Large-scale dynamics of horizontal transfers

Mobile Genetic Elements 2:3, 163-167; May/June, 2012; © 2012 Landes Bioscience

Large-scale dynamics of horizontal transfers

Luigi Grassi,[1] Jacopo Grilli,[2] and Marco Cosentino Lagomarsino[3,4,5,*]

[1]Dipartimento di Fisica, Sapienza Università di Roma; Rome, Italy; [2]Dipartimento di Fisica "G. Galilei" Università di Padova; CNISM and INFN; Padova, Italy; [3]Genomic Physics Group; UMR 7238 CNRS "Microorganism Genomics"; Paris, France; [4]University Pierre et Marie Curie; l'École de Médecine; Paris, France; [5]Dipartimento di Fisica; Università di Torino; Torino, Italy

The widespread exchange of genes between bacteria must have consequences on the global architecture of their genomes, which are being found in the abundant genomic data available today. Most of the expansion of bacterial protein families can be attributed to transfer events, which are positively biased for smaller evolutionary distances between genomes, and more frequent for classes that are larger, when summed over all known bacteria. Moreover, "innovation" events where horizontal transfers carry exogenous evolutionary families appear to be less frequent for larger genomes. This dynamic expansion of evolutionary families is interconnected with the acquisition of new biological functions and thus with the size and distribution of the genes' functional categories found on a genome. This commentary presents our recent contributions to this line of work and possible future directions.

Keywords: horizontal transfers, genomics, gene families, evolution, genome innovation

Submitted: 05/09/12

Accepted: 06/12/12

http://dx.doi.org/10.4161/mge.21112

*Correspondence to:
Marco Cosentino Lagomarsino;
Email: marco.cosentino-lagomarsino@upmc.fr

The current era of fully sequenced genomes and metagenomes confronts us with the challenge and the opportunity of integrating unprecedented amounts of biological data. With such an abundance of data, simplifying views are often useful for figuring out relevant biological phenomena. For example, one can characterize the content of a genome by partitioning it into classes of functional and evolutionary levels. In other words, a genome can be divided in subsets describing its different operative elements, such as genes and their functional and evolutionary, families, transposons and their families, noncoding RNAs, etc. Notably, studies following this approach revealed that some of these elementary features of the functional and evolutionary composition of a genome are often governed by simple quantitative laws.[1]

Considering the protein-coding part of genomes, it is often an advantage to focus on protein domains, rather than full proteins, because they are the building blocks of proteins and they follow similar trends.[2] The sizes of protein/protein-domain families have a fat-tailed distribution[2-7] whose slope depends on genome size.[8] The overall number of families represented by at least one member exhibits a slower than linear scaling trend with the total number of genes in a genome.[8]

Biologically, the growth of evolutionary families derives from combined processes of horizontal gene transfer, gene duplication, gene genesis and gene loss (**Figs. 1 and 2**). For prokaryotes, horizontal gene transfer (HGT), the acquisition of genetic material in a non-hereditary manner, is probably the main innovative force,[9-13] and there are systematic indications that HGT dominates gene family expansion.[14] The same process is presumably very important for the introduction of a new evolutionary family into an extant genome. Accordingly, theoretical models have been proposed that account for the observed power-law distribution of family sizes,[5,6,15-17] mostly using class-expansion/innovation/loss moves, abstractly mimicking basic evolutionary moves such as horizontal transfer, gene duplication and gene loss. A related model, in addition to family size distributions, also explains and successfully fits the scaling of the number of distinct gene families represented in a genome as a function of genome size.[18,19]

