Decoding Non-Coding DNA: Trash or Treasure?

Namrata Iyer

Non-coding DNA, once thought of as 'junk', represents a very large portion of an organism's genome. However, recent research has brought to light many functional elements present within non-coding DNA sequences and unravelled a fascinating array of functions performed by these elements. These findings have highlighted the nature of the evolutionary forces that led to the accumulation and retention of non-coding DNA. In this article, the various elements present within non-coding DNA, their functional relevance to the cell and the changing perspective of the scientific community towards this so-called 'junk' DNA have been described.

Namrata Iyer is a PhD student in the Department of Microbiology and Cell Biology, Indian Institute of Science, Bangalore. Her research interest is the molecular basis of host–pathogen interactions in human diseases.

Since the dawn of time, man has always been plagued by the question of the origin of life. What are the forces that govern the course of evolution? What are the elements that separate man from other forms of life? The discovery of DNA (deoxyribonucleic acid) as the genetic material in 1944 opened up new avenues to answer these questions. Surprisingly, the language of DNA comprises of only 4 letters, i.e., A,T,G,C which when read in groups of three (triplets) encode the information for the synthesis of proteins (by a process known as translation) which are the work-horses of a cell. Before translation can begin, the information on DNA is first copied into an intermediate known as mRNA (by a process known as transcription). Each cell hides within itself large stretches of DNA (i.e., genome of an organism), which contains within itself all the information required for the creation of an entire organism from a single cell. If only we could find a way to read this blueprint of life, we would be able to solve the mystery underlying our evolution!

With the invention of DNA Sequencing, it became possible to unravel the genome sequences of a wide variety of organisms – from bacteria to plants and even humans! It was believed that the

Keywords
Non-coding DNA, transposons, repeat elements, pseudogenes.

genome would reveal the genetic basis for the increasing complexity across these various organisms. The popular belief was that proteins, being the effector molecules of a cell, would determine the level of complexity of an organism and therefore increasing number and types of proteins coded by the genome would correlate with increasing biological complexity of the organism. However a detailed examination of the genome revealed a very different picture.

A close look at the genomes of various organisms ranging from bacteria to humans revealed that the entire genome does not code for proteins. In fact, a significant part of the genome apparently remains untranslated and sometimes even untranscribed (*Figure* 1). This part of the genome which does not code for any protein was hence known as non-coding DNA and thought to be of little or no use, and hence, *junk*! Furthermore, analysis of a variety of genomes brought forth two well-known paradoxes: (1) The *C value paradox* states that the amount of cellular DNA is not proportional to the biological complexity of the organism. In fact, organisms such as frogs and amoebae have much more DNA per cell compared to mammals. (2) In addition, the *G value paradox* states that the total number of protein-coding genes is also not proportional to biological complexity and even alternative splicing (the process by which the cell can combine different exons to generate a family of related proteins) cannot fully explain this

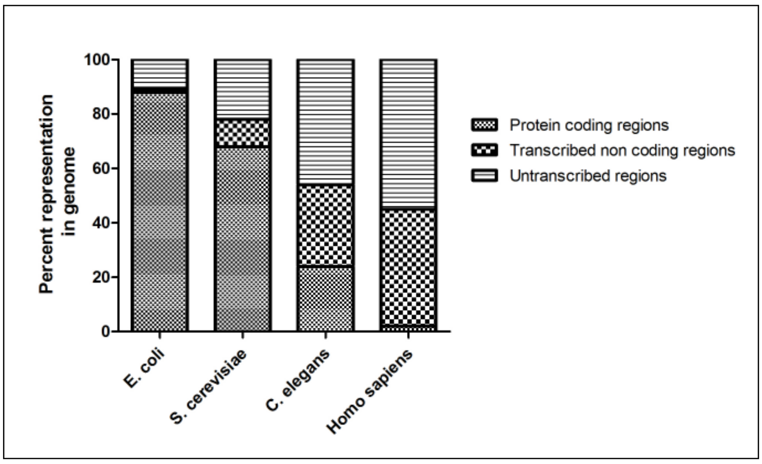

Figure 1. Genome organisation across various species.

Adapted from *Genome Biology*, Vol.5, p.105, 2004.

phenomenon. The fact that the proteomes (the entire set of proteins expressed by an organism) of mice and humans differ by only 1% raises the question: *What is the central factor which defines the level of biological complexity of an organism*?

