The selective force driving metabolic operon assembly

Marco Fondi[1], Francesco Pini[2], Christopher Riccardi[1], Pietro Gemo[3], and Matteo Brilli[3]

[1]Department of Biology, University of Florence, Italy
[2]Department of Biology, University of Bari, Italy
[3]Department of Biosciences, University of Milan, Italy

September 5, 2022

Abstract

The origin and evolution of operons have puzzled evolutionary biologists since their discovery. To date, many theories have been proposed to explain their evolution, among which reduced recombination rate within clustered genes, co-expression, simultaneous horizontal transfer and transcription/translation coupling. Most focus on the possible advantages provided by an already structured operon, while they all fall short in explaining the accretion of scattered genes into gene clusters and then operon. Here we argue that the way in which DNA replication and cell division are coupled in microbial species is a key feature in determining the clustering of genes on their chromosomes. More specifically, we start from the observation that bacterial species can accumulate several active replication forks by a partial decoupling of DNA replication and cytokinesis, which can lead to differences in copy numbers of genes that are found at distant loci on the same chromosome arm. We provide theoretical considerations suggesting that when genes belonging to the same metabolic process are far away on the chromosome, changes in the number of active replication forks result in perturbations to metabolic homeostasis, thereby introducing a selective force that promotes gene clustering. By deriving a formalization of the effect of active DNA replication on metabolic homeostasis based on Metabolic Control Analysis, we show that the above situation provides a selective force that can drive functionally related genes at nearby loci in evolution, which we interpret as the fundamental pre-requisite for operon formation. Finally, we confirmed that, in present-day genomes, this force is significantly stronger in those species where the average number of active replication forks is larger.

Introduction

Operons [17] are one of the hallmarks of prokaryotic gene regulation; in their most basic form they comprise a single promoter at the $5'$ end, followed by two or more

genes and a transcription terminator at the $3'$ end, therefore they are transcribed into polycistronic messenger RNAs. Additional regulatory sites (alternative and/or internal promoters, attenuators etc.) can be present to provide fine control over the gene expression levels [24, 45]; genes in an operon often participate to a same functional process. While common in Prokaryotes, they are exceptions [3, 27, 4] or peculiarities [5] in Eukaryotes, a fact that was explained by the smaller effective population sizes of many eukaryotes that - according to the *drift model*, would lead to operon disruption [29]. The evolution of operons is debated since their discovery; early ideas were often related to the origin of metabolic pathways, with a coupling in their evolutionary history. However, while metabolic pathways are often ancient, taxonomic variability in operon organization of the genes implementing them suggests a more recent evolutionary history e.g [12]. More plausible and general hypotheses focus on the advantages provided by operons; the Fisher model suggests that their compactness may reduce the chances of recombination events within [41] a recently revived idea [11]. The co-regulation model is intrinsically linked with the operon rationale: genes stay together to facilitate their coordinated expression [35], but others have shown that the formation of operons for the purposes of co-regulation is both unnecessary and implausible [26, 25] because independent promoters can evolve characteristics enabling co-expression of genes encoded at different loci. The *selfish operon model* [26] focuses on operons as easily mobilizable functional units, but it does not explain how an operon formed in the beginning, and contrasting evidences have been reported [32]. Another possible explanation for the evolution of operons is that the coupling of transcription and translation in Prokaryotes makes so that products are released in a relatively small volume of the cytoplasm, maximizing interactions and metabolic fluxes [10]. However, not all proteins encoded by the same operon interact, channeling of metabolic intermediates is not so widespread, and recent estimates, withstanding the reduced mobility of proteins in the cell with respect to pure water, suggest that two particles can find each other in the cell in a matter of seconds [37]. By using stochastic simulations of simple biochemical systems, evidences were found for noise reduction in the abundance of proteins encoded by operons, however this was limited to some type of interaction established by the products [36]. It should be added that the two hypotheses mentioned above strongly rely on the coupling of transcription and translation that might not be as general as previously assumed [16]. Genome size and/or its proxies were also suggested to be markers of the intensity of natural selection for operon organization, under the idea that since larger genomes have more complex genetic networks thanks to the presence of more transcription factors, they would endure weaker selection for operons than smaller genomes, where regulation alternatives are scarce [30].

The process of operon evolution can be split into two aspects: one is *operon assembly*, by which scattered genes become closely spaced on the genome enabling operon formation. The second is *operon maintenance* in evolution, which depends on the fitness differential provided by the operon with respect to the same genome with scattered genes. Since what makes an operon advantageous *once* it is formed, is not necessarily what drove the genes at nearby loci, *assembly* and *maintenance* may be consequences of very different forces. Most of the hypotheses presented above focus on *maintenance* and do not shed light about the selective forces for the *assembly* of operons. For instance, an operon may reduce noise in protein abundances as previously suggested [36],

but this advantage can be selected for by evolution only *after* the operon has formed, not during the intermediate steps. Since the probability that an operon originates from scattered genes in a single or only a few intermediate steps is extremely low, we postulate the existence of a selective pressure able to provide a drive to operon *assembly* in evolutionary time. We will show that this pressure is naturally active with varying intensity in many prokaryotic species. More specifically, we argue that the evolution of metabolic operons is related to active DNA replication through the effect it has on metabolic homeostasis.

