The Centrality of RNA

Leading Edge
Essay

Cell

The Centrality of RNA

Phillip A. Sharp[1,*]
[1]The Koch Institute for Integrative Cancer Research and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*Correspondence: sharppa@mit.edu
DOI 10.1016/j.cell.2009.02.007

New roles for RNAs in biology continue to emerge, and a glance at the history of RNAs may prepare molecular biologists for future discoveries about these powerful molecules. A striking new role for RNAs is their widespread involvement in the regulation of numerous genes, suggesting that there is much yet to discover about these amazing cellular components.

The centrality of RNA in cellular processes and gene regulation reflects its biochemical properties and early appearance during evolution. The linear sequence of RNA makes it a simple source of genetic information. The property of RNA to form secondary structure, shielding some sequences while exposing others for recognition, facilitates its interactions with other molecules. In a more complex fashion, RNA can assume tertiary structures that present surfaces for interactions and contain internal environments that create binding sites for metal ions that are sufficiently sequestered from solvent that they can promote catalytic reactions. No other macromolecule covers this range of chemical space and functions.

RNA and Gene Expression: The Early Years

With the discovery of the structure of DNA more than 50 years ago, the central mystery in molecular biology became how information flowed from this beautiful linear library to specify all of the components of the cell. In hindsight, one could argue that the early founders were a little slow to discover the existence of messenger RNA (mRNA). This short-lived species was hard to detect, as releasing it from ribosomes required specific conditions. At the same time, short RNAs termed tRNAs were recognized as the second genetic code relating the nucleic acid code to the amino acid code, with an RNA-containing machine termed the ribosome as the decoding device. The centrality of RNA in these critical life processes led to speculation about an RNA world, where RNA was the genetic material and RNA decoded RNA to synthesize proteins. It also produced the central dogma of molecular biology, as formulated by Francis Crick, that cellular information only flows from nucleic acid to proteins. (Crick was thoughtfully agnostic as to whether the nucleic acid form of the genetic information was RNA or DNA.) The notion of an RNA world has become generally accepted with the discovery of RNA catalysis and the more recent conclusions that the ribosome is an RNA machine.

RNA, Gene Regulation, and Transcription Factors

Once mRNA was defined as the medium by which information flows from DNA into the cell, the next great question in molecular biology was how this process was regulated. Variations in this information transmission allow single-cell organisms to respond to their environment. These variations also provide a means for individual cells carrying identical information to acquire different traits critical for their functions in multicellular organisms. The central model proposed in the 1960s posited that the utilization of genetic information was primarily controlled at the level of transcription of RNA from DNA. This control was effected by the actions of proteins, called transcription factors, which either inhibited or promoted access of the RNA polymerase to distinct regions termed promoter sites, from which transcription was initiated. This wonderful model seemed to satisfactorily explain most of biology and dominated the scene for decades. It has progressively evolved to take into account, for example, the roles of chromatin. Yet despite these modifications, the key feature of the model—that RNA production is at the core of gene regulation—has remained intact. In fact, with the recent discovery that differentiated somatic cells can be induced to become pluripotent stem cells by introduction of a combination of genes encoding transcription factors, the model wherein protein factors recognize DNA and serve as master regulators of gene expression is robustly supported once again (Takahashi and Yamanaka, 2006). Yet, as time passes, the notion of the cardinality of gene regulation by transcription is becoming progressively weakened. This dogma is rooted in studies in bacteria and thus handicapped by presupposition of the existence of a simple gene.

Regulating Split Genes

The importance of the transcription factor model of gene regulation progressively decreased as the concept of the nature of a gene became ambivalent. The split structure of genes in most nucleated cells added a new layer of regulation at the stage of RNA processing. The possibility of alternative RNA splicing meant that the specific information flowing from a gene-like segment to the cell could change depending upon cellular state or environment. Given that more than half of transcribed DNA segments in mammalian cells are expressed as alternatively spliced RNAs in different cells, the definition of a "gene" and how this entity is regulated has become staggeringly complicated (see for example Wang et al., 2008a). Although it is convenient to talk about 23,000 human genes, this should be considered scientific jargon and not precise. Paradoxically, with the logarithmic increase in our knowledge of the human genome sequence and the

transcripts it encodes, we have at a fundamental level been forced back to the original, abstract concept of the gene—a heritable trait, defined by a phenotype, that is best studied by classical genetic methods.

At a biochemical level, a gene is perhaps most easily described by the set of sequences that after transcription are available to be joined by RNA splicing to produce a final product. This leads to the concept of a transcription unit, the sequences transcribed by RNA polymerase into a contiguous segment that can serve as a substrate for *cis*-splicing processes. Obviously, this ignores the role of *trans*-splicing, which is important in certain cases but not generally. RNA as the product of a transcription unit then becomes the central feature in discussing gene regulation.

