Construction of the *gyrB* Database for the Identification and Classification of Bacteria

Hiroaki Kasai [1]
hkasai@kamaishi.mbio.co.jp
Kanako Watanabe [1]
kanakow@kamaishi.mbio.co.jp
Elizabeth Gasteiger [2]
Gasteiger@medecine.unige.ch
Amos Bairoch [2]
bairoch@cmu.unige.ch
Katsumi Isono [3]
isono@biol.kobe-u.ac.jp
Satoshi Yamamoto [4]
yamamotost@nichirei.co.jp
Shigeaki Harayama [1]
harayama@kamaishi.mbio.co.jp

[1] Marine Biotechnology Institute, Kamaishi Laboratories
3-75-1, Heita, Kamaishi, Iwate 026-0001, Japan
[2] Swiss Institute of Bioinformatics c/o Department of Medical Biochemistry
1, rue Michel Servet CH - 1211, Geneva 4, Switzerland
[3] Graduate School of Science and Technology, Kobe University
Rokkodai, Kobe 657-8501, Japan
[4] Food Science Laboratory, Research and Development Center, Nichirei Corporation,
Shinminato 9, Chiba 261-8545 Japan

Abstract

Nucleotide sequences of small-subunit rRNA (16S rRNA) are most commonly used for the identification and characterization of bacteria and their complex communities. However, 16S rRNA evolves slowly and is often not very convenient to resolve bacterial strains at the species level. We have therefore attempted to develop a rapid and more convenient system for bacterial identification using the *gyrB* gene sequences. We chose the *gyrB* gene, because (i) it is rarely transmitted horizontally, (ii) its molecular evolution rate is higher than that of 16S rRNA, and (iii) the gene is distributed ubiquitously among bacterial species. We PCR-amplified the 1.2 kb-long *gyrB* segments from about 1,000 bacterial species by using degenerate primers and determined their nucleotide sequences. The resultant data have been assembled into the *gyrB* database accessible via WWW.

1 Introduction

For the identification and classification of new bacteria, microbiologists have been using various taxonomic markers and performing comparative analysis of the data with those of type strains. Such studies are generally very laborious, and yet the results obtained are not necessarily conclusive. During the past decades, however, the method for taxonomy has evolved dramatically though the introduction of various nucleotide sequencing techniques and resultant massive amounts of data. Especially, rRNA gene sequence data have been proved to be useful in establishing the division of all living organisms into three domains, namely, *Archaea*, *Bacteria* and *Eucarya* [18]. Needless to say that the development in the taxonomy of living organisms, microorganisms in particular, has mainly been achieved through the analysis of rRNA sequence data. There are a variety of reasons why rRNA genes have been selected as standard genes for molecular taxonomy. First, rRNA is an essential constituent in all living organisms. Second, the existence of many conserved regions in the rRNA genes allows the alignment of their sequences derived from distantly related organisms, while their variable regions are useful for the distinction of closely related organisms. Furthermore, the horizontal transfer of rRNA genes is believed to be rare. Currently, more than 20,000 rRNA gene sequences are available in the Ribosomal Database Project (RDP; http://www.cme.msu.edu/RDP/) and GenBank at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). The RDP database is accessible via the internet and contains various analytical tools. However, in some cases, it seems that the resolution of rRNA-based analysis is too low due to the small numbers of substitutions between compared rRNA sequences.

Therefore, we felt that a gene(s) which is not frequently transmitted horizontally, evolves at a higher rate than the rRNA gene, and is found in most, if not all, bacterial species must be searched for and should be used as an alternative taxonomic marker. We have surveyed the protein-coding genes which satisfy these criteria, and found that several genes including *gyrB* appeared to be good candidates [4, 19, 20, 21, 22]. Therefore, we determined the nucleotide sequences of the *gyrB* and other genes from a wide variety of bacteria. We then construct the *gyrB* database for the identification and classification of bacteria which can be accessible via the internet (http://www.mbio.co.jp; Fig. 3-A).