Bacterial species can have a more or less strict coupling of DNA replication with cell division. Species like *Caulobacter crescentus* and *Staphylococcus aureus* lies at one extreme, because they implement a strict genetic program that determines cell division just after the chromosome has been replicated once [7, 34]. These species can contain one or two chromosomes at most. At the opposite extreme there are species like *E. coli* and *B. subtilis*, that normally replicate their chromosomes in excess with respect to the rate of cell division [8, 14] and can therefore contain several genome equivalents (up to 16 in *E. coli*) and active replication forks (Fig. 1a and b). The average number of active replication forks in the population is not fixed, and while an increased number of replication forks have been related to shorter division times in certain organisms [42], the relationship seems to fail significance when distant taxonomic groups are analyzed together [28]. This feature can be studied using genome sequencing data, and calculating the so-called ori/ter ratio, corresponding to the ratio of coverage around the *ori* (m_{ori} in Fig.1a,b) and *ter* loci. Whatever relationship the ori/ter ratio has in different taxa with respect to growth rate, it is a matter of fact that as a consequence, genes experience different copy numbers depending on their distance from the origin/terminus; a proof of this is the well-known influence on transcript abundance [9, 18]. Recent works indeed highlighted the well-known importance of the position of a locus on the ori/ter axis for modulating gene expression - summarized by interpreting this axis as a regulatory system itself [22].

Here, we provide theoretical considerations about the effect of active DNA replication on metabolic homeostasis, and show that this provides a selective force that can drive functionally related genes at nearby loci in evolution, which we interpret as the fundamental pre-requisite for operon formation.

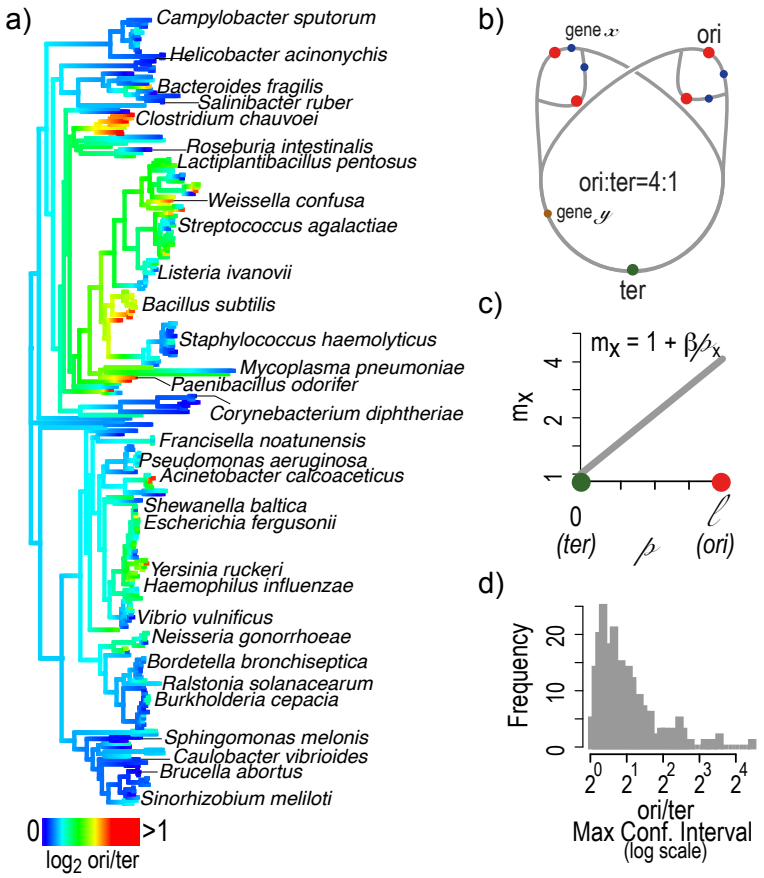

Figure 1: a) Mapping $log_2 ori/ter$ on a phylogenetic tree of Bacteria from many taxonomic groups. Only a few species are indicated for clarity and $log_2 ori/ter$ is truncated at 1 to highlight differences, whereas $(ori/ter)_{max} \approx 10$). b) Representation of a circular chromosome with three active replication forks, which translates in $m_{ori} = 4$. Genes x and y, are in a $4:1$ ratio but this can change since the number of replication forks is not constant; an *E. coli* cell can for instance be born with a single chromosome ($ori/ter = 1$) and after some time it can contain several genome equivalents and active replication forks. c) Explanation of the meaning of the parameter β used in the text: given the ori/ter ratio calculated from genome sequencing data and the length of one chromosome arm (i.e. distance of the origin from the terminus, ℓ), β is the slope of the line starting at $x = 0$, $y = 1$ (ter locus) and ending at $x = \ell$, $y = 4$ (ori locus), and enables to calculate the multiplicity m_x for all genes, given their position. d) Distribution of the maximum bound of the confidence interval estimated for the $log_2 ori/ter$ ratios available for each species.

Results

Metabolic consequences of chromosome replication

In this section we provide the theoretical foundations of our hypothesis by linking together chromosome replication and its possible effects on cellular homeostasis. We do this by integrating, for the first time, a widely adopted metric related to the average number of replication forks in cells from a population (hereinafter the *ori/ter ratio*, also known as Peak-to-Trough ratio [28]) with Metabolic Control Theory (MCT), an established theoretical framework focused on understanding the control of metabolic fluxes. If the ori/ter ratio was constant (Fig.1b), the copy number of two genes x and y located at distant loci on the genome would also be constant, but since the number of replication forks changes in time, those genes would not only experience different multiplicities in time but also varying relative abundance. For instance, in Fig.1b with three replication forks, x and y are in a 4 to 1 ratio, but with one only replication fork, they would be 2 to 1. Since the multiplicity of genes has an effect on the abundance of the products they code for, this equates to changing the relative expression level of genes, especially when the genes are separated by a large distance. The metabolic consequences of changes in the abundance of enzymes belonging to the same metabolic pathway is one of the fundamental results of MCT (Eq. 1) [38] providing an approximated relationship to the change in steady state flux of a linear pathway embedded in the metabolic network:

$$\frac{J^{r_i}}{J^0} = \frac{1}{1 - \sum_{i=j}^{m} C_{E_i}^{J^0} \frac{r_i - 1}{r_i}}\ , \tag{1}$$

