THE STATISTICAL PHYSICS OF MICROBIAL GENOMES: PART II. ORGANISATION OF CODING SEQUENCES INVOLVED IN THE CELL DIVISION OF *ESCHERICHIA COLI*

OANA ZAINEA, V.V. MORARIU*, O. POPESCU***

*Department of Molecular and Biomolecular Physics, National Institute for R&D of Isotopic and
Molecular Technology, P.O.Box 700, Cluj-Napoca 5, 400293, Romania,
E-mail: oanaz@L40.itim-cj.ro
**Faculty of Biology, “Babeş-Bolyai” University, Cluj-Napoca, Romania

Abstract. The cell division coding sequences of *Escherichia coli K12, O157:H7 Sakai and O157-H7 EDL933* strains were subject to a simple ranking method in connection to the origin and terminus of replication. These coding sequences were found to be distributed all along the chromosome. Data have a non-random organization. A clear grouping of genes around terminus and origin of replication is plainly observed.

Key words: bacteria, cell division, chromosome, coding sequences, *Escherichia coli*.

INTRODUCTION

In the previous part of this report [9] it has been shown that the length of coding sequences in the bacterial chromosome is a series of data having a non-random organization. Such a series represents a stochastic system with a moderate long-range correlation property [1]. The purpose of the present work is to investigate how the cell division coding sequences are organized in the genome of *Escherichia coli K12, O157:H7 Sakai* and *O157-H7 EDL933* strains. A fundamental characteristic of the bacterial chromosome organization is the universal scale-invariant property in respect to gene localization and orientation [1, 2]. This means that there is no characteristic dimension for the description of chromosome structure. Gene positions and orientations share a common scale-invariant correlation described by what is known as “long-range correlations” or by “power laws” or “fractal like” structures [13, 14]. This indicates that genes tend to assemble over any scale of observation greater than a few kilo bases [5]. The presence of base long-range correlation has also been found in coding and

Received June 2006;
in final form July 2006.

noncoding regions [4, 7]. A genomic tree of living organisms based on a fractal model has been proposed [15]. The fractal properties were proposed to explore and identify genes and intergenic regions in genomes [12]. Therefore, the fractal – or nonlinear – approach of the genome is a methodology applicable to most of the fundamental questions regarding such structures.

The bacterial cell division is a complex macromolecular machinery where at least ten well-conserved proteins take part. Many genes and proteins involved in the bacterial cell division were characterized by classical and molecular genetics using experimental models such as *Escherichia coli* – gram-negative bacteria – and *Bacillus subtilis* – gram-positive bacteria. However, the biochemical functions of only two proteins were elucidated: FtsZ and PBP (Penicillin-Binding Protein) [6]. In 1980 the ftsZ gene was identified for *E. coli* and it was demonstrated that it is involved in the cell division [8]. Other proteins, e. g. ZipA (*E. coli*) and probably EzrA (*B. subtilis*), are important for the cell division and bind the above mentioned cytosolic proteins to cell membrane. Other proteins involved in division (FtsK, FtsQ, FtsL, YgbQ, FtsW, FtsI, FtsN) are either integral membrane proteins or transmembrane proteins with their major domains outside the bacterial cell. Their functions are unknown [6].

The bacterial cell division has five stages:

a. the selection of the division place; as a general rule, in the middle of the cell between the replicated nucleoids and already segregated ones a little time before;

b. the assembly of the cytoplasmic apparatus which invariably involves FtsZ and, for the majority of the organisms, FtsA; these proteins recognize, bind and catalyze the hydrolysis of the triphosphate nucleosides offering the necessary energy for reshaping the cell;

c. the interaction with one or more proteins for the binding of the division complex to cell membrane;

d. the assembly of proteins to extracellular major domains, especially the specialized proteins of type PBP which participate in the direct synthesis of cell wall components;

e. the constriction and shutting of the septum.

Nowadays, it is clear that some aspects of the bacterial cell division mechanism are strongly conserved on the evolutionary scale while others, especially those of stages *d* and *e,* are very different among species.

In the present work the location of the cell division coding sequences is explored by a simple ranking method in relationship to the origin and terminus of replication.

MATERIALS AND METHODS

In this paper, we are interested in the organisation of coding sequence for the complete genome of *Escherichia coli K12* (4,639,675 bp), *O157:H7 Sakai* (5,498,450 bp) and *O157-H7 EDL933* (5,528,445 bp) strains, regarding only the cell division process. Our data were taken from the National Centre for Biotechnology Information (NCBI) website [10].

First, we searched under Genome category each bacterial genome of interest. The results for every search were displayed under Summary form, each of them being coded into different numbers (e.g. NC 000913 for *Escherichia coli K12*).

Second, by selecting the corresponding code for each genome of interest and then choosing COG (Clusters of Orthologous Groups of proteins) under Blast homologs, the complete list of proteins by COG functional categories was displayed.

