PERIODICUM BIOLOGORUM
VOL. 121–122, No 1–2, 3–6, 2020
DOI: 10.18054/pb.v121-122i1-2.10737

UDC 57:61
CODEN PDBIAD
ISSN 0031-5362

Spiridion Brusina Lecture 2020

Viewpoint

Evolutionary age of genes can assist in genome mining

IVAN MIJAKOVIC[1,2]

[1] Division of Systems and Synthetic Biology, Department of Biology and Biological Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden

[2] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Lyngby, Denmark

Correspondence:
ivan.mijakovic@chalmers.se

Received April 20, 2020
Revised May 19, 2020
Accepted May 21, 2020.

Abstract

The rate of sequencing microbial genomes is accelerating, with the hope of discovering new antibiotics, cures for various diseases or new industrial enzymes. However, about 25-30% of the genes in the sequenced microbial genomes do not have an assigned function. Predicting the functions of these "unknown" genes could unlock a considerable biological potential for biomedical and biotechnology applications, as well as further our understanding of the molecular tenets of life. Current methods for gene mining rely basically on comparison of primary sequences or 3D-structures to those of already characterized genes. The problem with such approaches is that unknown genes with no homology to the already characterized genes remain completely out of reach. Herein, I argue that evolutionary approaches, such as the genomic phylostratigraphy, can make a substantial contribution to genome mining – especially regarding genes with no homology to the characterized ones. My group has recently used genomic phylostratigraphy to discover new genes involved in sporulation of the bacterial model organism Bacillus subtilis. *These new sporulation genes exhibited no sequence homology with the known sporulation genes and were missed by all other genome mining approaches. They have been discovered solely based on their evolutionary age. Along these lines, I argue that phylostratigraphy should be integrated into genome mining pipelines and develop a brief example of how this could be done.*

Genome sequencing technologies have revolutionized modern life science and medicine *(1)*. Single cell genomic technologies have led to establishment of a human tumor atlas *(2)*, a platform that can be used for providing an unprecedented level of precision-medicine treatments for cancer patients. Sequencing technologies are extensively used in the humanity's combat against antibiotic-resistant bacteria, to increase our ability to detect and study antimicrobial resistance *(3)*. In the PubMed genome database, there are presently 246 954 fully sequenced bacterial genomes *(4)*, of which about 50 000 have been added in the last 12 months. Genomes of microbes isolated from various environments are getting sequenced at an increasing rate, in the hope that this biodiversity will help us find new antibiotics, cures for various diseases, new industrial enzymes, producers of biofuels and various other functions useful to humanity. However, about 25-30% of all the genes in the sequenced microbial genomes are still of unknown function. Predicting the functions of these "unknown" genes could unlock a tremendous potential for understanding the molecular tenets of life, as well as for biomedical and biotechnology applications. Methods currently used for discovering gene functions, i.e. for genome mining, are largely based on analyses of primary sequences or 3D-structures and searching for homology with already characterized