Js are steady state fluxes for the reference (^{0}) or the new steady state, induced by changing the abundance of $m - j$ enzymes in a pathway by different factors r_i. $C_{E_i}^{J^0}$ is the *flux control coefficient* (FCC) [19] of enzyme E_i on the pathway, defined in Eq 2.

$$C_{E_i}^{J} = \frac{\partial J}{\partial E_i}\frac{E_i}{J} = \frac{\partial lnJ}{\partial lnE_i}. \tag{2}$$

In practice, the FCC of enzyme i tells us about the fractional change in pathway flux elicited by a fractional change in enzyme abundance. FCCs are *systemic* quantities that can be measured only in the intact system, meaning their exact values are often unknown. Nevertheless, the so-called *summation theorem* [19]:

$$\sum_{i=1}^{n} C_{E_i}^{J} = 1, \tag{3}$$

constrains the possible values, at the same time introducing the fundamental concept that fluxes are not usually modulated by key *bottleneck* or *rate limiting* enzymes; instead, control is shared by many enzymes [19]. A consequence of Eq. 3, confirmed by *in vivo* measurements since early times is that the flux control coefficient of an enzyme with respect to a pathway is small on average [20, 38, 39]. Eq. 1 therefore shows that

when all (i.e. the Summation theorem holds) enzymes in a pathway are changed by the same factor $r_i = r$ the system relaxes to a new steady state where the flux is scaled by r and metabolite concentrations stay constant (perfect homeostasis). When the r_is are different, the ensuing change in flux instead depends both on the the FCCs and the r_is. In this case however, enzyme rates are not scaled proportionally, and therefore metabolite pools can also move to a new steady state. Eq. 8 shows the quantitative treatment for this situation; the relationship can be summarized by the observation that even enzymes with very low FCC on a certain pathway flux, can cause large perturbations to metabolite concentrations, if not modulated coordinately to the other enzymes [13].

Even with some limitation imposed by the FCCs (therefore by the system) cells normally manipulate fluxes as a function of the demand while keeping metabolite concentrations in-between acceptable limits. This can be achieved by using the same regulator to control the rate at which functionally related genes are transcribed - as for yeast's amino acid biosynthetic genes, that are controlled by *GCN4* [1]; alternatively a similar achievement results from organizing those genes in an operon. However, Prokaryotes have one major difference with respect to most Eukaryotes, and we are going to show that for this reason, the former strategy is much less advantageous for them. To better illustrate this, we put together the change in multiplicity of genomic loci, the position of genes along the ori/ter axis and the number of active replication forks into Eq. 1. We refer to a situation where there is one only *ter* locus and a certain number of *ori* loci that were previously replicated (Fig.1b). In this view, the multiplicity of the gene for enzyme i is:

$$m_i = 1 + \beta p_i, \tag{4}$$

where p_i is the distance of the gene from the terminus, β is obtained from sequencing data and corresponds to the slope of the multiplicity change along the chromosome in the population that underwent sequencing. Basically, by calculating an average coverage of *ori* and *ter* proximal regions using genome sequencing data, one can calculate the ori/ter ratio and, if ℓ is the ori-ter distance (Fig. 1c):

$$\beta = \frac{ori/ter - 1}{\ell}. \tag{5}$$

Eq. 4 assigns multiplicity $m_i = 1$ to loci at the terminus (having $p_i = 0$) and $m_i > 1$ elsewhere, with $m_{max} = 1 + \beta\ell$, at the origin. As the multiplicity of a gene in every single cell is an integer, we are here averaging the situation over the many cells of the population. Considerations made here remain however valid, as each single cell will experience different ori/ter ratios during its existence.

By considering the multiplicity of a gene as directly related to the abundance of its product in the cell, Eq.1 can now be written as:

$$\frac{J^\beta}{J^0} = \frac{1}{1 - \sum_{i=j}^{m} C_{E_i}^{J^0} \frac{\beta p_i}{1 + \beta p_i}}, \tag{6}$$

As expected, $\beta = 0$ corresponds to the situation where all genes have the same multiplicity in time, and correctly predicts no change in flux. On the converse, if $\beta > 0$, and genes are distant, J^β/J^0 can fluctuate around 1 and since the multiplicity of genes is unequal, metabolite pools will be also affected.

The *operon* case is when the p_is of all genes with control over a certain pathway flux (such that Eq. 3 holds) are very similar, therefore their abundances change by the same amount when β changes (given by Eq.4), and the flux undergoes scaling by the same factor:

$$\frac{J^\beta}{J^0} = \frac{1}{1 - \dfrac{\beta p}{1+\beta p}} = \frac{1}{\dfrac{1+\beta p - \beta c}{1+\beta p}} = 1 + \beta p \ , \tag{7}$$