Since we are interested only in cell division, we have chosen only the proteins with the D code, which involves cell cycle control, mitosis and meiosis. Here, under D COG category, for each bacterial genome regarding cell division process, useful information is given, such as strand (+ or –), coding sequence length, the function of proteins, etc. We imported each D COG gene file into an individual .doc file, in order to select the starting position for coding sequences (CDS). The programme Origin® version 6.1 was used to fit the genome data. The coding sequence location was further analyzed by a simple ranking method. The main finding refers to the fact that the cell division coding sequences were non-uniformly distributed all along the chromosome.

RESULTS AND DISCUSSION

The starting position of the coding sequence (CDS) for *Escherichia coli O157:H7 Sakai* strain in an ascending order, which is also the natural order of genes in the genome, is shown in Figure 1. The terminus – *ter* – and origin of replication – *ori* – are clearly indicated by an arrow pointing towards gene starting position of the coding sequence. By looking at the predicted positions [11] – 4716984 bp for the *ori* position; 2113730 bp for the *ter* position – and those we had observed (Fig. 1), a difference of around 70000 bp for *ori* and about 52000 bp for *ter* is observed. Since the entire genome length is over 5 Mbp, 70 Kbp or 52 Kbp are small compared with the complete genome base pairs.

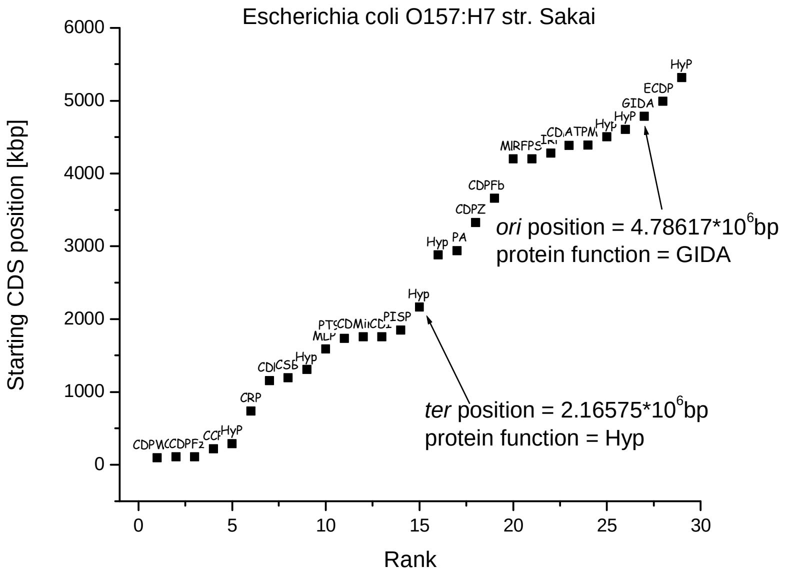

Fig. 1. The starting position of coding sequence in ascending order for *Escherichia coli O157:H7 Sakai* strain showing the terminus – *ter* – and origin – *ori* – of replication; with all cell division protein function codes (see Table 1).

So, we consider these qualitative positions fairly good for our statistical analysis.

Replication of the bacterial chromosomal DNA starts at the *origin* and continues up to *the terminus* region, as demonstrated in *Escherichia coli* [3]. The two replication forks need to reach the *ter* region at the same time to optimize the replication process [11]. Therefore, the terminus and origin of replication should be opposite to each other in order to guarantee a synchronous completion of the bi-directional chromosomal replication [11]. This is nicely illustrated in Figure 1 for *Escherichia coli O157:H7 Sakai* strain.

All protein functions for *Escherichia coli O157:H7 Sakai* strain included into the D COG cell division functional category are shown in detail in Table 1 and also as labels for starting position of the coding sequence in Figure 1.

Although prokaryotic genomes included in public data banks, such as NCBI, are well-annotated, an important number of proteins have unknown, hypothetical or predicted roles. This can be seen for both *Escherichia coli – O157:H7 Sakai* and *K12* – strains in Tables 1 and 2, which describe in detail all the cell division protein functions.