2 Molecular systematics based on *gyrB*

DNA topoisomerases are essential for DNA replication, transcription, recombination and repair and control the level of supercoiling by cleaving and resealing the phosphodiester bond of DNA. They are classified into type I (EC 5.99.1.2) and type II (EC 5.99.1.3) according to their enzymatic properties. The bacterial DNA gyrase is a type II topoisomerase that is capable of introducing negative supercoiling into a relaxed closed circular DNA molecule. This reaction is coupled with ATP hydrolysis. DNA gyrase can also relax supercoiled DNA without ATP hydrolysis. DNA gyrase consists of two subunit proteins in the quaternary structure of A2B2. The A protein (GyrA) has a molecular weight of approximately 100 kDa while the B protein (*gyrB*) has a molecular weight of either 90 kDa or 70 kDa. Comparison of the primary structures of the 90 kDa class and of the 70 kDa class revealed that the 90 kDa type has an insertion of about 170 amino acids at a site corresponding to residue 560 of the 70 kDa-type molecule. The N-terminal portion of the B protein was thought to catalyze ATP-dependent supercoiling of DNA while the C-terminal portion supports complex formation with the A protein and is involved in ATP-independent relaxation. The crystal structure of the N-terminal 43-kDa fragment of the B protein has been established. The 43-kDa protein monomer comprises two domains with the ATP-binding site being located at the center of the first domain.

Topoisomerase IV is a bacterial enzyme that appears to be closely related to DNA gyrase. It was identified through the analysis of mutations (*parE*) that let to a deficiency in the partition phase of bacterial replication. The role of this enzyme may be to link the catenated daughter chromosomes prior to partition into daughter cells. Sequence alignment shows that the *parE* gene product is very similar to *gyrB*. Topoisomerase IV cannot catalyze supercoiling of DNA, however. It catalyzes DNA relaxation by a mechanism that requires hydrolysis of ATP [11, 17]. Almost all bacteria are known to contain both DNA gyrase and topoisomerase IV. However, the gene for topoisomerase IV was not found in the total genome sequences of *Helicobacter pylori* [15]. *Mycobacterium tuberculosis* [2] and *Synechocystis* sp. PCC 6803 [5]. It is not known whether or not DNA gyrase in these bacteria acts also as topoisomerase IV.

We used the *gyrB* and its translated sequences for bacterial taxonomy. For this purpose, a set of universal primers were designed from two conserved amino acid sequences in *gyrB*s from *Escherichia coli*, *Pseudomonas putida* and *Bacillus subtilis*. The two conserved amino acid sequences were "reverse-translated", and a set of degenerate primers which contain 512 variations was designed. PCR amplification of *gyrB* was carried out by using the genomic DNA from bacteria of different taxonomic groups, and PCR products with a size predicted from the known *gyrB* sequences (1.2 kb) were amplified from various kinds of strain [19]. Thus, it was possible to amplify *gyrB*s from a broad range of bacteria: the α, β and γ subdivisions of *Proteobacteria*, the *Cytophaga/Flavobacterium/Bacteroides* complex, low and high G+C Gram-positive bacteria and some strains of *Cyanobacteria*. To avoid paralogous comparison, we have developed the *gyrB*-specific primers which do not amplify the genes for topoisomerase IV.

A validity of *gyrB* sequences as taxonomic markers was evaluated mainly from two points of view: the rate of their base substitutions and the consistency between the results of *gyrB*-based analysis and those of the DNA-DNA hybridization analysis. Protein-coding genes are thought to evolve faster than rRNA genes because synonymous substitutions mainly at the third positions of codons in the protein-coding genes are permitted without causing any changes in the amino acid sequences of their gene products. Actually, the average base-substitution rate of 16S rRNA genes was 1% per 50 million years, while that of *gyrB* at the synonymous sites was estimated to be 0.7 - 0.8% per one million years [20]. Therefore, some species with completely identical 16S rDNA sequences can be differentiated by using their *gyrB* gene sequences: four species belonging to *Mycobacterium tuberculosis* complex, *Mycobacterium kansasii* and *Mycobacterium gastri* (Kasai *et al.* manuscripts in preparation) are such examples. Theoretically, any protein-coding sequences can be used for phylogenetic analysis. However, many genes are known to spread horizontally among different bacterial species. In the case of *gyrB*, there was a high correlation between the phylogenetic distance based on the *gyrB* sequences and the total genome homology analyzed by DNA-DNA hybridization [22]. This observation suggests that the horizontal transfer of *gyrB* is, if anything, rare. Probably, the essential roles of the *gyrB* product in DNA replication and transcription functions against the introduction of foreign *gyrB*.