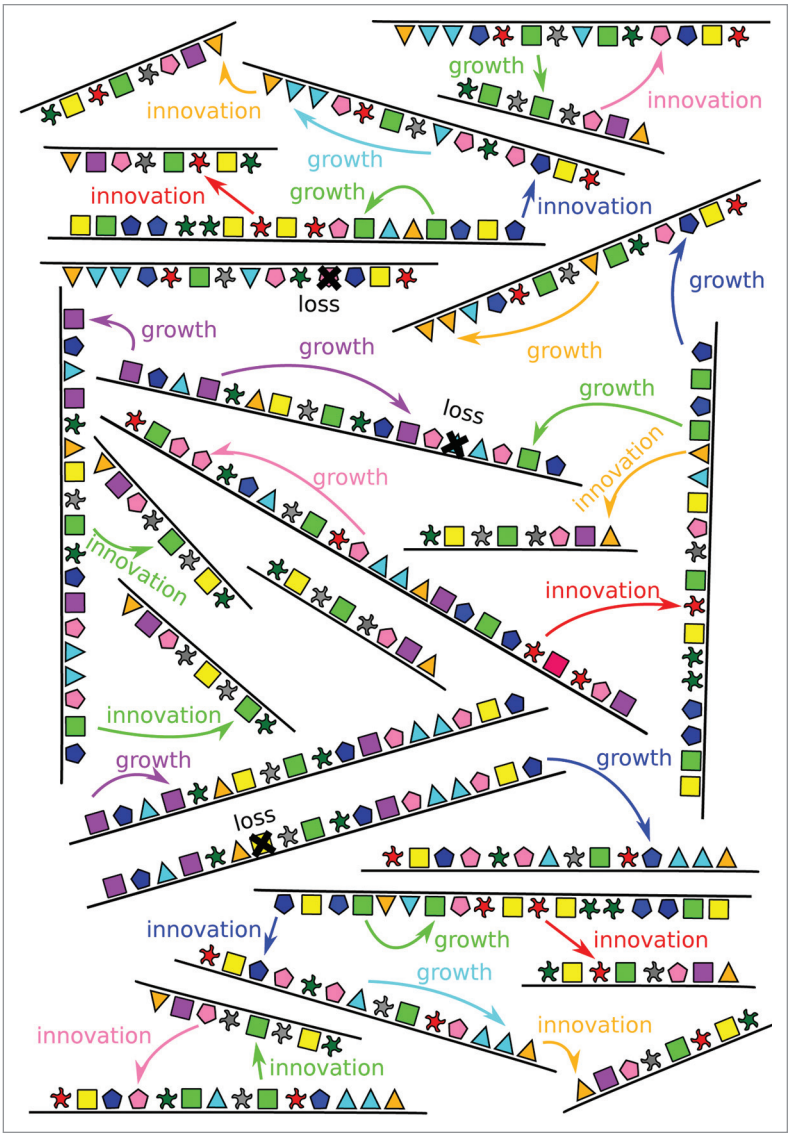

Figure 1. Reports a hypothetical model describing the evolutionary dynamics of protein domains. In this model, horizontal gene transfer can play a double role, on one hand causing the expansion of existing families, and on the other determining "innovation" through the foundation of new families for a specific lineage which did not possess it.

A related important observation is that horizontal gene transfer is reported to be generally biased toward a closer evolutionary lineage with respect to distantly related lineages.[20] This bias in transfer partners can create phylogenetic signals that are similar to shared ancestry but are not due to vertical inheritance. In other words, there exist HGT "exchange groups" of genomes, which are analogous to populations able to exchange alleles by recombination.

Data and models, taken together confronted us with two puzzles. First, the growth of the number of families with proteome size is sublinear, indicating that introduction of new families becomes relatively less likely than class expansion with genome size (both processes being presumably driven by HGT). Second, in order to reproduce the correct tails of the evolutionary family histograms, the models need to introduce a rich-get-richer principle for class expansion, where the probability of adding a new member to the class is proportional to the class size.

