mySCIENCE 1(1), 2006, 1–6

http://myscience.uni-mysore.ac.in

Review Article

Noncoding DNA: Junk, or a necessity for origin and evolution of biological complexity?

S. C. Lakhotia[‡]
Departments of Zoology & Molecular and Human Genetics,
Banaras Hindu University, Varanasi, India

(*Received January 2005; Accepted March 2005*)

Much of the remarkable progress in biological sciences during the past five decades following the unraveling of DNA structure has been based on the so-called "central dogma of molecular biology" which provides a formal basis to understand the flow of information from genes to phenotype. A strong faith in the "central dogma" has resulted in a common belief that any sequence of DNA or a gene is of relevance only if it has a protein-coding function and consequently, a significant proportion of molecular biological studies during the past few decades have been propelled by the concept that the noncoding DNA is "junk" or "selfish" or "parasitic". On the other hand, sequencing of genomes of large number of species, ranging from bacteria to human, has clearly re-established the earlier inference of classical geneticists and cytologists that much of the DNA in genomes of higher organisms does not carry typical "genes" or protein-coding genetic information. As the biological complexity has increased with evolution, the proportion of DNA in the genome that does not code for protein has also increased. Thus while noncoding DNA is almost non-existent in bacteria, it can make up as much as 90% or more of the genome in higher organisms like mammals.

Is the noncoding DNA in our genomes really "junk", whose accumulation to such high proportions reflects some kind of "failure" of natural selection, or is it a necessity for the biological complexity? It is now clear that the genetic differences between any two related species are mostly due to changes in the "noncoding" DNA rather than in the protein-coding genes. Thus while human genome has 25 fold more DNA compared to the fruit fly, the protein-coding genes appear to be only 2.5-fold greater. The evolutionary increase in biological complexity is thus not due to a greater variety of proteins but due to more complex regulatory circuits that allow a greater variety of combinations of similar numbers of proteins so that more complex structures and organizations can come into being. Some of the many possible roles that the noncoding DNA may have in biological organisation are briefly outlined in the following.

[‡] For correspondence Email: lakhotia@bhu.ac.in

DNA sequences that regulate expression of genes are essential for biological complexity

In many cases, the DNA sequences involved in cis- and trans-regulation of protein coding genes (promoters, enhancers, silencers, boundary elements etc) are actually longer than the transcribed parts. In fact, it is becoming clear that much of the evolution of differences in related species depends on modulating the regulation of the gene rather than the structure or function of the protein encoded by the given gene. Alterations in regulation of genes are brought about, rather rapidly, by small changes in the base sequences, which may create new (or eliminate existing) target sites for binding of the transcription factors or by insertion/mobilisation of transposable elements which affects expression of the given gene. Thus one of the very important functions of the noncoding DNA sequences is to provide for regulation of transcriptional activity of genes. Increasing biological complexity has to be associated with more complex regulatory networks.

Another important aspect, although little understood, is the physical organisation of the nuclear DNA in a cell-type specific architecture in the 3-dimensional nuclear volume. It is obvious that cell type specific regulation of genes is also facilitated by the way the genomic DNA is organised into chromatin and the way the chromatin is packaged in to the nuclear space. We do not yet know how much of the genomic DNA is required for providing this kind of "information".

Roles of noncoding but transcribed sequences in protein coding genes

Analysis of the human genome indicates that of the total of 3.2GB genomic DNA, only about 1.2GB accounts for protein-coding gene or gene related sequences, and of this only 48MB is actual coding region while the remaining DNA is present as pseudogenes, gene fragments, introns, 5'- and 3'- UTRs (untranslated regions), promoter/regulatory regions etc. Some of these noncoding regions, specially the introns and the UTRs have very significant roles in generation of protein diversity (through alternative splicing of introns) and in regulating the half-lives and locations of the transcripts in cells. In many cases, the length of DNA sequences present as UTR and/or introns is much larger than the protein coding exonic sequences. A particularly remarkable example of the great potential of introns in generating protein diversity is the *Dscam* gene of *Drosophila* which has the potential to generate as many as 38000 varieties of proteins because of alternative splicing and this potential diversity appears to be significant in guiding the different axons as they develop to reach their target sites.

Genes producing only noncoding RNA have vital functions

The more intriguing and often the more ignored component of the noncoding DNA is that which is transcribed but produces RNA species that are neither translated nor involved in any way with the process of translation. It has been known for several decades that the RNA species present in a eukaryotic nucleus are very diverse (one reason why these were designated as hnRNA or heterogeneous nuclear RNA) and that a majority of these never leave the nucleus. Unfortunately, however, these diverse RNA species did not receive the deserved attention because of the strong influence of the concept of selfish or junk DNA. Likewise although transcription in heterochromatin regions has been known in some cases from early 1970s, such reports did not receive serious and wider attention.

