changes are generally favored. This extrinsic selective force gives a plausible explanation for the excess of shorter, frame-preserving cassette exons in present mammalian genomes, among other possible mechanisms. Our observations are consistent with the work of Wen *et al.*, who suggested that AS events that introduce a short variable region might have a larger functional impact than expected [24]. Finally, our results support the notion that NMD is generally more a mechanism for quality control [17,25] rather than one for the regulation of gene expression [18].

Acknowledgements

This work was supported by NIH grants GM42699 (A.R.K.) and HG074688 (M.Q.Z.). The authors would like to thank Philip J. Smith and Andrew D. Smith for helpful comments on the manuscript.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2007.08.001.

References

1 Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* 72, 291–336
2 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
3 Johnson, J.M. *et al.* (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302, 2141–2144
4 Kampa, D. *et al.* (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* 14, 331–342
5 Modrek, B. and Lee, C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* 34, 177–180
6 Xing, Y. and Lee, C. (2006) Alternative splicing and RNA selection pressure - evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.* 7, 499–509
7 Zhang, X.H.F. and Chasin, L.A. (2006) Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc. Natl. Acad. Sci. U. S. A.* 103, 13427–13432
8 Sorek, R. *et al.* (2002) Alu-containing exons are alternatively spliced. *Genome Res.* 12, 1060–1067
9 Lev-Maor, G. *et al.* (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science* 300, 1288–1291
10 Dagan, T. *et al.* (2004) AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic Acids Res.* 32, D489–D492
11 Resch, A. *et al.* (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.* 32, 1261–1269
12 Sorek, R. *et al.* (2004) How prevalent is functional alternative splicing in the human genome? *Trends Genet.* 20, 68–71
13 Xing, Y. and Lee, C. (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl. Acad. Sci. U. S. A.* 102, 13526–13531
14 Xing, Y. and Lee, C.J. (2005) Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet.* 1, e34
15 Sorek, R. and Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* 13, 1631–1637
16 Sugnet, C.W. *et al.* (2006) Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput. Biol.* 2, e4
17 Pan, Q. *et al.* (2006) Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev.* 20, 153–158
18 Lewis, B.P. *et al.* (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U. S. A.* 100, 189–192
19 Thanaraj, T.A. *et al.* (2004) ASD: the alternative splicing database. *Nucleic Acids Res.* 32, D64–D69
20 Pritsker, M. *et al.* (2005) Diversification of stem cell molecular repertoire by alternative splicing. *Proc. Natl. Acad. Sci. U. S. A.* 102, 14290–14295
21 Sugnet, C. *et al.* (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.* 66–77
22 Stamm, S. *et al.* (2000) An alternative-exon database and its statistical analysis. *DNA Cell Biol.* 19, 739–756
23 Berget, S.M. (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem.* 270, 2411–2414
24 Wen, F. *et al.* (2004) The impact of very short alternative splicing on protein structures and functions in the human genome. *Trends Genet.* 20, 232–236
25 Maquat, L.E. and Carmichael, G.G. (2001) Quality control of mRNA function. *Cell* 104, 173–176

0168-9525/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2007.08.001

Large changes in regulome size herald the main prokaryotic lineages

Otto X. Cordero and P. Hogeweg

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

Using a large-scale reconstruction of ancestral gene content, we show that radical changes in regulome size occur at the origins of major prokaryotic lineages. Subsequently, the duplication and deletion of regulators slows down in most lineages, except proteobacteria, significantly reducing the scaling of regulators and keeping their average proportion lineage-specific. Our results also suggest that major transitions in prokaryote evolution are related to changes in regulatory capacity rather than proteome innovations.

Corresponding author: Cordero, O.X. (o.x.corderosanchez@bio.uu.nl).
Available online 10 August 2007.