Motivated by these questions, we recently performed a detailed analysis[21] of 20 genomes of Proteobacterial species,[22] evaluating the extent to which HGTs expand the genomes' domain repertoires (**Fig. 2**). As a control, we compared our results with those obtained with a database containing HGTs for hundreds of genomes.[23] We used these data to address the two main questions: (1) does a "rich-get-richer" principle hold for genome growth by HGT and (2) do horizontally transferred genes carry novel domains more than expected by chance?

Currently, there is no systematic quantification of how HGT success is correlated with the existing pool of gene classes in a genome. One possibility is that HGT could act effectively as a duplication move in a larger cross-genomic gene family pool (affected by the ecosystems where genes can be exchanged). In some cases, this pool may resemble the genome in question in terms of frequency of gene families, thereby causing a larger class-expansion rate for larger families, but in general this is not necessarily true. Furthermore, HGT may be more likely to be successful for domains that are rare (in the "metagenome" creating the community gene pool or in the receiving genome).

We found that horizontally transferred genes carry domains of exogenous families less frequently for larger genomes, although they might do it more than expected by chance. Additionally, protein domains that are more common in the total pool of genomes appear to have a proportionally higher chance to be transferred. Both features suggest that transfer events behave as if they were drawn randomly from a "cross-genomic" or metagenomic community gene pool, much like gene duplicates are drawn from a genomic gene pool. Since larger genomes will possess more domain classes, the first finding is also in agreement with the observation that the probability of true innovations will be smaller.[8]

Clearly, it is not obvious that the amount of transfers should behave as if they were drawn randomly from a common pool. Other scenarios are possible in which either a decrease of novelty in larger genomes or a rich-gets-richer class-expansion by horizontal transfer, or both, are not trivially expected. For example, gene gains could be sampled from a very large effective pool of families, or horizontal transfers could be dominated by specific

