FEBS Letters 476 (2000) 8–11 FEBS 23784

Minireview

Sulphur islands in the *Escherichia coli* genome: markers of the cell's architecture?

Eduardo P.C. Rocha[a,c], Agnieszka Sekowska[b,c], Antoine Danchin[b,c,*]

[a] *Atelier de BioInformatique, 12, rue Cuvier, 75005 Paris, France*
[b] *Hong Kong University Pasteur Research Centre, Dexter HC Man Building, 8, Sassoon Road, Pokfulam, Hong Kong, China*
[c] *Regulation of Gene Expression, URA 2171 CNRS, Institut Pasteur, 28, rue du Docteur Roux, 75724 Paris Cedex 15, France*

Received 5 May 2000

Edited by Gunnar von Heijne

Abstract Two highly contrasted images depict genomes: at first sight, genes appear to be distributed randomly along the chromosome. In contrast, their organisation into operons (or pathogenicity islands) suggests that, at least locally, related functions are in physical proximity. Analysis of the codon usage bias in orthologous genes in the genome of bacteria which diverged a long time ago suggested that some physical (architectural) selection pressure organised the distribution of genes along the chromosome. The metabolism of highly reactive species such as sulphur-containing molecules must be compartmentalised to escape the deleterious actions of diffusible reagents such as gases or radicals. We analysed the distribution of sulphur metabolism genes in the genome of *Escherichia coli* and found a number of them to be clustered into statistically significant islands. Another interesting feature of these genes is that the proteins they encode are significantly deprived of cysteine and methionine residues, as compared to the bulk proteins. We speculate that this clustering is associated to the organisation of sulphur metabolism proteins into islands where the sensitive sulphur-containing molecules are protected from reacting with elements in the environment such as dioxygen, nitric oxide or radicals.

Key words: Neighborhood; From structure to function; Protein complex; Hydrogen sulfide; Oxygen radical; Origin of life

1. Introduction

Sulphur is a ubiquitous element of the Earth's crust. It is one of the six atoms making the macromolecule core of living cells [1]. In *Escherichia coli*, sulphur is involved not only in well-known anabolic processes, in particular in cysteine and methionine biosynthesis, but also in the synthesis of many coenzymes or prosthetic groups (lipoate, coenzyme A, molybdopterin, thiamine, several kinds of iron–sulphur clusters) as well as in tRNA modification or in the synthesis of secondary metabolites [2]. Furthermore, a sulphur-containing molecule, *S*-adenosylmethionine, plays a central and essential role in many processes, not only by mediating methylation but also by donation of the two other groups of the sulphonium function of the molecule [3–5]. Sulphur is thus involved in the synthesis of polyamines [6]. It also undergoes several important recycling processes, notably the recycling of the first methionine residue of all proteins. Many processes are involved in the scavenging of sulphur from the environment [7]. Perhaps curiously, not much work has been devoted to the metabolic pathways involving this atom, as compared to those involving carbon or nitrogen. Fortunately, steady progress in DNA sequencing, in molecular genetics and in computer sciences has allowed the community of biologists to create a complete inventory of the genes in many organisms, their complete genome sequence, and to investigate the global function of specialised sets of genes. The genes involved in sulphur metabolism in *E. coli*, collected by Sekowska et al. [2], will be analysed in the present work.

It is generally admitted, without further thinking, that knowing the genome sequences will be enough to explain the most remote functions associated with life. However, nature does not proceed from structure – and even less from sequence – to function, but captures extant structures for whatever is needed [8]. This makes guessing the function from the sole sequence, in the absence of significant homology, an almost impossible task. Nevertheless, the knowledge of genome sequences is revolutionising biology, because it can be associated with other types of biological knowledge, allowing one to propose functions to gene sequences and prepare experiments for further understanding. For this, one has to find ways to put the type of knowledge we get from the study of living organisms in their natural environments using all kinds of physico-chemical tools, together with the knowledge of their genome sequence. To generate new knowledge from the sequence data, we have previously proposed an inductive reasoning approach, which explores neighbourhoods of biological entities, considering genes as starting points, and stressing that each entity exists in relation to other entities [9]. There, 'neighbour' was not simply a geometrical or structural notion, but had the largest possible meaning. Each neighbourhood was meant to shed a specific light on a gene with regard to its function, by bringing together genes belonging to the same class and sharing the same neighbourhood. A natural neighbourhood for genes is the proximity on the chromosome: operons or pathogenicity islands show that genes which are neighbours of each other can be functionally related [10,11]. Another interesting neighbourhood is the similarity between genes or gene products. The isoelectric point often gives a first idea of a gene product's compartmentalisation.

