Journal of Theoretical Biology 264 (2010) 613–622

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Essential molecular functions associated with the circular code evolution

Ahmed Ahmed, Gabriel Frey, Christian J. Michel *

Equipe de Bioinformatique Théorique, FDBT, LSIIT (UMR CNRS-ULP 7005), Université de Strasbourg, Pôle API, Boulevard Sébastien Brant, 67400 Illkirch, France

ARTICLE INFO

Article history:
Received 24 July 2009
Received in revised form
17 January 2010
Accepted 5 February 2010
Available online 11 February 2010

Keywords:
Circular code
Evolution
Evolutionary trinucleotides
Circular code stability
Trinucleotide stability
Comparative genomics
Database
Molecular functions
Essential genes

ABSTRACT

A circular code is a set of trinucleotides allowing the reading frames in genes to be retrieved locally, i.e. anywhere in genes and in particular without start codons, and automatically with a window of few nucleotides. In 1996, a common circular code, called *X*, was identified in large populations of eukaryotic and prokaryotic genes. Hence, it is believed to be an ancestral structural property of genes. A new computational approach based on comparative genomics is developed to identify essential molecular functions associated with circular codes. It is based on a quantitative and sensitive statistical method (FPTF) to identify three permuted trinucleotide sets in the three frames of genes, a flower automaton algorithm to determine if a trinucleotide set is a circular code or not, and an integrated Gene Ontology and Taxonomy (iGOT) database. By carrying out automatic circular code analyses on a huge number of gene populations where each population is associated with a particular molecular function, it identifies 266 gene populations having circular codes close to *X*. Surprisingly, their molecular functions include 98% of those covered by the essential genes of the DEG database (Database of Essential Genes). Furthermore, three trinucleotides GTG, AAG and GCG, replacing three trinucleotides of the code *X* and called "evolutionary" trinucleotides, significantly occur in these 266 gene populations. Finally, a new method developed to analyse and quantify the stability of a set of trinucleotides demonstrates that these evolutionary trinucleotides are associated with a significant increase of the stability of the common circular code *X*. Indeed, its stability increases from the 1502th rank to the 16th rank after the replacement of the three evolutionary trinucleotides among 9920 possible trinucleotide replacement sets.

1. Introduction

1.1. Minimal gene set

The study of essential genes and the analysis of their features are obviously of great interest in basic and applied researches. We recall a common definition of a minimal gene set. A minimal gene set, or essential genes, is a smallest possible group of genes that would be sufficient to sustain a functioning cellular life form under the most favourable conditions imaginable, that is, in the presence of a full complement of essential nutrients and in the absence of environmental stress (Cho et al., 1999; Hutchison et al., 1999; Mushegian, 1999; Nelson et al., 1999). The upper bound of a minimal gene set is the number of genes of the smallest known genome (Koonin, 2003).

The sequencing of the *Mycoplasma genitalium* which is the smallest known genome until now with a size of 580 kb and 470 predicted genes, allowed an upper bound to be determined (Fraser et al., 1995). This bacterium is capable of living independent of its host.

The analysis of a minimal gene set aims to identify those genes that are essential to support the cell life using comparative and experimental methods. However, a minimal gene set depends on environmental conditions of the cell (Koonin, 2000; Gerdes et al., 2006). Some researchers defined a set of functions that should be covered by a minimal gene set to maintain the cell integrity, including translation, transcription, replication, membrane-transport and energy conversion (Alberts et al., 2002; Gil et al., 2004).

A minimal gene set is an essential factor for further experimental and theoretical researches as the full-synthesize and semi-synthesize of a functional cell (Luisi et al., 2006; Forster and Church, 2006; Gibson et al., 2008) or the reconstruction of the last universal common ancestor (Lazcano and Forterre, 1999; Koonin, 2003). A specialized DEG database (Database of Essential Genes) is developed to gather all published essential genes. It organizes the genes according to its kingdom and allows search using gene criteria or BLAST (Zhang et al., 2004; Zhang and Lin, 2009).

1.1.1. Comparative genomic approach

The comparative genomic is a bioinformatics approach to identify essential genes. It consists of selecting those genes that

* Corresponding author.
E-mail addresses: ahmed@dpt-info.u-strasbg.fr (A. Ahmed), g.frey@dpt-info.u-strasbg.fr (G. Frey), michel@dpt-info.u-strasbg.fr (C.J. Michel).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.02.006

are shared between distantly related organisms (Mushegian and Koonin, 1996). Evidently, genes are not identically shared between genomes. Hence, the term "shared genes" is redefined by orthologue or homologue genes (Fitch, 2000). This comparative approach was used when the first two bacterial genome sequences of *H. influenzae* and *M. genitalium* were completed (Fraser et al., 1995). This comparison revealed 241 direct orthologue genes but although this gene set cannot maintain a viable cell. For certain substantial essential molecular functions, different organisms do not use genes that are orthologue or even homologue. Such genes are not identified by orthologue comparative approach, rather they are detected by examining non-orthologue genes for missing essential functions. Such genes between genomes are called non-orthologue gene displacement NOGD (Koonin et al., 1996). By adding these NODG genes of *H. influenzae* and *M. genitalium*, this minimal gene set reaches 256 genes.

Although only 15 NODG genes were identified in this bacterial study, large genome comparisons shown that the NODG genes are associated with most essential genes including transcription, translation and replication (Koonin and Galperin, 2002). These genetic findings show that the concept of essential gene functions is more appropriate than that of essential genes.

