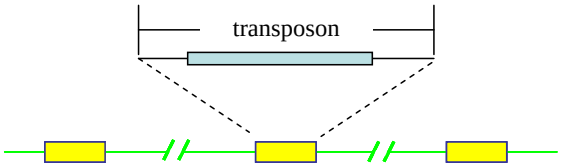
The Origin of Repeated Sequences in Genomes and (parenthetically) What is Life?

Jeff Elhai
Center for the Study of Biological Complexity

A great deal of attention has been lavished on the two billion nucleotides that make up the sequence of the human genome, but almost all that attention has focused on the 2% of the genome that encodes protein. The remaining 98% consists mostly of repeated DNA sequences, either dispersed repeats or tandem repeats. Dispersed repeats are segments of DNA that occur multiple times at more or less random positions in the genome. They are typically transposable elements, large segments that encode a protein responsible for the moving of the segment from one site to another. Most tandem repeats are small segments of DNA repeated one after another. For example, the trinucleotide CAG repeated hundreds of times is responsible for active Huntington Disease in humans.



Dispersed repeats

Tandem repeats

Bacteria, on the other hand have much smaller (0.6 to 11 million nt) and gene-dense genomes, with generally 70 to 80% of the genome devoted to protein-encoding genes. Many still possess transposons or other smaller dispersed repeats (typically 100 nt), but tandem repeats are rare. Recently a third type of repeated sequence has been discovered, called CRISPRs (for Clustered Regularly Interspaced Short Palindromic Repeats), consisting of short repeated regions (22 to 39 nt) separated by non-repetitive spacer regions of equal length. The mechanism by which they are formed is unknown.

CRISPRs

My colleagues and I have been examining the genome of the cyanobacterium *Nostoc punctiforme*, which is remarkable in many respects. At 9 million nt, its genome is amongst the largest of any bacterium. However, where it stands apart from other bacteria is in the quantity of repeated sequences of all types:

- Tandem repeats
  Much of the sequence between its genes consists of tandem 7-nt repeats, unique (so far) amongst bacteria. Some appear to function in genome rearrangements.
- Dispersed repeats
  Nostoc possesses large families of 24-nt repeated sequences. No repeat nearly this small has yet been reported. At least one sequence appears to insert itself randomly through an RNA intermediate.
- CRISPRs
  Nostoc has more CRISPRs in its genome than almost any other organism. By comparing sequences from *Nostoc* and related cyanobacteria, we have gained insight into how these remarkable sequences are formed.