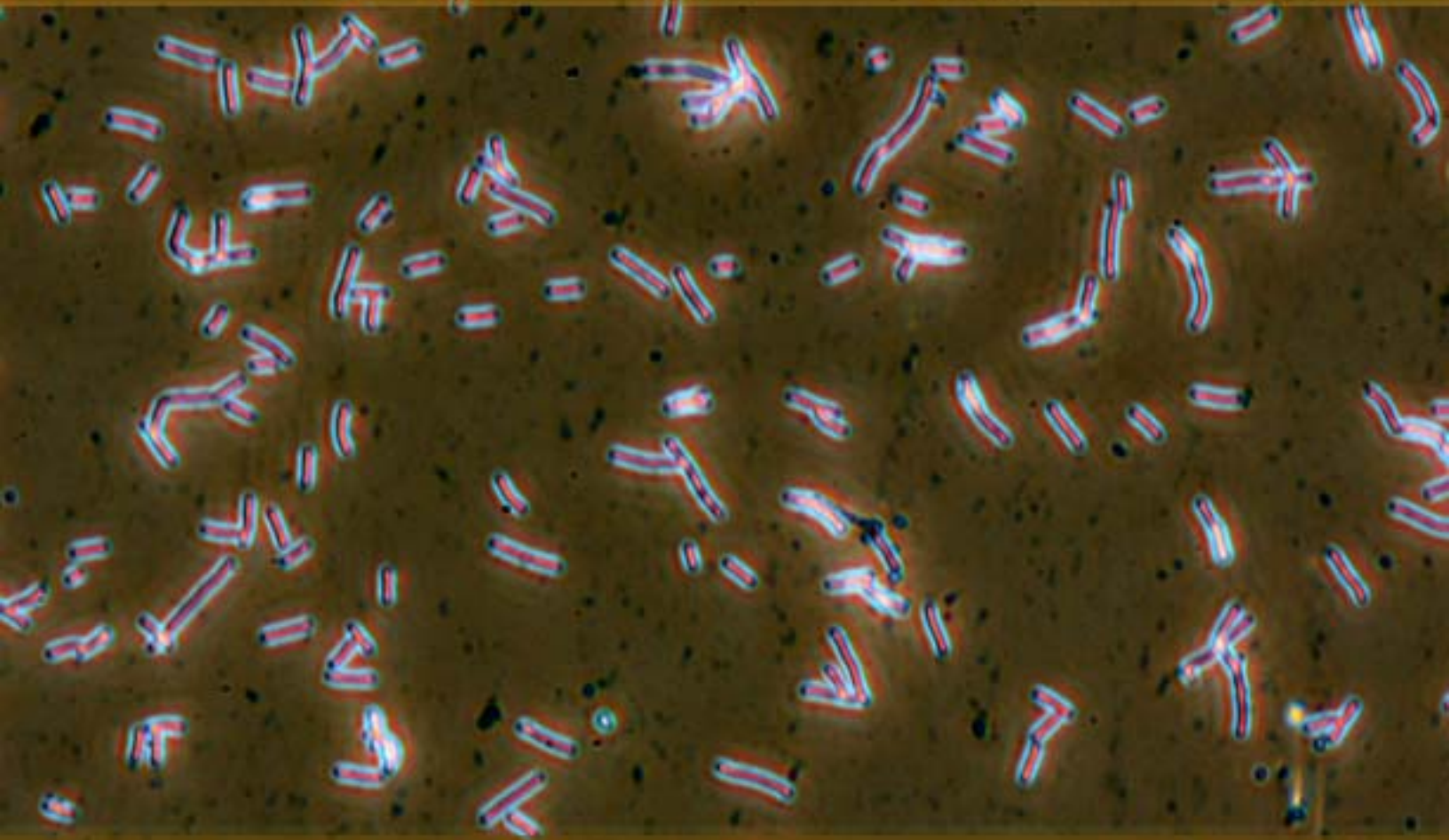
Important Concepts and Elements of Gene Expression & RNA Biology



Maragret Do

First Edition, 2012

ISBN 978-81-323-3311-1

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

Regulation of Gene Expression

Regulation of gene expression (or **gene regulation**) includes the processes that cells and viruses use to regulate the way that the information in genes is turned into gene products. Although a functional gene product may be an RNA or a protein, the majority of known mechanisms regulate protein coding genes. Any step of the gene's expression may be modulated, from DNA-RNA transcription to the post-translational modification of a protein.

Gene regulation is essential for viruses, prokaryotes and eukaryotes as it increases the versatility and adaptability of an organism by allowing the cell to express protein when needed. The first discovered example of a gene regulation system was the lac operon, discovered by Jacques Monod, in which protein involved in lactose metabolism are expressed by *E. coli* only in the presence of lactose and absence of glucose.

Furthermore, gene regulation drives the processes of cellular differentiation and morphogenesis, leading to the creation of different cell types in multicellular organisms where the different types of cells may possess different gene expression profiles though they all possess the same genome sequence.

Regulated stages of gene expression

Any step of gene expression may be modulated, from the DNA-RNA transcription step to post-translational modification of a protein. The following is a list of stages where gene expression is regulated, the most extensively utilised point is Transcription Initiation:

- Chromatin domains
- Transcription
- Post-transcriptional modification
- RNA transport
- Translation
- mRNA degradation

Modification of DNA

In eukaryotes, the accessibility of large regions of DNA can depend on its chromatin structure, which can be altered as a result of histone modifications directed by DNA methylation, ncRNA, or DNA-binding protein.

Chemical

Methylation of DNA is a common method of gene silencing. DNA is typically methylated by methyltransferase enzymes on cytosine nucleotides in a CpG dinucleotide sequence (also called "CpG islands" when densely clustered). Analysis of the pattern of methylation in a given region of DNA (which can be a promoter) can be achieved through a method called bisulfite mapping. Methylated cytosine residues are unchanged by the treatment, whereas unmethylated ones are changed to uracil. The differences are analyzed by DNA sequencing or by methods developed to quantify SNPs, such as Pyrosequencing (Biotage) or MassArray (Sequenom), measuring the relative amounts of C/T at the CG dinucleotide. Abnormal methylation patterns are thought to be involved in oncogenesis.

Structural

Transcription of DNA is dictated by its structure. In general, the density of its packing is indicative of the frequency of transcription. Octameric protein complexes called nucleosomes are responsible for the amount of supercoiling of DNA, and these complexes can be temporarily modified by processes such as phosphorylation or more permanently modified by processes such as methylation. Such modifications are considered to be responsible for more or less permanent changes in gene expression levels.

Histone acetylation is also an important process in transcription. Histone acetyltransferase enzymes (HATs) such as CREB-binding protein also dissociate the DNA from the histone complex, allowing transcription to proceed. Often, DNA methylation and histone deacetylation work together in gene silencing. The combination of the two seems to be a signal for DNA to be packed more densely, lowering gene expression.

Regulation of transcription

Regulation of transcription controls when transcription occurs and how much RNA is created. Transcription of a gene by RNA polymerase can be regulated by at least five mechanisms:

- **Specificity factors** alter the specificity of RNA polymerase for a given promoter or set of promoters, making it more or less likely to bind to them (i.e., sigma factors used in prokaryotic transcription).