A closer look at the genome composition revealed that while bacterial genomes have only 12% non-coding DNA, in humans it makes up 98% of the genome. This observation gave rise to a large number of questions. During the course of evolution, each organism is subjected to a wide variety of selection pressures, forcing each organism to optimise the efficiency of all its metabolic processes so as to attain maximum fitness. Thus, wasteful accumulation is very rarely encountered in nature. So how did we acquire such a large amount of apparently useless non-coding DNA and moreover why have we retained it? Is it just evolutionary junk that has not been eliminated due to the lack of selection pressure to streamline eukaryotic genomes? Or, are they simply a buffer to insulate coding DNA from accumulation of mutations? For a long time, these questions remained largely unexplored. However, recent research in this area has given us a better understanding of the world of non-coding DNA.

Decoding Non-Coding DNA

Non-coding DNA is composed of a number of elements such as transposable elements, intra- and inter-genic regions, *cis* acting elements (5' and 3' untranslated regions or UTRs), repeat elements and pseudogenes (*Table* 1).

Table 1. Genomic feature comparison across species.

Adapted from *Genome Biology*, Vol.5, p.105, 2004.

Species	**Genome Size Mb**	**CDS (%)**	**UTR (%)**	**Intron (%)**	**Inter-genic (%)**	**Repeats (%)**
C. elegans	100.3	25.3	2.2	30.4	42.3	12.9
D. melanogaster	132	16.6	4.9	29.1	49.4	12.3
G. galus	1054	2.4	0.5	32.7	64.4	9.9
M. musculus	2583	1.1	1	29.3	68.6	42.3
H. sapiens	2866	1.1	1.1	35.2	62.6	48.5

Introns

These are non-coding regions within a gene which separate the functional domains of a protein, encoded by exons, and are found mainly in eukaryotes. This neat separation of a protein into distinct, independent domains allows for the mixing and matching of different domains to give rise to entirely new proteins (exon shuffling), which gives a very significant evolutionary advantage. In addition, it also allows different types of proteins to be produced from a single large mRNA sequence based on the specific combination of exons included. It is in this process of alternative splicing that the role of these intra-genic sequences was first discovered. However, the signals required for splicing constitute only a very small part of the total intronic sequence. Furthermore, not all intron-containing genes undergo alternative splicing, which hints at some function for introns beyond splicing. Introns are now known to house a large variety of regulatory RNAs such as snRNAs (small nuclear RNAs) and some micro RNAs (small non-coding RNAs that regulate the transcript level of protein-coding target RNAs). They are also involved in nucleosome formation, chromatin organisation and contain scaffold/matrix attachment regions (S/MARs). Analysis of the variability in intron length has revealed a correlation between their length and the transcription level of the gene, i.e., the highly transcribed genes have fewer introns and *vice versa*. Thus, they have a role to play in tissue specific regulation of gene regulation.

This neat separation of a protein into distinct, independent domains allows for the mixing and matching of different domains to give rise to entirely new proteins (exon shuffling).

Transposable elements (TE)

Transposable elements are mobile genetic elements which were first discovered by Barbara McClintock[1], based on their mutagenic potential. They are characterised by the presence of long terminal repeats (LTRs) and sometimes code for proteins known as transposase which give them the ability to 'jump' from one location in the genome to another. The act of transposition may lead to gene disruption or chromosomal rearrangement, and hence is highly deleterious. DNA sequence elements in the TEs or the resulting SINEs (short interspersed elements) of a

[1] See *Resonance*, Vol.1, No.10, 1996.