Genome size and the regulome

It remains a great challenge in biology to understand the evolution of gene regulatory networks in relationship to the gene repertoire and ecology of an organism. How does gene regulation scale with genomic complexity and how

does it respond to new environmental challenges? [1] Larger genomes seem to harbor a greater ratio of transcription factors (TFs) and signal transduction genes than smaller ones [2–4], suggesting *a priori* that under complex lifestyle conditions, gene regulation and signal integration are strongly selected. Moreover, the number of genes per functional category scales as a power law, from which we can infer a constant scaling factor of ~2 for transcription regulation and signal transduction, unlike the case for most other categories [3]. This means that, on average, if the number of genes doubles, the number of regulators and signal transduction genes quadruples. The explanation proposed by van Nimwegen [3,5] is that the observed quadratic scaling results from average duplication rates that are twice as great as for the rest of the genome.

This has important implications in our understanding of the fundamental principles of genome evolution. If more complex organisms, for example with greater metabolic diversity, require a larger control machinery, there should be a theoretical maximum for the expansion of genomes [6]. If the regulome actually expands faster, we need to determine how new regulators couple with the existing machinery and whether certain kinds of regulation mechanisms are favored or not by these expansions.

Ancestral gene content reconstruction

To compare the evolution of a group of gene families (e.g. regulators, against changes in genome size), we performed a maximum parsimony reconstruction [7,8] of ancestral gene content along the prokaryotic tree of life, establishing the number of duplications, deletions and 'gains' – i.e. horizontal gene transfers (HGT) [9] – and *de novo* gene inventions (see section S1 in the supplementary material online). We used the COG (Cluster of Orthologous Groups) database [10] to determine putative numbers of regulators and total gene content for each of the 163 prokaryotic genomes present in the STRING v6.2 database [11]. The ancestor reconstruction was performed, per COG, on a recently published tree of life [12], rooted between eubacteria and archaea and pruned down to our 163 prokaryotic species.

Whole-genome sizes were reconstructed based on 4873 COGs (containing ~90% of the genes) and 15 460 NOGs (Nonsupervised Orthologous Groups). Our analysis is based on duplication cost $\lambda = 2$ and gain cost $\gamma = 3$ and validated with $\lambda = 1$, $\gamma = 2$, $\lambda = 2$, $\gamma = 4$ and $\lambda = 3$, $\gamma = 4$. We consider patterns that are consistent among reconstructions ranging from ~2% to ~12% of all events in COGs being HGT and the size of the last universal common ancestor (LUCA) ranging from 725 up to 1740 genes.

Scaling of regulators along evolutionary paths

Based on the ancestral genome reconstruction, we studied the evolutionary process leading to the observed quadratic scaling of regulators. To do this, we studied the scaling on the branches of the tree. For each branch on the tree going from an ancestor with R_A regulators and genome size G_A, to a descendant with R_D regulators and genome size G_D, we calculated the scaling exponent $\alpha = \log(R_D/R_A)/\log(G_D/G_A)$ and the change in proportion of regulators $\beta = (R_D/R_A)/(G_D/G_A)$ (see section S4 in the supplementary material online). We could then compare the obtained distribution of α values with the expected distribution obtained from a process where $\alpha = 2$ has been enforced. The expected distribution was calculated from a simulation of genome evolution that replayed the reconstructed history of events stochastically (see section S2 in the supplementary material online). Figure 1a shows the comparison between