Two cases of the existence of multiple copies of *gyrB* have been reported in the genomes of the *Streptomyces* family. In *Streptomyces sphaeroides*, two diverged *gyrB* genes were found, one encoding a novobiocin-sensitive and the other a novobiocin-resistant enzyme [13]. A similar case was found in the genome project of *Streptomyces coelicolor*. Here, one of the *gyrB* genes was located just upstream of *gyrA* in the oriC region as in the case of the *gyrB* genes found in almost all bacteria, while a second gene was located about 1.85 Mbp far from *oriC* [9]. The phylogenetic analysis of the products of these genes indicated that one of them is related with other *Streptomyces gyrB*s while the other was outside the *Streptomyces* cluster.

3 Comparison of the phylogenetic relationships deduced from *gyrB* with those from 16S rRNA sequences

While 6205 data of 16S rRNA are available from the RDP [6], and a 16S rRNA-based phylogenetic tree of bacteria has been constructed [7], the resolution of the 16S rRNA-based analysis is not so high as to distinguish very closely related bacteria. In general, it is said that organisms sharing more than 97% of identity in their 16S rRNA sequences might belong to one and the same species [16]. However, there are cases of bacteria exhibiting more than 99% identity in their 16S rRNA sequences, and yet belonging to two distinct species as revealed from DNA hybridization analysis. Evidently, due to the slow speed of divergent evolution of the 16S rRNA gene, the resolution of 16S rRNA-based analysis between closely related organisms is lower than that of DNA hybridization [12]. Furthermore, there are polymorphisms found amongst different rRNA genes in some bacteria [10], and consequently alignment of rRNA sequences requires some expertise. At the same time, it is often not very easy to obtain correct sequence data of rRNAs and their genes largely due to their highly ordered structure.

Yamamoto and Harayama [21] compared the phylogenetic relationships of 20 *Pseudomonas* strains deduced from the genes for 16S rRNA, RNA polymerase σ^{70} factor (*rpoD*) and *gyrB*. They showed that the phylogenetic trees based on *gyrB* and *rpoD* are congruent, but the topology of the phylogenetic tree based on the 16S rRNA sequences including their variable regions was not identical with those of the *gyrB*- and *rpoD*-based trees. When the variable regions were excluded from analysis, the topology of the phylogenetic tree based on the 16S rRNA sequences became less different from those of the *gyrB*- and *rpoD*-based trees, although the resolution of the 16S rRNA-based tree was much lower than that of the *gyrB*- and *rpoD*-based trees. They also showed that the genetic distances in the variable regions of 16S rRNA correlated poorly with the synonymous substitution distances in the *gyrB* and *rpoD* genes. From these observations, they proposed that many base substitutions in the variable regions may not be the results of successive point mutations, but may be caused by single-event mutations introducing multiple substitutions. Hence, the variable regions should not be included in the calculation of the genetic distance, and the base substitution rates outside such regions are too low to distinguish bacterial species.