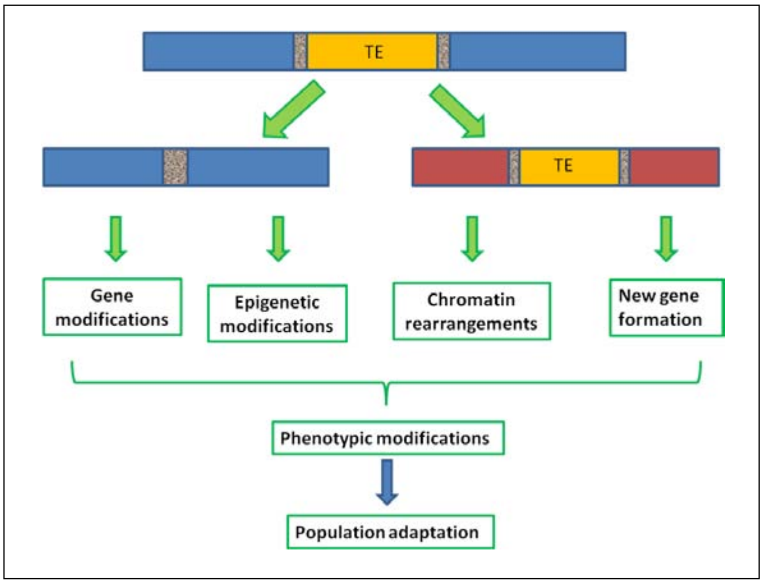

Figure 2. Evolutionary effects of Transposable elements.

Adapted from *Nature*, Vol.443, pp.521–524, 2006.

transposition event can also affect the expression and regulation of neighbouring genes. Their involvement in bringing about gene mutations and epigenetic changes (heritable changes in phenotype or gene expression caused by mechanisms other than changes in the underlying DNA sequence) has led to TEs being regarded as an important evolutionary force (*Figure* 2). Recent bioinformatics analyses show that TEs might have yet another function. Many TEs code for proteins such as reverse transcriptase and envelope proteins and domestication (loss of 'mobility' due to lack of or mutation in sequences required for transposition) of TEs may not necessarily be accompanied by a loss of coding potential. It is hypothesised that many host proteins such as RAGs (Recombination activating genes), and telomerase are derivatives of such proteins encoded by Transposable elements.

Simple Sequence repeats

SSRs are 1–6 bp long repeating units of DNA. While one may be tempted to dismiss these highly simplistic DNA elements as non-functional, the type (based on their sequence) and distribution of SSRs is known to be non-random. Many a times, genes which are

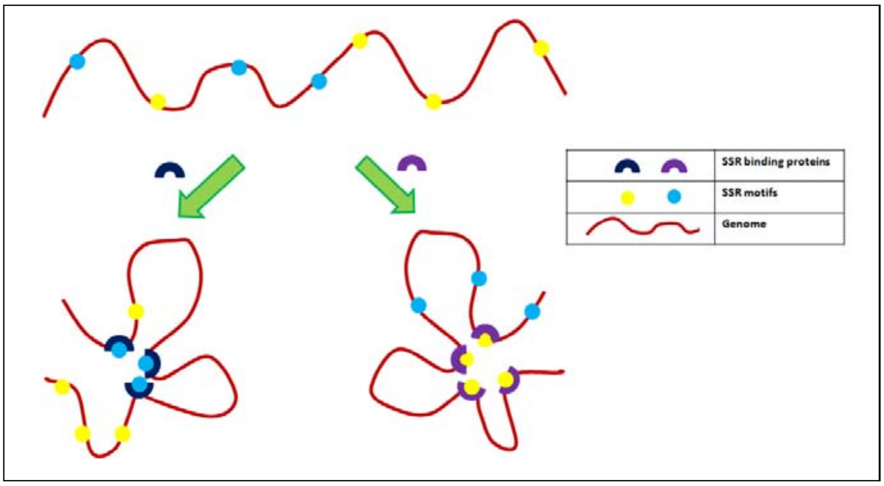

Figure 3. Long-range genome interactions mediated by action of SSRs and cognate SSR binding protein.

Adapted from *Bioessays*, Vol.32, pp.165–174, 2007.

regulated in a similar manner or are expressed under similar conditions are seen to be widely separated in the context of the genome. Synchronisation of the expression of such genes thus requires these genes to be brought spatially closer under appropriate conditions! SSR distribution is known to be associated with specific similarly-regulated loci. Research has now unearthed many SSR binding proteins which are believed to orchestrate these long range interactions between co-ordinately regulated loci (*Figure* 3). SSRs are highly prone to mutations such as insertions/ deletions and replication slippage leading to repeat number variability. SSRs present within the coding as well as *cis* regulatory elements can exert phenotypic effects on genes due to their repeat variability. For example, SSR repeat number variants (17 and 20) within the *per* gene of Drosophila (*period*, a gene that regulates the biological clock mechanism) leads to a change in the responsiveness of their circadian rhythm to temperature fluctuations. Thus SSRs provide a ready and virtually inexhaustible supply of new quantitative variation for rapid evolutionary adaptation; something akin to evolutionary tuning knobs!