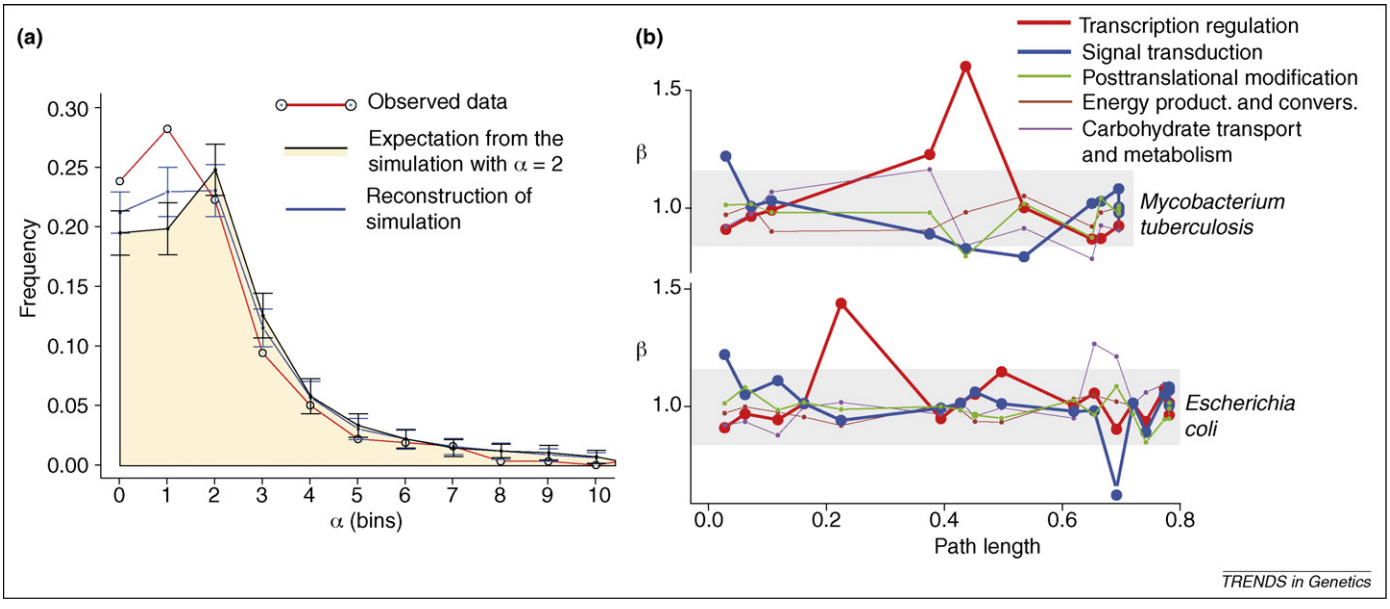

Figure 1. Distribution of scalings along the species tree. **(a)** Distribution of scaling exponents on the tree. The scaling exponents α are calculated per branch as $\alpha = \log(R_D/R_A)/\log(G_D/G_A)$, where R_D and R_A correspond to the number of regulators in the descendant and ancestor of the branch, respectively, and G_D and G_A represent the corresponding genome sizes. This shows that $\alpha = 1$ occurs more often than expected in a simulation with quadratic scaling, and that such a deviation cannot be explained by reconstruction bias alone. Error bars indicate standard deviation. **(b)** β along the evolutionary path of two species. Here β is also calculated for functional categories other than regulation. The figure illustrates that large changes in proportion of regulators and signal transduction genes are localized on few branches. The gray bars contain ~80% of all branches on the tree.

the observed and expected distributions. Surprisingly, we see that $\alpha \sim 1$ occurs in the data more often than expected for a constant quadratic scaling, or for the reconstruction of it. Looking at the changes in proportion along an evolutionary path reveals an interesting pattern that is consistent with the α distribution: large changes in the proportion of regulators and signal transduction genes (categories with quadratic scaling) are localized at a few transitions, whereas the rest of the evolutionary history is comparable to those for other functional categories.

