The evolutionary space of bacterial 16S rRNA gene

I. Korvigo[1], E. Pershina[1], A. Igolkina[1], A. Dolnik[2], G. Tamazyan[2], Y. Porozov[3]

[1]*All-Russia Research Institute for Agricultural Microbiology, St. Petersburg, Russia,*
ilia.korvigo@gmail.com

E. Andronov[1]

[2]*Saint Petersburg State University, St. Petersburg, Russia*

[3]*St. Petersburg National University of Information Technologies Mechanics and Optics, St. Petersburg, Russia*

Genetic sequence analyses have infiltrated biology to such an extent that there probably is not a single field left untouched, while the impact of this integration has been dramatic for many of them. First of all, intensive accumulation of environmental NGS data had revolutionized microbiology, particularly microbial ecology, by uncovering uncultured organisms that had hitherto been hidden [1]. The uncovered biodiversity was unparalleled in sheer numbers and possible physiological traits rendering the established phenotypic classification system inapplicable [2]. This resulted in a global shift from phenotypic to genotypic systematics relying on marker-gene sequences clustered into OTUs (Operational Taxonomic Units) upon defined sequence identity thresholds, with 16S rRNA being the de facto marker of choice. Although this approach effectively addresses aforementioned issues [3], its own caveats - mainly the requirement of sequence realignment once new data are added - lead to ever-increasing computational complexity [4] and results ambiguity, since clusters might be reshuffled. On the other hand sequence analysis unleashed molecular phylogenetics making evolutionary studies of morphologically uniform microorganisms possible. Once again, despite huge success of phylogenetics and tremendous leaps in tree-construction algorithms, classic tree-based evolutionary analysis is principally limited to pedigree reconstruction while having nothing to say about the process of evolution itself, e.g. it doesn't consider the process of extinction. Several attempts have already been made to recover evolutional patterns by adding dimensionality to phylogenetic reconstructions by means of multiple dimension scaling [5], yet such attempts suffer from inherited computational limitations [6]. It's obvious that both research areas could benefit of a united

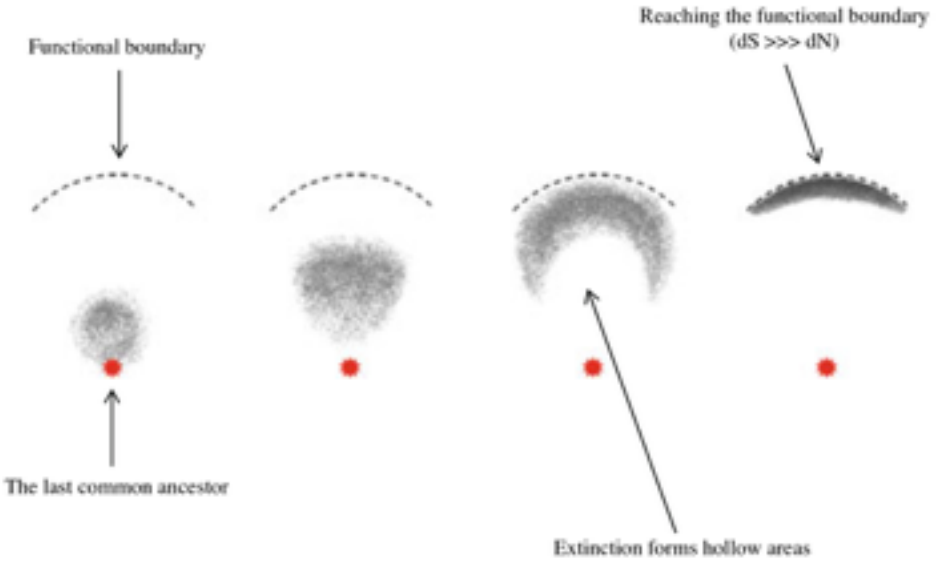

Figure 1. A sequence evolution model as reflected in the ES patterns. Unequal rates of direct and reverse evolution lead to sequence's space inflation up to a functional boundary of the sequence.

classification system where different states of a sequence could be unambiguously mapped in constant time. And here we propose a prototype of such a system we call the Evolutionary Space (ES) [7].

We develop the ES in order to visualize the variability of all currently existing and theoretically possible 16S rRNA sequences and to identify obscure evolutionary patterns behind enormous natural diversity of this gene. ES is a multidimensional space where each point represents a 16S rRNA sequence. We used the theory of regular simplexes to define spatial properties of the space (e.g. the number of dimensions/vertices of a simplex) so that it could fulfill our plans. The vertices represent a selection of sequence with pair-wise genetic distances meeting the definition of simplex edges. The result is a stable 13-dimensional sequence space invariant in regard to data mapped inside of it. Putting a random sequence inside the space takes roughly constant time and doesn't affect other sequences. In our experiment mapping a big 16S database in the ES uncovered a set of evolutionary patterns representing the gene's natural history (Fig. 1), including internal hollows left of ancient ancestral sequences (a side affect of the sequence's space inflation caused by unequal forward