Notwithstanding the prevailing bias, studies on an increasing number of noncoding RNA species have revealed them to be essential for some very basic but vital functions in cells. Some examples follow.

Noncoding RNA species regulate the chromatin organisation and transcriptional status of entire chromosomes

The sex-chromosomes carry, besides the genes directly involved in sex-determination/differentiation, many other genes that are required to be equally expressed in male as well as female individuals of the species. Dosage compensation, therefore, is a mechanism that compensates for the difference in the dosages of such sex-chromosome-linked genes between the homo- and hetero-gametic sexes. In mammals, dosage compensation is achieved by inactivation of one of the two X-chromosomes in somatic cells of females, while in *Drosophila,* the activity of the X-linked non-sex determining genes is equalised in males and females through hyperactivity of the single X-chromosome to the level of two X-chromosomes in females. Notwithstanding such opposing operative mechanisms, it is interesting that in both cases, noncoding RNA species are the key players that determine the activity level of the X-chromosome. In the case of human females, the long and nucleus-limited noncoding RNA of the *Xist* gene (or its homologue in other mammals) is produced only by one of the two X-chromosomes and the Xist transcript spreads in cis along the entire length of the X-chromosome and, thereby, provides the platform for binding of other proteins etc which keep that particular X-chromosome in an inactive state. It is interesting that the decision to allow Xist transcription from only one of the two X-chromosomes is also based on production of another noncoding RNA, the Tsix RNA, which is produced from the complementary strand of the *Xist* gene. In the case of Drosophila, the single X-chromosome of males becomes hyperactive because noncoding RNA species, the Rox1 and Rox2 "paint" the entire X-chromosome and thereby allow the assembly of protein

complexes (the Msl complex), which set the X-chromosome transcription at a higher level. Thus both in mammals and *Drosophila*, noncoding RNA provide a mechanism for recruiting the required proteins in an organised manner, which establish the chromatin organisation (condensed for inactivity or loose for hyperactivity in female mammalian X chromosome and male *Drosophila* X-chromosome, respectively.

Noncoding RNA species help sequester different families of proteins and thus may regulate their activity

A large variety of proteins are involved in processing (like splicing) and transport of the different protein-coding transcripts synthesized by the DNA templates. Since the cellular activities are extremely dynamic, the RNA processing activities also have to be equally dynamic. The different RNA processing and RNA transporting proteins are, therefore, required to dynamically toggle between active and inactive states. Besides, the cells may encounter sudden environmental stresses, which have drastic effect on RNA synthesis and processing activities. In the absence of new transcription, the diverse RNA processing proteins have no substrates and thus either their levels in the inactive pool rise substantially or these proteins need to be degraded. Since most of the RNA-processing proteins have long half-lives, the proteins need to be sequestered in the inactive compartment. Since the RNA-processing proteins would need some RNA molecules to which they must remain bound in their inactive state, a question that arises is what kinds or species of RNA molecules provide the platform for storage of inactive RNA-processing proteins. Insights into this dynamic state have been obtained through studies on stressed cells, since heat shock and similar other cellular stresses have been found to substantially inhibit new transcription and RNA processing without the components of the RNA synthesis and processing machinery being broken down. A large number of studies have shown that under these conditions, different classes of RNA processing proteins get sequestered in distinct nuclear compartments or speckles (with different names). Studies in our laboratory with *Drosophila* cells have shown that a species of noncoding RNA, the hsr?-n (heat shock RNA omega-nuclear) transcripts, is essential for organising a special nuclear compartment, the omega speckles, for storage/sequestration of the members of hnRNP (heterogeneous nuclear RNA binding proteins) family and other related proteins in normal as well as stressed cells. During development, this noncoding RNA is expressed, in a regulated manner, in almost all cells types of *Drosophila*. Both over and under-expression of this noncoding RNA has severe consequence for the organism.

A somewhat similar function seems to be carried out by the noncoding transcripts of the satellite III in human cells following heat shock. Heat shock

induces transcription of the satellite III sequences, located on centromeric heterochromatin of human chromosomes 9 and 11. A variety of RNA processing proteins, RNA polymerase II and heat shock transcription factor etc get sequestered with these transcripts as stress granules in heat shocked human cells.