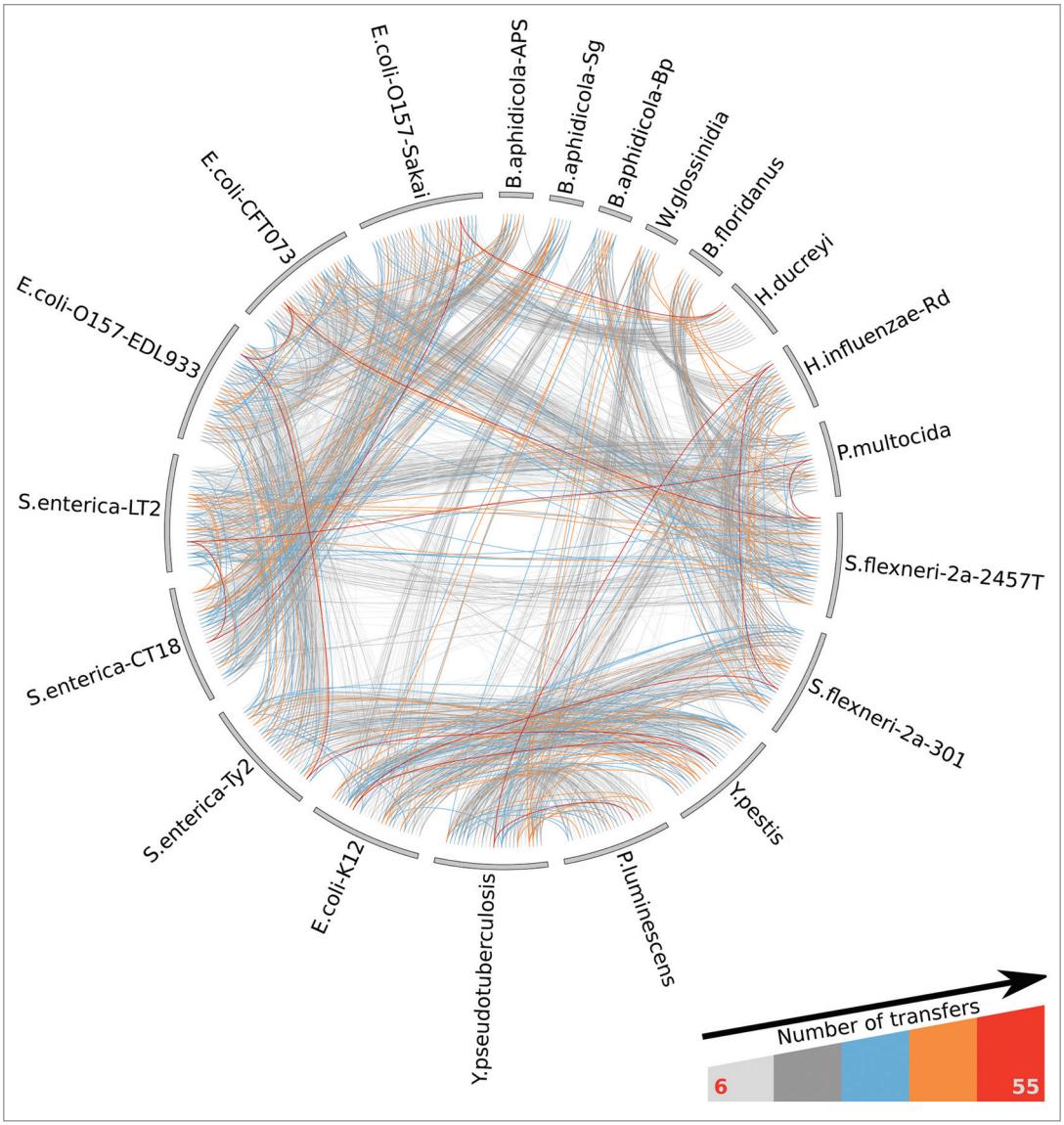

Figure 2. A representation of all species examined in reference 21 Each bar on the outer circle represents a studied genome and links represent protein domains. Different genomes are connected if they share a domain subject to HGT (in the cross-genomic gene pool formed by the union of the analyzed genomes). The color of the links reflects the number of transfers as shown in the legend.

ecological or functional mechanisms. We know that transfers are not random; protein length plays a role, for example,[24] and it is natural to expect that selective pressure favors the acquisition of specific traits.

However, our results suggest that when averaging over many transfer events there is a large contribution of purely combinatorial and statistical aspects to the "emergent" overall distributions of HGTs and their contribution to protein families, as typically happens in systems of many agents (such as crowds, the stock exchange, species in ecosystems[25-28]). In these cases the analysis tools and the modeling frameworks of statistical physics may prove useful, as they were developed having in mind closely related phenomenologies in physical and chemical systems.

Notably, since the class sizes within a single genome are similar to the corresponding sizes in the cross-genome gene pool, this also has the consequence that classes should grow according to a rich-get-richer principle. The latter has often been assumed, but is not justified in current models.[2,14,29] For gene duplications, a rich-get-richer principle follows from the null assumption that all genes of a given class are a priori equally likely to get duplicated. However, as we discussed, prokaryotes tend to add genes by HGT rather than by gene duplication.[11,12,14] This behavior also affects the statistics of (domain) functional categories, which in the case of domains are typically made of the sum of a number of evolutionary classes, and empirically grow as power laws[30] with genome size at a specific rate, termed "evolutionary potential."[2]

The joint partitioning of genes into functional and evolutionary classes also shows relevant universal quantitative trends,[29] and is connected to genome innovation by horizontal transfer.

Presumably, addition of new genes needs to follow correlated functional "recipes" where genes whose functions are related are added together. For example having in mind the classic "operon model"[31] (the general model of bacterial transcription control that partitions genes into specific regulatory genes and responding to metabolic cues, environment, and "structural" target genes performing specific tasks) it can be stated that addition of transcription factors needs to be related to the addition of a set of metabolic enzymes that are related by common metabolic pathways. The consequences of these statements have been explored recently using an integrated approach of data analysis and models,[32,33] and appear to explain very well the observed quantitative relationship between transcription factors and metabolic pathways, despite the fact that this might be subject to other constraints.[34] However, we still know very little about the nature and the very existence of these recipes, and gathering new insight into the process of how a prokaryotic genome builds new functions will be important for future studies, with evident implications for the applicability of synthetic biology.