*Corresponding author. Fax: (852)-2168 4427.
E-mail: adanchin@hkucc.hku.hk

PII: S0014-5793(00)01660-4

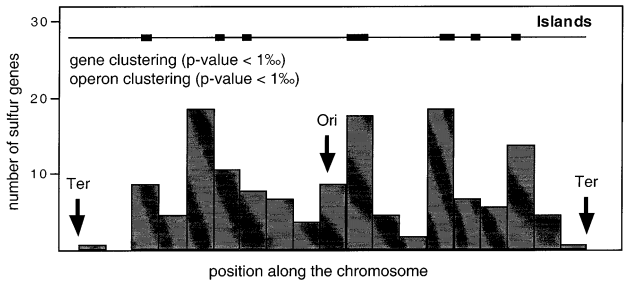

Fig. 1. Distribution of sulphur metabolism genes and location of the sulphur islands on the *E. coli* chromosome. Determination of the *P* values was done according to the serial randomness test referred to in the text and in [13,24]. Ori and Ter identify the origin and terminus of replication in the *E. coli* chromosome.

Also, the study of genes by scientists in laboratories all over the world creates another relation of proximity: a gene's neighbours will be other genes found together with it in the literature, 'in libro'. Finally, there may be more complex neighbourhoods, the study of which gives particularly revealing results: for example, two genes may be neighbours because they use the genetic code in the same way [12]. In the present work, we have studied the *E. coli* genes which belong to a metabolic neighbourhood, that of sulphur metabolism, and we have studied their distribution in a variety of other neighbourhoods, starting with proximity in the chromosome. An unexpected picture emerges from this analysis.

2. Analysis of sulphur metabolism gene neighbourhoods

2.1. Distribution of sulphur metabolism genes along the *E. coli* chromosome

Extracting information on sulphur metabolism-related genes in libro as well as in the SwissProt protein sequence library (http://www.expasy.ch/sprot/), it was found that the total number of genes likely to be involved in sulphur processes is ca. 150, a significant proportion of the *E. coli* genome [2]. Using the standard serial randomness statistics [13] we tested whether the distribution of the sulphur genes on the chromosome is random, uniform or aggregated. This was performed on two different kinds of data. First, we tested whether sulphur metabolism genes are clustered together. The result of the test is highly significant (Fig. 1), but this may simply be caused by the aggregation of genes into operons. Therefore we have defined 'operons' for the *E. coli* chromosome and tested whether these 'operons' were clustered together. In the absence of extensive and systematic information, we defined operons as sequences of co-oriented genes. This is a simplistic definition of an operon, but for our purpose it is a very conservative one, since the main source of error is to merge operons together. Therefore sets of co-oriented sulphur metabolism operons will be regarded as one single one and the test will not consider aggregation, even if it exists. In spite of this restriction, the test clearly demonstrates aggregation of sulphur metabolism operons in the chromosome (Fig. 1).

Taking into account the two criteria (clustering of genes together into operons, and clustering of operons together) the 150 or so sulphur metabolism-related genes are not evenly distributed in the chromosome. Half of the corresponding genes make short clusters, composed of isolated genes or short operons, interspersed with many genes of unknown function, so that it is difficult to say whether they can be grouped into larger units, similar to pathogenicity islands. However, we tried to identify statistically significant 'islands' of such genes. For this, we tested whether the operon aggregation observed in the chromosome was statistically significant using a test based on the binomial distribution. Statistics about finite sets are notoriously difficult to justify, especially when they give borderline results. Retaining only those observations that are significant to a *P* value lower than 5% (but they are frequently lower than 5‰), we identified a significant number of sulphur metabolism-related genes clustered into seven major 'islands' representing a total of 64 sulphur metabolism genes, in a highly significant way (Table 1).

Remarkably, a quarter of the *E. coli* chromosome encompassing the region of the terminus of replication is almost entirely lacking identified sulphur metabolism genes. It may also be significant that, apart from a couple of islands located in the vicinity of the origin of replication, there is a symmetrical cluster of these genes of either side of the origin, about one quarter of the chromosome away from the origin (Fig. 1).