- **Repressors** bind to non-coding sequences on the DNA strand that are close to or overlapping the promoter region, impeding RNA polymerase's progress along the strand, thus impeding the expression of the gene.
- **General transcription factors** position RNA polymerase at the start of a protein-coding sequence and then release the polymerase to transcribe the mRNA.
- **Activators** enhance the interaction between RNA polymerase and a particular promoter, encouraging the expression of the gene. Activators do this by increasing the attraction of RNA polymerase for the promoter, through interactions with subunits of the RNA polymerase or indirectly by changing the structure of the DNA.
- **Enhancers** are sites on the DNA helix that are bound to by activators in order to loop the DNA bringing a specific promoter to the initiation complex. Enhancers are much more common in eukaryote than prokaryotes, where only a few examples exist (to date).

Post-transcriptional regulation

After the DNA is transcribed and mRNA is formed, there must be some sort of regulation on how much the mRNA is translated into proteins. Cells do this by modulating the capping, splicing, addition of a Poly(A) Tail, the sequence-specific nuclear export rates, and, in several contexts, sequestration of the RNA transcript. These processes occur in eukaryotes but not in prokaryotes. This modulation is a result of a protein or transcript that, in turn, is regulated and may have an affinity for certain sequences.

Regulation of translation

The translation of mRNA can also be controlled by a number of mechanisms, mostly at the level of initiation. Recruitment of the small ribosomal subunit can indeed be modulated by mRNA secondary structure, antisense RNA binding, or protein binding. In both prokaryotes and eukaryotes, a large number of RNA binding proteins exist, which often are directed to their target sequence by the secondary structure of the transcript, which may change depending on certain conditions, such as temperature or presence of a ligand (aptamer). Some transcripts act as ribozymes and self-regulate their expression.

Examples of gene regulation

- Enzyme induction is a process in which a molecule (e.g., a drug) induces (i.e., initiates or enhances) the expression of an enzyme.
- The induction of heat shock proteins in the fruit fly *Drosophila melanogaster*.
- The Lac operon is an interesting example of how gene expression can be regulated.
- Viruses, despite having only a few genes, possess mechanisms to regulate their gene expression, typically into an early and late phase, using collinear systems regulated by anti-terminators (lambda phage) or splicing modulators (HIV).

Developmental biology

A large number of studied regulatory systems come from developmental biology. Examples include:

- The colinearity of the Hox gene cluster with their nested antero-posterior patterning
- It has been speculated that pattern generation of the hand (digits - interdigits) The gradient of Sonic hedgehog (secreted inducing factor) from the zone of polarizing activity in the limb, which creates a gradient of active Gli3, which activates Gremlin, which inhibits BMPs also secreted in the limb, resulting in the formation of an alternating pattern of activity as a result of this reaction-diffusion system.
- Somitogenesis is the creation of segments (somites) from a uniform tissue (Pre-somitic Mesoderm, PSM). They are formed sequentially from anterior to posterior. This is achieved in amniotes possibly by means of two opposing gradients, Retinoic acid in the anterior (wavefront) and Wnt and Fgf in the posterior, coupled to an oscillating pattern (segmentation clock) composed of FGF + Notch and Wnt in antiphase.
- Sex determination in the soma of a Drosophila requires the sensing of the ratio of autosomal genes to sex chromosome-encoded genes, which results in the production of sexless splicing factor in females, resulting in the female isoform of doublesex.

Circuitry

Up-regulation and down-regulation

Up-regulation is a process that occurs within a cell triggered by a signal (originating internal or external to the cell), which results in increased expression of one or more genes and as a result the protein(s) encoded by those genes. On the converse, **down-regulation** is a process resulting in decreased gene and corresponding protein expression.

- Up-regulation occurs, for example, when a cell is deficient in some kind of receptor. In this case, more receptor protein is synthesized and transported to the membrane of the cell and, thus, the sensitivity of the cell is brought back to normal, reestablishing homeostasis.

- Down-regulation occurs, for example, when a cell is overstimulated by a neurotransmitter, hormone, or drug for a prolonged period of time, and the expression of the receptor protein is decreased in order to protect the cell.

Inducible vs. repressible systems

Gene Regulation can be summarized as how they respond:

- Inducible systems - An inducible system is off unless there is the presence of some molecule (called an inducer) that allows for gene expression. The molecule is said to "induce expression". The manner by which this happens is dependent on the control mechanisms as well as differences between prokaryotic and eukaryotic cells.

- Repressible systems - A repressible system is on except in the presence of some molecule (called a corepressor) that suppresses gene expression. The molecule is said to "repress expression". The manner by which this happens is dependent on the control mechanisms as well as differences between prokaryotic and eukaryotic cells.

Theoretical circuits

- Repressor/Inducer: an activation of a sensor results in the change of expression of a gene
- negative feedback: the gene product downregulates its own production directly or indirectly, which can result in
    - keeping transcript levels constant/proportional to a factor
    - inhibition of run-away reactions when coupled with a positive feedback loop
    - creating an oscillator by taking advantage in the time delay of transcription and translation, given that the mRNA and protein half-life is shorter
- positive feedback: the gene product upregulates its own production directly or indirectly, which can result in
    - signal amplification
    - bistable switches when two genes inhibit each other and have both positive feedback
    - pattern generation

Methods

In general, most experiments investigating differential expression used whole cell extracts of RNA, called steady-state levels, to determine which genes changed and by how much they did. These are, however, not informative of where the regulation has occurred and may actually mask conflicting regulatory processess, but it is still the most commonly analysed (QPCR and DNA microarray).

When studying gene expression, there are several methods to look at the various stages. In eukaryotes these include:

- The chromatin conformation of the region can be determined by ChIP-chip analysis by pulling down RNA Polymerase II, Histone 3 modifications, Trithorax-group protein, Polycomb-group protein, or any other DNA-binding element to which a good antibody is available.

- Epistatic interactions can be investigated by synthetic genetic array analysis
- Due to post-transcriptional regulation, transcription rates and total RNA levels differ significantly. To measure the transcription rates nuclear run-on assays can be done and newer high-throughput methods are being developed, using thiol labelling instead of radioactivity.
- Only 5% of the RNA polymerised in the nucleus actually exists, and not only introns, abortive products, and non-sense transcripts are degradated. Therefore, the differences in nuclear and cytoplasmic levels can be see by separating the two fractions by gentle lysis.
- Alternative splicing can be analysed with a splicing array or with a tiling array.
- All in vivo RNA is complexed as RNPs. The quantity of transcripts bound to specific protein can be also analysed by RIP-Chip. For example, DCP2 will give an indication of sequestered protein; ribosome-bound gives and indication of transcripts active in transcription (although it should be noted that a more dated method, called polysome fractionation, is still popular in some labs)
- Protein levels can be analysed by Mass spectrometry, which can be compared only to QPCR data, as microarray data is relative and not absolute.
- RNA and protein degradation rates are measured by means of transcription inhibitors (actinomycin D or α-amanitin) or translation inhibitors (Cycloheximide), respectively.

Chapter- 2

Piwi-Interacting RNA

Piwi-interacting RNA (**piRNA**) is the largest class of small RNA molecules that is expressed in animal cells. piRNA forms RNA-protein complexes through interactions with Piwi proteins. These piRNA complexes have been linked to transcriptional gene silencing of retrotransposons and other genetic elements in germ line cells, particularly those in spermatogenesis. They are distinct from miRNA in size (26–31 nt rather than 21–24 nt), lack of sequence conservation, and increased complexity .