The approach followed in our study disregarded relevant ecological aspects, which will be important to explore in future studies, such as population sizes, by assuming that a given domain has a certain occurrence just based on genomic sequences. For a specific ecosystem, total domain occurrence should ideally be derived from a weighted average, where weights are empirical population sizes. Results from individual ecosystems should then be averaged over all the ecosystems concerning the considered set of species, weighted by their relevance in evolutionary terms. We believe this can be addressed in future studies, as data of this kind starts to become available.[35] Overall, we believe there are great potentials and great unmet challenges in genomics and metagenomics studies addressing the reciprocal roles of ecology and evolution.[36]

Acknowledgments

We thank M.J. Lercher, S. Maslov, M. Caselle, F. Bassetti and B. Bassetti for useful discussions.

References

1. Koonin EV. Are there laws of genome evolution? PLoS Comput Biol 2011; 7:e1002173; PMID:21901087; http://dx.doi.org/10.1371/journal.pcbi.1002173
2. Molina N, van Nimwegen E. The evolution of domain-content in bacterial genomes. Biol Direct 2008; 3:51; PMID:19077245; http://dx.doi.org/10.1186/1745-6150-3-51
3. Huynen MA, van Nimwegen E. The frequency distribution of gene family sizes in complete genomes. Mol Biol Evol 1998; 15:583-9; PMID:9580988; http://dx.doi.org/10.1093/oxfordjournals.molbev.a025959
4. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature 2002; 420:218-23; PMID:12432406; http://dx.doi.org/10.1038/nature01256
5. Dokholyan NV, Shakhnovich B, Shakhnovich EI. Expanding protein universe and its origin from the biological Big Bang. Proc Natl Acad Sci U S A 2002; 99:14132-6; PMID:12384571; http://dx.doi.org/10.1073/pnas.202497999
6. Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol 2001; 313:673-81; PMID:11697896; http://dx.doi.org/10.1006/jmbi.2001.5079
7. Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biol 2002; 3:H0040, h0040, 7; PMID:12186647; http://dx.doi.org/10.1186/gb-2002-3-8-research0040
8. Cosentino Lagomarsino M, Sellerio AL, Heijning PD, Bassetti B. Universal features in the genome-level evolution of protein domains. Genome Biol 2009; 10:R12; PMID:19183449; http://dx.doi.org/10.1186/gb-2009-10-1-r12
9. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 1999; 96:3801-6; PMID:10097118; http://dx.doi.org/10.1073/pnas.96.7.3801
10. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000; 405:299-304; PMID:10830951; http://dx.doi.org/10.1038/35012500
11. Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res 2008; 36:6688-719; PMID:18948295; http://dx.doi.org/10.1093/nar/gkn668
12. Pál C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 2005; 37:1372-5; PMID:16311593; http://dx.doi.org/10.1038/ng1686
13. Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol 2005; 3:679-87; PMID:16138096; http://dx.doi.org/10.1038/nrmicro1204
14. Treangen TJ, Rocha EP. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet 2011; 7:e1001284; PMID:21298028; http://dx.doi.org/10.1371/journal.pgen.1001284
15. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV. Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol Biol 2002; 2:18; PMID:12379152; http://dx.doi.org/10.1186/1471-2148-2-18
16. Kamal M, Luscombe N, Qian J, Gerstein M. 2006; in *Power Laws, Scale-Free Networks and Genome Biology*, eds Koonin E, Wolf Y, Karev G (Spinger, New York), pp 165–193. [17] Durrett R, Schweinsberg J Power laws for family sizes in a duplication model. *Ann. Probab.* 2005; 33:2094–2126