In the case of regulation, when we plot the extreme values of β on the species phylogeny (Figure 2), we see that many of the branches that precede the last common ancestor (LCA) of a major lineage show a large change in the proportion of regulators. Table S1 in the supplementary material online shows that this is an exclusive property of regulation, where the statistical significance of the co-occurrence is at least in the order of 10^{-3} for all studied λ and γ combinations. By contrast, for signal transduction the significance is never below ~10%. The *P* value of the co-

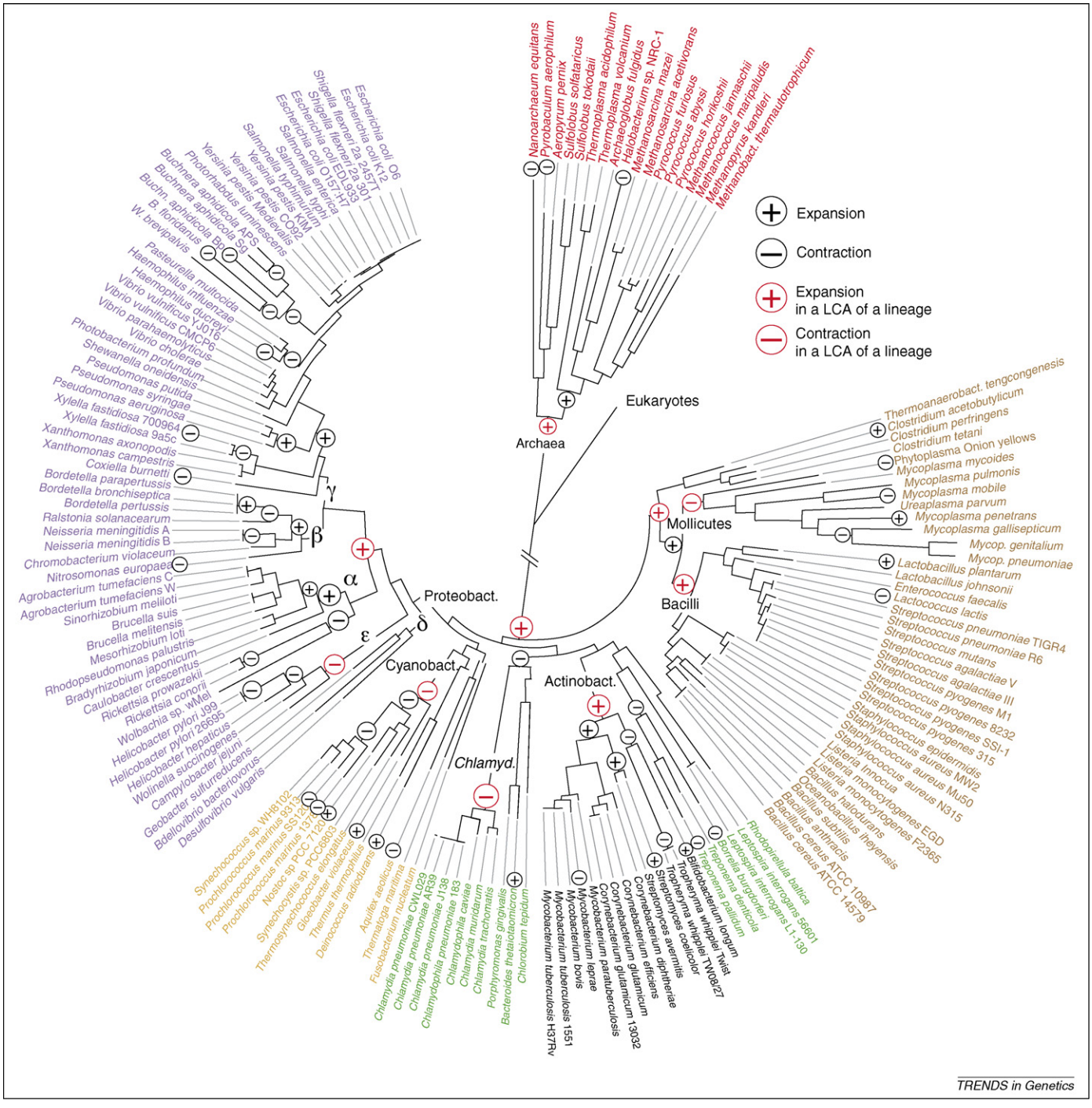

Figure 2. Localization of large changes in proportion of regulators within the species tree. Branches marked with + and − correspond to the 20% largest changes in proportion measured as the 10% tails of the β and β^{-1} distributions (see section S5 in supplementary material online). Points in red correspond to major lineage LCAs as described in Table S1. Projected on the tree of life as published in Ref. [12].

occurrence of major lineage LCAs with large changes in genome size and large number of gains is 0.47 and 0.1, respectively. Thus large changes in the proportion of regulators are better correlated with major lineage LCAs. This leads to the surprising conclusion that at the origin of major lineages a large expansion or contraction of the regulatory machinery occurs.

