Addendum to the paper "Codon usage domains over bacterial chromosomes" PLoS Computational Biology Vol. 2, No. 4, e37

Marc Bailly-Bechet, Antoine Danchin, Mudassar Iqbal
Matteo Marsili, Massimo Vergassola.

October 18, 2006

An issue left unexplained in the paper [1] is the striking quantitative difference between *E. coli* and *B. subtilis*. This is clearly visible in Fig. 6 of [1], where it is shown the probability that two genes at distance ℓ belong to the same cluster of codon usage. Clusters are characterized by a similar codon bias and were identified using a novel information-based clustering method. While both curves decay on distances sizably longer than what could be accounted by operons, *B. subtilis* curve manifestly features much longer correlations. It is hard to develop a biologically well-founded explanation for such a striking difference between the two organisms. This observation and discussions with Dr. Morten Kloster (Princeton Univ.) spurred us to reconsider the issue and further pursue our analysis of the clusters. The purpose of this addendum is to describe this analysis, which allows us to point out an incorrect statement made in the paper [1], and provide an explanation for the aforementioned difference between the two organisms. The conclusion is that clusters of *B. subtilis* and *E. coli* not biased in GC content display now the same behaviour, with correlations of codons usage of the same order, roughly three times the length of the average operon.

Contrary to what was previously stated, the GC content of the various clusters is not quite homogeneous and the correct values are reported in Table A1.

Cluster	1	2	3	4
GC%	.527	.443	.541	.522

Cluster	1	2	3	4	5
GC%	.439	.358	.450	.470	.436

Table A1: The GC percentage for the four and the five clusters of *E. coli* (left) and *B. subtilis* (right), respectively, which were identified on the basis of their codon usage.

The demonstration given in [1] that clusters are biologically significant still holds and does not depend on the GC content. In particular, the third cluster of *B. subtilis*, which was shown to feature an over-representation of anabolic genes and lagging-strand transcriptional orientation, does not show any particular deviation from the average genomic GC content. The clusters which most significantly deviate from the average are clusters 2, both in *E. coli* and *B. subtilis*. The two clusters are enriched in AT and have been shown in [1] to be enriched in horizontally transferred genes. The higher AT% shown in Table 1 is in agreement with the well-known observation that horizontally transferred genes tend to be AT rich (see [2] and references therein).

The GC percentage resolves the aforementioned observation of the different correlation lengths in Fig. 6 of [1] for *E. coli* and *B. subtilis*. To demonstrate this, we considered the same correlation functions plotted in Fig. 6, but for each individual cluster, to highlight the contribution of the various groups. Specifically, we measured the probability that two genes, g and $g+\ell$, belong to the same cluster ($s_g = s_{g+\ell}$), with the additional constraint that $s_g = S$ ($S = 1, \ldots, 4$ for *E. coli* and $S = 1, \ldots, 5$ for *B. subtilis*):

$$\mathcal{P}_2^{(S)}(\ell) = \frac{1}{N_S} \sum_g \delta(s_g\,, s_{g+\ell}) \delta(s_g, S)\,. \tag{1}$$

Here, δ is the Kronecker delta-function and the function $P_2^{(S)}$ is normalized by the total number of genes N_S belonging to the S-th cluster. The function can also be interpreted as the histogram of the distances among genes belonging to the same cluster. The resulting curves for the various clusters are shown in Fig. A1 for *E. coli* and Fig. A2 for *B. subtilis*, with the value at large distances subtracted for more clarity.

A first observation is that the curves are more noisy than in Fig. 6 of [1]. This is quite natural as each group contains less genes and was our reason for grouping all the clusters together to produce Fig. 6. Some statistically robust and informative behaviors are still clearly discernible, though. In particular, it is quite evident that the cluster of *B. subtilis* having the longest correlations is the fourth one. The correlation length of the cluster is clearly dominant over all the others and it is comparable to the decay length observed in Fig. 6 of [1]. Table A1 indicates that the fourth cluster deviates from the average GC percentage of the whole *B. subtilis* genome and is GC enriched (with respect to the genomic average). This suggests that the dominant contribution to the anomalous decay length observed in Fig. A2 is due to the correlations in the GC content of the *B. subtilis* genome. It is important, though, to remark that groups not biased in their (relative to the average) GC content also feature extended correlations, longer than what could be accounted by operons. Furthermore, the effects are now comparable in *E. coli* and *B. subtilis*. A contribution to those correlations might be driven by the advantage of recycling rare tRNAs to tame stallings in the translation process and ensure

a coordinated expression of a set of neighboring genes, as discussed in the conclusions of [1]. The importance of pauses is translation is also highlighted by the large number of tmRNAs typically present in the cell [3, 4].

