PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

A census of *E. coli* sequence repeats

ArticleInfo		
ArticleID	:	3560
ArticleDOI	:	10.1186/gb-2000-1-1-reports026
ArticleCitationID	:	reports026
ArticleSequenceNumber	:	51
ArticleCategory	:	Paper report
ArticleFirstPage	:	1
ArticleLastPage	:	4
ArticleHistory	:	RegistrationDate : 2000–2–1 Received : 2000–2–1 OnlineDate : 2000–4–27
ArticleCopyright	:	BioMed Central Ltd2000
ArticleGrants	:	
ArticleContext	:	130591111

Rachel Brem

Abstract

A statistical analysis of simple repeat motifs in DNA across the *Escherichia coli* genome is described.

Significance and context

Many genomes contain kilobases of DNA sequence that do not code for protein or RNA. Much of this DNA is simple sequence repeats (SSRs), long repeats of simple nucleotide patterns. Recent biochemical work suggests that SSRs might have a role in the regulation of gene expression, histone binding or DNA replication. Because the environment of one base in an SSR is a lot like that of its neighbors, the DNA replication machinery can confuse one base with another, skip a base or repeat replication of a base. As a consequence, SSRs are more likely to mutate than are other DNA sequences. Here, Gur-Arie *et al*. identify putative SSRs in the genome of *E. coli*. They then compare SSR sequences from several strains of *E. coli*, and report the polymorphisms - the sequence differences - between strains. The proposed SSRs are of interest as the basis for further biochemical and diagnostic work with bacteria.

Key results

Gur-Arie *et al*. first identify tracts of one-, two-, three- and four-nucleotide patterns in the *E. coli* genome that might be genuine SSRs. They weed out artifacts from this set in two ways. First they compare against control genomes. These are computer-generated random strings of sequence with the same nucleotide composition as the real *E. coli* genome. From this experiment they find several thousand mononucleotide repeats of up to eight bases (for example, AAAAAA) and about 2300 trinucleotide repeats (for example GAT) which appear more often in the *E. coli* genome than expected by chance. A and T appear more often in these repeats than G and C. The putative SSRs were not randomly distributed in the genome. Many appeared in noncoding regions; of these, 50% lay within 200 bases of the start site of a gene. Finally, the authors compare the sequences of 14 SSRs among 23 strains of *E. coli*. They find four mononucleotide SSRs that vary in length among the strains, some by as many as six bases.

Methodological innovations

The authors' newly developed SSR identifying software is available by FTP (named ssr.exe).

Links

Gur-Arie *et al*. obtained the *E. coli* genome from the *Escherischia coli* WWW homepage.

Conclusions

The authors hypothesize that, as has been found in other organisms, SSRs in *E. coli* might have a role in the rapid mutation of the regulation of gene expression or histone binding. They postulate that the SSRs they find close to genes are good candidates for such regulators. They also suggest that the polymorphisms they have discovered might be used to make probes to pick out infectious strains of *E. coli*, for example in screening packaged food for contamination.

Reporter's comments

The list of new SSRs is a useful starting point for constructing hypotheses to be tested by experiment. Gur-Arie *et al*. argue that the SSRs they find near gene sequences are biologically important, but this needs to be confirmed by mutational analysis *in vitro*. Similarly, their polymorphism results might be diagnostically useful as probes to distinguish between *E. coli* strains, but probes to long repeat sequences can cross-react, so it might be difficult to put this idea into practice.

Table of links

Genome Research

Escherischia coli WWW homepage

SSR identifying software

References

1. Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y: Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. Genome Res. 2000, 10: 62-71. 1088-9051