Cell PRESS

Regulome size in Prokaryotes: universality and lineage-specific variations

Otto X. Cordero and Paulien Hogeweg

Theoretical Biology and Bioinformatics, University of Utrecht, Padualaan 8, 3584 CH Utrecht, The Netherlands

Molina and van Nimwegen [1] recently reported scaling exponents for different functional groups of genes and analysed the distribution of these scaling coefficients across clades and lifestyles. On the basis of the overlap between the 99% confidence intervals of their fits, they conclude that, for almost all functional groups, the exponents are equal ('universal') for all Prokaryotes. They place particular emphasis on transcription regulation and conclude that the scaling exponent over all clades and lifestyles is ~2. This differs from our earlier study of regulome evolution across clades [2], in which we found abrupt expansions and contractions of relative regulome size at the origin of major lineages and lower exponents within clades than in the whole dataset. Molina and van Nimwegen suggest that this discrepancy is because of their larger dataset, improved annotation (i.e. GO annotated domains rather than COG) and a better fitting procedure, and conclude that the differences we reported can be explained by 'noise'. Here, we question their conclusions on technical and conceptual grounds and discuss the possible mechanisms underlying the apparent constraints and variation of regulome size in prokaryotes.

On a technical level, the main problem with Molina and van Nimwegen's cross-annotation comparison is that the 'transcription regulation' category in the COG database collects one-component systems (1CS), which combine substrate-binding and DNA-binding domains in a single protein [3], rather than two-component systems (2CS), which are present in the 'signal transduction' category. Therefore, by comparing the 'transcription regulation' COG category with their GO-PFAM annotation, which bundles 1CS and 2CS together, they compare different biological systems. In our previous *Trends in Genetics* paper [2], we showed that, although 1CS and 2CS have similar scaling exponents when fitting through all species, for the specific case of 1CS, the scaling became near-linear in large clades such as *Bacilli*, a result confirmed by Molina and van Nimwegen (see their supplementary material online). Figure 1a shows the results of combining 1CS and 2CS with COG annotation: as expected, the exponent of *Bacilli* is somewhat greater than for 1CS only. The fitted trends are actually consistent with the results obtained when we annotate proteins based on PFAM and not on GO-PFAM annotation (which misses ~20% of all proteins that do have PFAM domains) as Molina and van Nimwegen do (see Figure 1 and supplementary material online for details).

Molina and van Nimwegen see the lower spread in the GO-PFAM data as a sign of 'better' annotation. However, a reduction of the log-errors is a natural consequence of having more proteins on the y-axis (i.e. 1CS + 2CS), simply because what is on the y-axis is a subset of what is on the x-axis. (See supplementary material online, which also shows that this is independent of 'natural categories'.) The category-size-dependent trend can also be seen in Figure 1c–k of Molina and van Nimwegen's paper.

On a more conceptual level, Molina and van Nimwegen claim that their results support the 'simple' hypothesis of a universal constant for evolution. Currently, no biologically plausible hypothesis (simple or otherwise) is available either for universal constants or for the causes of variation. The hypothesis proposed in earlier work by van Nimwegen [4] – a faster duplication and deletion rate – seems unlikely because of the role that horizontal transfer has in bacterial evolution and the substantial differences in the scaling exponents observed within closely related groups (Figure 1c), contravening a mutational or causal explanation. Evidently, not disproving a common distribution ('universality') does not mean that the observed variation is biologically meaningless. Data fitting alone cannot distinguish between measurement noise, neutral variation, evolution contingencies or functional or adaptive variation.

In the absence of a hypothesis, a first step to identify possible mechanisms for convergence or variation is to investigate the patterns of variation. Interestingly, Molina and van Nimwegen state that variation among lifestyles is less than that across lineages, suggesting that genome organization (possibly caused by evolutionary contingency), rather than direct adaptation, underlies variation. It would be particularly interesting to study if different features of the genome organization vary in consort. The Molina and van Nimwegen data show that *Bacilli*, which take an extreme position with respect to regulome scaling, do so in several functional categories (including those for which the authors concede that there is clade variation). Moreover, Figure 1d shows a lineage-specific pattern of regulatory network organization which coincides with the observed variation in exponents: the *Bacilli*, which have a low scaling exponent and high offset, have also one of the lowest percentages of divergently transcribed regions in the genome. At the other extreme, *Archaea*, which show a lower number of regulators and high exponent, has a large percentage of such divergent regions. An analysis of regulatory networks indicates that divergently transcribed gene pairs tend to be co-regulated, possibly because of their overlapping upstream domains [5]. Thus, scaling features seem to co-vary with incidence of co-regulation,

DOI of original article: 10.1016/j.tig.2009.04.004.

Corresponding author: Hogeweg, P. (P.Hogeweg@uu.nl).