This effect on metabolites can be quantified by a relationship similar to Eq. 1 that focuses on the deviation of metabolite concentrations when the abundance of one enzyme in a pathway is scaled by r [39]:

$$\frac{S^r}{S^0} = \frac{1 - (C_E^J - C_E^S)\frac{r-1}{r}}{1 - \frac{r-1}{r} C_E^J} \tag{8}$$

where C_E^S is the *concentration control coefficient* (CCC) of the enzyme over metabolite S, defined similarly to the FCCs [19] and:

$$\sum_{i=1}^{n} C_{E_i}^S = 0, \tag{9}$$

is the summation theorem for the CCC [15].We point out that when $C_{E_i}^J \approx 0$ (i.e. the enzyme has negligible control on the flux) Eq. 8 gives [13]:

$$\frac{S^r}{S^0} = 1 + \frac{r-1}{r} C_E^S, \tag{10}$$

which shows - together with the fact that Eq. 9 does not limit the absolute value of the coefficients - that enzymes with negligible control over the flux of a pathway can perturb metabolite concentrations by an arbitrarily large factor when changed in isolation. Since significant reduction or increase of metabolite concentrations in the cell can break down cellular homeostasis and that this can also be caused by functionally related genes being positioned at large distances on actively replicated chromosomes, we hypothesize that a possible solution worked out by evolution in this scenario could be the construction of gene clusters and therefore operons.

Gene proximity minimizes variations in metabolite homeostasis

Let us now introduce a toy pathway to discuss more thoroughly the points raised above:

$$X_{in} \overset{E_1}{\rightleftharpoons} S \overset{E_2}{\rightleftharpoons} X_{out}, \tag{11}$$

where X_{in} and X_{out} are external metabolites, and E_1, E_2 two enzymes coded for by genes located at distances p_1 and p_2 from the terminus. Using mass action we write this simple system as:

$$\frac{dS}{dt} = E_1 k_1 X_{in} - E_2 k_2 S, \tag{12}$$

which can be solved analytically at the steady state (when $dS/dt = 0$):

$$S_{ss} = \frac{E_1}{E_2}\frac{k_1}{k_2} X_{in}. \tag{13}$$

Enzyme abundances (E_is) can be replaced by the relation introduced in Eq. 4 and by derivation we can calculate the scaled sensitivity of the concentration of metabolite S with respect to β i.e. how changes in the number of active replication forks affect metabolite concentration in this system:

$$\frac{\partial S}{\partial \beta}\frac{\beta}{S} = \frac{k_1}{k_2}\frac{X_{in}}{S}\frac{p_1 - p_2}{(1+\beta p_2)^2}\beta, \tag{14}$$

Limit cases are: (i) if $p_1 = p_2$, genes are at the same locus, then varying β has no effect on metabolite concentration since all genes change of the same quantity; (ii) if $p_1 > p_2$, when β increases, the first enzyme changes more than the second, $\sigma_\beta > 0$ and therefore S will increase, and (iii) if $p_1 < p_2$, that is when E_1 abundance grows less than E_2's if β changes, then $\sigma_\beta < 0$ indicating that in this case metabolite S gets depleted when β increases. This provides additional theoretical basis for an effect of changes in the number of replication forks active in the genome (here a change in β) on homeostasis, when genes participating to the same process are scattered over the chromosome. It is therefore plausible that during evolution, events leading to the minimization of those perturbations could contribute positively to the fitness of a cell and therefore increase its probability of fixation in the population. One of those mechanism, as the above considerations suggest, is to group functionally related genes in a compact locus. To better show this idea, we refer to our toy model, and by scanning many positions for our two genes in a virtual $2Mbp$ linear chromosome, we calculate a measure of the ensuing variation in S. Figure 2a therefore confirms our theoretical prediction that when β changes over a certain interval metabolite homeostasis is maintained if the genes are kept close on the genome, and also shows that the induced perturbation increases when genes are farther.

Evolving operons *in silico*

The study of a very simple system of two only genes provided support to the idea that in species where the number of active replication forks is partially decoupled from cell division, the operon might be especially important for metabolic homeostasis.

To further test our hypothesis we introduce a more realistic metabolic model that we previously developed [2], comprising reactions collectively annotated as *carbon metabolism* in *E. coli* i.e. Glycolysis, Pentose-phosphate pathway, Krebs cycle, Glyoxylate shunt and Acetate excretion/import. The model (Figure 2c) has 34 enzymatic

reactions, and 26 variable metabolites. It is encoded as a linear approximation called *linlog*, that was introduced by [43, 44] as an alternative to classical *Michaelis-Menten*-like rate functions; these are non linear, complicating parameter identification from the data; additionally, systems of even a few non-linear equations are usually not solvable analytically. In the linlog, each rate is modeled as a linear combination of logarithms of normalized concentrations, providing an analytical solution for the steady state (Eq. 16); the original parameters corresponds to the elasticity coefficients from Metabolic Control Theory. Normalization of metabolites and fluxes is performed by their reference value in the original formulation, but here we include the reference level into the parameters to work on absolute quantities. Using this approximation, rates are modeled as:

$$\boldsymbol{v} = \boldsymbol{E}\left(\boldsymbol{A} + \boldsymbol{B} log \boldsymbol{x}\right), \tag{15}$$

where $\boldsymbol{E}_{r\times r}$ is a diagonal matrix of enzyme abundances (modeled as in Eq. 4), $\boldsymbol{A}_{r\times 1}$ and $\boldsymbol{B}_{r\times m}$ are matrices of parameters derived from elasticities and reference steady state, and $\boldsymbol{x}_{m\times 1}$ are concentration of metabolites. Given the stoichiometry matrix $\boldsymbol{N}_{m\times r}$, we can solve analytically the system for the steady state:

$$log \boldsymbol{x} = -\left(\boldsymbol{NEB}\right)^{-1}\boldsymbol{NEA}. \tag{16}$$

and

$$\boldsymbol{J} = \boldsymbol{E}\left(\boldsymbol{A} - \boldsymbol{B}\left(\boldsymbol{NEB}\right)^{-1}\boldsymbol{NEA}\right). \tag{17}$$