Scaling of regulators 'slows down' within most phyla

Figure 3a shows that for most major lineages the fitted power law relationship $R = \lambda G^{\theta}$ plotted in log-log scale results in lines with a slope that deviates significantly from the quadratic trend. In fact, the chance of finding an equal or smaller slope for each group is no greater than 5% for all groups except proteobacteria, for which the chance is 57% (see section S8 in the supplementary material online). This is consistent with our previous observation of many $\alpha = 1$ cases and extreme β values localized mostly at the root of the lineage subtrees. Moreover, the fact that the lines occur at different heights also supports our observations and shows that the average ratio of regulators to genome size is lineage specific. In contrast to regulation, for signal transduction (Figure 3b), where the localization of nonlinearities on major lineage LCAs does not hold, we see that the intralineage scalings keep elevated values and that their lines overlap.

A prototype example of the 'slow-down' in the scalings is found in the group of Bacilli, which after a large regulome expansion in their LCA keeps a fixed, large ratio of regulators relative to genome size during their further evolution, in which their genome sizes change considerably. In accordance with this, the fitted scaling factor θ for Bacilli species is 1.06, that is, effectively linear. The major exception to the decrease in scaling exponent within lineages is the proteobacteria group, which has an exponent $\theta = 2.18$, in fact similar to what we found when grouping all bacteria together. This is not only because of the nonlinear transitions near the LCA of the major proteobacterial subdivisions (which are grouped together in Figure 3); superlinear scaling is also seen within these major sublineages. Moreover, their lines occur at the same height, falling on top of the line of the whole group (see section S7 in the supplementary material online). Indeed, as seen in Figure 2, proteobacterial subdivisions contain many internal nonlinear transitions [e.g. split of Pasteurellaceae (containing *Pasteurella multocida*, H*aemophilus influenzae* and *Haemophilus ducreyi*) in the γ subdivision, or Rickettsiales (containing *Rickettsia prowazekii* and *Rickettsia conorii*) in the α subdivision]. Accordingly, the scaling for the major proteobacterial subdivisions is superlinear ($\sim$2). In summary, the results shown in Figure 3 are consistent with the reconstruction results and with our hypothesis that large changes in proportions of regulators occur at the onset of major lineages.

Finally, by simulating genome evolution (see section S2 in supplementary material online) we can test to what extent the different hypotheses predict the observed data. Figure 4 shows that a model where the quadratic scaling has been enforced on the evolutionary process leads to underestimates of the number of regulators and does not show the observed lineage scalings. By contrast, a model of inflationary evolution, with an explosion of regulators at the last common ancestors of major lineages, followed by a scaling of $\sim$2 for proteobacteria and $\sim$1 for all other lineages (an oversimplification of the observation so far reported), corrects the intercepts of the lines in double logarithmic scale without changing the quadratic scaling, leading to results that approximate closer to the data, at the level of the whole dataset, at the level of the lineages and at the level of the α distribution.