It remains unclear how piRNAs are generated, but potential methods have been suggested, and it is certain their biogenesis pathway is distinct from miRNA and siRNA, while rasiRNAs are a piRNA subspecies.

Characteristics

Proposed piRNA structure

piRNAs have been identified in both vertebrates and invertebrates, and although biogenesis and modes of action do vary somewhat between species, a number of features are conserved. piRNAs have no clear secondary structure motifs, the length of a piRNA is, by definition, between 26 and 31 nucleotides, and the presence of a 5’ uridine is common to piRNAs in both vertebrates and invertebrates. piRNAs in *C. elegans* have a 5’ monophosphate and a 3’ modification that acts to block either the 2’ or 3’ oxygen, and this has also been confirmed to exist in *D.melanogaster*, zebrafish, mice and rats . This 3’ modification is likely to be a 2’-O-methylation, but the reason for this modification is not known. It is thought that there are many hundreds of thousands of different piRNA species found in mammals. Thus far, over 50,000 unique piRNA sequences have been discovered in mice and more than 13,000 in *D.melanogaster*.

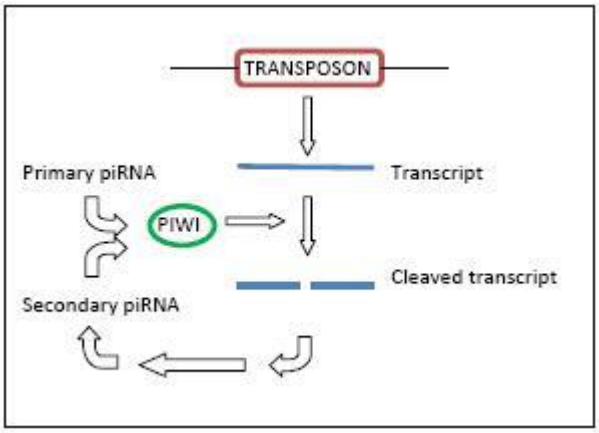
Location

piRNAs are found in clusters throughout the genome; these clusters may contain as few as ten or up to many thousands of piRNAs and can vary in size from one to one hundred kb . While the clustering of piRNAs is highly conserved across species, the sequences are not. While *D.melanogaster* and vertebrate piRNAs have been located in areas lacking any protein coding genes, piRNAs in *C. elegans* have been identified amidst protein coding genes .

In mammals, piRNAs are found only within the testes, with an estimated one million copies per cell in spermatocytes and spermatids. In invertebrates, piRNAs have been detected in both the male and female germlines, but in no other cell types.

At the cellular level, piRNAs have been found within both nuclei and cytoplasm, suggesting that piRNA pathways may function in both of these areas and, therefore, may have multiple effects.

Biogenesis



piRNA ping pong mechanism

The biogenesis of piRNAs is not yet fully understood, although possible mechanisms have been proposed. piRNAs show a significant strand bias, that is, they are derived from one strand of DNA only, and this may indicate that they are the product of long single stranded precursor molecules. A primary processing pathway is suggested to be the only pathway used to produce pachytene piRNAs; in this mechanism, piRNA precusors are transcribed resulting in piRNAs with a tendency to target 5' uridines. Also proposed is a 'Ping Pong' mechanism wherein primary piRNAs recognise their complementary targets and cause the recruitment of Piwi proteins. This results in the cleavage of the transcript at a point ten nucleotides from the 5' end of the primary piRNA, producing the secondary piRNA. These secondary piRNAs are targeted toward sequences that possess an adenine at the tenth position. Since the piRNA involved in the ping pong cycle directs its attacks on transposon transcripts, the ping pong cycle acts only at the level of transcription. One

or both of these mechanisms may be acting in different species; *C. elegans*, for instance, does have piRNAs, but does not appear to use the ping pong mechanism at all.

A significant number of piRNAs identified in zebrafish and *D.melanogaster* contain adenine at their tenth position, and this has been interpreted as possible evidence of a conserved biosynthetic mechanism across species. piRNAs are expressed through unique pathways, which are dissimilar to the expression pathways utilised by other small RNAs. Neither Argonaute proteins, RNA-dependent RNA polymerases, nor other proteins required for most small-RNA biosynthesis are necessary for piRNA expression. However, while PIWI proteins are not required for the biosynthesis of piRNAs, evidence suggests that PIWI proteins must be present to stabilise piRNAs and, thus, facilitate their accumulation. No mechanism for the control of piRNA propagation has yet been established.

Function

The wide variation in piRNA sequences and PIWI function over species contributes to the difficulty in establishing the functionality of piRNAs. However, like other small RNAs, piRNAs are thought to be involved in gene silencing, specifically the silencing of transposons. The majority of piRNAs are antisense to transposon sequences, suggesting that transposons are the piRNA target. In mammals it appears that the activity of piRNAs in transposon silencing is most important during the development of the embryo, and in both *C. elegans* and humans, piRNAs are necessary for spermatogenesis.

RNA Silencing

piRNA has a role in RNA silencing via the formation of an RNA-induced silencing complex (RISC). piRNAs interact with Piwi proteins that are part of a family of proteins called the Argonautes. These are active in the testes of mammals and are required for germ-cell and stem-cell development in invertebrates. Three Piwi subfamily proteins - MIWI, MIWI2 and MILI - have been found to be essential for spermatogenesis in mice. piRNAs direct the Piwi proteins to their transposon targets. A decrease or absence of PIWI protein expression is correlated with an increased expression of transposons. Transposons have a high potential to cause deleterious effect on their host, and, in fact, mutations in piRNA pathways are found to reduce fertility in *D.melanogaster*. However, piRNA pathway mutations in mice do not demonstrate reduced fertility; this may indicate redundancies to the piRNA system. Further, it is thought that piRNA and endogenous small interfering RNA (endo-siRNA) may have comparable and even redundant functionality in transposon control in mammalian oocytes.

piRNAs appear to have an impact on particular methyltransferases that perform the methylations which are required to recognise and silence transposons, but this relationship is not well understood.

Epigenetic Effects

piRNAs can be transmitted maternally, and based on research in *D.melanogaster*, piRNAs may be involved in maternally derived epigenetic effects. The activity of specific piRNAs in the epigenetic process also requires interactions between Piwi proteins and HP1a, as well as other factors.

Recent discovery also show, the existence of snoRNA, microRNA, piRNA characteristics in a novel non-coding RNA: x-ncRNA and its biological implication in Homo sapiens.

Investigation

Due to their small size, expression and amplification of small RNAs can be challenging, so specialised PCR-based methods have been developed in response to this difficulty.

Chapter- 3

RasiRNA and Small Interfering RNA

RasiRNA

Repeat associated small interfering RNA (rasiRNA) is a class of small RNA that is involved in the RNA interference (RNAi) pathway. RasiRNA is a subclass of Piwi-interacting RNAs (piRNAs), which are small RNA molecule which interact with Piwi proteins. Piwi proteins are an Argonaute subfamily. RasiRNA associate with both the Ago and Piwi Argonaute protein subfamily while piRNA only associates with the Piwi Argonaute subfamily. In the germline, RasiRNA is involved in establishing and maintaining heterochromatin structure, controlling transcripts that emerge from repeat sequences, and silencing transposons and retrotransposons.