The model is not parameterized with respect to experimental data, but has biologically meaningful parameters; as a consequence it is here used as a benchmark to check our hypothesis in a more realistic situation. Since matrix $\boldsymbol{E}$ integrates the abundance of enzymes, we can simulate how the position of genes evolves in time when metabolic homeostasis is fixed as an evolutionary priority. Deviations from homeostasis can be summarized by the variation in metabolite concentrations when the β changes. In brief, we simulate a situation with a varying average number of replication forks per cell, a chromosome containing the genes encoding the enzymes of the metabolic model and we ask how metabolite pools are affected when we change gene order:

$$\mathcal{F} = \frac{1}{m \times b}\sum_{i=1}^{m}\sum_{j=1}^{b}\frac{\sigma_{ij}}{\mu_{ij}}. \tag{18}$$

Here μ_{ij} and σ_{ij} are the mean and the standard deviation of the logarithm of the concentration of metabolite i for β_j and $\mathcal{F}$ is therefore the average coefficient of variation of metabolite concentrations in the system when β is changed. To simulate evolution, we use an optimization tool in R (nlminb) asking for the combination of gene positions minimizing the objective function. In Fig. 2, we show the outcome of 100 optimization runs, with several genes consistently forming compact clusters at convergence (same color) as expected by our theoretical considerations. Clusters formed at convergence of different optimizations are not always exactly the same, with some of the genes ending in different or no cluster, suggesting the existence of alternative, similarly optimal solutions, or a possible effect from the starting conditions of some genes.

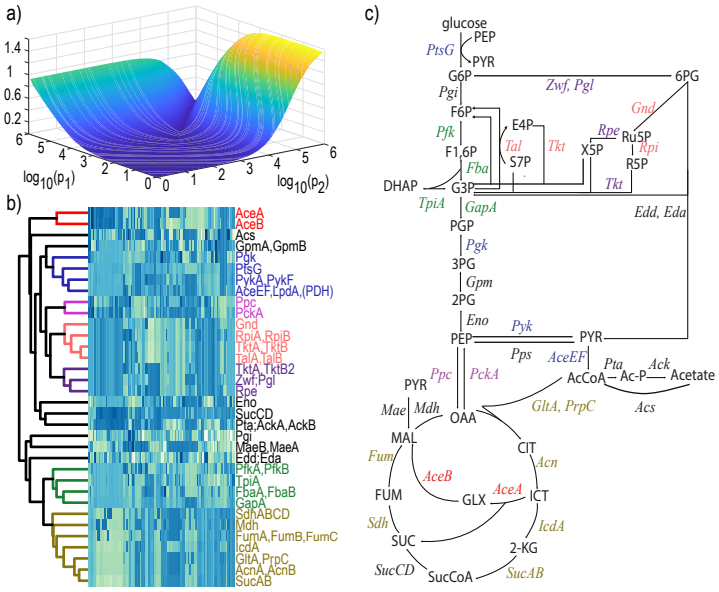

Figure 2: a) Scanning of the genomic positions for the two genes in the toy model (p_1 and p_2, expressed as a distance from the terminus). The z-axis is the coefficient of variation (standard deviation / mean) of metabolite S at steady state for 20 values of β sampled uniformly in the range $[10^{-6}, 10^{-4}]$. When the two genes are close on the genome (along the diagonal in the plot), perturbations to the homeostasis of S consequent to the changes in multiplicity of the genes induced by the change in β are strongly reduced. b) The results of 100 independent optimizations (columns) as described in the text and started from random gene positions. The chromosome is partitioned into 50 windows, and color goes from light for terminus proximal loci, to dark for origin proximal ones. Rows correspond to enzymes in the model, and the heatmap colormap indicates their position at convergence of the optimization (similar color in the same column means the two genes are close on the chromosome). The dendrogram shows the existence of groups of genes with very similar profiles across the simulations, suggesting that gene organization is not evolving randomly but towards the formation of similar gene clusters (operons and überoperons [23]) in different optimizations when the target is the minimization of metabolite deviations from homeostasis. Gene names are provided for reference to the well-known metabolic network shown in c).

The above considerations would suggest that the selective pressure driving the formation of compact gene clusters and operons is a function of the ori/ter ratio, therefore species might experience different levels of selection towards the formation of operons, depending on how they regulate genome replication.

Degree of compaction of functionally related genes correlates to the ori/ter ratio