RNA splicing is critical for expression of most genes, and small nuclear RNA (snRNA) species play an important role in this process. Excision of an intron depends upon formation of a spliceosome by processes that recognize the 5′ spliced site, 3′ spliced site, branch site, and other sequences. The local sequences constituting splice and branch sites are recognized by snRNAs in small nuclear ribonucleoprotein particles (snRNPs), and proteins that bind to the nascent RNA promote or inhibit snRNP and spliceosome assembly. There are probably hundreds if not thousands of RNA-binding proteins that regulate RNA splicing. Their activities are regulated in different cell types and environments by levels of expression and posttranslational modifications. Thus, the spectrum of RNA-binding proteins expressed in a nucleus and the extent of modification of these proteins are important in alternative splicing and gene regulation.

The advent of deep sequencing, coupled with new methods to isolate RNA-protein complexes, is facilitating a new depth of understanding about the RNA species in a cell and how proteins control splicing, transport, translation, and the subcellular location of these species (Licatalosi et al., 2008). Indeed, deep-sequencing data from mammalian cells have recently provided evidence that over 90% of all transcription units are spliced in more than one pattern (Wang et al., 2008a). Whether each of the products resulting from this variation encodes functionally different proteins is not clear. But there are enough examples where distinct functions have been assigned to particular isoforms that the role of alternative splicing in gene regulation cannot be ignored.

Gene expression is integrated from reading of the information encoded in the DNA by the binding of transcription factors through processing of the nascent RNA, through transport and translation, and perhaps even to where a particular protein is localized in the cell. The transcription initiation complex forms under control of transcription factors, and its particular composition may vary between different promoters and cell states. In conjunction with initiation, the carboxy-terminal domain of polymerase II is modified by phosphorylation. Proteins that bind to transcription factors, the 7-methylguanine cap, the modified carboxy-terminal domain, or other components of the transcription complex are thought to interact with factors that promote steps in RNA splicing, cleavage, and polyadenylation. RNA processing also may be coupled to transcriptional processes directing chromatin structure and modifications. In fact, these processes can direct genes to specific locations in the nucleus, such as association with a nuclear pore complex, and thereby influence gene expression. Thus, sequences at some distance from a gene could influence the structure of complexes associated with the processed RNA and control its ultimate fate in cells.

Another indication of the interdependence of the various stages of RNA processing is that, as the RNA is spliced, a junction complex is placed on the RNA some 24 nucleotides upstream of the intron-excision site (Le Hir et al., 2000). This complex is part of a signal to transport the mRNA to the cytoplasm and to facilitate the pioneer round of translation. The complex also facilitates screening for a nonsense codon located upstream of an intron: a signal for nonsense-mediated decay of the mRNA. Thus, the concept that factors controlling transcription can also regulate posttranscriptional processes even in the cytoplasm is important. Conversely, it is likely that factors that bind to nascent RNA to direct splicing and other processing reactions can influence the nature of transcription along the gene through changes in pausing, processivity, and termination (Lacadie et al., 2006).

Gene Regulation by Small RNAs

About 10 years ago, scientists were stunned by the generality of gene regulation by RNA revealed through the discoveries of RNA interference (RNAi) (Fire et al., 1998) and microRNAs (Lee et al., 1993; Wightman et al., 1993). In both cases, the common result is that small RNAs silence gene expression by recognizing mRNAs and either directing their destruction or inhibiting their utilization in translation or both. In RNAi, any nucleic acid sequence can be converted into a *trans*-acting regulatory factor by conversion into double-stranded RNAs followed by processing to small RNAs by an enzyme termed Dicer. These small RNAs, or siRNAs, are then assembled into Argonaute complexes that can (1) direct cleavage of complementary RNA, (2) recognize mRNA through partially complementary sequences, direct translational silencing, and mRNA degradation, or (3) recognize nascent RNAs in the nucleus and direct assembly of heterochromatin, leading to silencing at the level of transcription.

