GENE TRANSFER AGENTS: THEIR EVOLUTION AND POTENTIAL ROLE IN MICROBIAL COOPERATION. M. Shakya[1], D. P. Birnbaum[1,2], T. B. Neely[1], and O. Zhaxybayeva[1,3], [1]Department of Biological Sciences, Dartmouth College, Hanover, NH 03755 [2]Present address: Broad Institute, Cambridge, MA 02142 [3]Department of Computer Science, Dartmouth College, Hanover, NH 03755 (olgazh@dartmouth.edu)

The existence of cooperation is one of the greatest puzzles in evolutionary biology. Cooperation allows individual organisms (or cells) to build complex multicellular consortia that communicate with each other and share resources, and is likely to have evolved early in Earth's history. Genes could be viewed as a valuable resource to share, because they may encode traits that could increase a recipient's fitness under certain environmental conditions. Both Bacteria and Archaea (prokaryotes) routinely exchange genes via horizontal gene transfer (HGT). However, if and how HGT affects a population of cooperating individuals is not known. One specific mechanism of HGT, mediated by Gene Transfer Agents (GTAs), promises to provide insights into the evolution of cooperation.

GTAs are virus-like entities encoded within a host genome, and are controlled by the host and environmental factors (reviewed in [1]). However, unlike viruses, GTAs package seemingly random pieces of host DNA. Moreover, GTA particles are too small to incorporate all of the genes necessary for their production, precluding GTAs from easily propagating themselves across a population. When GTAs are produced, host cell lyses and GTA particles are released. The particles can deliver the packaged DNA to other cells, which may help propagating novel adaptive traits throughout the population.

To date, genetically unrelated GTAs have been found only in four divergent taxonomic groups of prokaryotes: bacterial classes of α-proteobacteria, δ-proteobacteria and Spirochaetia, and archaeal class Methanococci. However, genes encoding GTAs have indisputable homology to the viral counterparts. Moreover, the majority of prokaryotes harbor at least one viral-like region in their genomes. This raises the possibility that GTAs might be more widespread among prokaryotes than currently perceived.

Here, we trace origin and evolution of GTAs using bacterial class of α-proteobacteria as a "model" system. One of the best-studied GTAs, *RcGTA,* is hosted by a free-living marine α-proteobacterium *Rhodobacter capsulatus*. A sequence similarity survey of 258 α-proteobacterial genomes revealed that many members of α-proteobacterial orders have RcGTA gene homologs, but are not known to produce GTAs. Therefore, GTAs may be more widespread than currently perceived. Alternatively, due to the shared ancestry between *bona fide* GTA genes and their viral counterparts, traditional sequence similarity search methods (for example, BLAST) might not be able to differentiate between the two types. Therefore, some of the detected GTA homologs may simply represent prophages (*i.e.*, viruses embedded in the host genomes). To address this problem, we are developing a machine learning method that would be able to classify virus-like sequences in a prokaryotic genome as either "GTA" or "prophage". Our approach is to use amino acid *k*-mers from known GTA genes and their viral homologs as features (so-called "bag of words" model [2]) and a support vector machine as a classification algorithm [3]. Here, we are presenting the results of the method's cross-validation, as well as classification tests of known α-proteobacterial prophages.

Given the wide distribution of RcGTA gene homologs across α-proteobacterial orders, how early did GTA arise in the evolution of this bacterial class? In other words, did the α-proteobacterial last common ancestor already encode a GTA? Or is *RcGTA* a recent invention that has spread within the class via HGT? To address these questions, we are reconstructing evolutionary histories of GTA gene homologs throughout α-proteobacteria, and comparing them to the histories of α-proteobacterial core genes. In addition, we are investigating biases in GC content, dinucleotide frequencies, and codon usage of GTA genes in comparison to the α-proteobacterial core genes.

If a GTA is advantageous to the host, its genes are expected to be under selective pressure. Moreover, this pressure would be distinct from that acting on the prophage genes. We are examining α-proteobacterial GTA genes for signatures of selection and comparing the inferred patterns to those observed in the conserved housekeeping genes, as well as in genes from α-proteobacterial viruses and prophages.

Our work is bringing insights into the mode of propagation of GTA genes within α-proteobacteria, and ultimately it will help us understand how GTAs originated and why they are maintained.

References:

[1] Lang, A. S., Zhaxybayeva, O., and Beatty, J. T. (2012) *Nat Rev Micro*, ***10***, 472-482.

[2] Forman, G. (2003) *The Journal of Machine Learning Research*, ***3***, 1289-1305.

[3] Hearst, M. A., *et al.* (1998) *Intelligent Systems and Their Applications, IEEE*, ***13***, 18-28.