One prediction of this hypothesis is that species with a larger ori/ter ratio would experience a more effective selective pressure for clustering functionally related genes. Two major factors might confound this signal: horizontal gene transfer and the evolutionary rate of the ori/ter ratio itself. Metabolic operons are easily transferable and self-contained modules optimized for fixation in the host and this should be true even in species that are not experiencing significant pressure toward operon formation. Additionally, when we measure the ori/ter ratio, we record a snapshot of a present-day organism, and even by ignoring how fast the average ori/ter ratio can change in evolution, we expect it to be much faster than operon formation. Therefore, while the degree of gene clustering in a genome reflects an old history of selection, the ori/ter ratio might have changed recently. With these drawbacks in mind, we decided to test our hypothesis by deriving a specific *Proximity Score* (PS) that summarizes the degree of compaction of functionally related metabolic genes. For each organism and functional category in KEGG , we retrieved the genes, sorted them by position and recorded the distance separating consecutive ones. After processing of all pathways we calculated the 20th percentile; our PS is the logarithm base 2 of the inverse of this number, such that larger values correspond to more compact configurations. The PS was then compared to the ori/ter ratio by using both linear models and t.tests. Fig. 3a shows the existence of possible covariation of the two traits across the phylogenetic tree in a qualitative way; Fig. 3b confirms the presence of a significant relationship among the PS and the $log_2 ori/ter$ ratio, when using linear regression models. Regression coefficients are significant when considering all organisms, or the Proteobacteria, but not the Firmicutes. When significant, the model explain around 10% of the total variance in the data, which is a consistent fraction if we think about the many additional forces that act on genome organization on the short evolutionary time. In Fig.3c we strengthen this idea by showing that species with an ori/ter ratio significantly larger than 2 (at $p = 0.01$) also have significantly larger PS. Since the Firmicutes have PS and ori/ter ratios that are significantly larger than in the Proteobacteria, we additionally show in Fig.3d that the difference is significant even if we limit the test to Proteobacteria.

The situation of the Firmicutes may seem in contrast with our hypothesis. However, when comparing their PS and ori/ter ratio to the Proteobacteria, we found they are both significantly larger ($p < 2.2e-16$ for PS and $p < 0.00038$ for ori/ter). This suggests that the Firmicutes likely have a tendency toward high ori/ter ratio since their common ancestor, which may give time for a thorough optimization of gene organization. Indeed the existence of a linear relationship of PS with $log_2 ori/ter$ is a strong assumption, as it would predict that at very large ori/ter ratio the distance of genes would reduce to nothing, which is clearly impossible. We therefore speculate that a better relationship would be a saturable function with its asymptote at around the average size of a gene i.e. 1000nt. Once this is reached, there's no way to additionally improve the situation by getting genes closer, which ends up in breaking the correlation of ori/ter and PS i.e. being both close to optimal, they randomly fluctuate around this optimum and therefore there's no more correlation.

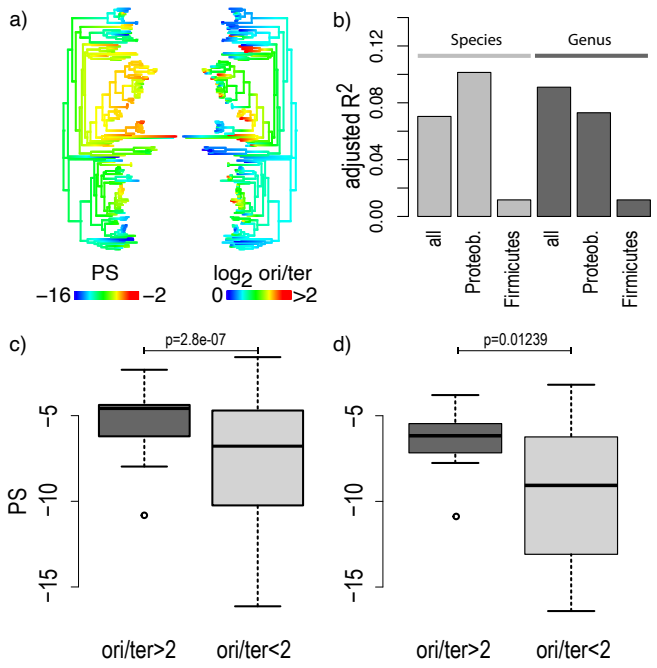

Figure 3: a) Face to face phylogenetic trees with mapped ancestral reconstructions of Proximity Score (PS) and $log_2 ori/ter$. b) Barplot of the adjusted R^2 for linear regression models using PS as a predictor of $log_2 ori/ter$. All regression coefficients for PS are significant except for Firmicutes (data not shown). *Species* correspond to models exploiting all the organisms in the dataset and *Genus* means that models are build on averaged values across species belonging to the same genus. We also built models for Proteobacteria ($N = 46$) and Firmicutes ($N = 29$) as they cumulatively represent 86% of the available genomes; c) boxplot showing that species having $p \leq 0.01$ when testing for $ori/ter \geq 2$ tend to have significantly larger PS. Firmicutes have significantly larger PS ($p < 2.2e-16$) and ori/ter ratios ($p = 0.0003766$) with respect to the Proteobacteria, and this may affect the above pvalue, but d) which is limited to Proteobacteria, shows that the test is still significant without the Firmicutes.

Conclusions

In the last decades, a number of hypotheses have been formulated concerning the origin and evolution of operons. However, most of them only focus on the molecular mechanisms that might affect the evolutionary persistence of operons once they are formed, not the intermediate steps during their formation. Indeed, selective pressures that might favor the intermediate steps in operon evolution, and the potential gain in fitness provided by the very same operon once it is formed, might even be two different things. In this paper we provide evidence that compact metabolic gene clusters may evolve to face homeostatic perturbations introduced by DNA replication. By integrat-