1.1.2. Experimental approach

The experimental approach to identify essential genes existed before the comparative genomic one. It is simply based on gene-knockouts to determine lethal genes. The first experimental study generated 79 random gene-knockouts in the bacterial genome *B. subtilis* and identified 73 genes which did not kill the cell and six lethal genes (Itaya, 1995). It did not list the essential genes but only their proportion regarding to the total number of genes. This proportion (six essential genes out of 79 genes, i.e. about 8%) is very close to that obtained by comparative genomic approach (256 essential genes out of 4100 genes of the studied organism, i.e. about 6%).

Genetic methods used by this approach include transposon-insertion mutagenesis (Judson and Mekalanos, 2000), plasmid-insertion mutagenesis (Vagner et al., 1998) and gene inactivation using antisense RNAs (Ji et al., 2002).

The study of full wide genomes using experimental approaches showed many limitations. A first problem appeared because about half of gene-knockouts tend to be interrupted during experiences, and hence, a full list of essential genes is difficult to obtain. A second problem is the difficulty to identify essential cell functions performed by multiple proteins. Indeed, the disruption of one gene may not be lethal as gene redundancy exists, even if the function is essential to the cell. Another limitation of this experimental approach is that it does not consider the notion of gene evolution at all, in contrast to the comparative approach which is based on common orthologue genes.

The common circular code which was identified in large populations of prokaryotic and eukaryotic genes, is considered as an ancestral structural property of genes, and hence, as an essential property. It satisfies the principle of the comparative genomic approach. In this paper, we study the essential molecular functions using a comparative genomic approach based on this common circular code and its evolution. Hence, essential molecular functions will be analysed according to this circular code property.

In the next section, we recall the definition of the common circular code identified in genes of eukaryotes and prokaryotes and its main properties which will be used in the different methods developed.

1.2. The common circular code

1.2.1. Identification

In 1996, a simple occurrence study of the 64 trinucleotides $\mathbb{T} = \{AAA, \ldots, TTT\}$ in the three frames of genes showed that the trinucleotides are not uniformly distributed in these three frames. By excluding the four trinucleotides with identical nucleotides $\mathbb{T}_{id} = \{AAA, CCC, GGG, TTT\}$ and by assigning each trinucleotide to a preferential frame (frame of its highest occurrence frequency), the same three subsets X_0, X_1 and X_2 of 20 trinucleotides are found in the frames 0, 1 and 2, respectively, of two large and different gene populations (protein coding regions) of eukaryotes (26,757 sequences, 11,397,678 trinucleotides) and prokaryotes (13,686 sequences, 4,709,758 trinucleotides) (Arquès and Michel, 1996) (Table 1). Note: A 2010 statistical study of archaeal genes (150 genomes, 85,804 sequences, 70,411 kb) shows that this set X_0 is only partially retrieved in these genes (AAT is replaced by ATA, CAG by GCA, CTG by GCT, GGT by GTG and TTC by CTT). By convention, the reading frame established by a start codon {ATG, GTG, TTG} is the frame 0, and the frames 1 and 2 are the reading frame 0 shifted by 1 and 2 nucleotides in the 5′–3′ direction, respectively. These three trinucleotide subsets present several strong biomathematical properties, particularly the fact that they are circular codes.

Before to give a few basic definitions of circular codes, the principle and the biological importance of circular codes which could construct the coding sequences are described. In present-day genes, the principle of decoding of a DNA sequence is "not complicated": by knowing the beginning of a sequence (a start codon), a complex translation apparatus (ribosome with ARNs and proteins) allows a reliable reading of nucleotides three by three, each trinucleotide being a word coding an amino acid (except the stop codons). This efficient translation process has no frameshifting (except for the particular cases of frameshift genes). In contrast, reading a DNA sequence from a random initial position in the sequence is a "more complicated" problem. Indeed, the correct reading frame must be retrieved among the three potential frames. Is the randomly selected nucleotide the 1st, the 2nd or the 3rd nucleotide of a codon?

Circular codes are sets of words. Trinucleotide circular codes, i.e. circular codes which are subsets of the genetic code, have interesting synchronizing properties. Any DNA sequence composed of a concatenation of words of a circular code can be read (decoded) from any position randomly chosen in the sequence as well as automatically (no need for a translation apparatus). Indeed, circular words sufficiently long always synchronize (at least 13 nucleotides with the common circular code). In other words, the reading frame of the sequence can always be retrieved. In addition to this random position property, circular words have polymorphic patterns. Indeed, a maximal trinucleotide circular code (see Definition 3 and Remark 3 below) leads to 20^n synchronizing (circular) words of codon length $n > 4$ (≥ 13 nucleotides). Note: there also exist short synchronizing words of codon length $n \leq 4$ (not detailed).

Table 1
The common circular code X identified in both eukaryotic and prokaryotic genes: X_0, X_1 and X_2 are the preferential sets of 20 trinucleotides in frames 0, 1 and 2 of genes, respectively.

X_0	AAC AAT ACC ATC ATT CAG CTC CTG GAA GAC GAG GAT GCC GGC GGT GTA GTC GTT TAC TTC
X_1	ACA ATA CCA TCA TTA AGC TCC TGC AAG ACG AGG ATG CCG GCG GTG TAG TCG TTG ACT TCT
X_2	CAA TAA CAC CAT TAT GCA CCT GCT AGA CGA GGA TGA CGC CGG TGG AGT CGT TGT CTA CTT