Table 1

Protein function description and code for *Escherichia coli O157:H7 Sakai* strain

Rank	Protein function description (NCBI Data Bank)	Protein function code
1	cell division protein; ingrowth of wall at septum	CDPW
2	cell division protein	CDP
3	cell division protein FtsZ	CDPFz
4	cell cycle protein	CCP
5	hypothetical protein	HyP
6	camphor resistance protein CrcB	CRP
7	cell division protein	CDP
8	condesin subunit B	CSB
9	hypothetical protein	Hyp
10	Maf-like protein	MLP
11	putative tail component of prophage CP-933X	PT933
12	cell division topological specificity factor MinE	CDMin
13	cell division inhibitor, a membrane ATPase, activates minC	CDI
14	putative intracellular septation protein	PISP
15	hypothetical protein	Hyp
16	hypothetical protein	Hyp
17	putative ATPase	PA
18	cell division protein ZipA	CDPZ
19	cell divison protein FtsB	CDPFb
20	Maf-like protein	MLP
21	regulator of ftsI, penicillin binding protein 3, septation function	RFPS
22	induced in stationary phase, recognized by rpoS, affects cell division	IRr
23	cell division membrane protein	CDMP
24	ATP-binding component of a membrane-associated complex involved in cell division	ATPM
25	hypothetical protein	Hyp
26	hypothetical protein	HyP
27	glucose-inhibited division protein A	GIDA
28	essential cell division protein	ECDP
29	hypothetical protein	HyP

At NCBI Data bank from where we took our data, same D code COG functional category file (table 2) is available for both *Escherichia coli O157:H7 Sakai* and *O157-H7 EDL933* strains. Therefore, a similar graph such as that shown in Figure 2 is valid for *Escherichia coli O157:H7 EDL933* strain too. It is interesting that the above mentioned strains, although they have diverse genome base pairs lengths, their COGs functional category for cell division is identical. Obviously, the terminus and origin of replication for *Escherichia coli O157:H7 EDL933* are located at different positions along the coding sequence. The predicted *ori* and *ter* are 4786037 bp and 1974813 bp, respectively [11].

For all investigated *Escherichia coli* strains, *K12, O157:H7 Sakai* and *O157-H7 EDL933*, the coding sequences starting positions are distributed among *ter* and *ori*. Interesting enough, one can clearly see the grouping of genes around terminus and origin of replication. Since they are vital for cell division, it is compulsory that this assembly takes place. A surprising observation is that there is no grouping of genes regarding protein functions. Flanked by our three analysed *Escherichia coli* strains are the following genes: FtsA, FtsZ, CcrB, FtsK, Maf, MinE, FtsB, ZipA and GidA. Their role is cell division protein (FtsA, FtsK, FtsZ, FtsB and ZipA), camphor resistance protein CrcB, maf-like protein, cell division topological specificity factor MinE and glucose-inhibited division protein A. For all our explored *Escherichia coli* strains, GidA gene that results into glucose-inhibited division protein A is located at the origin of replication, which is another nice observation. The terminus of replication is unique for each *Escherichia coli* strain explored: hypothetical protein for *O157:H7 Sakai* and *O157:H7 EDL933* strains and toxin of the RelE-RelB toxin-antitoxin system for *K12* strain.

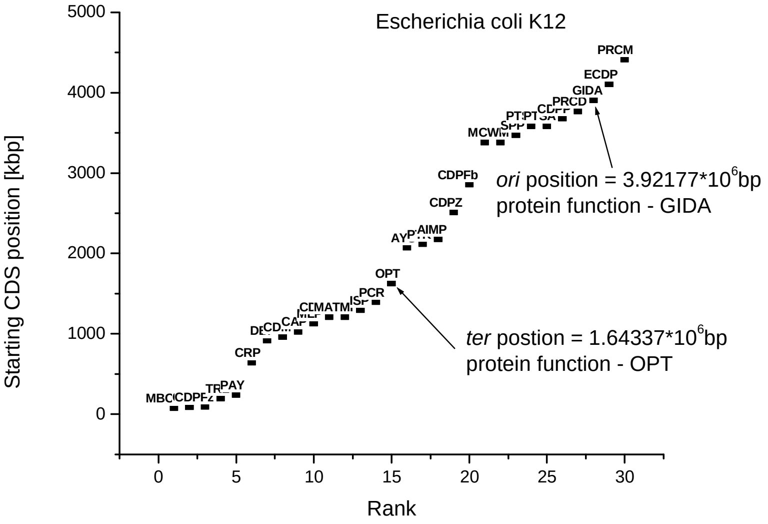

Fig. 2. The starting position of coding sequence in ascending order for *Escherichia coli K12 strain* showing the terminus – *ter* – and origin – *ori* – of replication; with all cell division protein function codes (see Table 2); the predicted position for *ori* is 3921168 bp and for *ter* 1606253 bp.