Classification

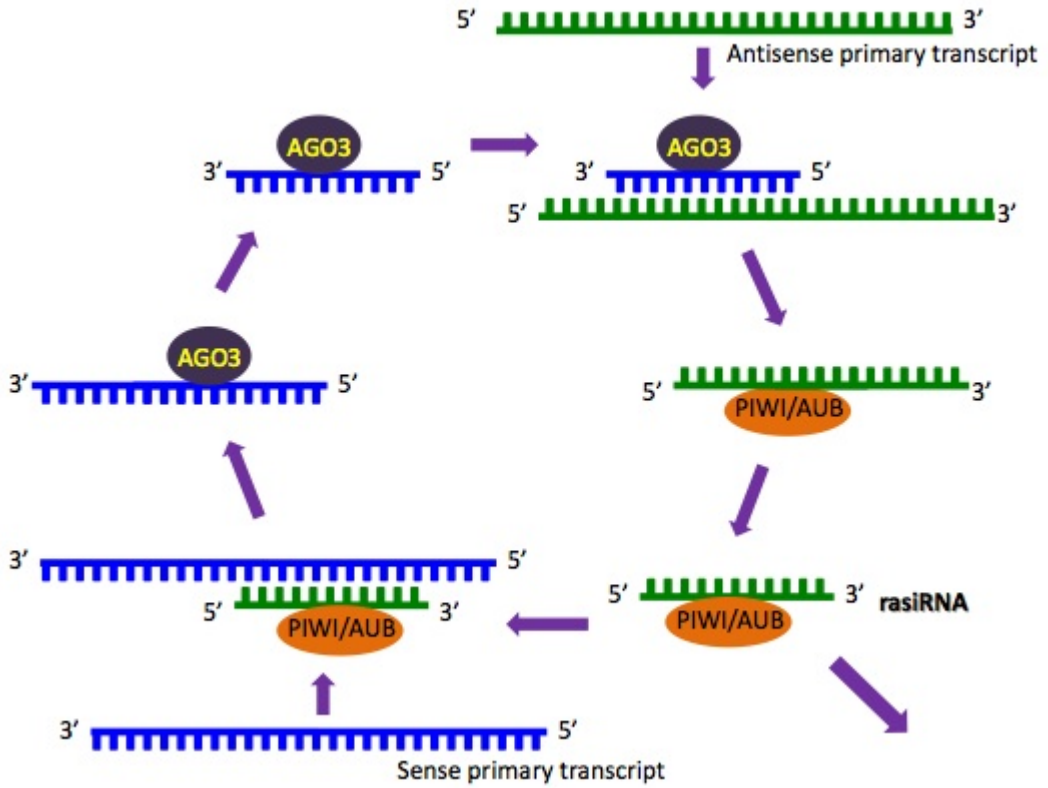
There are at least three Argonaute subfamilies that have been found in eukaryotes. Unlike the Ago subfamily which is present in animals, plants, and fission yeast, the Piwi subfamily has only been found in animals. RasiRNA has been observed in *Drosophila* and some unicellular eukaryotes but its presence in mammals has not been determined unlike piRNA which has been observed in many species of invertebrates and vertebrates including mammals; however, since proteins which associate with rasiRNA are found in both vertebrates and invertebrates, it is possible that active rasiRNA exist and have yet to be observed in other animals. RasiRNAs have been observed in Schizosaccharomyces pombe, a species of yeast, as well in some plants both of which have not been observed to contain the Piwi Argonaute protein subfamily. It has been observed that both rasiRNA and piRNA are maternally linked, but more specifically it is the Piwi protein subfamily that is maternally linked and therefore leads to the observation that rasiRNA and piRNA are maternally linked.

RasiRNA are Distinct from other RNAi pathways

RasiRNA is distinct from other RNAi pathways such as microRNA (miRNA) and small interfering RNA (siRNA) as well as from piRNA. Unlike miRNA and siRNA which function through the Ago Argonaute protein subfamily, RasiRNA function through the Piwi Argonaute protein subfamily. RasiRNA is also distinct in its size. Contrary to miRNAs which are 21-23 nucleotides in length, siRNAs which are 20-25 nucleotides in

length, and piRNAs which are 24-31 nucleotides in length, rasiRNAs are 24-29 nucleotides in length depending on the organism it came from. Unlike siRNA which are derived from both the sense and antisense strand, rasiRNA are derived from the antisense. Interestingly, while miRNA requires Dicer-1 for its production, and siRNA requires Dicer-2, rasiRNA does not require either; however, in some plants there are Dicer-like (Dcl) proteins that have been identified where Dcl1 produces 24 nucleotide miRNA and siRNA while Dcl2 produces 24 nucleotide rasiRNA. This research shows that not only is rasiRNA production distinct from miRNA and siRNA, but that rasiRNA may be found in plants while Piwi proteins are not.

RasiRNA Biogenesis



The ping-pong mechanism for the biogenesis of the 5' end of rasiRNA

It is presumed that the source of rasiRNA is double stranded RNA produced by annealing of sense and antisense related transposable elements. The biogenesis of rasiRNA is independent of Dicer, but does require the Argonaute proteins Argonaute 3 (Ago 3), Piwi and Aubergine which is a Piwi-like protein. The mechanism for rasiRNA biogenesis is a ping-pong mechanism. The Piwi/Aub associated RNA is the rasiRNA. The rasiRNAs match the antisense strand of retrotransposons and repetitive sequence elements (hence the name rasiRNA). The Ago3 associated RNAs are derived from the sense strand. The ping-pong mechanism which is observed in this image is the mechanism for the generation of the 5’ end of rasiRNA while the generation of the 3’ end of rasiRNA is still unknown.

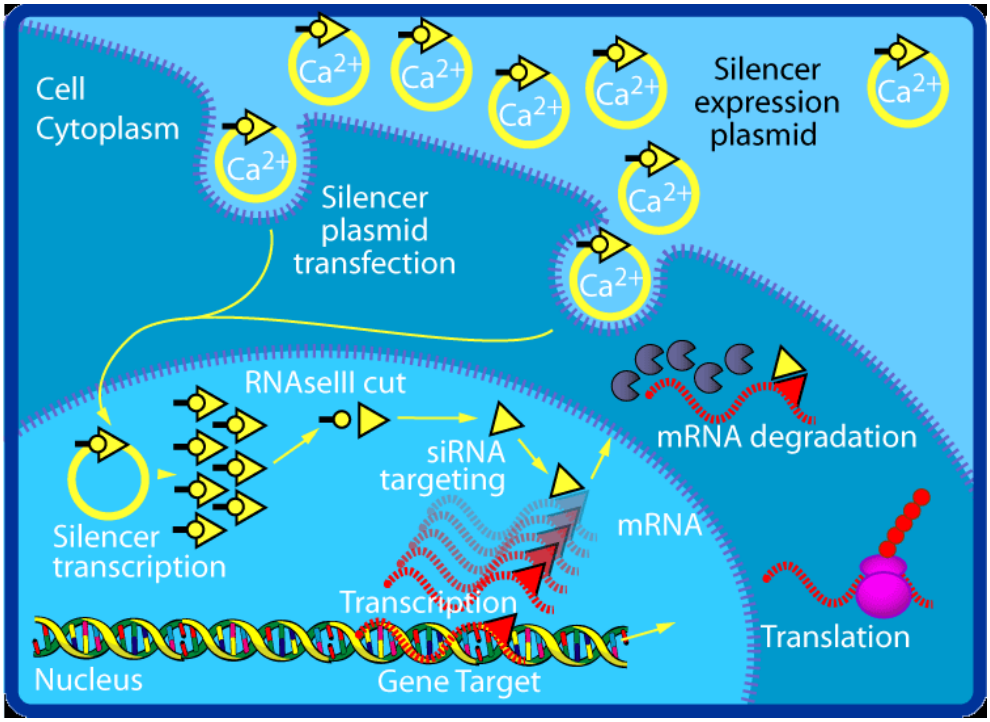
Importance of rasiRNA

While miRNA act in translational repression and mRNA cleavage and siRNA act in mRNA cleavage, rasiRNA act to regulate chromatin structure and transcriptional silencing. In Drosophila, mutations in the Piwi proteins that associate with rasiRNA lead to sterility and loss of germline cells in both males and females. Transposon repression is not affected by the loss of Dicer within the germline cells revealing that this is the target of the rasiRNA pathway. Similar to miRNA and siRNA, the rasiRNA silencing pathway is evolutionarily conserved and homology dependent. When the rasiRNA pathway is not present, germline cells may undergo retrotransposition which are sensed as DNA damage and signal the cell to apoptosis. RasiRNA is key to the regulatory mechanism of many organisms as part of the RNA interference pathway.