2.2. Codon usage bias in sulphur metabolism genes

This surprising finding could be accounted for by many processes, including horizontal transfer of genes, known to be responsible for the formation of pathogenicity islands, for example [11,14]. Horizontally transferred genes in *E. coli* K12 have been identified by several criteria. It was found that

Table 1
Genes in sulphur metabolism islands

Start	End	NopS	Nop	*P* value	No. genes	Genes
146 314	193 429	5	10	0.00043	15	*panD/panC panB/yadS yadT pfs/yaeH yaeI dapD glnD map/rpsB tsf pyrH frr*
434 858	443 739	3	4	0.00189	5	*thiL pgpA/thiI/thiJ apbA*
805 221	818 970	3	4	0.00189	12	*ybhC ybhB bioA/bioB bioF bioC bioD/moaA moaB moaC moaD moaE*
2 180 055	2 201 931	3	4	0.00189	7	*yegW yegX thiD thiM/mrp/metG yehI*
2 871 410	2 889 921	2	3	0.0526	6	*cysC cysN cysD/cysH cysI cysJ*
3 132 887	3 137 608	2	2	0	3	*pitB gsb/yghU*
4 106 414	4 225 090	9	29	3.75 E-05	16	*sbp/metJ/metB metL metF/murB birA/coaA/thiH thiG thiF thiE thiC/yjaB/metA/metH*

'Start' corresponds to the position of the beginning of the first gene known to be involved in sulphur metabolism. 'End' corresponds to the stop codon of the last sulphur metabolism gene. 'NopS' is the number of putative sulphur operons, 'Nop' is the number of likely or known operons in the island. '*P* value' is the probability of the existence of the island computed using a binomial test. 'No. genes' is the number of sulphur metabolism genes in the island. 'Genes' lists the names of the relevant genes (genes located inside different operons are separated by a slash).

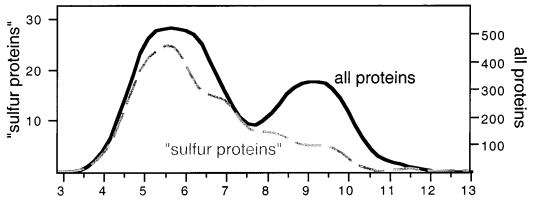

Fig. 2. Distribution of p*I* values in all *E. coli* proteins (continuous line) and in the sulphur metabolism proteins (dashed line).

E. coli harbours three well-defined classes of genes: the core metabolism genes, the genes which are highly expressed under exponential growth conditions, and horizontally exchanged genes [12,15–17]. Interestingly, the 'sulphur island' genes do not belong to the latter class (only two members of this class were found, a very low proportion, since this class includes more than 15% of the *E. coli* genes [12,18]). Moreover, the codon usage bias in these genes is remarkably self-consistent (i.e. with only little variation from gene to gene), with a strong counterselection against codons ATA, AGA and TCA (data not shown), indicating that they are subjected to a common type of selection pressure. It has been found that the genes encoded by the leading strand as compared to those encoded by the lagging strand differ in base composition, in codon usage and in amino acid composition of the corresponding proteins [19,20]. We did not, however, find major differences in the sulphur metabolism gene distribution in the leading or lagging DNA strand, nor significant differences in the average gene length as compared to the average *E. coli* gene length (data not shown).

2.3. Amino acid composition of sulphur metabolism-related proteins

However, several specific features appear to distinguish the proteins encoded by these genes from the average *E. coli* protein. The *E. coli* proteins are divided into two major isoelectric point domains, with proteins with a p*I* of about 7.6 highly selected against (Fig. 2). The 'sulphur island' proteins belong mostly to the low p*I* group and are significantly more acidic than the bulk (average p*I* 6.5 against 7.0, respectively). This is because they contain more aspartate and glutamate residues than the average *E. coli* protein. This corresponds to the low p*I* group of *E. coli* proteins. They do, however, contain the average number of the basic amino acids lysine, arginine and histidine (data not shown). In contrast, sulphur metabolism-related proteins are relatively deficient in serine residues ($P < 0.001$). More significantly perhaps, putting aside the first methionine residue of each protein, which is always present and would alter the statistical analysis, sulphur metabolism-related proteins are poorer in sulphur-containing amino acids than the average *E. coli* protein ($P < 0.002$) (Table 2).

