Novel small repetitive sequences in bacterial genomes, and how biologists can build novel small nonrepetitive tools to study new phenomena

Jeff Elhai, Center for the Study of Biological Complexity
Virginia Commonwealth University

A large fraction of the human genome is made up of repeated sequences -- some tandem repeats, but mostly dispersed transposons and retroposons. Repetitive sequences are generally much less abundant in bacterial genomes. The genome of the cyanobacterium *Nostoc punctiforme* represents an extreme case. About 7.5% of its intergenic sequences is taken up by related families of tandem repeats found only in *Nostoc* and its relatives. Even more surprising are the large families of small dispersed repeats, ranging from 21 to 28 nucleotides, that that occur hundreds of times in the Nostoc genome. They appear to function through RNA intermediates and, in the case of some families, target specific sequences in the genome. No known mechanism of dispersal appears capable of explaining the sometimes rapid propagation of these sequences.

The discoveries associated with these repeated sequences occurred by chance, according to the following pattern: (1) Look more or less at random at lots of sequences, (2) Notice something that looks peculiar, (3) Build a quick computational tool to examine whether the peculiarity is worthy of further interest, and (4) If so, build computational tools to characterize in depth the nature of the phenomenon. This path to discovery is not available to the great majority of biologists, who do not know how to program a computer and do not care to learn. Since mass biological information is becoming increasingly difficult to avoid, and since computational tools are necessary for its analysis, most biologists have come to rely on a small number of inflexible tools that appeal to a general audience. These tools are useful in a narrow sense, but unfortunately direct discovery towards well-worn pathways and tend to filter out unexpected phenomena.

I will present a programming environment that my colleagues and I are developing that was designed to appeal to the biologist who does not want to learn a programming language. I will show how it was used to obtain biologically interesting results.