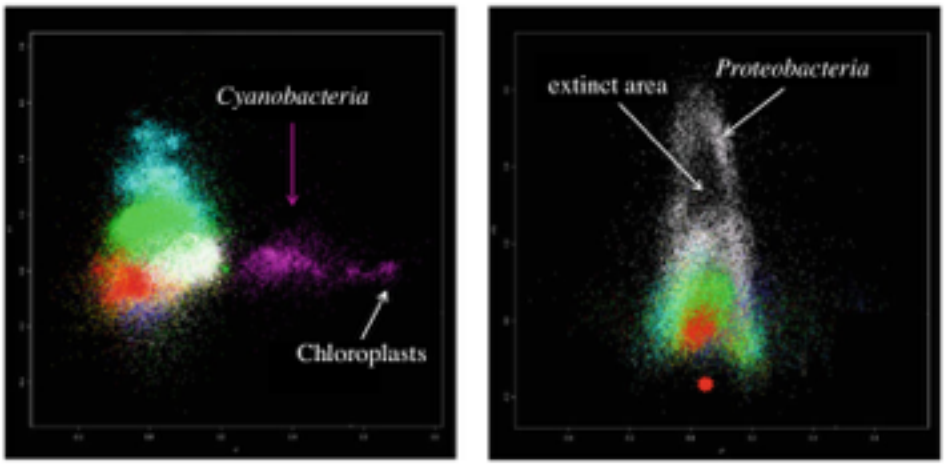

Figure 2. Two large-scale 2-dimensional slices of the ES inflated by various bacterial 16S rRNA sequences. Different colours represent different phyla. The red asterisk on the right picture represents a probable location of the last common ancestor of bacteria.

and reverse evolution rates); similarly different bacterial phyla formed well-defined areas within the space (Fig. 2). We believe that community pattern analysis can be applied in several research areas beyond evolutionary studies, including environmental research, particularly the analysis of microbial community succession. To show this we examined the dynamics of two microbiomes, exposed to salt stress in natural and artificial conditions. ES

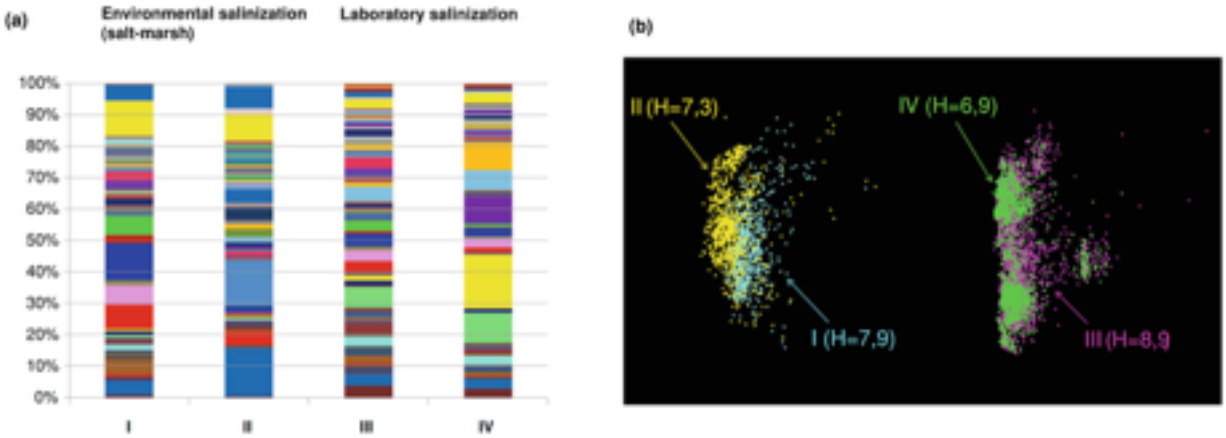

Figure 3. ES applied to an environmental study of naturally and artificially salinized soils: *I.* normal soil nearby a salty marsh; *II.* soil of the salty marsh; *III.* normal soil nearby a salty marsh *IV.* artificially salinized soil III.
(a) Taxonomic profile of the soils, different colours represent different bacterial families.
(b) Communities as visualized by the ES; H - Shannon diversity index (entropy).

helped us to study these distant microbiomes as mathematical objects possessing several geometric characteristics including shape, density, trajectory, and the vector of community development (Fig. 3).

It's clear that some essential biological questions could benefit from ES, yet its practical implementation requires collaboration of many scientists. Therefore we seek to draw attention to the ES in order to intensify its development.

Reference list

1. D.E. Dykhuizen (1998) Santa Rosalia revisited: why are there so many species of bacteria? *Antonie Van Leeuwenhoek* 73:25–33
2. Lilburn TJ, Garrity GM (2004) Exploring prokaryotic taxonomy. *Int J Syst Evol Microbiol* 54:7–13
3. X. Hao, R. Jiang, T. Chen (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* 5:611–618
4. T.Z. DeSantis, P. Hugenholtz, K. Keller, E.L. Brodie, N. Larsen, Y.M. Piceno, R. Phan, G.L. Andersen (2006) NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes, *Nucleic Acids Res* 34:W394–W399
5. T. Hughes, Y. Hyun, D.A. Liberles (2004) Visualizing very large phylogenetic trees in three dimensional hyperbolic space. *BMC Bioinform* 5:48
6. G.M. Garrity, T.G. Lilburn (2002) Mapping taxonomic space: an overview of the road map to the second edition of Bergey's manual of systematic bacteriology. *WFCC Newsl* 35:5–15
7. E.V. Pershina, A.S. Dolnik, G.S. Tamazyan, K.V. Vyatkina, Yu.B. Porozov, A.G. Pinaev, S.O. Karimov, N.A. Provorov, E.E. Andronov (2014) The Evolutionary Space Model to be Used for the Metagenomic Analysis of Molecular and Adaptive Evolution in the Bacterial Communitiesm In: *Evolutionary Biology: Genome Evolution, Speciation, Coevolution and Origin of Life* (Springer International Publishing Switzerland)