History and Discovery

Small RNAs guiding RNA silencing pathways were first discovered in 1993 in Caenorhabditis elegans and since then have been observed in many organisms. RasiRNA were discovered sometime after 1993 and observed in both Drosophila melanogaster and Schizosaccharomyces pombe and since have been observed in some single celled eukaryotes and some plants.

Small interfering RNA



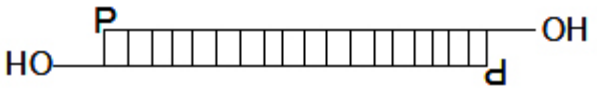
Mediating RNA interference in cultured mammalian cells

Small interfering RNA (**siRNA**), sometimes known as **short interfering RNA** or **silencing RNA**, is a class of double-stranded RNA molecules, 20-25 nucleotides in length, that play a variety of roles in biology. Most notably, siRNA is involved in the RNA interference (RNAi) pathway, where it interferes with the expression of a specific gene. In addition to their role in the RNAi pathway, siRNAs also act in RNAi-related pathways, e.g., as an antiviral mechanism or in shaping the chromatin structure of a genome; the complexity of these pathways is only now being elucidated.

siRNAs were first discovered by David Baulcombe's group at the Sainsbury Laboratory in Norwich, England, as part of post-transcriptional gene silencing (PTGS) in plants. The group published their findings in *Science* in a paper titled "A species of small antisense RNA in posttranscriptional gene silencing in plants". Shortly thereafter, in 2001, synthetic siRNAs were shown to be able to induce RNAi in mammalian cells by Thomas Tuschl, and colleagues in a paper published in *Nature*. This discovery led to a surge in interest in harnessing RNAi for biomedical research and drug development.

Structure

siRNAs have a well-defined structure: a short (usually 21-nt) double strand RNA (dsRNA) with 2-nt 3' overhangs on either end:



Schematic representation of a siRNA molecule: a ~19-21basepair RNA core duplex that is followed by a 2 nucleotide 3' overhang on each strand. OH: 3' hydroxyl; P: 5' phosphate.

Each strand has a 5' phosphate group and a 3' hydroxyl (-OH) group. This structure is the result of processing by dicer, an enzyme that converts either long dsRNAs or small hairpin RNAs into siRNAs. siRNAs can also be exogenously (artificially) introduced into cells by various transfection methods to bring about the specific knockdown of a gene of interest. Essentially any gene of which the sequence is known can thus be targeted based on sequence complementarity with an appropriately tailored siRNA. This has made siRNAs an important tool for gene function and drug target validation studies in the post-genomic era.

RNAi induction using siRNAs or their biosynthetic precursors

Dicer protein colored by protein domain

Transfection of an exogenous siRNA can be problematic because the gene knockdown effect is only transient, particularly in rapidly dividing cells. One way of overcoming this challenge is to modify the siRNA in such a way as to allow it to be expressed by an appropriate vector, e.g., a plasmid. This is done by the introduction of a loop between the two strands, thus producing a single transcript, which can be processed into a functional siRNA. Such transcription cassettes typically use an RNA polymerase III promoter (e.g., U6 or H1), which usually directs the transcription of small nuclear RNAs (snRNAs) (U6 is involved in gene splicing; H1 is the RNase component of human RNase P). It is assumed (although not known for certain) that the resulting siRNA transcript is then processed by Dicer.

RNA activation

It has recently been found that dsRNA can also activate gene expression, a mechanism that has been termed "small RNA-induced gene activation" or RNAa. It has been shown that dsRNAs targeting gene promoters induce potent transcriptional activation of associated genes. RNAa was demonstrated in human cells using synthetic dsRNAs,

termed "small activating RNAs" (saRNAs). It is currently not known whether RNAa is conserved in other organisms.

Challenges: avoiding nonspecific effects

Because RNAi intersects with a number of other pathways, it is not surprising that on occasion nonspecific effects are triggered by the experimental introduction of an siRNA. When a mammalian cell encounters a double-stranded RNA such as an siRNA, it may mistake it as a viral by-product and mount an immune response. Furthermore, because structurally related microRNAs modulate gene expression largely via incomplete complementarity base pair interactions with a target mRNA, the introduction of an siRNA may cause unintended off-targeting.

Innate immunity

Introduction of too much siRNA can result in nonspecific events due to activation of innate immune responses. Most evidence to date suggests that this is probably due to activation of the dsRNA sensor PKR, although retinoic acid inducible gene I (RIG-I) may also be involved. The induction of cytokines via toll-like receptor 7 (TLR7) has also been described. One promising method of reducing the nonspecific effects is to convert the siRNA into a microRNA. MicroRNAs occur naturally, and by harnessing this endogenous pathway it should be possible to achieve similar gene knockdown at comparatively low concentrations of resulting siRNAs. This should minimize nonspecific effects.

Off-targeting

Off-targeting is another challenge to the use of siRNAs as a gene knockdown tool. Here, genes with incomplete complementarity are inadvertently downregulated by the siRNA (effectively, the siRNA acts as a miRNA), leading to problems in data interpretation and potential toxicity. This, however, can be partly addressed by designing appropriate control experiments, and siRNA design algorithms are currently being developed to produce siRNAs free from off-targeting. Genome-wide expression analysis, e.g., by microarray technology, can then be used to verify this and further refine the algorithms. A 2006 paper from the laboratory of Dr. Khvorova implicates 6- or 7-basepair-long stretches from position 2 onward in the siRNA matching with 3'UTR regions in off-targeted genes.

Possible therapeutic applications and challenges

Given the ability to knock down essentially any gene of interest, RNAi via siRNAs has generated a great deal of interest in both basic and applied biology. There are an increasing number of large-scale RNAi screens that are designed to identify the important genes in various biological pathways. Because disease processes also depend on the activity of multiple genes, it is expected that in some situations turning off the activity of a gene with an siRNA could produce a therapeutic benefit.

However, applying RNAi via siRNAs to living animals, especially humans, poses many challenges. Experimentally, siRNAs show different effectiveness in different cell types in a manner as yet poorly understood: some cells respond well to siRNAs and show a robust knockdown, whereas others show no such knockdown (even despite efficient transfection).

Phase I results of the first two therapeutic RNAi trials (indicated for age-related macular degeneration, aka AMD) reported at the end of 2005 that siRNAs are well tolerated and have suitable pharmacokinetic properties.