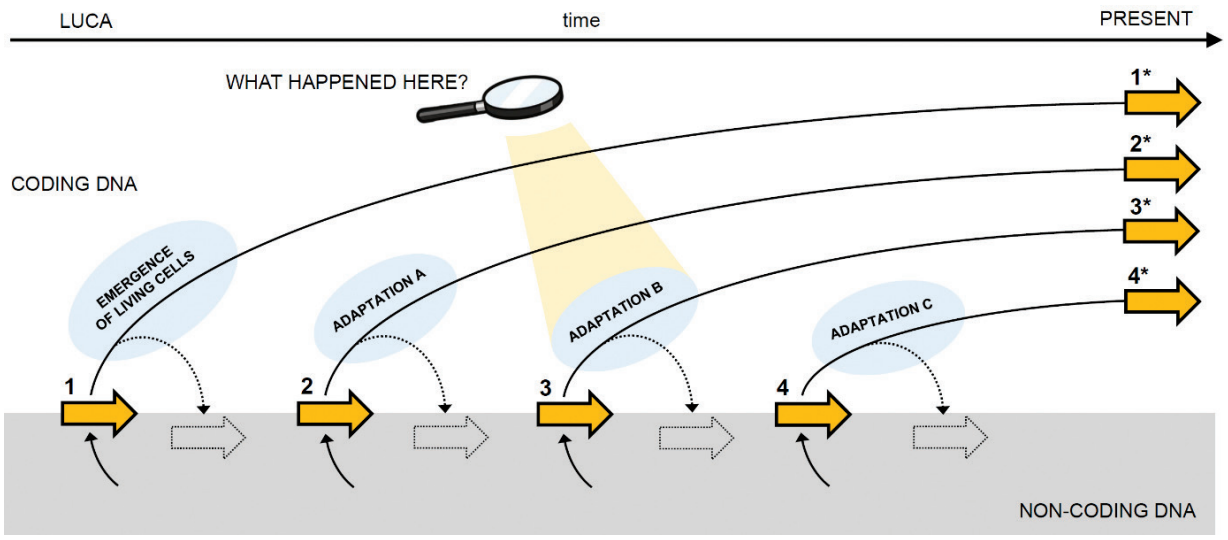

Figure 1. The concept of genomic phylostratigraphy and discovery of gene function. *Non-coding DNA space is represented with a grey box, and the coding DNA is shown as the white space above. The timeline spans the evolutionary history from emergence of the last universal common ancestor (LUCA) to extant species. A certain set of genes (orange arrow 1) emerged ("gene birth") with LUCA and kept evolving via adaptive evolution until present times (1*). At different time-points throughout evolution, major adaptations took place (A, B, C), which led to the emergence of new sets of genes (orange arrows 2, 3, 4). These new genes then also underwent adaptive evolution, and kept evolving until present times (2*, 3*, 4*). Each emerging gene that did not survive biological selection mutated into a pseudogene and "sunk" again into the non-coding DNA space (process of "gene death" shown by dotted arrows). The essence of our approach is shown by the magnifying glass: if a certain gene emerged at the time point of a specific major adaptation, does it have a higher probability to be functionally involved in that adaptation?*

genes *(5,6)*. Such approaches have a serious limitation: completely "unknown" genes, with no homology shared with the "known" genes, remain completely out of reach. By consequence, current genome mining technologies have advanced the discovery of new biosynthetic pathways only very modestly. The focus in this respect has so far been on connecting available genomics (homology-based methods) and metabolomics data *(7)* and expressing silent biosynthetic clusters *(8)*. Clearly, conceptually new methods for discovering gene functions are needed.

Recently, my group published a study where we used genomic phylostratigraphy *(9)* to discover new genes involved in sporulation of the bacterial model organism *Bacillus subtilis (10)*. To introduce the concept of genomic phylostratigraphy, let us first consider that genes have a "life cycle", as proposed by Neme and Tautz *(11)* (Figure 1). This means that non-coding DNA sequences within genomes can be altered by random mutations into coding sequences. By starting to code for proteins, these genes are no longer subject to only stochastic evolution (random mutations). This is the step of "gene birth", shown in Figure 1 as emergence of genes and their transition to a status of coding sequence. The mutations in those genes are now subject also to natural selection, and this process is called adaptive evolution. Through further random mutations and natural selection, functions can be gained, modified or lost. Mutations can also lead to "gene death". A gene which has lost its function "sinks" again from adaptive into stochastic evolution, shown as the non-coding sequence space in Figure 1. Since genes have a life cycle, with "gene birth" and "gene death", the evolutionary tree of life will contain some genes that are very old (e.g. those that emerged in the last universal common ancestor - LUCA), and other genes that have gradually been appearing throughout evolution, some of which very recently. If this description of adaptive evolution is generally true, then one should be able to correlate emergence of new genes to emergence of new features of life, i.e. new functions brought about by adaptation. Phylostratigraphy is a computational method for studying genome evolution and it operates with a timeline defined by an evolutionary tree. As a first step, a consensus phylogeny from the perspective of the species to be analyzed is constructed following phylogenetic literature. Each node in the evolutionary tree is named a phylostratum, by analogy to strata (layers) found in paleontology, where deeper layers of soil contain older fossils. Phylostratigraphy analysis can be performed on an entire genome and it classifies individual genes into phylostrata, each populated with genes whose founder genes emerged at a specific node in a reference evolutionary tree. In a phylostratigraphy map, all the genes from a given genome get distributed into phylostrata. So far, phylostratigraphy approach has been successfully used to explain macro-evolutionary phenomena. If a function emerged at a given time-point in evolution, genes related to that function tend to be significantly enriched in the corresponding phylostratum. Based on this principle, phylostratigraphy maps have been used e.g. to trace the evolutionary origin of over 1000 cancer-related genes to the evolutionary time-point of emergence of

Metazoa *(12)*, for evolutionary stratification of developmental phenomena *(13)* and for studying the emergence of new genes from non-genic regions *(14)*.