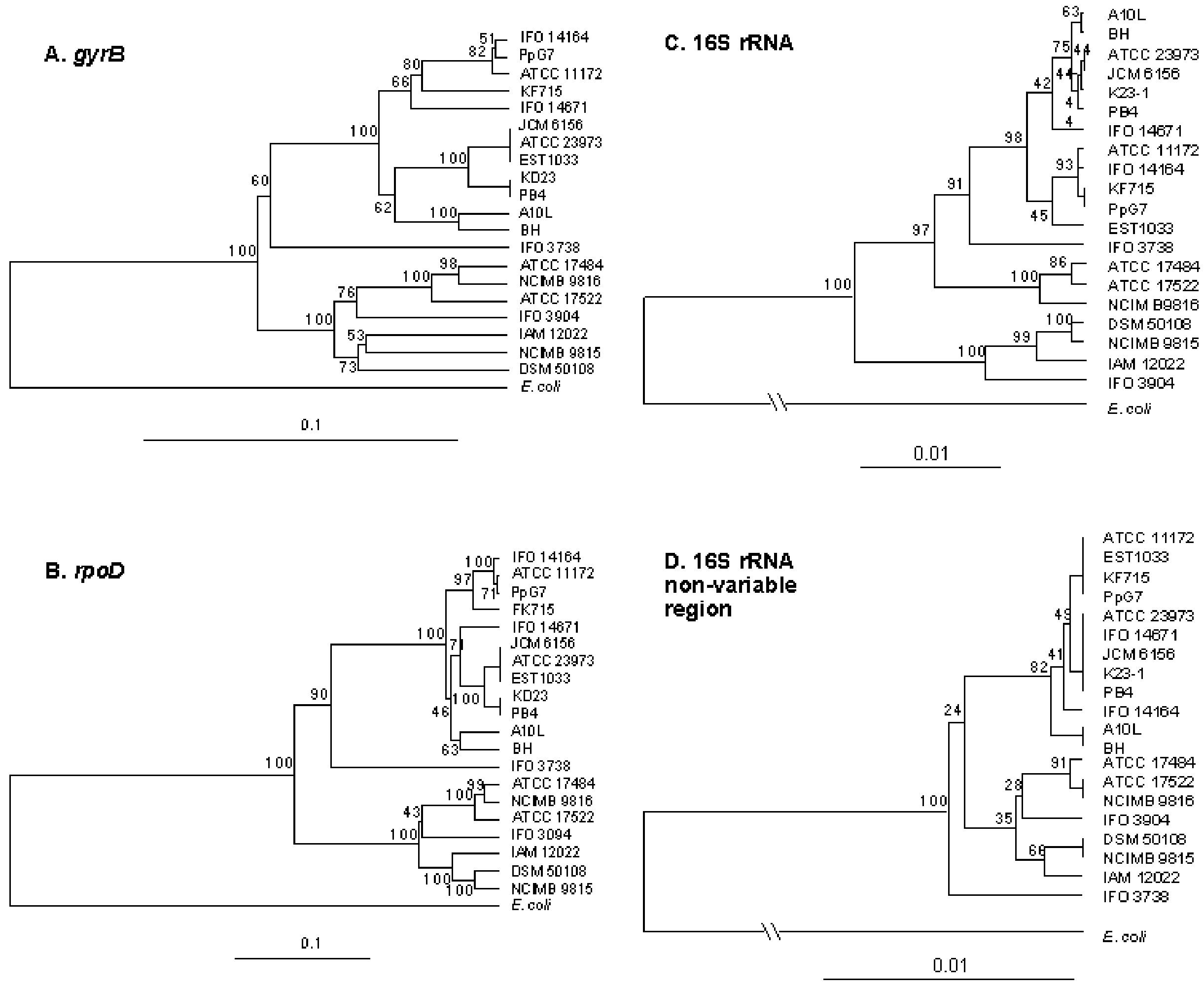


Figure 1: Phylogenetic trees of 20 *Pseudomonas* strains based on the nucleotide sequences of *gyrB* (A), *rpoD* (B) and 16S rRNA genes (C, D). Two types of phylogenetic trees constructed from the 16S rRNA sequences are presented: a tree based on the whole 16S rRNA sequences including the variable regions (C) and another based on the partial 16S rRNA sequences without the variable regions (D). Trees were constructed by using the UPGMA method [8]. Bars indicate the genetic distance of either 0.1 or 0.01. Numbers at individual nodes indicate the percent bootstrap values of 1000 trials. The corresponding *Escherichia coli* K-12 sequence was treated as the outgroup in each case.