Pseudogenes

These are defunct gene copies generated during gene duplication events by TEs or by reverse transcription of the corresponding mRNA followed by recombination of the cDNA into the genome.

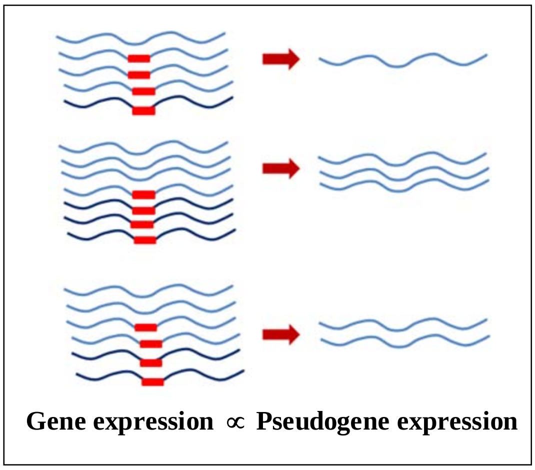

Figure 4. Regulation of gene expression by sequestration of specific miRNA by Pseudogene mRNA.

Adapted from *Nature*, pp.1033–1038, 2010.

Considering that they are no longer functional, one would assume that their coding sequence would have undergone large scale mutation in the absence of any selection pressure. However, it has been shown that the coding sequences of pseudogenes are relatively conserved. Pseudogenes are homologous in sequence to the transcripts (i.e., mRNA) of their functional gene counterparts and therefore have the potential to act as decoys for gene-specific micro RNAs. In fact, a recent report has shown the ability of PTENP1, a pseudogene, to compete for the miRNA against the tumor suppressor gene PTEN. The level of PTENP1 transcript in the cell was shown to regulate the degree of suppression of PTEN expression by its cognate micro RNA. Thus the RNA of the pseudogene PTENP1 regulates the net cellular levels of PTEN and hence the growth suppressive effect exerted by PTEN on the cell (*Figure* 4).

Conclusion

While all the elements described above are functional, they still account only for a small part of the total non-coding DNA in an organism. The question still remains as to why the genomes have accumulated so much non-coding DNA. Different groups have adopted different approaches to address this issue. Some take a function independent approach, looking at the accumulation of non-coding DNA as a consequence of processes that contribute to growth in genome size such as spontaneous insertions, transposable elements, etc. The net change in the genome size has an effect on fitness and hence is subjected to selection pressure. The degree of accumulation of non-coding DNA is therefore defined by the equilibrium between forces leading to decrease and increase in the genome size as a whole. On the other hand in the function dependent approach, non-coding DNA is assumed to confer an adaptive advantage to the organism. The adaptive theory looks at non coding DNA as contributory to nuclear and cellular volumes. These in turn regulate adaptive traits such as doubling time which are subject to selection pressure.

Non-coding DNA, once considered junk, now appears to house a huge regulatory network which is responsible for the execution of complex developmental programs and therefore is the key to the evolution of biological complexity.

Addressing these questions represents a challenge because of the inherent variability in the non-coding DNA. Since the selection pressure acting on non-coding DNA is lower than that on coding DNA, the bioinformatics approach of conservation as an indication of function is less applicable. Also the sequence-independent nature of some of the non-coding DNA segments adds another level of complexity. Despite these constraints, comparative genomics tools along with transcriptome analysis hold enormous potential in uncovering the mysteries buried within the non-coding DNA.

Thus, non-coding DNA, once considered junk, now appears to house a huge regulatory network which is responsible for the execution of complex developmental programs and therefore is the key to the evolution of biological complexity.

Suggested Reading

[1] S A Shabalina and N A Spiridonov , The mammalian transcriptome and the function of non-coding DNA sequences, *Genome Biology*, Vol.5, p.105, 2004.

[2] R J Taft, M Pheasant and J S Mattick, The relationship between non-protein-coding DNA and eukaryotic complexity, *BioEssays*, Vol.29, pp.288–299, 2007.

[3] C Biemont and C Vieira, Junk DNA as an evolutionary force, *Nature*, Vol.443, pp.521–524, 2006.

Address for Correspondence
Namrata Iyer
Microbiology and Cell Biology
Indian Institute of Science
Bangalore 560 012, India.
Email:
namrataiyer@mcbl.iisc.ernet.in