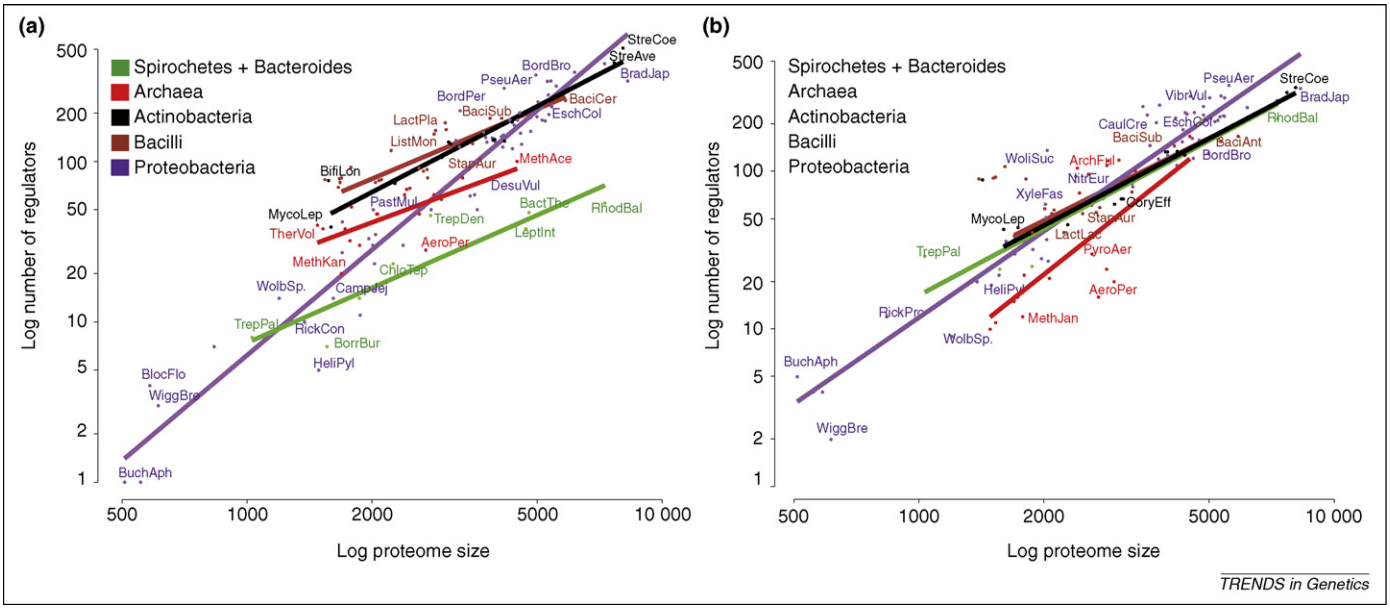

Figure 3. Scaling of regulation and signal transduction for different lineages. Cyanobacteria, mollicutes and chlamydiae were omitted because of insufficient independent data points for a meaningful regression. **(a)** Scaling of regulators. The slopes are: Spirochetes + Bacteroides: 1.13 ± 0.21; Archaea: 0.96 ± 0.24; Actinobacteria: 1.34 ± 0.11; Bacilli: 1.06 ± 0.11; Proteobacteria: 2.18 ± 0.08. **(b)** Scaling of signal transduction genes. We see larger scaling factors and lines at similar heights, in contrast with the case of regulators. The slopes are: Spirochetes + Bacteroides: 1.4 ± 0.22; Archaea: 2.08 ± 0.5; Actinobacteria: 1.38 ± 0.10; Bacilli: 1.31 ± 0.11; Proteobacteria: 1.81 ± 0.07. See sections S7–S10 in supplementary material online for further analysis and a full list of the species abbreviations used.