In our study *(10)*, we went beyond using phylostratigraphy to only explain evolutionary events, we used it to predict functions of completely uncharacterized genes, based solely on their evolutionary age. We commenced by performing a genome-wide phylostratigraphy analysis of *B. subtilis (15)*. This revealed that most of the >300 genes known to be involved in development of *B. subtilis* spores (extremely resistant forms of bacterial cells) tend to cluster in only a few phylostrata *(10)*. Put simply, this means that sporulation most probably first emerged in phylostratum 8, corresponding to separation of Bacilli from other bacterial genera. The process co-opted a certain number of older genes (phylostrata 1 and 2) and was modified/enhanced with addition of new genes/functions in phylostrata 9, 10 and 13. Based on this, we hypothesized that genes of still unknown function belonging to sporulation-enriched phylostrata have a higher probability to be involved in sporulation, compared to genes of unknown function in other phylostrata. Next, we individually knocked out (inactivated) a selection of genes of unknown function in the concerned phylostrata. Then we examined whether the resulting strains with gene knockouts can sporulate normally. 16 out of 37 (43%) of tested knockout strains exhibited a sporulation phenotype. The 16 newly discovered sporulation genes have been shown to participate in transcriptional regulation of sporulation functions, maintaining structural integrity of the spore coat and signaling between the forespore and the mother cell. These results clearly confirmed that phylostratigraphy can indeed be used to predict genes involved in sporulation with considerable success. It should be emphasized that the newly discovered sporulation genes bore no sequence homology with the known sporulation genes. They were missed by all other genome mining approaches and were identified solely based on their evolutionary age. Based on these findings, I would like to argue that phylostratigraphy should be integrated into genome mining pipelines, e.g. when searching for uncharacterized biosynthetic genes in microbial genomes.

In the following paragraph I will outline how this integration could work, by describing a putative computational workflow to predict biosynthetic gene clusters that would integrate evolutionary history, genomic, transcriptomic and metabolomic data (Figure 2). Firstly, the biosynthetic gene clusters of bacterial isolates should be identified using antiSMASH *(16)*. The resulting biosynthetic gene clusters should be clustered into gene families. A BLAST-based phylostratigraphic approach should then be employed to estimate the evolutionary age of biosynthetic gene cluster families and to assess their clustering along evolutionary age *(13)*. Each biosynthetic gene will be assigned to a phylostratum (PS), representing the oldest phylogenetic node to which the gene can be traced. With the predicted age of each gene, the evolutionary transcriptome age of any specific growth condition can then be inferred by the transcriptome age index methods (the weighted sum of genes ranked by their expression level in each phylostratum). With available metabolomics data, the sequence similarity networks should then be associated to chemical similarity networks of natural products, which provide a link between gene families and natural products. Finally, this information should be used to pre-