18. Cosentino Lagomarsino M, Sellerio AL, Heijning PD, Bassetti B. Universal features in the genome-level evolution of protein domains. Genome Biol 2009; 10:R12; PMID:19183449; http://dx.doi.org/10.1186/gb-2009-10-1-r12
19. Angelini A, Amato A, Bianconi G, Bassetti B, Cosentino Lagomarsino M. Mean-field methods in evolutionary duplication-innovation-loss models for the genome-level repertoire of protein domains. Phys Rev E Stat Nonlin Soft Matter Phys 2010; 81:021919; PMID:20365607; http://dx.doi.org/10.1103/PhysRevE.81.021919
20. Andam CP, Gogarten JP. Biased gene transfer in microbial evolution. Nat Rev Microbiol 2011; 9:543-55; PMID:21666709; http://dx.doi.org/10.1038/nrmicro2593
21. Grassi L, Caselle M, Lercher MJ, Lagomarsino MC. Horizontal gene transfers as metagenomic gene duplications. Mol Biosyst 2012; 8:790-5; PMID:22218456; http://dx.doi.org/10.1039/c2mb05330f
22. Lercher MJ, Pál C. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol 2008; 25:559-67; PMID:18158322; http://dx.doi.org/10.1093/molbev/msm283
23. Garcia-Vallve S, Guzman E, Montero MA, Romeu A. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res 2003; 31:187-9; PMID:12519978; http://dx.doi.org/10.1093/nar/gkg004
24. Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 2005; 3:711-21; PMID:16138099; http://dx.doi.org/10.1038/nrmicro1234
25. de Vladar HP, Barton NH. The contribution of statistical physics to evolutionary biology. Trends Ecol Evol 2011; 26:424-32; PMID:21571390; http://dx.doi.org/10.1016/j.tree.2011.04.002
27. Castellano C, Fortunato S, Loreto V. Statistical physics of social dynamics. Rev Mod Phys 2009; 81:591-646; http://dx.doi.org/10.1103/RevModPhys.81.591
28. Banavar JR, Maritan A, Volkov I. Applications of the principle of maximum entropy: from physics to ecology. J Phys Condens Matter 2010; 22:063101; PMID:21389359; http://dx.doi.org/10.1088/0953-8984/22/6/063101
29. Grilli J, Bassetti B, Maslov S, Cosentino Lagomarsino M. Joint scaling laws in functional and evolutionary categories in prokaryotic genomes. Nucleic Acids Res 2012; 40:530-40; PMID:21937509; http://dx.doi.org/10.1093/nar/gkr711
30. van Nimwegen E. Scaling laws in the functional content of genomes. Trends Genet 2003; 19:479-84; PMID:12957540; http://dx.doi.org/10.1016/S0168-9525(03)00203-8
31. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 1961; 3:318-56; PMID:13718526; http://dx.doi.org/10.1016/S0022-2836(61)80072-7
32. Maslov S, Krishna S, Pang TY, Sneppen K. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. Proc Natl Acad Sci U S A 2009; 106:9743-8; PMID:19482938; http://dx.doi.org/10.1073/pnas.0903206106
33. Pang TY, Maslov S. A toolbox model of evolution of metabolic pathways on networks of arbitrary topology. PLoS Comput Biol 2011; 7:e1001137; PMID:21625566; http://dx.doi.org/10.1371/journal.pcbi.1001137
34. Itzkovitz S, Tlusty T, Alon U. Coding limits on the number of transcription factors. BMC Genomics 2006; 7:239; PMID:16984633; http://dx.doi.org/10.1186/1471-2164-7-239

35. McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, et al. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol Lett 2007; 10:995-1015; PMID:17845298; http://dx.doi.org/10.1111/j.1461-0248.2007.01094.x

36. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics 2011; 12:32; PMID:21232151; http://dx.doi.org/10.1186/1471-2164-12-32