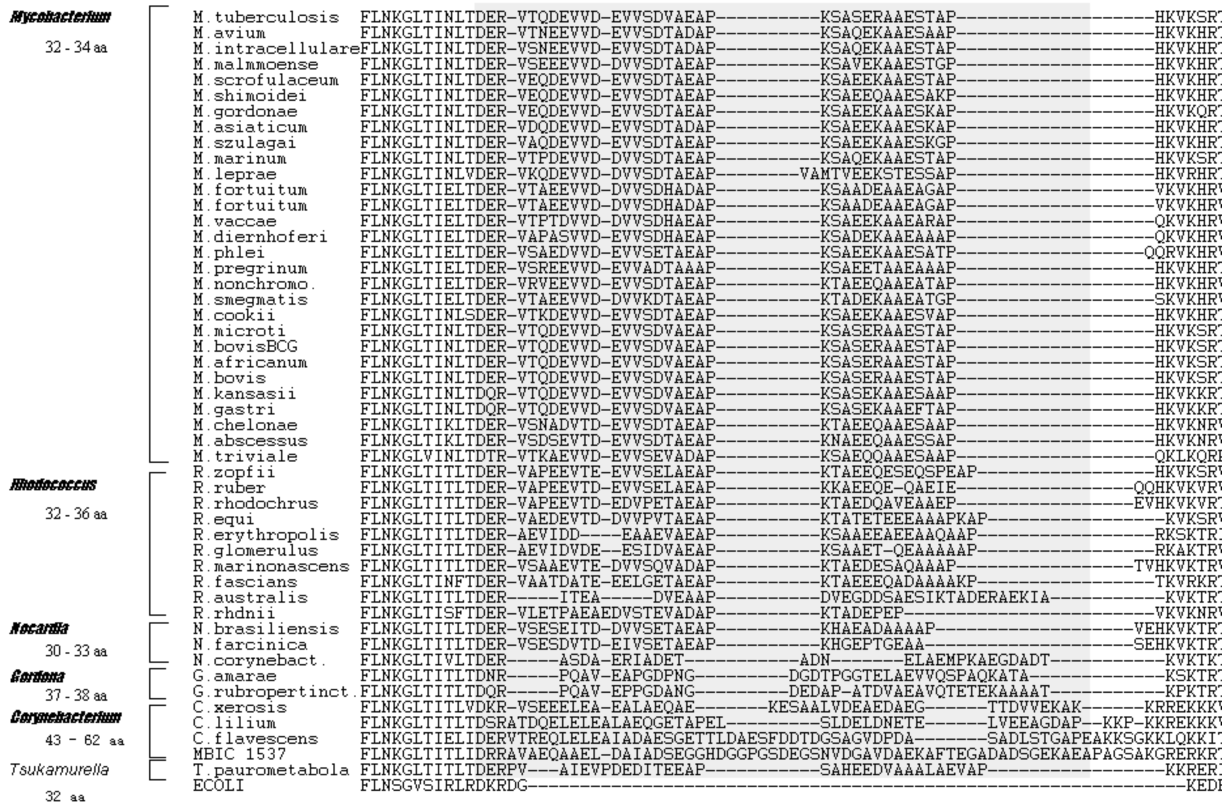


Figure 2: Signature sequences found in the GyrB sequences of *Mycobacteria* and relatives. 349 GyrB sequences of high G+C Gram-positive bacteria were aligned by running ClustalW 1.7. Consequently, their GyrB sequences were found to contain a stretch of insertion sequence (shaded) as indicated.

4 Detection of signatures in multiply aligned GyrB sequences

We have been analyzing the *gyrB* of high G+C Gram-positive bacteria of so called actinomycetes. 349 GyrB sequences were aligned by using ClustalW 1.7. As the results, the GyrB sequences of this group were found to contain a stretch of insertion sequence which is not present in other taxonomic groups. The site of insertion is located between the seventh and eighth beta sheet in the first domain based on the X-ray analysis of the *Escherichia coli* GyrB fragment [17]. The length of the inserted sequences appeared to be variable: 3 to 4 amino acid residues in *Micromonosporaceae*, 10 to 15 in *Streptomyces* and *Pseudonocardiaceae* except for a mycolic acid-containing group which possesses insertions of 25 to 55 amino acid residues (Kasai and Harayama, in preparation). The insertion sequence found in mycolic acid-containing group, *Mycobacterium tuberculosis* for example, was predicted to form an alpha helix through the analysis of a secondary structure estimation tool, PREDATOR [3]. Biological meanings of these insertions are unknown, although the nucleotide substitution rates and the codon usage pattern in the insertion sequences are comparable with those of other regions. This observation suggests that these insertion sequences are not generated by the insertion of IS (insertion sequences) or transposon. Consequently, they would serve as useful signature sequences for the identification of bacteria belonging to high G+C Gram-positive bacteria.