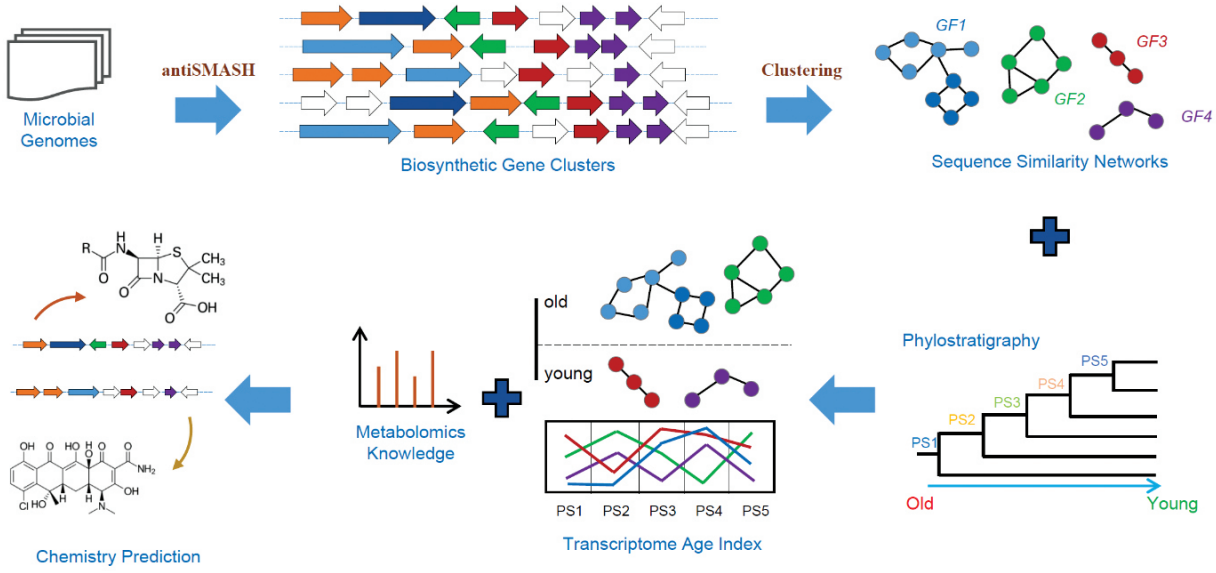

Figure 2. Proposed pipeline for systematic exploration of biosynthesis gene clusters from large-scale omics data. *Biosynthetic gene clusters identified by antiSMASH are clustered into gene families (GFs) and then stratified into phylostrata (PS) according to evolutionary age. Transcriptome age index and metabolomics data are used to link GFs to chemical similarity networks, leading to a functional link to specific natural products, e.g. novel antibiotics.*

dict the potential products of biosynthetic gene clusters in microbial isolates using machine learning methods by integration of gene similarity, gene age and transcriptional information.

It is exciting to imagine phylostratigraphy and other evolutionary methods as components of future genome mining pipelines. It should be mentioned that some authors claimed that phylostratigraphy tends to underestimate age of certain categories of genes, such as short coding sequences with fast evolution rates *(17)*. To address this putative bias of phylostratigraphy, a strategy that involves exclusion of error-prone genes has been proposed *(18)*. This and other limitations should be kept in mind when implementing phylostratigraphy in genome mining. It could also be argued that phylostratigraphy has been developed for species with predominant vertical evolution, and one could speculate that extensive lateral gene transfer known to occur in bacteria might limits its use *(19)*. However, the success of phylostratigraphy in predicting new sporulation genes in *B. subtilis (10)* may be taken as a counter argument. In conclusion, evolutionary biology has much to offer and all biologists would be well advised to keep that in mind. For example, experimental evolution *(20)* has been used with notable success in metabolic engineering and synthetic biology *(21)*. Hopefully, phylostratigraphy-based approaches will have similarly broad impact on other fields of biology.

Acknowledgments: *In preparation of this manuscript IM was supported by funding from the Novo Nordisk foundation (grant NNF10CC1016517) and the Danish Research Council DFF (grant 9040-00075A). Dr. Lei Shi is acknowledged for help in preparing Figure 2.*