Table 2
Methionine and cysteine residues are avoided in sulphur metabolism proteins

		Observed (%)	Expected (%)	*P* value
↗	Asp	5.5	5	< 0.002
↗	Glu	6.4	5.8	< 0.002
↘	Ser	5.3	5.9	< 0.001
↘	Cys+Met[a]	3.3	3.8	< 0.002

P values were calculated using the Wilcoxon test, a non-parametric test for the comparison of means in two sets. We tested whether the mean composition of these amino acids in the sulphur metabolism proteins was identical, higher or lower than the average *E. coli* protein.
[a]The first methionine residue of each protein was omitted for the calculation.

3. Discussion

We have used a standard approach to explore whether genes related to sulphur metabolism were randomly distributed in the chromosome. We find that sulphur metabolism genes and operons are significantly clustered together in *E. coli*. In fact, many of these genes are grouped into 'islands' comprising 3–16 sulphur metabolism genes, distributed into groups of 2–29 operons (Table 1). The terminus of replication is almost entirely lacking in such genes. This may provide a leading thread to explain these unexpected findings. Indeed, it has been found in many instances that the terminus of replication region is prone to accept foreign DNA [21–23] and is less sensitive to the general constraints of genome organisation than the regions nearer the origin of replication [24,25]. In addition, we have observed that the number of genes in sulphur metabolism explicitly obtained through horizontal gene transfer is extremely low, if it exists. This strongly suggests – as would indeed be expected from their role in metabolism – that these genes are part of the core gene system of *E. coli*.

Why should they be clustered together? Operons in general, and pathogenicity islands as well, are made of genes for proteins working together for a more or less specific function. These genes usually have similar codon usage biases, and the corresponding proteins are usually organised into multisubunit complexes. More generally, the existence of similarly strong biases in codon usage for the synthesis of orthologous proteins in distantly related bacteria has been interpreted as the landmark of a relationship between the architecture of the chromosome and that of the cell [25]. We therefore suggest that sulphur metabolism proteins are assembled into particular compartments of the cell, where they build up functional complexes. Several theories have suggested that sulphur is involved in the initial metabolic processes built up around an iron–sulphur pyrite-like core [26–28]. The extant bacterial cell architecture might still preserve some of this ancestral organisation (but see the discussion between S. Benner and G. Wächtershäuser on this topic [29,30]). Despite the lack of experimental identification of many gene functions in other bacteria, in silico analysis of the distribution of genes likely to be involved in sulphur metabolism in other bacteria suggests that, in their case also, they are clustered together [2], substantiating this hypothesis.

Once an architecture has been established, the constraints imposed by the organisation of functional (RNA) protein complexes into 'nanosomes' some 10–50 nm in width or into a network with similar local dimensions ('reticulosomes') permitting small molecule channelling and organised distribution of substrates into the cell [31], impose a strong selection pressure on the cell objects. Do there exist forces leading to the formation of such complexes? The bias in amino acid composition of the sulphur metabolism proteins may hint at an interesting explanation. Evolution results from founder

effects as well as mutational drift and selection pressure. Serine, cysteine and methionine, the amino acids less abundant than expected in the proteins considered (Table 2), are strongly related in their metabolic pathways (sulphur and one carbon incorporation into amino acids [32]). An important effect is linked to the chemical nature of sulphur. This atom is extremely prone to all kinds of oxido–reduction changes [1,32]. In particular, many of the sulphur metabolism proteins use sulphur in a reduced form (including the H_2S gas) as intermediates in the reactions they catalyse. Because gases or radicals diffuse extremely rapidly in the cell, it is important that these substrates both are confined where they are to be used, and are protected against the action of gases such as dioxygen or nitric oxide. To be assembled in tight complexes is an expedient way to perform this protective function.

4. Conclusion

These results of an in silico analysis of the *E. coli* genome must be experimentally validated in vivo (using reverse genetics, physiological biochemistry and microscopy), in order to further explore the hypothesis of compartmentalisation. Even if our working hypothesis is somewhat speculative, it is both precise and original enough to allow us to propose new hypotheses and discover new functions of orphan genes, for which the usual 'functional analysis' approaches are ineffective. Indeed, if sulphur metabolism islands suggest the existence of multi-component complexes, it is likely that the genes present in the islands will encode proteins which are members of the complexes, even if they are not directly related to sulphur metabolism. The results obtained with model organisms will serve as blueprints to build up an integrated study of pathogens and may help us to understand better the function and organisation of another important type of gene islands, the pathogenicity islands.