ing Metabolic Control Theory with a simple model of gene multiplicity during replication we provide the theoretical basis for our hypothesis, which turned out to be strongly supported both by simulations with a realistic metabolic model and by comparative genomics analysis. Introducing the ori/ter bias as a controller of enzyme abundances, the structure of a *virtual* genome spontaneously evolved clusters of functionally related genes *in silico* when the maintenance of homeostasis was considered as an evolutionary objective. Strong support for our hypothesis also comes from genomic analysis: by deriving a measure of the degree of compaction of functionally related genes for many species, we were able to highlight the predicted association with the ori/ter bias. We are fully aware that our hypothesis has several important repercussions on the way we conceive metabolic operon evolution. Not only because all tests performed seems to corroborate that the selective pressure toward gene clustering would be a consequence of better homeostatic control, but also because this would suggest that species experience selective pressures whose intensity is a function of the number of replication forks (summarized here by the ori/ter ratio, or the β). The ori/ter ratio might therefore create a continuous range of selective pressures in different species with those who are markedly *operon formers*, since they have a strong ori/ter bias, and species that are instead not experiencing the very same pressures for operon formation as they don't replicate the genome in parallel. Nonetheless, non-operon-formers species might have significant advantages obtaining operons by horizontal gene transfer, even if they don't have pressures for the intermediate steps in their construction. Additionally, not all the cellular processes might benefit from the clustering of their genes into operons. Indeed, in this work, we have provided compelling evidences that this likely the case for metabolic operons but other pathways involved in different cellular processes (e.g. signalling, transcription/translation, etc.) might behave differently.

DNA replication is since long known to have had a significant role in genome evolution, but the provided mechanistic explanations were always case-specific. In this work we show that DNA replication could provide a selective force able to select intermediate steps during metabolic operon formation - when fixing a simple and plausible evolutionary objective, represented by maintenance of homeostasis.

Methods

ori/ter calculation and optimization-based simulations

Identification of origin and terminus for each genome was done with function *oriloc* from package seqinr [6]. Coverage around the two loci was obtained by first extracting a 50kb region centered on ori (ter) locus, and then mapping publicly available genome sequencing data from SRA using Salmon [33] in mapping mode. Only species with at least 5 sequencing libraries available were considered. Coverage was not normalized because we focus on the ratio at the two loci. The R function t.test was used to test for log2(*ori/ter*) larger than 0 (*ori/ter*>1) or larger than 1 (*ori/te*r>2), and also provided 95% confidence intervals. The optimization-based simulation was carried out using the R function *nlminb*. The two gene-system was implemented using deSolve [40].

Gene clustering analysis and proximity score (PS) calculation

KEGG, the Kyoto Encyclopedia of Genes and Genomes [31], provides a collection of *manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks* for a number or biologically relevant areas of research, such as Metabolism or Cellular Processes. Each KEGG pathway also contains manually defined functionally tight gene sets of different types and scopes. Metabolic modules are assigned to the *Pathway module* category, on which we focus our attention in this work. For every species under analysis (181), we obtained all gene-to-gene distances within each Pathway module and then calculated the first quartile after processing all modules. This number represents the gene-to-gene distance such that 25% of all distances are smaller. The logarithm base 2 of the reciprocal of this number is the Proximity Score (PS) of a species that we contrast to the ori/ter ratio.

References

[1] Gerd Albrecht, Hans-ulrich Mo, Bernd Hoffmann, Ueli Reusser, and Gerhard H Braus. Monitoring the Gcn4 Protein-mediated Response in the Yeast Saccharomyces cerevisiae. *The Journal of Biological Chemistry*, 273(21):12696–12702, 1998.

[2] S. Berthoumieux, Matteo Brilli, H. de Jong, Daniel Kahn, and E. Cinquemani. Identification of metabolic network models from incomplete high-throughput datasets. *Bioinformatics*, 27(13), 2011.

[3] Thomas Blumenthal, Donald Evans, Christopher D. Link, Alessandro Guffanti, Daniel Lawson, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Lu Chiu, Kyle Duke, Moni Kiraly, and Stuart K. Kim. A global analysis of Caenorhabditis elegans operons. *Nature*, 417(6891):851–854, 2002.

[4] Svetlana Boycheva, Laurent Daviet, Jean-Luc Wolfender, and Teresa B Fitzpatrick. The rise of operon-like gene clusters in plants. *Trends in Plant Science*, 19(7):447–459, 2014.

[5] Giovanni Bussotti, Laura Piel, Pascale Pescher, Malgorzata A. Domagalska, K. Shanmugha Rajan, Smadar Cohen-Chalamish, Tirza Doniger, Disha-Gajanan Hiregange, Peter J. Myler, Ron Unger, Shulamit Michaeli, and Gerald F. Späth. Genome instability drives epistatic adaptation in the human pathogen Leishmania. *Proceedings of the National Academy of Sciences*, 118(51):https://doi.org/10.1101/2021.06.15.448517, 2021.

[6] D. Charif and J.R. Lobry. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York, 2007. ISBN : 978-3-540-35305-8.

[7] Justine Collier. Regulation of chromosomal replication in Caulobacter crescentus. *PLASMID*, 67(2):76–87, 2012.

[8] Stephen Cooper and Charles E Helmstetter. Chromosome replication and the division cycle of Escherichia coli Br. *Journal of Molecular Biology*, 31(3):519–540, 1968.

[9] Etienne Couturier and Eduardo P.C. Rocha. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Molecular Microbiology*, 59(5):1506–1518, 2006.

[10] Thomas Dandekar, Berend Snel, Martijn Huynen, and Peer Bork. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324–328, 1998.

[11] Gang Fang, Eduardo P.C. Rocha, and Antoine Danchin. Persistence drives gene clustering in bacterial genomes. *BMC Genomics*, 9:4, 2008.

[12] Renato Fani, Matteo Brilli, and Pietro Liò. The origin and evolution of operons: The piecewise building of the proteobacterial histidine operon. *Journal of Molecular Evolution*, 60(3):378–390, 2005.

[13] David A. Fell. Enzymes, metabolites and fluxes. *Journal of experimental botany*, 56(410):267–272, 2005.