Proof of concept trials have indicated that Ebola-targeted siRNAs may be effective as post-exposure prophylaxis in humans, with 100% of non-human primates surviviing a lethal dose of Zaire Ebolavirus, the most lethal strain.

Chapter- 4

Transcriptional Regulation

Transcriptional regulation is the change in gene expression levels by altering transcription rates.

Regulation of transcription

Regulation of transcription controls when transcription occurs and how much RNA is created. Transcription of a gene by RNA polymerase can be regulated by at least five mechanisms:

- **Specificity factors** alter the specificity of RNA polymerase for a given promoter or set of promoters, making it more or less likely to bind to them (i.e. sigma factors used in prokaryotic transcription).
- **Repressors** bind to non-coding sequences on the DNA strand that are close to or overlapping the promoter region, impeding RNA polymerase's progress along the strand, thus impeding the expression of the gene.
- **General transcription factors** These transcription factors position RNA polymerase at the start of a protein-coding sequence and then release the polymerase to transcribe the mRNA.
- **Activators** enhance the interaction between RNA polymerase and a particular promoter, encouraging the expression of the gene. Activators do this by increasing the attraction of RNA polymerase for the promoter, through interactions with subunits of the RNA polymerase or indirectly by changing the structure of the DNA.
- **Enhancers** are sites on the DNA helix that are bound to by activators in order to loop the DNA bringing a specific promoter to the initiation complex.

Regulatory protein

Regulatory protein is a term used in genetics to describe a protein involved in regulating gene expression. It is usually bound to a DNA binding site which is sometimes located near the promoter although this is not always the case. Regulatory proteins are often needed to be bound to a regulatory binding site to switch a gene on (activator) or to shut off a gene (repressor). Generally, as the organism grows more sophisticated, their cellular protein regulation becomes more complicated and indeed some human genes can be controlled by many activators and repressors working together.

Prokaryotes vs. eukaryotes

In prokaryotes, regulation of transcription is needed for the cell to quickly adapt to the ever-changing outer environment. The presence or the quantity and type of nutrients determines which genes are expressed; in order to do that, genes must be regulated in some fashion. In prokaryotes, repressors bind to regions called operators that are generally located downstream from and near the promoter (normally part of the transcript). Activators bind to the upstream portion of the promoter, such as the CAP region (completely upstream from the transcript). A combination of activators, repressors and rarely enhancers (in prokaryotes) determines whether a gene is transcribed.

In eukaryotes, transcriptional regulation tends to involve combinatorial interactions between several transcription factors, which allow for a sophisticated response to multiple conditions in the environment. This permits spatial and temporal differences in gene expression. Eukaryotes also make use of enhancers, distant regions of DNA that can loop back to the promoter. A major difference between eukaryotes and prokaryotes is the fact the eukaryotes have a nuclear envelope, which prevents simultaneous transcription and translation. RNA interference also regulate gene expression in most eukaryotes, both by epigenetic modification of promoters and by breaking down mRNA.

Examples

Examples:

- When E. coli bacteria are subjected to heat stress, the σ^{32} subunit of its RNA polymerase changes such that the enzyme binds to a specialized set of promoters that precede genes for heat-shock response proteins.

- When a cell contains a surplus amount of the amino acid tryptophan, the acid binds to a specialized repressor protein (tryptophan repressor). The binding changes the structural conformity of the repressor such that it binds to the operator region for the operon that synthesizes tryptophan, preventing their expression and thus suspending production. This is a form of negative feedback.

- In bacteria, the lac repressor protein blocks the synthesis of enzymes that digest lactose when there is no lactose to feed on. When lactose is present, it binds to the repressor, causing it to detach from the DNA strand.

Inducible vs. repressible systems

Gene Regulation can be summarized as how they respond:

- Inducible systems - An inducible system is off unless there is the presence of some molecule (called an inducer) that allows for gene expression. The molecule is said to "induce expression". The manner in which this happens is dependent on

the control mechanisms as well as differences between prokaryotic and eukaryotic cells.

- Repressible systems - A repressible system is on except in the presence of some molecule (called a corepressor) that suppresses gene expression. The molecule is said to "repress expression". The manner in which this happens is dependent on the control mechanisms as well as differences between prokaryotic and eukaryotic cells.

Regulation of transcription machinery

In order for a gene to be expressed, several things must happen. First, there needs to be an initiating signal. This is achieved through the binding of some ligand to a receptor. Activation of g-protein-coupled receptors can have this effect; as can the binding of hormones to intra- or extracellular receptors.

This signal gives rise to the activation of a protein called a transcription factor, and recruits other members of the "transcription machine." Transcription factors generally simultaneously bind DNA as well as an RNA polymerase, as well as other agents necessary for the transcription process (HATs, scaffolding proteins, etc.). Transcription factors, and their cofactors, can be regulated through reversible structural alterations such as phosphorylation or inactivated through such mechanisms as proteolysis.

Transcription is initiated at the promoter site, as an increase in the amount of an active transcription factor binds a target DNA sequence. Other proteins, known as "scaffolding proteins" bind other cofactors and hold them in place. DNA sequences far from the point of initiation, known as enhancers, can aid in the assembly of this "transcription machinery." Histone arms are acetylated, and DNA is transcribed into RNA.

Frequently, extracellular signals induce the expression of immediate early genes (IEGs) such as c-fos, c-jun, or AP-1. These are in and of themselves transcription factors or components thereof, and can further influence gene expression.