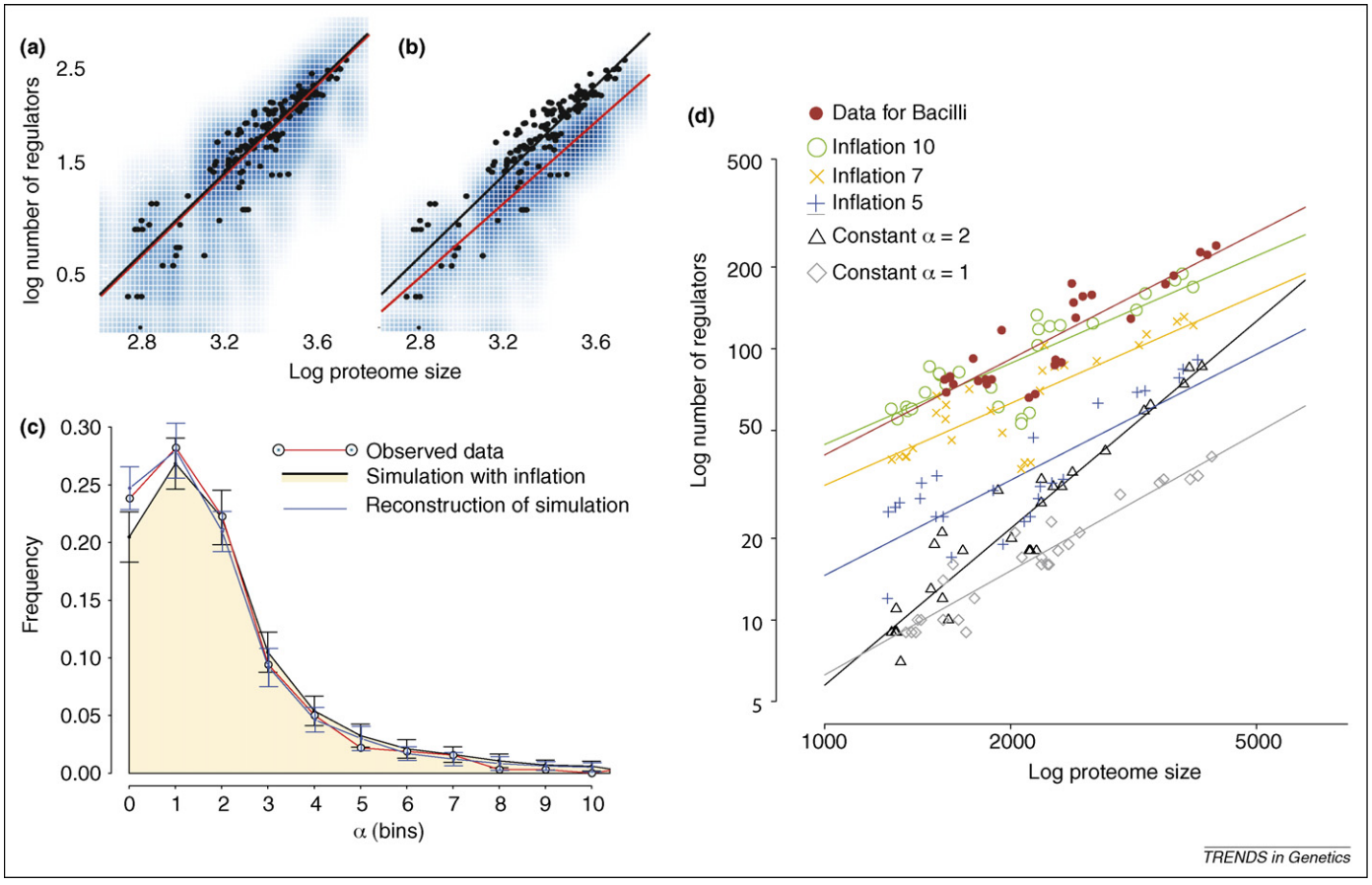

Figure 4. The constant scaling model versus the inflation model. In **(a)** the data are compared against 100 simulations of an inflation model ($\alpha = 7$ at major lineage LCA followed by $\alpha = 1$ everywhere except in proteobacteria), whereas in **(b)** the data are compared against 100 simulations with a constant $\alpha = 2$ scaling. In both panels the red line corresponds to the fitted power law relationship for the simulation and the black line corresponds to the fitted power law relationship for the data. The blue background indicates a smoothed 2D histogram of the simulated species data. We see that the inflation model ($\log_{10}$ error −0.008) reflects the trend observed on the whole dataset much better than the $\alpha = 2$ model ($\log_{10}$ error 0.3, *P* value of such a large error in dataset <0.01), which in fact underestimates the number of regulators. **(c)** The α distribution of the inflation model matches the distribution seen in the data. Expected α distributions are measured directly from the simulation of the inflation model (black shaded curve) and the parsimony reconstruction of it (blue line). **(d)** Example of how the models correspond to data of Bacilli. The graph shows examples of simulated lineage species getting closer to real data for different 'inflation' sizes. Inflation sizes are changed by imposing a large α only on the major lineage LCAs. Here, inflations of size $\alpha = 5$, 7 and 10 are shown.

Concluding remarks

Using comparative genome analysis, Madan Babu *et al.* [13] showed that TFs in prokaryotes change faster in evolution than their target genes, and suggested that organisms might evolve new TFs to sense new signals in the environment. Given our results, this suggests that the evolution of regulators operates in large initial explosions followed by fast sequence divergence.

We have shown that large changes in the ratio of regulators and genome size occurred at the origins of the major prokaryotic lineages. Except in the case of proteobacteria, once established, this ratio is conserved within the lineages despite large changes in genome size. Our results suggest that major transitions in the evolution of prokaryotes are related to changes in regulatory capacity rather than to innovations in the proteome.

Acknowledgements

We thank Berend Snel for useful discussion and suggestions concerning the contents of this manuscript. We are also grateful to the anonymous referees for their important contribution to the final version of the manuscript. This research was funded by the Computational Life Sciences program of the Netherlands Organization for Scientific Research, grant NWO 635.100.001.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2007.07.006.

