INFORMATION-PROCESSING GENES

K. TAHIR SHAH
International Centre for Theoretical Physics,
P.O. Box 586, 34100 Trieste, Italy

ABSTRACT

DNA computing is an active interdisciplinary area of research. Our theoretical work and a recent laboratory experiment by Adleman clearly demonstrate that nucleotide sequences are capable of solving NP-complete (or computationally intractable) problems. We reached this conclusion through an analysis of molecular fossils.

There are an estimated 100,000 genes in the human genome of which 97% is non-coding. On the other hand, bacteria have little or no non-coding DNA. Non-coding region includes introns, ALU sequences, satellite DNA, and other segments not expressed as proteins. Why it exists? Why nature has kept non-coding during the long evolutionary period if they it no role in the development of complex life forms? Does complexity of a species somehow correlated to the existence of apparently useless sequences? What kind of capability is encoded within such nucleotide sequences that is a necessary, but not a sufficient condition for the evolution of complex life forms, keeping in mind the C-value paradox and the omnipresence of non-coding segments in higher eukaryotes and also in many archea and prokaryotes.

The physico-chemical description of biological processes is hardware oriented and does not highlight algorithmic or information processing aspect. Any information-processing system has two independent components: the algorithm and its hardware implementation. However, an algorithm without its hardware implementation is useless as much as hardware without its capability to run an algorithm. The nature and type of computation an information-processing hardware can perform depends only on its algorithm and the architecture that reflects the algorithm. Given that enormously difficult tasks such as high fidelity replication, transcription, editing and regulation are all achieved within a long

linear sequence, it is natural to think that some parts of a genome are involved in these tasks. If some complex algorithms are encoded within these parts, then it is natural to think that non-coding regions contain processing-information algorithms.

A comparison between well-known automatic sequences and sequences constructed out of motifs in found in all species proves the point: noncoding regions are a sort of "hardwired" programs, i.e., they are linear representations of information-processing machines. Thus in our model, a noncoding region, e.g., an intron contains a program (or equivalently, it is an automaton) while an exon contains acceptable data in the language of this automaton.

There are many important questions that we shall discuss in a monograph in preparation. These are:

Critical segments and their computational significance;
Long-range correlation in introns and their possible relevance to fractal nature of biological systems;
Anti-sense strategy and enzyme-directed therapy;
Evolution in the RNA world and molecular fossils;
Evolution of cellular communication system; and
The C-value paradox.