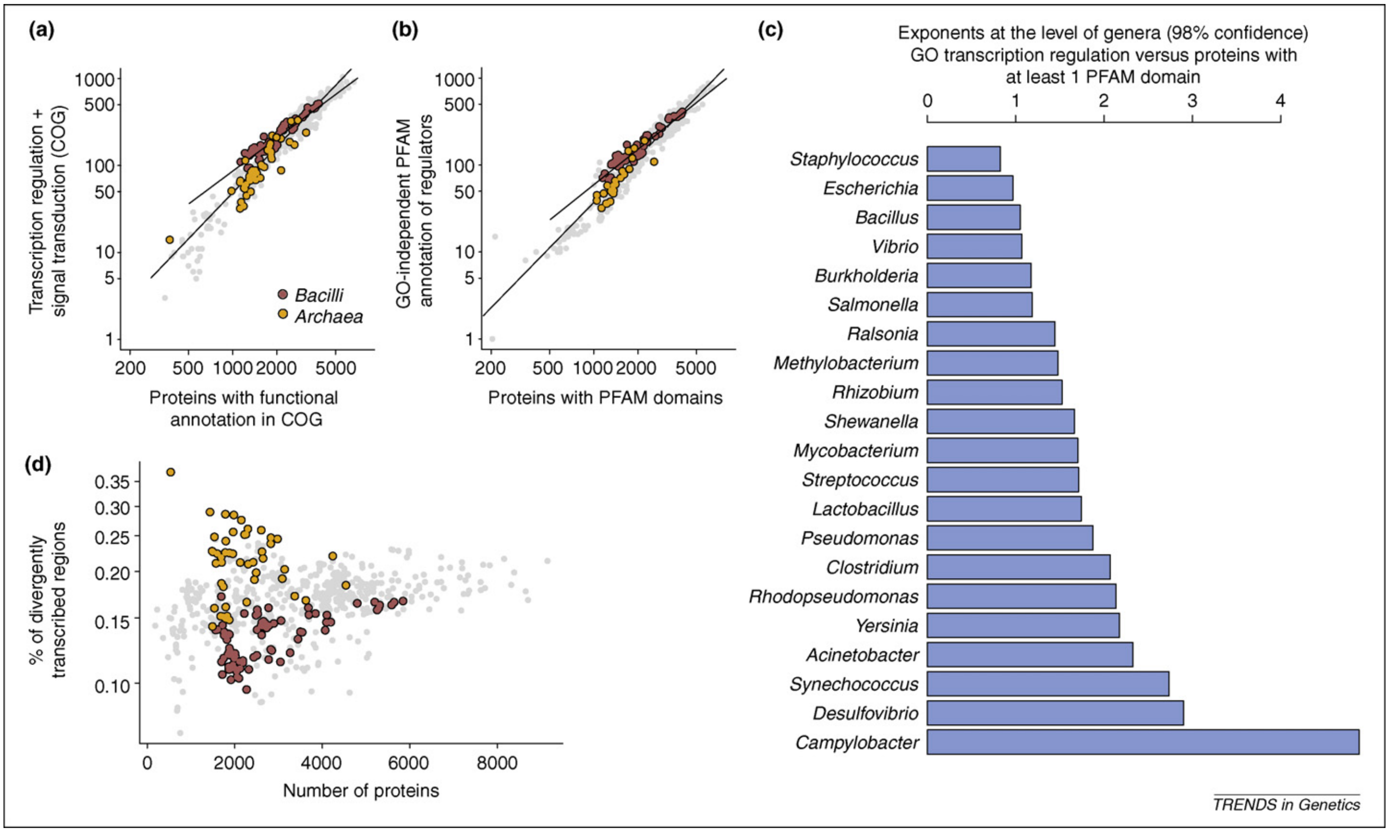

Figure 1. Is there a universal constant underlying the evolution of regulators? **(a,b)** This figure shows the scaling of 1CS + 2CS on species with at least 65% annotation in COG (577 species) and PFAM (558 species) data. Both annotation systems converge to sub-quadratic exponents (1.75 linear fit, 1.85 maximum likelihood [ML] fit) and to low scaling exponents within *Bacilli* (COG: 1.27 linear fit, 1.36 ML fit. PFAM: 1.35 linear fit, 1.42 ML fit). To obtain a GO-independent set of regulatory domains, we detected those non-GO-annotated PFAM domains that appear frequently in proteins of which the text annotation in the String7.1 [6] database indicates transcriptional regulatory function. The predicted function of those non-GO-annotated regulatory domains was verified manually against Interpro (www.ebi.ac.uk/interpro/) annotation (see supplementary material online for full list of domains). **(c)** No evidence of universal quadratic law within short evolutionary time-scales. The figure shows all within-genera fitted slopes with P <0.02, using GO-PFAM 'regulation of transcription' against proteins with PFAM domains. We can see that well-sampled genera such as *Bacillus*, *Escherichia* and *Salmonella* do not comply with the hypothesis of quadratic universality. **(d)** The number of proteins in the genome versus percentage of divergent regions. The variation observed for *Bacilli* and *Archaea* is consistent with the variation observed in parts (a,b) and suggests that lineage-specific patterns of network organization, and not annotation noise, explain the differences in scaling.

which might reflect a causal relationship. These observations give initial suggestions about the links between the scaling properties as studied by Molina and van Nimwegen and other properties of genome organization. Such links might ultimately reveal mechanisms that constrain scaling factors, and/or reveal interesting lineage specific differences, and should be studied further.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2009.05.001.

References

1 Molina, N. and van Nimwegen, E. (2009) Scaling laws in functional genome content across prokaryotic clades and life-styles. *Trends Genet.* 25, 243–247

2 Cordero, O.X. and Hogeweg, P. (2007) Large changes in regulome size herald the main prokaryotic lineages. *Trends Genet.* 23, 488–493

3 Ulrich, L.E. *et al.* (2005) One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol.* 13, 52–56

4 Van Nimwegen, E. (2003) Scaling laws in the functional content of genomes. *Trends Genet.* 19, 479–484

5 Warren, P.B. and Ten Wolde, P.R. (2004) Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*. *J. Mol. Biol.* 342, 1379–1390

6 Von Mering, C. *et al.* (2007) STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic. Acids. Res.* 35, D358–D362

0168-9525/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2009.05.001 Available online 18 June 2009