5 The *gyrB* database for the identification and classification of bacteria

We have thus created a *gyrB* sequence database to facilitate easier Identification and Classification of Bacteria (for this reason, we have named the database the ICB database). The ICB database offers useful data and information concerning bacterial phylogeny. It also provides tools for the identification and classification of newly isolated bacteria by performing comparative analyses of the nucleotide sequences and/or their derived amino acid sequences against the stored *gyrB* sequence database. The database contains not only the nucleotide and amino acid sequence data of the *gyrB* gene and its protein products from a large variety of microorganisms, but also other information such as useful hints for experimental protocols and trouble shootings so as to reduce the time and efforts in the identification and classification of bacteria. The WWW pages and related cgi-scripts have been designed for these purposes as described below.

In the data browsing pages, the *gyrB* sequence data are listed in the alphabetical order of genus and species names. In the list, a check box is provided for each entry that can be used for retrieving the corresponding sequence data in either Genbank or fasta format. Other information concerning each species can also be accessed and retrieved by clicking species names. Furthermore, the database can be searched for any desired species and/or group of species by entering key words.

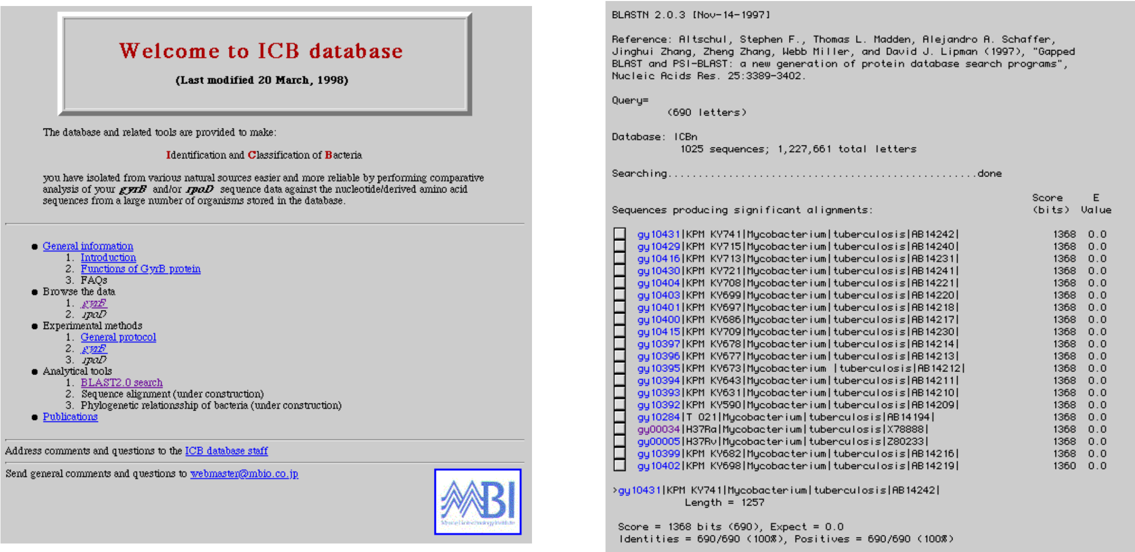
Currently, more than 1,000 *gyrB* sequences are stored in the database (some of them are open only to authorized users). The data have been obtained from bacteria of various taxa. 501 *gyrB* sequences are originated from *Proteobacteria*, 389 from Gram-positive bacteria, 64 from *Cytophaga/Flavobacterium/ Bacteroides* complex, and 52 from other taxonomic groups. The numbers of genera included are 61 *Proteobacteria*, 38 Gram-positive bacteria, 20 *Cytophaga/Flavobacterium/Bacteroides* complex, 14 other groups and 36 unidentified species (the data were scored at the end of July 1998). We believe that there is no other comparative database containing data of protein-coding genes from such a large number of bacteria of taxonomical variety. In addition, the data have been found to contain almost no sequence errors, because it is quite easy to find errors, if any, of a new sequence by comparing its amino acid translation against the database.