[14] Solveig Fossum, Elliott Crooke, and Kirsten Skarstad. Organization of sister origins and replisomes during multifork DNA replication in Escherichia coli. *EMBO Journal*, 26(21):4514–4522, 2007.

[15] Reinhart Heinrich and Tom A. Rapoport. A Linear SteadyâĂŘState Treatment of Enzymatic Chains: General Properties, Control and Effector Strength. *European Journal of Biochemistry*, 42(1):89–95, 1974.

[16] Mikel Irastortza-Olaziregi and Orna Amster-Choder. Coupled Transcription-Translation in Prokaryotes: An Old Couple With New Surprises, 2021.

[17] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356, 1961.

[18] Joanna Jaruszewicz-Błońska and Tomasz Lipniacki. Genetic toggle switch controlled by bacterial growth rate. *BMC Systems Biology*, 11(1):1–11, 2017.

[19] Henrik Kacser and James A Burns. The control of flux. *Symp. Soc. Exp.*, 27:65–104, 1973.

[20] Henrik Kacser and James A Burns. The molecular basis of dominance. *Enzyme*, (5):2411–2414, 1981.

[21] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2012.

[22] Kosmas Kosmidis, Kim Philipp Jablonski, Georgi Muskhelishvili, and Marc Thorsten Hütt. Chromosomal origin of replication coordinates logically distinct types of bacterial genetic regulation. *npj Systems Biology and Applications*, 6(1):1–9, 2020.

[23] Warren C Lathe III, Berend Snel, and Peer Bork. Gene context conservation of a higher order than operons. *Trends in biochemical sciences*, 13(1998):25388–25392, 2000.

[24] Jeffrey G. Lawrence. Gene organization: selection, selfishness, and serendipity. *Annual Review of Microbiology*, 57(1):419–440, 2003.

[25] Jeffrey G. Lawrence and Howard Ochman. Molecular archaeology of the Escherichia coli genome. *Proceedings of the national Academy of Sciences*, 95(16):9413, 1998.

[26] Jeffrey G. Lawrence and J R Roth. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143(4):1843–1860, 1996.

[27] E B Lewis. A gene complex controlling segmentation in Drosophila, 1978.

[28] Andrew M. Long, Shengwei Hou, J. Cesar Ignacio-Espinoza, and Jed A. Fuhrman. Benchmarking microbial growth rate predictions from metagenomes. *ISME Journal*, 15(1):183–195, 2021.

[29] Michael Lynch. Streamlining and Simplification of Microbial Genome Architecture. *Annual Review of Microbiology*, 60(1):327–349, 2006.

[30] Pablo A. Nuñez, Héctor Romero, Marisa D Farber, and Eduardo P.C. Rocha. Natural Selection for Operons Depends on Genome Size. *Genome Biology and Evolution*, 5(11):2242–2254, 2013.

[31] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.

[32] Csaba Pál and Laurence D Hurst. Evidence against the selfish operon theory. *Trends in Genetics*, 20(6):232–234, 2004.

[33] Rob Patro, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, 2017.

[34] Mariana G. Pinho, Morten Kjos, and Jan Willem Veening. How to get (a)round: Mechanisms controlling growth and division of coccoid bacteria. *Nature Reviews Microbiology*, 11(9):601–614, 2013.

[35] Morgan N. Price, Katherine H. Huang, Adam P. Arkin, and Eric J. Alm. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Research*, 15(6):809–819, 2005.

[36] J Christian J Ray and Oleg A. Igoshin. Interplay of Gene Expression Noise and Ultrasensitive Dynamics Affects Bacterial Operon Organization. *PLoS Computational Biology*, 8(8):e1002672, 2012.

[37] Paul E Schavemaker, Arnold J Boersma, and Bert Poolman. How Important Is Protein Diffusion in Prokaryotes? *Frontiers in Molecular Biosciences*, 5(November):1–27, 2018.

[38] J. Rankin Small and Henrik Kacser. Responses of metabolic systems to large changes in enzyme activities and effectors: 1. The linear treatment of unbranched chains. *European Journal of Biochemistry*, 213(1):613–624, 1993.

[39] J. Rankin Small and Henrik Kacser. Responses of metabolic systems to large changes in enzyme activities and effectors 2. The linear treatment of branched pathways and metabolite concentrations. Assessment of the general non-linear case. *European Journal of Biochemistry*, 213(1):625–640, 1993.

[40] Karline Soetaert, Thomas Petzoldt, and R. Woodrow Setzer. Solving differential equations in R: Package deSolve. *Journal of Statistical Software*, 33(9):1–25, 2010.

[41] Franklin W Stahl and Noreen E Murray. THE EVOLUTION OF GENE CLUSTERS AND GENETIC CIRCULARITY IN MICROORGANISMS. *Genetics*, 53(3):569–576, 1966.

[42] Damian Trojanowski, Joanna Hołówka, and Jolanta Zakrzewska-Czerwińska. Where and when bacterial chromosome replication starts: A single cell perspective. *Frontiers in Microbiology*, 9(NOV):1–9, 2018.

[43] Diana Visser and Joseph J Heijnen. Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. *Metabolic Engineering*, 5(3):164–176, 2003.

[44] Diana Visser, Joachim W Schmid, Klaus Mauch, Matthias Reuss, and Joseph J Heijnen. Optimal re-design of primary metabolism in Escherichia coli using linlog kinetics. *Metabolic Engineering*, 6(4):378–390, 2004.

[45] Yu Zheng, Joseph D. Szustakowski, Lance Fortnow, Richard J. Roberts, and Simon Kasif. Computational Identification of Operons in Microbial Genomes. *Genome Research*, 12(8):1221–1230, 2002.