Clustering and Evolutional Analysis of E. coli Proteins

Sivasundaram Suharnan [†]
suharu-s@is.aist-nara.ac.jp
Takeshi Itoh [†]
t-itoueb@bs.aist-nara.ac.jp
Hidemi Watanabe[‡]
hidwatall@lab.nig.ac.jp
Jun-ichi Takeda[†]
j-takeda@bs.aist-nara.ac.jp
Keiko Takemoto[*]
ktakemot@virus.kyoto-u.ac.jp
Hideo Matsuda[††]
matsuda@ics@es.osaka-u.ac.jp
Hirotada Mori[†]
hmori@gtc.aist-nara.ac.jp

[†]Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-01, Japan
[‡]National Institute of Genetics
1111 Yata, Mishima, Shizuoka 411, Japan
[*]Institute for Virus Research, Kyoto Univ.
Shogoin-Kawahara, Sakyo, Kyoto, 606-01, Japan
[††]Faculty of Engineering Science, Osaka Univ.
1-3 Machikaneyama, Toyonaka, Osaka 560. Japan

1 Introduction

As a model organism, Escherichia coli has been playing significant roles in the establishment of a number of basic concepts in molecular biology. In addition, determination of the whole genome sequence of E. coli is undoubtedly awaited for further insight into various biological phenomena. In E. coli the DNA sequences of the 0 - 40 minutes[1]and 69 - 100 min region[2] had been determined and at present the region between 40 to 69 minutes is under proceeding both by F. Blattner*s group and the Japan E.coli sequencing group. At present a contiguous sequence of E. coli, covering almost 90predict ORFs and cluster all ORFs on the basis of amino acid sequence similarities and analyze evolution of E. coli genes.

2 Methods

We used two main methods to predict coding regions. To assign ORFs, we first selected potential ORFs which were composed of more than 50 consecutive seance codons. All of these ORFs were translated into amino acid sequences in three frames in both strands and then subjected to similarity analysis against the SWISS-PROT and PIR protein databases by the BLASTP program[3]. The second method is a combination of the GeneMark program[4] and N-terminal prediction described by following steps;

1. Search ATG, GTG or TTG as a candidate of an initiation codon from 5'-terminal.
2. Calculate efficiency of ribosome binding by the scoring matrix[5].
3. If the score is not less than a certain threshold, go to 5.
4. Go to 1, while the ORF is not shorter than 50 amino acids length.
5. if the ORF overlaps with neighbor ones, and if the shorter one is longer than 2/3 of the longer one, or the half of the shorter one overlaps with the longer one, the shorter one is discarded.

All ORFs of E. coli were clustered by the FASTA[6] program in order to find out domain structures and then we intend to elucidate internal structural and functional units of genes as traces of ancestral small genes by multiple alignment analysis. Furthermore, to infer physiological functions and biochemical properties of newly identified ORFs, we have a plan of experimental approaches (e.g. gene disruption experiment etc.) based on computational approaches as described above.

3 Progress

The determination of the complete genome sequence of E.coli is now in progress, and this will be finished until the end of 1996. All genes coded on E.coli chromosome are predicted, classified by the similarity and analyzed by the theory of evolution. Through these approaches, our objectives are the identification of the common ancestral genes and of the structural and the functional units of genes. As described in method, the ORFs are predicted by the combination of the GeneMark program and SD based ORF prediction method developed by ourselves. All ORFs predicted are subjected to similarity analysis in all combination by FASTA2.0. The ORFs are divided into groups by the single linkage method at the similarity score of 100 of FASTA2.0. In each ORF groups, they are divided into subgroups at the higher score of FASTA. The ORFs in the subgroup are aligned and identified the common conserved regions. Through these steps, the units which construct genes are predicted. Further more, the correspondence between those units and the 3D structures will be identified. The results from our approaches as described above from the latest sequence data of E.coli will be shown and discussed.

References

[1] Ohshima, T. et. al. (1996) DNA Research, 3, 137-155.

[2] Plunkett, G. et al. (1994) GenBank, entry name: ECOUW67, ECOUW76, ECOUW82, ECOUW85, ECOUW87, ECOUW89, ECOUW93

[3] Altschul, S. F et. Al.(1991) J. Mol. Bioll., 215, 403-410.

[4] Borodovsky, M. and Mclnil'ch, (1993) J. Comput. Chem. 17 , 123-133.

[5] Barrick, D., et. Al. (1994) Nucleic Acids Res., 22, 1287-1295.

[6] Lipman, D.J. and Pearson, W.R. (1985) Science, 227, 1435-1441