REFERENCES

1. LAPPALAINEN T, SCOTT AJ, BRANDT M, HALL IM 2019 Genomic analysis in the age of human genome sequencing. Cell 177: 70–84. http://dx.doi.org/10.1016/j.cell.2019.02.032
2. ROZENBLATT-ROSEN O, REGEV A, OBERDOERFFER P, NAWY T, HUPALOWSKA A, ROOD JE, ASHENBERG O, CERAMI E, COFFEY RJ, DEMIR E, DING L, ESPLIN ED, FORD JM, GOECKS J, GHOSH S, GRAY JW, GUINNEY J, HANLON SE, HUGHES SK, HWANG ES, IACOBUZIO-DONAHUE CA, JANÉ-VALBUENA J, JOHNSON BE, LAU KS, LIVELY T, MAZZILLI SA, PE'ER D, SANTAGATA S, SHALEK AK, SCHAPIRO D, SNYDER MP, SORGER PK, SPIRA AE, SRIVASTAVA S, TAN K, WEST RB, WILLIAMS EH 2020 The human tumor atlas network: charting tumor transitions across space and time at single-cell resolution. Cell 181: 236–249. http://dx.doi.org/10.1016/j.cell.2020.03.053
3. BOOLCHANDANI M, D'SOUZA AW, DANTAS G 2019 Sequencing-based methods and resources to study antimicrobial resistance. Nat Rev Genet 20: 356–370. http://dx.doi.org/10.1038/s41576-019-0108-4
4. https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/
5. GILES TC, EMES RD 2017 Inferring function from homology. Methods Mol Biol 1526: 23–40. http://dx.doi.org/10.1007/978-1-4939-6613-4_2
6. WEI Q, MCGRAW J, KHAN I, KIHARA D 2017 Using PFP and ESG protein function prediction web servers. Methods Mol Biol 1611: 1–14. http://dx.doi.org/10.1007/978-1-4939-7015-5_1
7. MEDEMA MH, FISCHBACH MA 2015 Computational approaches to natural product discovery. Nat Chem Biol 11: 639–648. https://doi.org/10.1038/nchembio.1884
8. FOULSTON L 2019 Genome mining and prospects for antibiotic discovery. Curr Opin Microbiol 51: 1–8. http://dx.doi.org/10.1016/j.mib.2019.01.001
9. DOMAZET-LOSO T, BRAJKOVIĆ J, TAUTZ D 2007 A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet 23: 533–539. http://dx.doi.org/10.1016/j.tig.2007.08.014
10. SHI L, DEROUICHE A, PANDIT S, RAHIMI S, KALANTARI A, FUTO M, RAVIKUMAR V, JERS C, MOKKAPATI VRSS, VLAHOVIČEK K, MIJAKOVIC I 2020 Evolutionary analysis of the *Bacillus subtilis* genome reveals new genes involved in sporulation. Mol Biol Evol pii: msaa035. http://dx.doi.org/10.1093/molbev/msaa035
11. NEME R, TAUTZ D 2014 Evolution: dynamics of de novo gene emergence. Curr Biol 24: R238–R240. http://dx.doi.org/10.1016/j.cub.2014.02.016
12. DOMAZET-LOSO T, TAUTZ D 2010 Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biol 8: 66. https://doi.org/10.1186/1741-7007-8-66
13. DOMAZET-LOSO T, TAUTZ D 2010 A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature 468: 815–818. http://dx.doi.org/10.1038/nature09632
14. CARVUNIS AR, ROLLAND T, WAPINSKI I, CALDERWOOD MA, YILDIRIM MA, SIMONIS N, CHARLOTEAUX B, HIDALGO CA, BARBETTE J, SANTHANAM B, BRAR GA, WEISSMAN JS, REGEV A, THIERRY-MIEG N, CUSICK ME, VIDAL M 2012 Proto-genes and de novo gene birth. Nature 487: 370–374. http://dx.doi.org/10.1038/nature11184
15. RAVIKUMAR V, NALPAS NC, ANSELM V, KRUG K, LENUZZI M, ŠESTAK MS, DOMAZET-LOŠO T, MIJAKOVIC I, MACEK B 2019 In-depth analysis of *Bacillus subtilis* proteome identifies new ORFs and traces the evolutionary history of modified proteins. Sci Rep 8: 17246. http://dx.doi.org/10.1038/s41598-018-35589-9
16. MEDEMA MH, BLIN K, CIMERMANCIC P, DE JAGER V, ZAKRZEWSKI P, FISCHBACH MA, WEBER T, TAKANO E, BREITLING R 2011 antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39: W339–W346. http://dx.doi.org/10.1093/nar/gkr466
17. MOYERS BA, ZHANG J 2017 Further simulations and analyses demonstrate open problems of phylostratigraphy. Genome Biol Evol 9: 1519–1527. https://doi.org/10.1093/gbe/evx109
18. MOYERS BA, ZHANG J 2018 Toward reducing phylostratigraphic errors and biases. Genome Biol Evol 10: 2037–2048. http://dx.doi.org/10.1093/gbe/evx109
19. BHANDARI V, NAUSHAD HS, GUPTA RS 2012 Protein based molecular markers provide reliable means to understand prokaryotic phylogeny and support Darwinian mode of evolution. Front Cell Infect Microbiol 2: 98. http://dx.doi.org/ 10.3389/fcimb.2012.00098
20. GOOD BH, MCDONALD MJ, BARRICK JE, LENSKI RE, DESAI MM 2017 The dynamics of molecular evolution over 60,000 generations. Nature 551: 45–50. http://dx.doi.org/10.1038/nature24287
21. PORTNOY VA, BEZDAN D, ZENGLER K 2011 Adaptive laboratory evolution--harnessing the power of biology for metabolic engineering. Curr Opin Biotechnol 22: 590–594. http://dx.doi.org/10.1016/j.copbio.2011.03.007