MicroRNAs are expressed from genomic sequences forming hairpins. These RNA hairpins are recognized by the related RNase III enzymes Drosha and Dicer (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). In many cases, genomic sequences encoding microRNAs are conserved through vertebrate evolution and into other species such as insects and worms. By using evolutionary conservation of target sites as a criterion, microRNAs have been estimated to regulate roughly 50% of all mRNAs in vertebrates (Friedman et al., 2009). This regulation is primarily mediated through sequence recognition within the 3′ untranslated region (UTR) of mRNAs. Conserved sequences complementary to the microRNA "seed" sequence (bases occupying positions 2–7 from the 5′ end of the microRNA) are present at a 3-fold higher frequency than expected by any normalization protocol. In general, about half of all mRNAs appear to have short 3′UTRs and are not targets for this type of regulation, whereas those with longer 3′UTRs

are generally targets. Those mRNAs with longer 3′UTRs, on average, contain four conserved microRNA target sites per 3′UTR. Thus, there is extensive regulation of gene expression by microRNAs in these systems. Recent evidence also indicates that many diseases, such as cancer and autoimmune disorders, are associated with alterations in regulation by microRNAs.

Small-interfering RNAs (siRNAs) can target transcriptional regulation in a number of organisms. For example, in the fission yeast *Schizosaccharomyces pombe*, siRNAs specify heterochromatin formation by binding to nascent RNA and directing association of histone-modifying activities and factors important for this compact chromatin structure (Iida et al., 2008). Similarly, in plants, both chromatin structure and DNA modifications are directed in part by processes involving small RNAs (Herr et al., 2005). Again, these small RNAs appear to target nascent transcripts to specify local chromatin modifications and silencing. Thus small RNAs can become transcription factors under certain conditions.

Another type of small RNA—Piwi-interacting RNA (piRNA)—is generated in the germlines of many animals. Excitingly, these RNAs, and perhaps other types in conjunction, appear to control the expression of mobile repetitive elements in the genome by directing silencing processes, some of which are epigenetic in nature (Brennecke et al., 2007). Interestingly, such germline silencing seems to depend upon the combination of transcription generating two strands of complementary RNA and the concentrations of piRNAs and target sequences in complementary RNA (Brennecke et al., 2008). This system, as well as the processes described above for plants, provides an adaptive immunity to parasitic mobile agents that use transcription to generate resistance (Ketting and Plasterk, 2000). Although in theory any sequence in a genome might come under such epigenetic control if these conditions are met due to either mutation or induction of atypical gene expression in germline tissue, to date, this has not been observed for any endogenous mammalian gene. But it is clear that in the germline, silencing processes involving small RNAs are part of the explanation as to how parasitic mobile DNA agents, whose remnants constitute over half of all genomic sequences, can coexist with a stable genome.

Gene Regulation by Large Noncoding RNAs

The world of regulatory RNAs continues to expand with the emerging recognition that many large noncoding RNAs are expressed from the human genome (Katayama et al., 2005). Further, many of these RNAs contain sequences conserved in different species, a strong indication of an important function. The prototype large noncoding RNA, *Xist*, is associated with the inactive X chromosome in mammals (Brown et al., 1992). Other large noncoding RNAs include *U19* RNA, which is associated with imprinting, and *HOTAIR*, which is involved in the regulation of a *HOX* gene cluster (Rinn et al., 2007). There is a growing impression that epigenetic chromatin silencing in mammals can be influenced by RNA recognition. For example, transcriptional suppressor complexes that epigenetically regulate genes might have RNA as components. However, the biochemical nature of these processes is still poorly understood.

Future Insights about RNA

The biochemical properties of RNA are easily observable in the genetic material of many viruses, in the interactions of snRNAs with sequences at splice sites, and with catalysis of the peptidyl bond during translation. These processes certainly date from the RNA world, and it is highly likely that some aspects of RNA splicing do as well. It also seems very likely that RNA had multiple roles in gene regulation in this early stage of evolution. Here, small segments of RNA could pair with the RNA genome to control duplication, processing, and translation. It is possible that some of the activities closely identified with RNAi processes could reflect biochemical reactions that evolved before the appearance of the DNA genome. For example, the central catalytic domain in the Argonaute family of proteins is related to RNase H, an enzyme that specifically recognizes the intermediate between the RNA and the DNA worlds, an RNA-DNA hybrid (Song et al., 2004; Wang et al., 2008b). The Dicer and Drosha enzymes are members of the RNase III family, which recognizes double-stranded RNA and processes one of the oldest molecules, ribosomal RNA (Bernstein et al., 2001). The RNA-dependent RNA polymerase activities related to RNAi processes are of the same superfamily as those that transcribe DNA in contemporary cells and probably coevolved with DNA (Salgado et al., 2006). Thus, we can perhaps view RNAi as a snapshot of the transition from the RNA world to the DNA world.

