterial genes that had been published at that time. The number of genomic sequences deposited into WGS has been increasing, and the number of WGS entries exceeded the total number of the other INSDC entries in 2005. However, finding useful genes in the WGS entries has not been straightforward due to heterogeneity of the annotations. Therefore, we will carry out GTPS analysis for the WGS sequences to extract genes with consistent reliability scores. With the emergence of new sequencing technologies such as the 454 sequencing system,\textsuperscript{35} meta-genome data will be produced and submitted to INSDC in succession. GTPS will also be useful for annotating a large number of possible genes in the submitted meta-genome data.

The results of re-annotation of complete microbial genomes are freely available at http://gtps.ddbj.nig.ac.jp/.

Acknowledgements: The authors express their sincere thanks to Yasumasa Shigemoto of Fujitsu Co., Ltd., Yoshikazu Kuwana of Tokai Soft Co., Ltd., Shigetaka Sakamoto of HOLONICS Co., Ltd, and Mitsui Knowledge Industry Co., Ltd for their part in the development of the GTPS database. This work was supported in part by the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency and the Japanese Bioportal Project of Special Coordination Funds for Promoting Science and Technology. Computation was done with the RIKEN Super Combined Cluster System.

Supplementary data: Supplementary data are available online at www.dnaresearch.oxfordjournals.org

REFERENCES

1. Fleischmann, R. D., Adams, M. D., White, O., et al. 1995, Whole-genome random sequencing and assembly of \textit{Haemophilus influenzae} Rd, \textit{Science}, \textbf{269}, 496–512.
2. Lander, E. S. and Waterman, M. S. 1988, Genomic mapping by fingerprinting random clones: a mathematical analysis, \textit{Genomics}, \textbf{2}, 231–239.
3. Fumoto, M., Miyazaki, S., and Sugawara, H. 2002, Genome Information Broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more, \textit{Nucleic Acids Res.}, \textbf{30}, 66–68.
4. Amann, R. I., Ludwig, W., and Schleifer, K. H. 1995, Phylogenetic identification and in situ detection of
individual microbial cells without cultivation, Microbiol. Rev., 59, 143–169.
5. Hugenholtz, P. and Pace, N. R. 1996, Identifying microbial diversity in the natural environment: a molecular phylogenetic approach, Trends Biotechnol., 14, 190–197.
6. Curtis, T. P., Sloan, W. T., and Scannell, J. W. 2002, Estimating prokaryotic diversity and its limits, Proc. Natl Acad. Sci. USA., 99, 10494–10499.
7. Blattner, F. R., Plunkett, G. III, Bloch, C. A., et al. 1997, The complete genome sequence of Escherichia coli, K-12, Science, 277, 1453–1474.
8. Riley, M., Abe, T., Arnaud, M. B., et al. 2006, Escherichia coli K-12: a cooperatively developed annotation snapshot—2005, Nucleic Acids Res., 34, 1–9.
9. Johnsson, A., Heldin, C. H., Wasteson, A., et al. 1984, The c-sis gene encodes a precursor of the B chain of platelet-derived growth factor, EMBO J., 3, 921–928.
10. Costa, C., Gise, W., Miller, W., Evans, E. W., and Lipman, D. J. 1990, Basic local alignment search tool, Mol. Biol., 215, 403–410.
11. Mittenhuber, G. 2001, Comparative genomics of prokaryotic GTP-binding proteins [the Era, Ogb, EngA, ThdF (TrmE), YchF and YihA families] and their relationship to eukaryotic GTP-binding proteins (the DRG, ARF, RAB, RAN, RAS and RHO families), J. Mol. Microbiol. Biotechnol., 3, 21–35.
12. Constantinidis, K. T. and Tiedje, J. M. 2005, Towards a genome-based taxonomy for prokaryotes, J. Bacteriol., 187, 6258–6264.
13. Bubeck, G., Afghani, S. L., Yang, J. W., and John, J. P. 2005, Searching for hypothetical proteins: theory and practice based upon original data and literature, Prog. Neurobiol., 77, 90–127.
14. Tavares, A. H., Silva, S. S., Bernardes, V. V., et al. 2005, Virulence insights from the Paracoccidioides brasiliensis transposome, Genet. Mol. Res., 4, 372–389.
15. Roberts, R. J. 2004, Identifying protein function—a call for community action, PLoS Biol., 2, E42.
16. Elphick, M. R. and Egertova, M. 2005, The phylogenetic distribution and evolutionary origins of endocannabinoid signalling, Handb. Exp. Pharmacol., 283–297.
17. Raskin, D. M., Seshadri, R., Pukatzki, S. U., and Mekalanos, J. J. 2006, Bacterial genomics and pathogen evolution, Cell, 124, 703–714.
18. Fraser-Liggett, C. M. 2005, Insights on biology and evolution from microbial genome sequencing, Genome Res., 15, 1603–1610.
19. Hotopp, J. C., Lin, M., Madupu, R., et al. 2006, Comparative genomics of emerging human ehrlichiosis agents, PLoS Genet., 2, e21.
20. 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.
21. 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.
22. 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.
23. 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.
24. Aggarwal, G. and Ramaswamy, R. 2002, Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER, J. Biotechnol., 27, 7–14.
25. Bairach, A., Apweiler, R., Wu, C.H., et al. 2005, The Universal Protein Resource (UniProt), Nucleic Acids Res., 33, D154–D159.
26. Suzuki, 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.
27. Kunst, F., Ogasawara, N., Moszer, I., et al. 1997, The complete genome sequence of the gram-positive bacterium Bacillus subtilis, Nature, 390, 249–256.
28. Ouyang, Z., Zhu, H., Wang, J., and She, S. Z. 2004, Multivariate entropy distance method for prokaryotic gene identification, J. Bioinformatics and Comp. Biol., 2, 353–373.
29. Apweiler, R., Attwood, T. K., Bairach, A., et al. 2001, The InterPro database, an integrated documentation resource for protein families, domains and functional sites, Nucleic Acids Res., 29, 37–40.
30. Kawabata, T., Fukuchi, S., Homma, K., et al. 2002, GTOP: a database of protein structures predicted from genome sequences, Nucleic Acids Res., 30, 294–298.
31. Hayashi, K., Morooka, N., Yamamoto, Y., et al. 2006, Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110, Mol. Syst. Biol., 2, 0007.
32. Li, W., Jaroszewski, L., and Godzik, A. 2001, Clustering of highly homologous sequences to reduce the size of large protein databases, Bioinformatics, 17, 282–283.
33. Li, W., Jaroszewski, L., and Godzik, A. 2002, Tolerating some redundancy significantly speeds up clustering of large protein databases, Bioinformatics, 18, 77–82.
34. Venter, J. C., Remington, K., Heidelberg, J. F., et al. 2004, Environmental genome shotgun sequencing of the Sargasso Sea, Science, 304, 66–74.
35. Margulies, M., Egholm, M., Altman, W. E., et al. 2005, Genome sequencing in microfabricated high-density picolitre reactors, Nature, 437, 376–380.