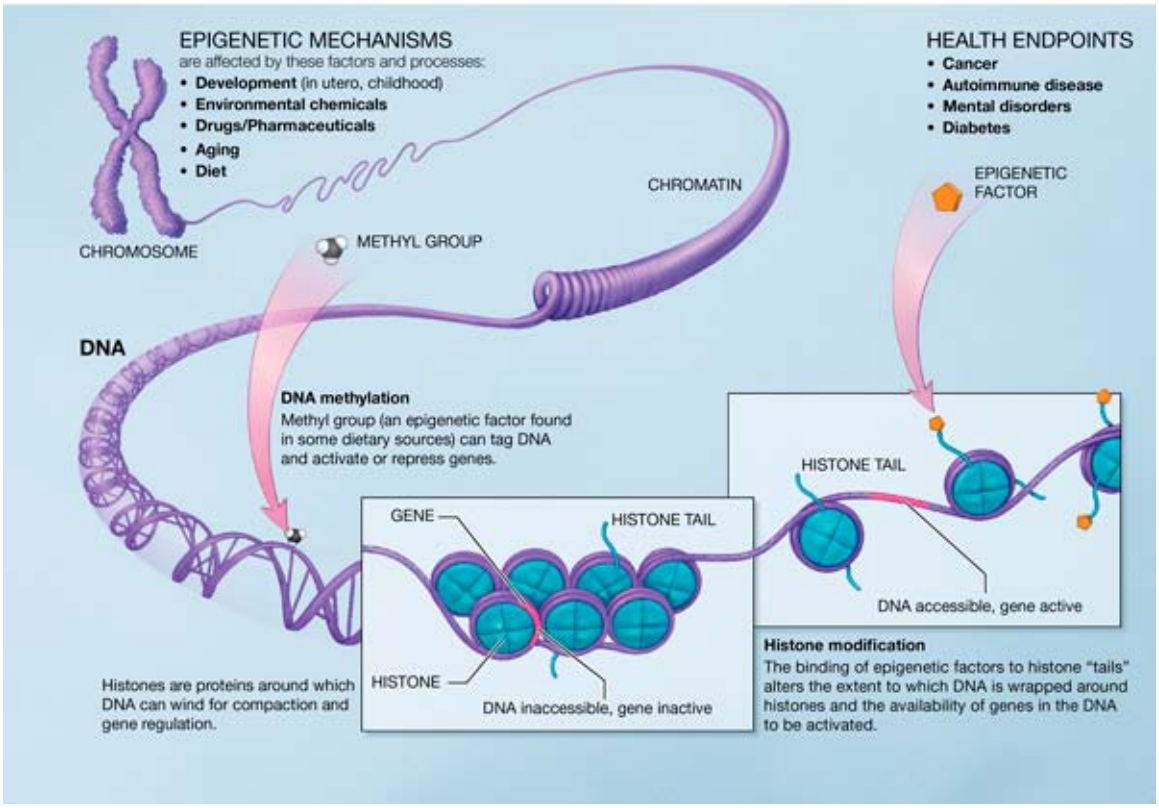
Chapter- 5

Epigenetics

In biology, and specifically genetics, **epigenetics** is the study of heritable changes in phenotype (appearance) or gene expression caused by mechanisms other than changes in the underlying DNA sequence, hence the name *epi-* (Greek: *επί-* over, above) *-genetics*. These changes may remain through cell divisions for the remainder of the cell's life and may also last for multiple generations. However, there is no change in the underlying DNA sequence of the organism; instead, non-genetic factors cause the organism's genes to behave (or "express themselves") differently.

The best example of epigenetic changes in eukaryotic biology is the process of cellular differentiation. During morphogenesis, totipotent stem cells become the various pluripotent cell lines of the embryo which in turn become fully differentiated cells. In other words, a single fertilized egg cell – the zygote – changes into the many cell types including neurons, muscle cells, epithelium, blood vessels etc. as it continues to divide. It does so by activating some genes while inhibiting others.

Etymology and definitions



Epigenetic mechanisms

Epigenetics (as in "epigenetic landscape") was coined by C. H. Waddington in 1942 as a portmanteau of the words *genetics* and *epigenesis*. *Epigenesis* is an old word which has more recently been used to describe the differentiation of cells from their initial totipotent state in embryonic development. When Waddington coined the term the physical nature of genes and their role in heredity was not known; he used it as a conceptual model of how genes might interact with their surroundings to produce a phenotype.

Robin Holliday defined epigenetics as "the study of the mechanisms of temporal and spatial control of gene activity during the development of complex organisms." Thus *epigenetic* can be used to describe anything other than DNA sequence that influences the development of an organism.

The modern usage of the word in scientific discourse is more narrow, referring to heritable traits (over rounds of cell division and sometimes transgenerationally) that do not involve changes to the underlying DNA sequence. The Greek prefix *epi-* in *epigenetics* implies features that are "on top of" or "in addition to" genetics; thus *epigenetic* traits exist on top of or in addition to the traditional molecular basis for inheritance.

The similarity of the word to "genetics" has generated many parallel usages. The "epigenome" is a parallel to the word "genome", and refers to the overall epigenetic state of a cell. The phrase "genetic code" has also been adapted—the "epigenetic code" has been used to describe the set of epigenetic features that create different phenotypes in different cells. Taken to its extreme, the "epigenetic code" could represent the total state of the cell, with the position of each molecule accounted for in an *epigenomic map*, a diagrammatic representation of the gene expression, DNA methylation and histone modification status of a particular genomic region. More typically, the term is used in reference to systematic efforts to measure specific, relevant forms of epigenetic information such as the histone code or DNA methylation patterns.

The psychologist Erik Erikson used the term *epigenetic* in his theory of psychosocial development. That usage, however, is of primarily historical interest.

Molecular basis of epigenetics

The molecular basis of epigenetics is complex. It involves modifications of the activation of certain genes, but not the basic structure of DNA. Additionally, the chromatin proteins associated with DNA may be activated or silenced. This accounts for why the differentiated cells in a multi-cellular organism express only the genes that are necessary for their own activity. Epigenetic changes are preserved when cells divide. Most epigenetic changes only occur within the course of one individual organism's lifetime, but, if a mutation in the DNA has been caused in sperm or egg cell that results in fertilization, then some epigenetic changes are inherited from one generation to the next. This raises the question of whether or not epigenetic changes in an organism can alter the basic structure of its DNA, a form of Lamarckism.

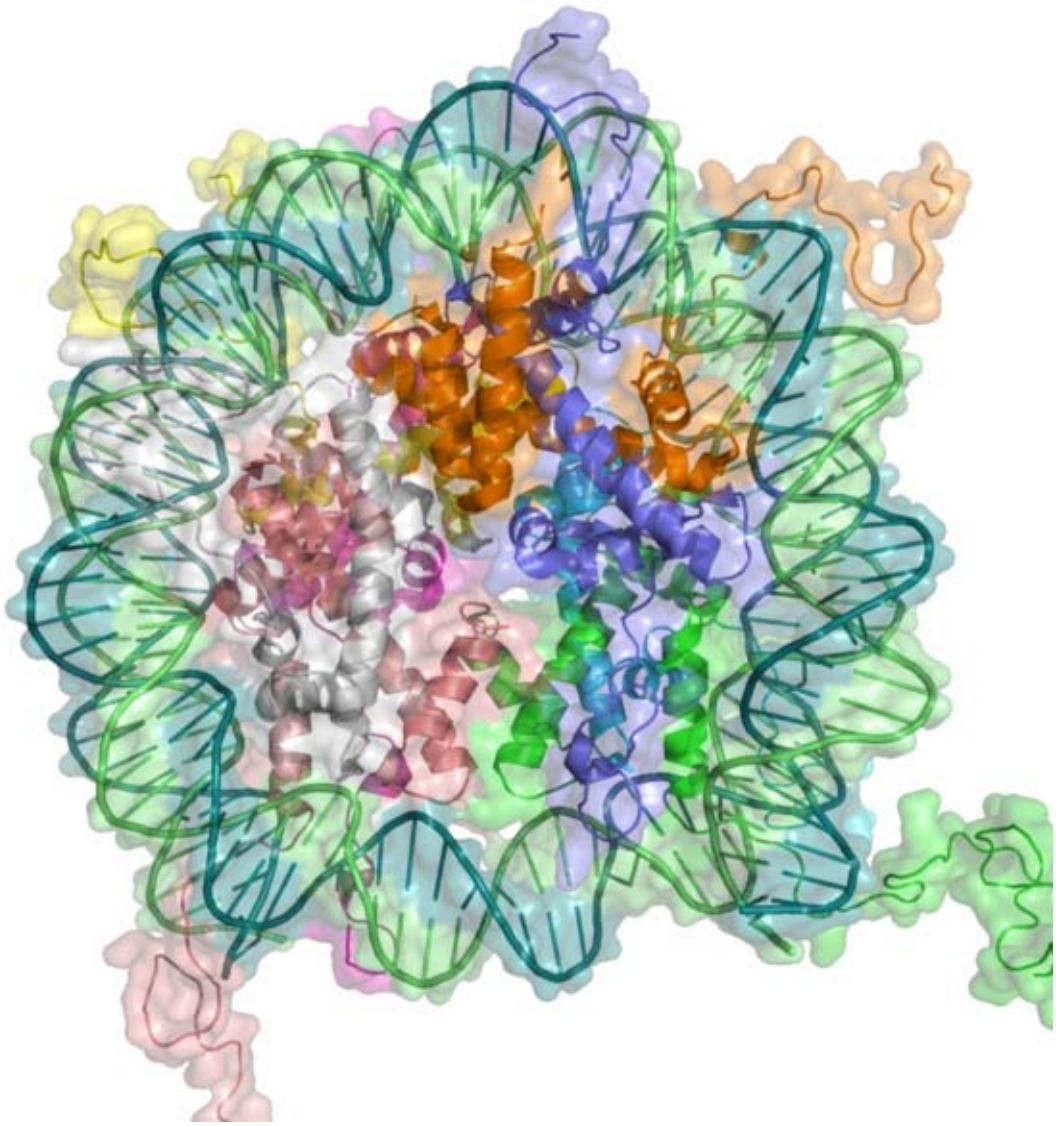
Specific epigenetic processes include paramutation, bookmarking, imprinting, gene silencing, X chromosome inactivation, position effect, reprogramming, transvection, maternal effects, the progress of carcinogenesis, many effects of teratogens, regulation of histone modifications and heterochromatin, and technical limitations affecting parthenogenesis and cloning.

