Computational Biology and Chemistry xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Conserved patterns in bacterial genomes: A conundrum physically tailored by evolutionary tinkering

Ivan Junier[a,b,*]

[a] Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain
[b] Universitat Pompeu Fabra (UPF), Barcelona, Spain

ARTICLE INFO

Article history:
Accepted 11 July 2014
Available online xxx

Keywords:
Organization of bacterial genomes
Evolutionary genomics
Co-evolution of amino acids
Gene clustering
Chromosome structuring
Integrated functioning

ABSTRACT

The proper functioning of bacteria is encoded in their genome at multiple levels or scales, each of which is constrained by specific physical forces. At the smallest spatial scales, interatomic forces dictate the folding and function of proteins and nucleic acids. On longer length scales, stochastic forces emerging from the thermal jiggling of proteins and RNAs impose strong constraints on the organization of genes along chromosomes, more particularly in the context of the building of nucleoprotein complexes and the operational mode of regulatory agents. At the cellular level, transcription, replication and cell division activities generate forces that act on both the internal structure and cellular location of chromosomes. The overall result is a complex multi-scale organization of genomes that reflects the evolutionary tinkering of bacteria. The goal of this review is to highlight avenues for deciphering this complexity by focusing on patterns that are conserved among evolutionarily distant bacteria. To this end, I discuss three different organizational scales: the protein structures, the chromosomal organization of genes and the global structure of chromosomes.

"*What distinguishes a butterfly from a lion, a hen from a fly, or a worm from a whale is much less a difference in chemical constituents than in the organization and the distribution of these constituents*"
François Jacob, Evolution and tinkering, 1977

"*there cannot be any general law of evolution that accounts for increasing complexity at all levels.[...] The rules of the game differ at each level*"
Ibid.

1. Introduction

The breadth of bacterial evolution can be appreciated from simply considering diversity in genome size, which ranges across species from 500 to more than 10,000 genes (Lynch, 2007; Abby and Daubin, 2007; Koonin and Wolf, 2008; Rocha, 2008). It can also be observed in an individual bacterium by the diversity of genetic functions encoded. This diversity reflects both the random nature of mutations and genome modifications, as well as the degeneracy of possible evolutionary solutions. In contrast, conserved patterns reflect the irreversibility and convergence properties of evolution. These patterns can be seen both at the level of genomic content (conserved orthologous genes) and in the genomic organization of chromosomes, e.g. in groups of genes that have remained proximal along the chromosomes throughout billions years of evolution (Lathe et al., 2000).

The tension between pattern randomization and pattern conservation occurs at all physical scales of genome organization, from the internal structure of proteins to the cellular structure of chromosomes. In this regard, the conservation of a protein domain, gene, a group of genes, or global chromosomal structure, may result from different evolutionary constraints on fitness, meaning that conserved patterns may reflect very distinct aspects of evolution (Jacob, 1977). Added to this the absence of a one-to-one relationship between the scale of a pattern and the scale of the structure it can affect (Fig. 2), reductionism has remained a particularly challenging issue in biology.

In this review, I thus aim at highlighting avenues for deciphering the complex organization of genomes related to the integrated functioning of bacteria (Fig. 1). To this end, I discuss physical mechanisms that occur at different scales of cellular organization and provide insight into the evolutionary pressure associated with these mechanisms. Specifically, I present physical aspects of three types of conserved patterns. In the first section, I discuss the physics

* Correspondence to: Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.
E-mail address: ivan.junier@crg.eu