References

[1] Bailly-Bechet M, Danchin A, Iqbal M, Marsili M, Vergassola M (2006) Codon usage domains over bacterial chromosomes. PLoS Computational Biology, Vol. 2, No. 4, e37.

[2] Rocha EPC, Danchin A (2002) Base composition bias might result from competition for metabolic resources. Trends in Genetics 18(6):291–294.

[3] Altuvia S, Weinstein-Fischer D, Zhang A, Postow L, Storz G (1997) A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator. Cell 90:43–53.

[4] Moore S D, Sauer R T (2005) Ribosome rescue: tmRNA tagging activity and capacity in *Escherichia coli*. Molecular Microbiology 58(2):456–466.

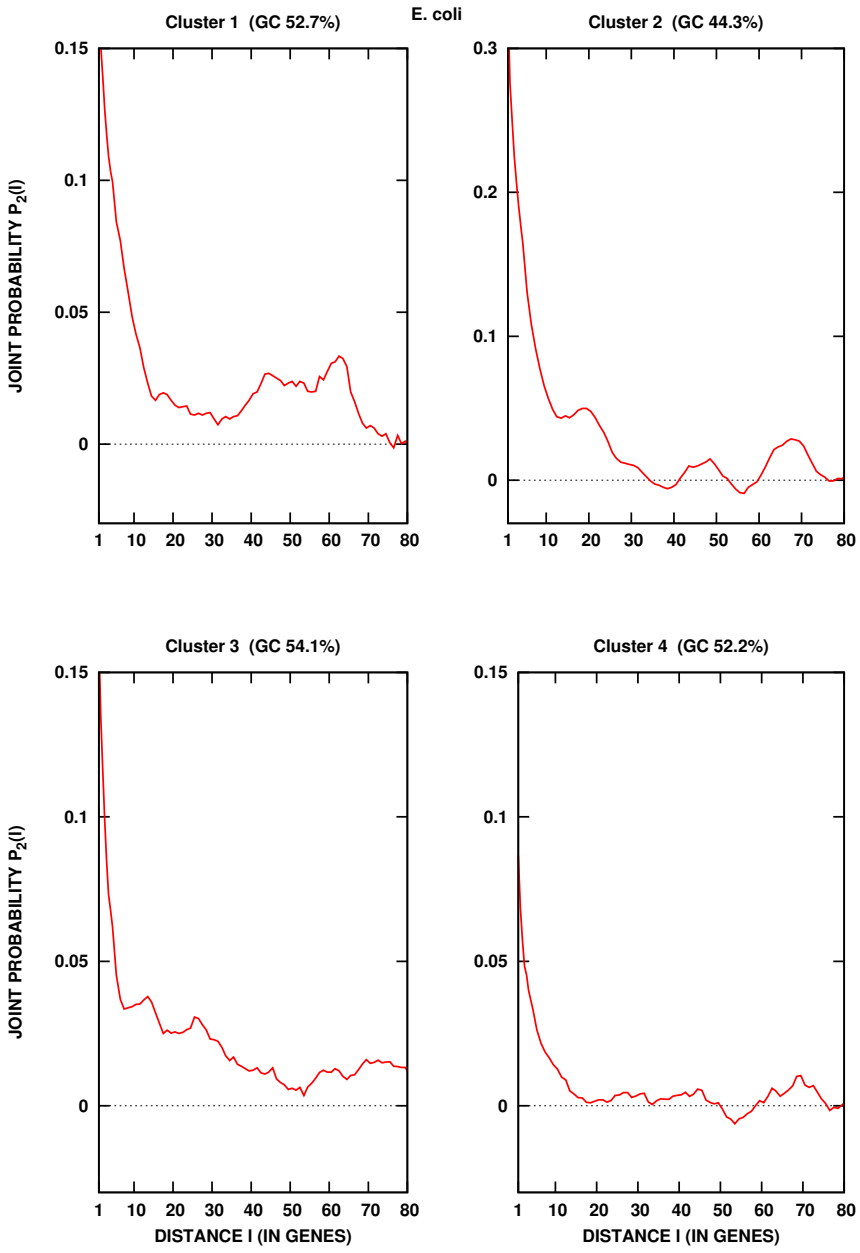

Figure 1: The probability distribution (1) of cluster membership for the four clusters identified in *E. coli*.