In both the above cases, the hsr -n or the satellite III noncoding transcripts serve an important function through sequestering specific classes of proteins when they are not required to be active. Such a storage function may also regulate the availability of specific members of these protein families for activity. It is highly likely that many other such noncoding RNA species, which are involved in regulating activities of other kinds of cellular proteins involved in RNA metabolism, remain to be discovered.

A large variety of very small noncoding RNAs are important riboregulators

One specific aspect of the role of RNA as riboregulators has actually received considerable attention in recent years. This relates to the large variety of small RNA species (21-23 nucleotide long) designated as micro RNA, silencing RNA, small temporal RNA and interfering RNA etc. These small RNAs, which affect gene expression at the levels of chromatin structure, RNA degradation and translation, obviously have very significant roles in establishing the biological complexity of multicellular organisms.

Concluding remarks

Understanding gene function in terms of protein synthesizing activity has been a major achievement of modern Biology. It is also clear now that the effective proteome in any species can be much larger than that originally encoded by the genome due to the versatile processes like alternative splicing, post-translational modifications of proteins etc. Yet, it is obvious that the enormous diversity in structure and organisation of the multicellular organisms cannot be explained only in terms of the proteome. While a positive correlation between the proportion of noncoding DNA and biological complexity by itself suggests an important role for such DNA sequences (through their cis-regulatory roles and through the production of noncoding RNAs), elucidation of functions of many noncoding RNA species in recent years has established a case for the need for more in depth studies on the noncoding DNA component in any genome. It will be erroneous to ignore this as "junk" or "selfish" DNA. Like in the primitive "RNA-world", the ribonucleic acid molecules can be, and indeed are, functionally versatile even in the modern "DNA-world".

(This article is based on a lecture delivered during the 70th Annual Meeting of the Indian Academy of Sciences, Bangalore, at Banaras Hindu University, Varanasi. The accompanying presentation can be seen at `http://www.ias.ac.in/meetings/annmeet/70am_talks/sclakhotia/img0.html`)

References

Eddy S. R. 2002. *Cell* **109:** 137–140.

Gaffney D. J. and Keightley P. D. 2004. *Trends Genet.* **20:** 332–337.

Heard E. 2004. *Curr. Opin. Cell Biol.* **16:** 247–255.

Kindel D. and Amrein H. 2003. In: *Noncoding RNAs: Biology and molecular medicine* Eds. Barciszewski J. and Erdmann V. E. Kluwer/Plenum Publ. pp 67–83.

Lakhotia S. C. 2003. In: *Noncoding RNAs: Biology and molecular medicine* Eds. Barciszewski J. and Erdmann V. E. Kluwer/Plenum Publ. pp 202–219.

Lewin B. 2004. *Genes VIII,* Pearson/Prentice-Hall.

Metz A., Soret J., Vourc'h C., Tazi J. and Jolly C. 2004. J. *Cell Sci.* **117:** 4551–4558.

Nelson P., Kiriakidou M., Sharma A., Maniataki E., and Mourelatos Z. 2003. *Trends Biochem. Sci.* **28:** 534–540.

Plath K., Mlynarczyk-Evans S., Nusinow D. A. and Panning B. 2002. *Ann. Rev. Genet.* **36:** 233–78.

Prasanth K. V., Rajendra T. K., Lal A. K. and Lakhotia S. C. 2000. J. *Cell Sci.* **113:** 3485–3497.

Rizzi N., Denegri M., Chiodi I., Corioni M., Valgardsdottir R., Cobianchi F., Riva S., and Biamonti G. 2004. *Mol. Biol. Cell* **15:** 543–551.

Schattner P. 2002. *Nucl. Acids Res.* **30:** 2076–2082.

Shabalina S. A. and Spiridonov N. A. 2004. *Genome Biol.* **5:** 105.

Szymanski M. and Barciszewski J. 2002. *Genome Biol.* (Reviews) **3:** 0005.1–0005.8

Szymanski M., Erdmann V. A. and Barciszewski J. 2003. *Nucl. Acids Res.* **31:** 429–431.

Tijsterman M. and Plasterk R. H. A. 2004. *Cell* **117:** 1–4.

Wutz A. 2003. In: *Nncoding RNAs: Biology and molecular medicine* Eds. Barciszewski J. and Erdmann V. E. Kluwer/Plenum Publ. pp 49–65.

Yuan G., Klambt C., Bachellerie J. P., Brosius J. and Huttenhofer A. 2003. *Nucl. Acids Res.* **31:** 2495–2507.

Zuckerkandl E. 2002. *Genetica* **115:** 105–129.