In the pages showing experimental procedures, how to obtain materials for *gyrB* sequence and how to perform actual analysis are described in details. These include methods for bacterial DNA preparation, a list of universal PCR primers as well as genus- and/or group-specific PCR primers, conditions for successful PCR amplification for each bacteria, technical details of DNA sequencing, and so on. Trouble shootings and know-how's for designing PCR primers are also described.

The ICB database also provides users with tools for the identification of their bacterial isolates by using its *gyrB* gene sequence data. A typical analytical procedure will include the following steps: 1) perform a similarity search of the obtained DNA/amino acid sequence data against the *gyrB* database; 2) select several sequences showing higher similarity scores; 3) align the selected sequences; and, 4) find the most closely related species by analyzing the aligned sequence data. At step 1, we use the blast 2.0 program [1], in which both BLASTN and BLASTP are made available. At step 2, the results of blast search will be listed together with a check box (Fig. 3-B). The checked data and query data will then be analyzed by ClustalW ver. 1.7 [14] at step 3. Users can choose various options for ClustalW analysis at this step. The results of multiple alignment of selected data along with a treefile will then be presented. The aligned data can easily be subjected to construction of a phylogenetic tree by running a suitable program.

6 Plans for future development of the ICB database

How to obtain a reasonable multiple alignment is one of the key questions for phylogenetic analysis based on gene sequences. To solve various problems encountered in multiple alignment, we are currently seeking the way to incorporate the secondary structure information based on the data obtained with *E. coli gyrB*



A. B.

Figure 3: Presentation of data of the ICB database via internet. The first page of the ICB database is shown in panel A. Blast search of query sequence against the whole data stored in ICB database is available. 20 resultant sequences will then be listed with check boxes for further analysis as shown in panel B and the selected sequences will be subjected to ClustalW analysis with various selectable options.

protein to evaluate and modify the multiple alignment data produced by ClustalW. However, even if the resultant multiple alignment appears reasonable, the taxonomical relationship based on it could be biased, since only one protein-coding gene has been used for the phylogenetic analysis. Therefore, we are preparing to incorporate the data for some other protein- coding genes into the database as well, so that multiple alignment of more than one protein-coding gene will become possible. Thus in the near future, the data for *rpoD* as well as *fliC* (the structural gene for flagellin) will be incorporated into the ICB database. In addition, eukaryotic type II topoisomerase genes from yeast strains will also be included for the comparison of fungal microorganisms. The whole genomic nucleotide sequence data of as many as fifteen bacterial species have become available to date. However, the data are restricted to only model organisms or organisms of some pathogenic significance. In the natural environment, millions of bacterial species are said to exist of which only less than 1% are currently known to be cultivated. The ICB database, which stores the nucleotide sequence information for several protein-coding genes from a wide variety of bacteria, is thus expected, at least more efficiently, to cope with the analysis of the extremely diverse bacterial world.

Acknowledgments

We are grateful to Tohru Hamada, Makoto Suzuki and Shoko Nakamura for their efforts to accumulate *gyrB* data, and to Yuhka Takahashi, Yukiko Itazawa, Nakako Hayashi and Ikuko Hiramatsu for their technical assistance. This work was supported by Industrial Science and Technology Frontier Program from New Energy and Industrial Technology Development Organization.

References

[1] Altschul, S.F., Madden, T.L, Schaffer, A.A, Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*,

25:3389–3402, 1997.

[2] Cole, S. T., Brosch, R., Parkhill, J., *et al.*, Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence, *Nature*, 393: 537-544, 1998.

[3] Frishman, D. and Argos, P., Knowledge-based secondary structure assignment, *Proteins: structure, function and genetics*, 23:566–579, 1995.

[4] Harayama, S. and Yamamoto, S., Phylogenetic Identification of *Pseudomonas* Strains Based on a Comparison of *gyrB* and *rpoD* Sequences, In *Molecular Biology of Pseudomonads*, edited by T. Nakazawa, K. Furukawa, D. Haas, S. Silver, ASM Press, Washington, D.C., 250–258, 1996.