References

[1] Ehrlich, H.L. (1996), pp. 717, Marcel Dekker, New York.
[2] Sekowska, A., Kung, H.-F. and Danchin, A. (2000) J. Mol. Microbiol. Biotech. 2, 145–177.
[3] Stipanuk, M.H. (1986) Annu. Rev. Nutr. 6, 179–209.
[4] Chiang, P.K., Gordon, R.K., Tal, J., Zeng, G.C., Doctor, B.P., Pardhasaradhi, K. and McCann, P.P. (1996) FASEB J. 10, 471–480.
[5] Grogan, D.W. and Cronan Jr., J.E. (1997) Microbiol. Mol. Biol. Rev. 61, 429–441.
[6] Cohen, S.S. (1998), pp. 595, Oxford University Press, Oxford.
[7] Cook, A.M., Laue, H. and Junker, F. (1998) FEMS Microbiol. Rev. 22, 399–419.
[8] Danchin, A. (1999) Curr. Opin. Struct. Biol. 9, 363–367.
[9] Nitschké, P., Guerdoux-Jamet, P., Chiapello, H., Faroux, G., Hénaut, C., Hénaut, A. and Danchin, A. (1998) FEMS Microbiol. Rev. 22, 207–227.
[10] Neidhardt, F.C. and Savageau, M.A. (1996) in: *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology (Neidhardt, F.C., Ingraham, J.L., Lin, E.C.C., Brooks Low, K., Magasanik, B., Reznikoff, W.S., Riley, M., Schaechter, M. and Umbarger, H.E., Eds.), pp. 1310–1324, ASM Press, Washington, DC.
[11] Groisman, E.A. and Ochman, H. (1997) Trends Microbiol. 5, 343–349.
[12] Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A. (1991) J. Mol. Biol. 222, 851–856.
[13] Zar, J.H. (1996) Prentice Hall, Englewood Cliffs, NJ.
[14] Marcus, S.L., Brumell, J.H., Pfeifer, C.G. and Finlay, B.B. (2000) Microbes Infect. 2, 145–156.
[15] Ochman, H. and Groisman, E.A. (1994) Exs 69, 479–493.
[16] Lawrence, J.G. and Ochman, H. (1997) J. Mol. Evol. 44, 383–397.
[17] Munoz, R., Garcia, E. and Lopez, R. (1998) J. Mol. Evol. 46, 432–436.
[18] Lawrence, J.G. and Ochman, H. (1998) Proc. Natl. Acad. Sci. USA 95, 9413–9417.
[19] Lobry, J.R. (1996) Mol. Biol. Evol. 13, 660–665.
[20] Rocha, E.P., Danchin, A. and Viari, A. (1999) Mol. Microbiol. 32, 11–16.
[21] Carlson, C.R. and Kolsto, A.B. (1994) Mol. Microbiol. 13, 161–169.
[22] Romling, U., Schmidt, K.D. and Tummler, B. (1997) J. Mol. Biol. 271, 386–404.
[23] Kunst, F. et al. (1997) Nature 390, 249–256.
[24] Rocha, E.P.C., Guerdoux-Jamet, P., Moszer, I., Viari, A. and Danchin,A. (2000) J. Biotechnol., in press.
[25] Danchin, A., Guerdoux-Jamet, P., Moszer, I. and Nitschké, P. (2000) Phil. Trans. R. Soc. Lond. B Biol. Sci. 355, 179–190.
[26] Danchin, A. (1989) Prog. Biophys. Mol. Biol. 54, 81–86.
[27] Wächtershäuser, G. (1992) Prog. Biophys. Mol. Biol. 58, 85–201.
[28] Edwards, M.R. (1996) J. Theor. Biol. 179, 313–322.
[29] Benner, S.A., Ellington, A.D. and Tauer, A. (1989) Proc. Natl. Acad. Sci. USA 86, 7054–7058.
[30] Wächtershäuser, G. (1997) J. Theor. Biol. 187, 483–494.
[31] Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) Nature 402, C47–C52.
[32] Michal, G. (1999), pp. 277, John Wiley and Sons, New York, Spectrum Akademischer Verlag, Heidelberg.