What makes a gene? A synthetic biology perspective

We asked a simple question. What makes a gene? How did nature determine that she must covert a “specific piece of DNA” into a coding region? Did she sample all the possibilities, retained some and retired the rest? What kind of “algorithm” was used to determine the best fit among a large number of possibilities? What happened to retired sequences? Where are they now?

To answer the first question, we decided to explore the possibility of artificially constructing genes from non-coding regions. Our aim was to develop a method that was simple and scalable. The problem was to covert a non-coding sequence into a coding sequence? To meet this key challenge, we invented a novel method that consisted of the following steps (a) amplify a given sequence, (b) find the right vector that supplies basic infrastructure for expression (c) paste the amplified sequence into this vector (d) transfer the artificial gene cassette to the cell (e) confirm the expression and (f) hope for a novel phenotype.

Six E.coli intergenic regions with no history of transcription were randomly picked up. All the sequences were computationally translated and matched against non-redundant NCBI database to ensure that we did not end up creating a known natural equivalent of these user-defined proteins. Sequences were amplified and inserted into pBAD topo vector and expressed in E.coli MG 1655 cells. Protein expression was confirmed by Western blotting. The intracellular expression of one of the proteins resulted in the cell growth inhibition. The growth inhibition was completely rescued by culturing cells in the inducer-free medium. Computational structure prediction suggested globular tertiary structure for two of the six non-natural proteins synthesized. We called these artificially constructed genes EKA (ekam - first in sanskrit). To our best knowledge, this is the first report that describes making genes from junk DNA.

These findings lead us to revisit the first question - what makes a gene? Can we extend this theme to pseudogenes, repetitive sequences, introns and subsets of exons, and so on ? What is the best-case scenario and boundary condition of converting non-coding to coding regions? Having provided the proof of the concept, we are studying features that are common to a broad set of coding regions. Once patterns are identified, our questions will move towards understanding of the emergence of genes from evolutionary perspective, and 'extract' gene-like regions from a huge mass of non-coding DNA regions. Preliminary results point to the new classification system and an evolutionary understanding that could emerge from these studies.