Given that the RNA world and the DNA world seem so interconvertible, we might anticipate that an organism with an RNA genome could be discovered in some contemporary niche. For example, perhaps some cells growing in hyperthermic environments might use RNA as genetic material for a period and then convert to a DNA genome under other conditions. An analogy might be the transitions by micronuclei in germline and macronuclei in zygotic states in *Tetrahymena* and other ciliates (Aronica et al., 2008). However, even though the transition between these two states depends upon the RNA sequences within the cell and involves RNAi-related processes, the end product genetic material is DNA in both cellular states.

Many small RNA families have been identified, and elucidation of their roles in gene regulation is rapidly advancing. With the availability of deep-sequencing technology, most of the types of small RNAs less than 20–30 nucleotides probably have been discovered. However, the technology to explore the range of long noncoding RNAs within cells is in its infancy. It is easy to speculate three general functions for these species. First, they can provide a sequence-specific matrix for clustering components near genes or other sites on the genome to facilitate regulation. A large noncoding RNA could be the platform for organizing regulatory signals located at large distances from a gene or the clustering of groups of genes. Second, RNAs can be transported to different locations in cells such as poles of embryos, synapses, and specific subcellular compartments such as P-bodies and stress granules. In this capacity, noncoding RNAs could serve both as a carrier and a regulatory

signal for certain proteins and mRNAs. Third, RNA can be a structural component providing a track for movement of a cargo. Several thousand nucleotides in a noncoding RNA could span across a nucleus or over a significant fraction of the cytoplasm. Although these proposed functions are speculative, given existing examples of RNA's involvement in biological processes, it would not be surprising if RNA had these types of functions and many others.

Interestingly, the one thing that regulatory RNA has never been demonstrated to do under normal conditions is recognize DNA sequences directly by invasion of the helix. Perhaps the resulting RNA-DNA structure might be too recombinogenic or mutagenic. Such a structure is only commonly found in replicative intermediates where the RNA primes or templates DNA synthesis. Nonetheless, the ability of RNA to form both catalytic and secondary structures has made this molecule the most interesting nucleic acid in the cell. The myriad processes that are mediated or facilitated by RNA provide a window into an RNA world that existed before the evolution of its more stable sibling, DNA. The most surprising aspect of all of this is how late in the study of cell biology the importance and ubiquitous nature of RNA in gene regulation became widely recognized.

ACKNOWLEDGMENTS

The author thanks J. Neilson and A.C. Seila for help in editing the article.

REFERENCES

Aronica, L., Bednenko, J., Noto, T., DeSouza, L.V., Siu, K.W., Loidl, J., Pearlman, R.E., Gorovsky, M.A., and Mochizuki, K. (2008). Genes Dev. *22*, 2228–2241.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Nature *409*, 363–366.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Cell *128*, 1089–1103.

Brennecke, J., Malone, C.D., Aravin, A.A., Sachidanandam, R., Stark, A., and Hannon, G. (2008). Science *322*, 1387–1392.

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., and Willard, H.F. (1992). Cell *71*, 527–542.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Nature *391*, 806–811.

Friedman, R.C., Farh, K.K., Burge, C.B., and Bartel, D.P. (2009). Genome Res. *19*, 92–105.

Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. (2005). Science *308*, 118–120.

Iida, T., Nakayama, J., and Moazed, D. (2008). Mol. Cell *31*, 178–189.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., RIKEN Genome Exploration Research Group, Genome Science Group (Genome Network Project Core Group), FANTOM Consortium, et al. (2005) Science *309*, 1564–1566.

Ketting, R.F., and Plasterk, R.H. (2000). Nature *404*, 296–298.

Lacadie, S.A., Tardiff, D.F., Kadener, S., and Rosbash, M. (2006). Genes Dev. *20*, 2055–2066.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Science *294*, 853–858.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). Science *294*, 858–862.

Le Hir, H., Izaurralde, E., Maquat, L.E., and Moore, M.J. (2000). EMBO J. *19*, 6860–6869.

Lee, R.C., and Ambros, V. (2001). Science *294*, 862–864.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). Cell *75*, 843–854.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). Nature *456*, 464–469.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., and Chang, H.Y. (2007). Cell *129*, 1311–1323.

Salgado, P.S., Koivunen, M.R., Makeyev, E.V., Bamford, D.H., Stuart, D.I., and Grimes, J.M. (2006). PLoS Biol. *4*, e434. 10.1371/journal.pbio.0040434.

Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. (2004). Science *305*, 1434–1437.

Takahashi, K., and Yamanaka, S. (2006). Cell *126*, 663–676.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008a). Nature *456*, 470–476.

Wang, Y., Juranek, S., Li, H., Sheng, G., Tuschl, T., and Patel, D.J. (2008b). Nature *456*, 921–926.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Cell *75*, 855–862.