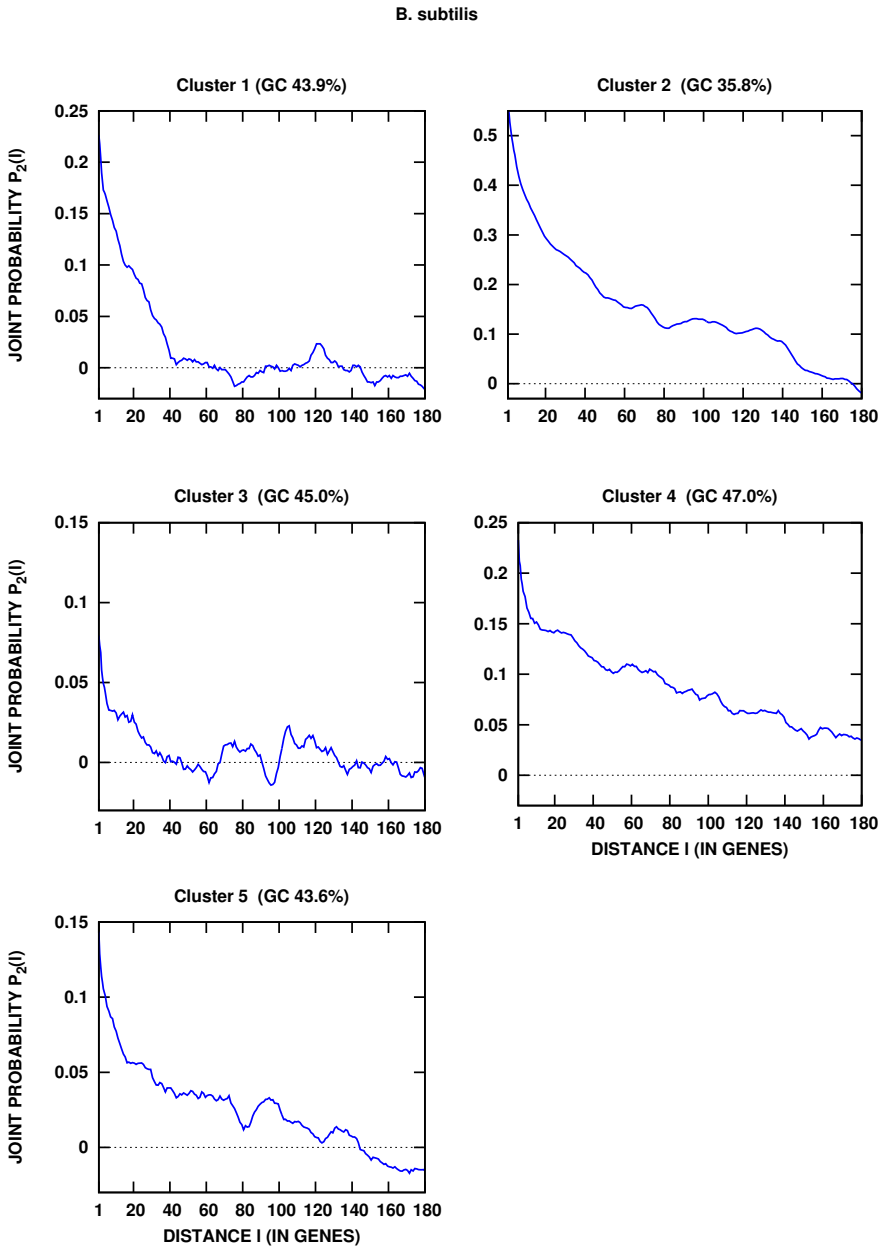

Figure 2: The probability distribution (1) of cluster membership for the five clusters identified in *B. subtilis*.