References

1 Aravind, L. *et al.* (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol. Rev.* 29, 231–262
2 Cases, I. *et al.* (2003) Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol.* 11, 248–253
3 van Nimwegen, E. (2003) Scaling laws in the functional content of genomes. *Trends Genet.* 19, 479–484
4 Konstantinidis, K.T. and Tiedje, J.M. (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 3160–3165
5 van Nimwegen, E. (2005) Scaling laws in the functional content of genomes: Fundamental constants of evolution? In *Power Laws, Scale-Free Networks and Genome Biology* (Koonin, E. *et al.*, eds), pp. 236–253, Springer
6 Ranea, J.A. *et al.* (2005) Microeconomic principles explain an optimal genome size in bacteria. *Trends Genet.* 21, 21–25
7 Mirkin, B.G. *et al.* (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* 3, 2
8 Swofford, D.L. (1998) *Phylogenetic Analysis Using Parsimony (PAUP), Version 4.0b10.* Sinauer

9 Gogarten, J.P. and Townsend, J.P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* 3, 679–687
10 Tatusov, R.L. *et al.* (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33–36
11 Snel, B. *et al.* (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* 28, 3442–3444
12 Ciccarelli, F.D. *et al.* (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311, 1283–1287
13 Madan Babu, M. *et al.* (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J. Mol. Biol.* 358, 614–633

0168-9525/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2007.07.006

AGORA initiative provides free agriculture journals to developing countries

The Health Internetwork Access to Research Initiative (HINARI) of the WHO has launched a new community scheme with the UN Food and Agriculture Organization.

As part of this enterprise, Elsevier has given hundreds of journals to Access to Global Online Research in Agriculture (AGORA). More than 100 institutions are now registered for the scheme, which aims to provide developing countries with free access to vital research that will ultimately help increase crop yields and encourage agricultural self-sufficiency.

According to the Africa University in Zimbabwe, AGORA has been welcomed by both students and staff. ''It has brought a wealth of information to our fingertips'', says Vimbai Hungwe. ''The information made available goes a long way in helping the learning, teaching and research activities within the University. Given the economic hardships we are going through, it couldn't have come at a better time.''

For more information, visit www.aginternetwork.org