Clustering of all known and predicted open reading frames of *Escherichia coli* K12

Takeshi Itoh[1]
t-itou@bs.aist-nara.ac.jp

Minoru Yano[1]
m-yano@bs.aist-nara.ac.jp

Keiko Takemoto[2]
ktakemot@virus.kyoto-u.ac.jp

Miwako Kajihara[1]
m-kaziha@bs.aist-nara.ac.jp

Hirotada Mori[1]
hmori@gtc.aist-nara.ac.jp

[1] Research and Education Center for Genetic Information,
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-01, Japan

[2] Institute for Virus Research, Kyoto Univ.
Syougoin-Kawahara, Sakyo, Kyoto 606-01, Japan

Abstract

At present, the non redundant contig sequences of E.coli which covers about 70% of the whole chromosome are constructed. We predicted ORF's (Open Reading Frames) from 2,554,518 bp contig sequences on the basis of Shine-Dalgarno (ribosome binding) sequence. All ORF's were classified according to the structural similarities. Through examining the homology of ORF's in each group in detail, some structural units were revealed.

1 Method

To construct contig sequences, overlapped sequences were removed from *E.coli* K12 entries of GenBank release 83 and DNA sequences from *E.coli* genome project. The total length of *E.coli* contig sequences was 2,554,518bp. The known genes were identified by the BLAST[1] search against SWISSPROT and PIR protein databases. ORF's from contig sequences were predicted described as follows;

1. ORF's containing 75 or more consecutive sense codons are searched in 3 phrases for both orientation.

2. The candidates of initiation condons, ATG, GTG and TTG are searched from N terminal end of each ORF's.

3. To search ribosome binding sequence (3 consecutive nucleotides of A or G) in the appropriate position (8 to 13 bp upstream to each candidate of initiation codon).

[1]伊藤 剛、矢野 実、梶原美和子、森 浩禎：奈良先端科学技術大学院大学遺伝子教育研究センター、〒 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

4. To calculate a score of each ribosome binding site according to Barrick's matrix[2]. The candidate which showed higher score than a certain threshold was considered as a plausible initiation codon. However, the amino acids sequence of consecutive codons whose length from plausible initiation codon to termination codon was less than 66 amino acids length was not defined as an ORF.

5. The whole set of amino acids sequences of consecutive codons is subjected to each step from 2 to 4.

6. When two ORF's are overlapped with each other, if the shorter one is equal to or less than 2/3 length of the longer one, or more than half of the shorter one overlaps with the longer one, the shorter one was discarded as a false ORF.

2 Results and Discussion

To test the accuracy of this prediction method, we compared the length of 974 ORF's predicted by our method to the length of corresponding data from protein database. 86.1% of the predicted ORF's were assigned to *E.coli* genes when the threshold of the ribosome binding score was defined as 4.0, though some of them were predicted as longer ORF's than corresponding genes. 2.5% of genes were completely failed to predict. Without step 6, the accuracy of the prediction was not changed but a lot of false ORF's have still remained (data not shown). Finally, we got 2,471 ORF's from our prediction method as mentioned above. The number of predicted ORF's in our way was 2471 from 2.5 Mb contig sequence and it is consistent with the average length of gene of Bacteria (estimated approximately 1.1- 1.2 kb). There are still remained the problems to be solved in the condition of prediction and the accuracy of prediction is expected to be improved. It is expected that the combination of our method and the method which is based on consideration of bias of nucleotide composition in coding region makes it possible to much more accurate prediction. The whole set of predicted ORF's were then subjected to homology analysis using FASTA[3] program to each other. The predicted ORF's were then clustered on the basis of similarity which makes equal to or more than 100 FASTA score. 1011 ORF's were divided into 196 clusters. 1460 ORF's showed no homology against each other. Most of ORF's belong to the small clusters which have less than 10 ORF's. Few clusters, however, consist of a large number of ORF's. We examined the FASTA results and Motif search results of ORF's in each group in detail. Some Aminoacyl-tRNA synthetase formed a group and two motifs were identified. It is suggested that two conserved domains containing a motif exist. It is clearly observed that ORF's of *araC* family and methyltransferase family were constructed by multi domain structures. *araC* family consists mainly of regulator proteins, such as repressor, and ORF's of this group have HTH motif on the C terminus. Some proteins which belong to the same cluster but have different functions like *rob* gene which participates in chromosomal replication shows a different structure and HTH motif is found on N terminus. Through these analysis as mentioned above, it is possible to construct a database of all kinds of functional and structural motifs of *E.coli* proteins. We expected that this makes us possible not only to predict the function of newly sequenced region but to elucidate the initial set of functional peptides which form proteins.

References

[1] Altschul, S.F., Gish, W., Miller, W., and Myers, E.W. (1991) *J. Mol. Biol.* **215** 403-410

[2] Barrick, D., Villanueba, K., Childs, J., Kalil, R., Schneider, T.D., Lawrence,C.E., Gold, L. and Stormo, G.D. (1994) *Nucleic Acids Res.* **22** 1287-1295

[3] Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci.* **85** 2444-2448