Table 2

Protein function description and code for *Escherichia coli K12* strain

Rank	Protein function description (NCBI Data Bank)	Protein function code
1	membrane bound cell division protein at septum containing leucine zipper motif	MBCD
2	cell division protein	CDP
3	cell division protein FtsZ	CDPFz
4	tRNA(Ile)-lysidine synthetase	TRL
5	predicted antitoxin of the YafO-YafN toxin-antitoxin system	PAY
6	camphor resistance protein CrcB	CRP
7	DNA-binding membrane protein required for chromosome resolution and partitioning	DBP
8	cell division protein MukB	CDM
9	cryptic autophosphorylating protein tyrosine kinase Etk	CAP
10	Maf-like protein	MLP
11	cell division topological specificity factor MinE	CDMin
12	membrane ATPase of the MinC-MinD-MinE system	MATMi
13	intracellular septation protein A	ISP
14	predicted C32 tRNA thiolase	PCR
15	Qin prophage; toxin of the RelE-RelB toxin-antitoxin system	OPT
16	antitoxin of the YoeB-YefM toxin-antitoxin system	AYS
17	protein-tyrosine kinase	PTK
18	antiporter inner membrane protein	AIMP
19	cell division protein ZipA	CDPZ
20	cell divison protein FtsB	CDPFb
21	Maf-like protein	MLP
22	cell wall structural complex MreBCD, actin-like component MreB	CWM
23	stationary-phase protein, cell division	SPP
24	predicted transporter subunit: membrane component of ABC superfamily	PTSM
25	predicted transporter subunit: ATP-binding component of ABC superfamily	PTSA
26	cell division protein (chromosome partitioning ATPase) (pseudogene)	CDPP
27	protease with a role in cell division	PRCD
28	glucose-inhibited division protein A	GIDA
29	essential cell division protein	ECDP
30	predicted regulator of cell morphogenesis and cell wall metabolism	PRCM

To conclude, the coding sequence location for *Escherichia coli* was analyzed by a simple ranking method. The main findings refer to the fact that the cell division coding sequences were non-uniformly distributed all along the chromosome and there is an evident grouping of these genes before terminus and origin of replication.

Acknowledgements. This study was supported by a PNCD *CERES* grant from the Romanian Research Agency.

REFERENCES

1. ALLEN, T.E., N.D. PRICE, A.R. JOYCE, B.O. PALSSON, Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization, *PLoS Computational Biology,* 2006, **2**, 1–9.
2. AUDIT, B., C.A. OUZOUNIS, From genes to genomes: universal scale-invariant properties of microbial chromosome organization, *J. Mol. Biol.*, 2003, **332**, 617–633.
3. BAKER, T.A., Replication Arrest, *Cell*, 1995, **80,** 521–524.
4. BULDYREV, S.V., N.V. DOKHOLYAN, A.L. GOLDBERGER, S. HAVLIN, C.-K. PENG, H.E. STANLEY, G.M. VISWANATHAN, Analysis of DNA sequences using methods of statistical physics, *Physica A,* 1998, **249**, 430–438.
5. BULDYREV, S.V., A.L. GOLDBERGER, S. HAVLIN, R.N. MANTEGNA, E. MATSA, C.-K. PENG, M. SIMONS, H.E. STANLEY, Long-range correlations properties of coding and noncoding DNA sequences: GenBank analysis, *Phys. Rev. E,* 1995, **51,** 5084–5091.
6. ERRINGTON, J., R.A. DANIEL, D.-J. SCHEFFERS, Cytokinesis in bacteria, *Microbiology and Molecular Biology Reviews,* 2003, **67,** 52–65.
7. LI, W., K. KANEKO, Long-range correlations and partial 1/f spectrum in a non-coding DNA sequence, *Europhys. Lett*., 1992, **17,** 655.
8. LUTKENHAUS, J.F., H. WOLF-WATZ, W.D. DONACHIE, Organization of genes in the *ftsA-envA* region of the *Escherichia coli* genetic map and identification of a new *fts* locus (*ftsZ*), 1980, *J. Bacteriol.*, **142**, 615–620.
9. MORARIU, V.V., O. ZAINEA, A. BENDE, O. POPESCU, The statistical physics of microbial genomes: Part I. Organisation of coding sequences in the chromosome of *Escherichia coli*, *Romanian J. Biophysics***,** 2006, **16**, 2, 103–110.
10. National Centre for Biotechnology Information (NCBI), website http://www.ncbi.nih.gov.
11. SONG, J., A. WARE, S.-L. LIU, Wavelet to predict bacterial *ori* and *ter*: a tendency towards a physical balance, *BMC Genomics,* 2003, **4(1)**, 17.
12. TSONIS, A.A., P.A. TNOSIS, Exploring nonlinearity to identify genes and intergenic regions in genomes, *Physica A,* 2005, **348,** 339–348.
13. YU, Z.G., V. ANH, B. WANG, Correlation property of length sequences based on global structure of the complete genome, *Phys .Rev. E,* 2001, **63,** 011903.
14. YU, Z.G., V. ANH, Time series model based on global structure of complete genome, *Chaos, Solitons and Fractals,* 2001, **12,** 1827–1834.
15. YU, A.G., V. ANH, K.S. LAU, K.H. CHU, The genomic tree of living organisms based on a fractal model, *Physics Letter A*, 2003, **317,** 293–302.