[5] Kaneko, T., Sato, S., Kotani, H., *et al.*, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions, *DNA Res.*, 3:109–136, 1996.

[6] Maidak, B. L., Olsen, G.J., Larsen, N., Overbeek, R., McCaughey, M.J. and Woese, C.R., The RDP (Ribosomal Database Project), *Nucleic Acids Res.*, 25:109–111, 1997.

[7] Olsen, G.J., Woese, C.R., and Overbeek, R., The winds of (evolutionary) change: breathing new life into microbiology, *J. Bacteriol.*, 176:1–6, 1994.

[8] Prager, E.M. and Wilson, A.C., Construction of phylogenetic trees for proteins and nucleic acids: empirical evaluation of alternative matrix methods, *J. Mol. Evol.*, 11:129–142, 1978.

[9] Redenbach, M., Kieser, H.M., Denapaite, D., Eichner, A., Cullum, J., Kinashi, H. and Hopwood, D.A., A set of ordered cosmids and a detailed genetic and physical map for 8 Mb *Streptomyces coelicolor* A3(2) chromosome, *Mol. Microbiol.*, 21:77–96, 1996.

[10] Reischl U., Feldmann K., Naumann L., Gaugler B.J., Ninet, B., Hirschel, B., and Emler, S., 16S rRNA sequence diversity in *Mycobacterium celatum* strains caused by presence of two different copies of 16S rRNA gene, *J. Clin. Microbiol.*, 36:1761–1764, 1998.

[11] Roca, J., The meanisms of DNA topoisomerases, *Trends. Biochem. Sci.*, 20: 156-160, 1995.

[12] Stackebrandt, E. and Goebel, B.M., Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species difinition in bacteriology, *Int. J. Syst. Bacteriol.*, 37:463–464, 1994.

[13] Thiara, A.S. and Cundliffe, E., Expression and analysis of two *gyrB* genes from novobiocin producer, *Streptomyces sphaeroides*, *Mol. Microbiol.*, 8: 495–506, 1993.

[14] Thompson, J.D., Higgins, D.G., and Gibson, T.J., Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 22:4673–4680, 1994.

[15] Tomb, J.F., White, O., Kerlavage, A.R., *et al.*, The complete genome sequence of the gastric pathogen *Helicobacter pylori*, *Nature*, 388:539–547, 1997.

[16] Waynae, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E., Starr, M.P., and Trüper, H.G., Report of the ad Hoc committee on reconciliation of approaches to bacterial systematics, *Int. J. Syst. Bacteriol.*, 37:463–464, 1987.

[17] Wigley, D.B., Structure and mechanism of DNA topoisomerases, *Ann. Rev. Biomol. Struct.*, 24:185–208, 1995.

[18] Woese, C.R., Kandler, O. and Wheelis, M.L., Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, *Proc. Natl. Acad. Sci. USA*, 87:4576–4579, 1990.

[19] Yamamoto, S. and Harayama, S., PCR Amplification and Direct Sequencing of *gyrB* Genes with Universal Primers and Their Application to the Detection and Taxonomic Analysis of *Pseudomonas putida* Strains, *Appl. Environ. Microbiol.*, 61:1104–1109, 1995.

[20] Yamamoto, S. and Harayama, S., Phylogenetic Analysis of Acinetobacter Strains Based on the Nucleotide Sequences of *gyrB* Genes and on the Amino acid Sequences of Their Products, *Int. J. Syst. Bacteriol.*, 46:506–511, 1996.

[21] Yamamoto, S. and Harayama, S., Phylogenetic relationships of *Pseudomonas putida* strains deduced from the nucleotide sequences of *gyrB*, *rpoD* and 16S rRNA genes, *Int. J. Syst. Bacteriol.*, 48:813–819, 1998.

[22] Yamamoto, S., Bouvet, P.J.M., and Harayama, S., Phylogenetic structures of the genus Acinetobacter based on the *gyrB* sequences: Comparison with the grouping by DNA-DNA hybridization, *Int. J. Syst. Bacteriol.*, (in press), 1998.