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.017
1476-9271/

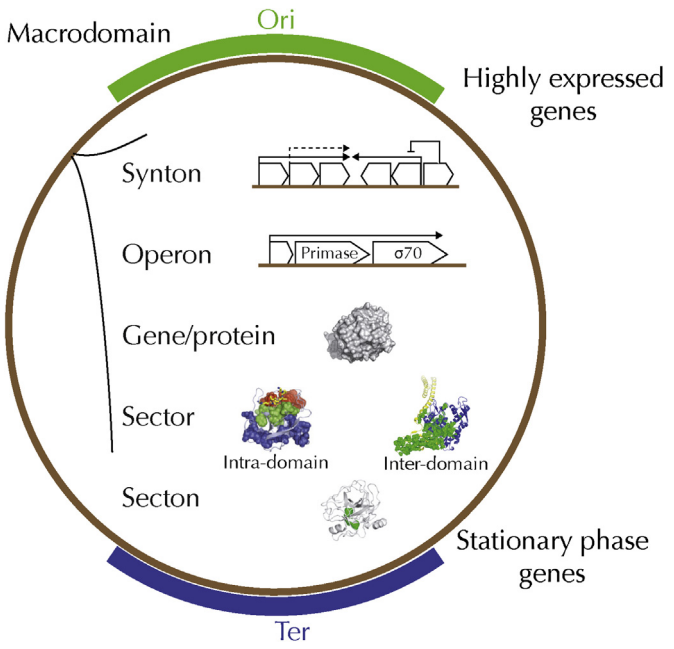

Fig. 1. Schematic representation of the multi-scale organization of conserved patterns in bacterial genomes. At the largest scale, the chromosomes of *E. coli* and *B. subtilis* are organized into macrodomains inside which genes make frequent contacts in space (Niki et al., 2000; Valens et al., 2004). Remarkably, these phylogenetically distant bacteria display similar macrodomains around the origin of replication (Ori, in green) and around the terminus of replication (Ter, in blue) – see Fig. 5 for a precise location of Ori and Ter in *E. coli*. Highly expressed genes are usually found close to the origin (Abby and Daubin, 2007; Rocha, 2008), whereas stationary phase responsive genes have a tendency, in Enterobacteria, to locate close to the terminus (Sobetzko et al., 2012). At a smaller scale, operons are fundamental transcriptional units containing several related genes that are transcribed into a single mRNA. Operons can also contain genes that have different functions. A striking example concerns the primase and the sigma factor σ^{70} that are found in the same operon across most bacteria. Remarkably, genes and operons have been shown to remained clustered together across various bacteria (Lathe et al., 2000). This is all the more remarkable that bacterial genomes are highly dynamic, with a neutral rate of 10^{-2} to 10^{-4} recombination events per genome per generation (Rocha, 2008). In this context, units of synteny, called syntons, can be defined for various degree of conservation (Junier and Rivoire, 2013) and may harbor complex regulatory relationships (Fischbach and Voigt, 2010). At small scales, genes are not the smallest functional units of genomes (besides non-coding RNAs). Indeed, single-domain proteins can be further partitioned into "sectors" on the basis of the co-evolution of their amino acids (indicated in blue, red, and green on the left protein). Sectors are networks of physically connected amino acids and correspond to independent functional features of proteins (Halabi et al., 2009). From an evolutionary perspective, they transcend the notion of protein domains, as shown for the two-domain Hsp70 molecular chaperone (Smock et al., 2010) – on the right protein, the sector in green extends through the two domains of Hsp70. Finally, small fundamental groups of co-evolving amino, named "sectons" (Rivoire, 2013), have been defined on the basis of the direct coupling between amino acids (Weigt et al., 2009). These refine sectors by identifying conserved sets of "indivisible" residues. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of protein structuring by focusing on the co-evolution of amino acids, both within and between genes. In the second section, I present several physical and fitness constraints associated with the tendency of genes to cluster in specific groups along chromosomes. In the last section, I describe the large-scale structuring of chromosomes in the context of the coordination of DNA replication, gene expression and cell division.

2. The structural basis of protein function: evolution of amino acid networks

Recent work shows that structural and functional properties of proteins can be inferred from the evolution of their sequences (Marks et al., 2012). At the most basic level, this evolution is characterized by an important heterogeneity, both for the average mutation rates between genes and for the mutation rate of amino acids within proteins. Understanding the origin of these heterogeneities requires consideration of the physical principles of protein folding, and perhaps less obviously, consideration of both the constraints imposed by gene and protein interactions (Pazos and Valencia, 2008) as well as the mechanisms by which the physicochemical properties of environments are encoded in genomes (Denamur and Matic, 2006). In this context, I first review a few aspects of gene evolution, focusing in particular on the relationship that exists between the evolution rate of a gene and its adaptive capacity. Following this line, I next present the relationships that have been highlighted between the structural organization of proteins and their adaptive capabilities. From a physical point of view, I discuss the possibility of deciphering the structural basis of protein functions by investigating the co-evolution of amino acids. I also discuss the implication of local concentration effects in the co-evolution of interacting proteins.

2.1. Heterogeneities of evolutionary rates: in search of causal relationships

The evolution rate of individual genes, i.e. the average rate of amino acid substitution, is negatively correlated with their expression level and, quite remarkably, barely to their essentiality (Rocha and Danchin, 2004; Charlesworth and Eyre-Walker, 2006) – see (Pál et al., 2006) for yeast. Correlations between evolutionary rates are thus expected for genes that are co-expressed, such as operons (Tenaillon et al., 2012), and for genes that belong to a common biological pathway, as has been shown explicitly for yeast species (Clark et al., 2012). In this regard, the high level of correlation between gene expression, genome organization, and biological pathways (Bork et al., 1998) hinders our ability to precisely determine the dominant constraints driving this heterogeneity.

A particularly informative example concerns amino acid substitutions that stem from physical collisions of the transcription and replication machineries (Mirkin and Mirkin, 2005). These collisions are more frequent for lagging-strand genes due to the head-on configuration of the machineries (Rocha and Danchin, 2003). In this context, Paul et al. (2013) have recently shown, in *Bacillus subtilis*, that for genes with an equal rate of neutral mutation (as measured by the rate of synonymous mutation), amino acid substitutions in core genes, i.e. in ~800 genes that are conserved in very divergent strains of *B. subtilis* (Paul et al., 2013), are more frequent along the lagging strand. In parallel, they have shown that the few core genes on the lagging strand are mostly stress-responsive. As proposed by the authors, this suggests that *B. subtilis* exploits the high mutation rate of the lagging strand in order to generate, at the population level, a broad spectrum of responses to environmental changes.

In light of these results, it is tempting to hypothesize that a main factor of mutation rate heterogeneity between genes is the extent to which genes participate in adaptation. Although it may not capture the whole story, this hypothesis has the advantage of being experimentally testable. For instance, it could be done by studying the adaptive trajectories of individual genes within a population during the shift to adaptive conditions (see e.g. the morbidostat experiments described by Toprak et al. which quantitatively follow the evolutionary trajectories of genes within a population of bacteria as it develops drug resistance, Toprak et al., 2013). Note also that the routes of adaptation may be as diverse as the perturbation possibilities (Touchon et al., 2009). An interesting case related to that matter concerns the mutational response to antibiotics that deviate metabolic fluxes toward the production of protective pathways (Lee et al., 2010). Indeed, it has been shown that the response in *Escherichia coli* strongly depends on the operational mode of the antibiotics (Toprak et al., 2012). Specifically, ribosome-targeting

Please cite this article in press as: Junier, I., Conserved patterns in bacterial genomes: A conundrum physically tailored by evolutionary tinkering. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.08.017