Construction and analysis of Escherichia coli genome database

Takeshi Itoh[1] t-ito@bs.aist-nara.ac.jp
Minoru Yano[1] m-yano@bs.aist-nara.ac.jp
Keiko Takemoto[2] ktakemot@virus.kyoto-u.ac.jp
Yutaka Akiyama[3] akiyama@kuicr.kyoto-u.ac.jp
Hirotada Mori[1] hmori@gtc.aist-nara.ac.jp

[1] Nara Institute of Science and Technology Ikoma 630-01, Japan
[2] Institute for Virus Research, Kyoto University Kyoto 606-01, Japan
[3] Institute for Chemical Research, Kyoto University Gokasho, Uji 606, Japan

Abstract

It is possible to elucidate whole genome structure by current technique. The genome projects of some species, C.elegans, Yeast, Escherichia coli, Bacillus subtilis, Arabidopsis, rice and human are now running. In Escherichia coli, two lines of large scale sequencing have emerged. One by the Wisconsin group in U.S.A. and the another by the collaborative research group in Japan. To make a non redundant sequence database is essential not only for effective promotion of sequencing project but for whole genome analysis and reference by biologists. We determine the sequences as one of the research group in Japan and make a non redundant DNA sequence database for effective promotion of genome project and analysis of genome structure. In Genome Workshop meeting 1993, we reported the construction of Escherichia coli genome database on Genomatica system. We update our E.coli genome database by incorporating of E.coli new entries of GenBank and from genome project research groups. The contiguous sequence data were then used to predict possible open reading frames. The translated amino acid sequences from these ORFs were subjected to homology analysis against the PIR and the SWISSPROT protein database. The whole sets of plausible ORF's were further classified by similarities between ORF's and those of gene organizations. It may be possible to detect rearrangements of chromosome through its own evolution by that analyses.

1 Methods and Progress

First, Escherichia coli K12 entries were selected from bacterial sequence data set gbbct.seq of GenBank Release 83. All entries were divided into reference and sequence parts. Checksum field was added into the head line of sequence part in order to distinguish whether sequences of entries possessing same accession number were revised easily. We revised Escherichia coli database as follows,

1. select entries possessing the same accession numbers in previous Escherichia coli database.
2. check entries whether sequences were revised or not by comparing checksum.
3. revised entries were then replaced and realigned into contiguous sequence.

[1]伊藤 剛，矢野 実，森 浩禎：奈良先端科学技術大学院大学，〒 630-01 奈良県生駒市高山町 8916-5
[2]竹本 経緯子：京都大学ウイルス研究所，〒 606-01 京都市左京区聖護院川原町
[3]秋山 泰：京都大学化学研究所，〒 611 宇治市五ヶ庄

4. others were treated as new entries for E.coli database except for plasmid sequences or other than K12 strains.

5. new entries were subjected to blast [1] analysis against E.coli contig data sets to find contig which belong to.

6. realigned each members and new entry sequence to make a new contig.

We assorted all entries in Escherichia coli K12 data sets from gbbct.seq into 1) not necessarily revised, 2) necessarily replaced and 3) necessarily added sequences. We also developed alignment editor as joint development with SDC Co. and we applied this tool to making contig. We will also introduce this editor in this meeting. Each contig should be assigned their own physical positions on Escherichia coli chromosome. We determined their position by Mapsearch program developed by K.Rudd et al [2]. This program assigned fragment's physical position deduced from Escherichia coli physical map, and we adopted Kohara's physical map after slight modification [3]. As mentioned the above, we revised data sets of Escherichia coli and rebuild into Genomatica system except the gene locations. We used genes' location from Barbara Bachmann's data [4] in previous database, however, we adopted the positions of each gene from orf analysis in order to make more accurate correspondence between genes and Kohara's lambda clones. Contig sequences were next subjected to orf analysis. Open Reading Frames of 75 or more consecutive sense codons were translated from each contig sequence in 3 phases for both orientations and subjected to identity analysis against Protein database of Escherichia coli for picking up known proteins and similarity analysis against PIR and SWISSPROT protein databases. Some orf's predicted in the newly sequenced region from Japanese genome research group showed considerable homology with those of known proteins from various organisms. A contiguous nearly 580 kb sequence from 0 min was determined except for a few small gaps, and analyses of 0 - 2.4 min and 2.4 - 4.1 min region were reported [3, 5], and the next 4.1 - 6.0 min region will be published in near future [6]. We are analyzing next 6.0 - 12.0 min region and will report the results in this meeting. For analyzing newly sequenced region from Japanese genome project, we develop a tool for prediction of orfs on the basis of SD sequence similarity. We applied to pick up plausible open reading frames from functional unknown region and developed further. We also introduce some aspects of this tool.

References

[1] Altschul,SF., Gish,W., Miller,W., and Myers,EW. (1991) Basic Local alignment search tool. J. Mol. Biol. 215, 403-410

[2] Rudd,K.E., Miller,W., Werner,C., Ostell,J.,Tolstoshev,C. and Satterfield,S.G. (1991) Mapping sequenced E.coli genes by computer: software, strategies and examples. Nucleic Acids Res. 19, 637-647

[3] Fujita, N., Mori, H., Yura, T., Matsubara, K. and Ishihama, A. (1994) Systematic Sequencing of the Escherichia coli Genome: Analysis of the 2.4-4.1 Min (110,917-193,643 bp) Region. Nucleic Acids Res. Vol.22, 1637-1639

[4] Bachmann,B.J. (1990) Linkage Map of Escherichia coli. Microbiological Reviews, 54, 130-197

[5] Yura, T, Mori,H, Nagai,H, Nagata,T, Ishihama,A, Fujita,N, Isono,K, Mizobuchi,K and Nakata A. (1992) Systematic Sequencing of the Escherichia coli Genome: Analysis of the 0 - 2.4 min region, Nucleic Acids Res. Vol. 20, 3305-3308

[6] Takemoto,K. et al. Systematic sequencing of the Escherichia coli genome:Analysis of the 4.1 - 6.0 min region. in preparation.