Epigenetic research uses a wide range of molecular biologic techniques to further our understanding of epigenetic phenomena, including chromatin immunoprecipitation (together with its large-scale variants ChIP-on-chip and ChIP-seq), fluorescent in situ hybridization, methylation-sensitive restriction enzymes, DNA adenine methyltransferase identification (DamID) and bisulfite sequencing. Furthermore, the use of bioinformatic methods is playing an increasing role (computational epigenetics).

Mechanisms

Several types of epigenetic inheritance systems may play a role in what has become known as cell memory:

DNA methylation and chromatin remodeling



DNA associates with histone proteins to form chromatin.

Because the phenotype of a cell or individual is affected by which of its genes are transcribed, heritable transcription states can give rise to epigenetic effects. There are several layers of regulation of gene expression. One way that genes are regulated is through the remodeling of chromatin. Chromatin is the complex of DNA and the histone proteins with which it associates. Histone proteins are little spheres that DNA wraps around. If the way that DNA is wrapped around the histones changes, gene expression can change as well. Chromatin remodeling is accomplished through two main mechanisms:

1. The first way is post translational modification of the amino acids that make up histone proteins. Histone proteins are made up of long chains of amino acids. If

the amino acids that are in the chain are changed, the shape of the histone sphere might be modified. DNA is not completely unwound during replication. It is possible, then, that the modified histones may be carried into each new copy of the DNA. Once there, these histones may act as templates, initiating the surrounding new histones to be shaped in the new manner. By altering the shape of the histones around it, these modified histones would ensure that a differentiated cell would stay differentiated, and not convert back into being a stem cell.

2. The second way is the addition of methyl groups to the DNA, mostly at CpG sites, to convert cytosine to 5-methylcytosine. 5-Methylcytosine performs much like a regular cytosine, pairing up with a guanine. However, some areas of genome are methylated more heavily than others and highly methylated areas tend to be less transcriptionally active, through a mechanism not fully understood. Methylation of cytosines can also persist from the germ line of one of the parents into the zygote, marking the chromosome as being inherited from this parent (genetic imprinting).

The way that the cells stay differentiated in the case of DNA methylation is clearer to us than it is in the case of histone shape. Basically, certain enzymes (such as DNMT1) have a higher affinity for the methylated cytosine. If this enzyme reaches a "hemimethylated" portion of DNA (where methylcytosine is in only one of the two DNA strands) the enzyme will methylate the other half.

Although histone modifications occur throughout the entire sequence, the unstructured N-termini of histones (called histone tails) are particularly highly modified. These modifications include acetylation, methylation, ubiquitylation, phosphorylation and sumoylation. Acetylation is the most highly studied of these modifications. For example, acetylation of the K14 and K9 lysines of the tail of histone H3 by histone acetyltransferase enzymes (HATs) is generally correlated with transcriptional competence.

One mode of thinking is that this tendency of acetylation to be associated with "active" transcription is biophysical in nature. Because it normally has a positively charged nitrogen at its end, lysine can bind the negatively charged phosphates of the DNA backbone. The acetylation event converts the positively charged amine group on the side chain into a neutral amide linkage. This removes the positive charge, thus loosening the DNA from the histone. When this occurs, complexes like SWI/SNF and other transcriptional factors can bind to the DNA and allow transcription to occur. This is the "cis" model of epigenetic function. In other words, changes to the histone tails have a direct affect on the DNA itself.

Another model of epigenetic function is the "trans" model. In this model changes to the histone tails act indirectly on the DNA. For example, lysine acetylation may create a binding site for chromatin modifying enzymes (and basal transcription machinery as well). This Chromatin Remodeler can then cause changes to the state of the chromatin. Indeed, the bromodomain — a protein segment (domain) that specifically binds acetyl-

lysine — is found in many enzymes that help activate transcription, including the SWI/SNF complex (on the protein polybromo). It may be that acetylation acts in this and the previous way to aid in transcriptional activation.

The idea that modifications act as docking modules for related factors is borne out by histone methylation as well. Methylation of lysine 9 of histone H3 has long been associated with constitutively transcriptionally silent chromatin (constitutive heterochromatin). It has been determined that a chromodomain (a domain that specifically binds methyl-lysine) in the transcriptionally repressive protein HP1 recruits HP1 to K9 methylated regions. One example that seems to refute this biophysical model for acetylation is that tri-methylation of histone H3 at lysine 4 is strongly associated with (and required for full) transcriptional activation. Tri-methylation in this case would introduce a fixed positive charge on the tail.

It has been shown that the histone lysine methyltransferase (KMT) is responsible for this methylation activity in the pattern of histones H3 & H4. This enzyme utilizes a catalytically active site called the SET domain (Supressor of variegation, Enhancer of zeste, Trithorax). The SET domain is a 130-amino acid sequence involved in modulating gene activities. This domain has been demonstrated to bind to the histone tail and causes the methylation of the histone.

Differing histone modifications are likely to function in differing ways; acetylation at one position is likely to function differently than acetylation at another position. Also, multiple modifications may occur at the same time, and these modifications may work together to change the behavior of the nucleosome. The idea that multiple dynamic modifications regulate gene transcription in a systematic and reproducible way is called the histone code.

DNA methylation frequently occurs in repeated sequences, and helps to suppress the expression and mobility of 'transposable elements': Because 5-methylcytosine is chemically very similar to thymidine, CpG sites are frequently mutated and become rare in the genome, except at CpG islands where they remain unmethylated. Epigenetic changes of this type thus have the potential to direct increased frequencies of permanent genetic mutation. DNA methylation patterns are known to be established and modified in response to environmental factors by a complex interplay of at least three independent DNA methyltransferases, DNMT1, DNMT3A and DNMT3B, the loss of any of which is lethal in mice. DNMT1 is the most abundant methyltransferase in somatic cells, localizes to replication foci, has a 10–40-fold preference for hemimethylated DNA and interacts with the proliferating cell nuclear antigen (PCNA). By preferentially modifying hemimethylated DNA, DNMT1 transfers patterns of methylation to a newly synthesized strand after DNA replication, and therefore is often referred to as the ‘maintenance' methyltransferase. DNMT1 is essential for proper embryonic development, imprinting and X-inactivation.

Histones H3 and H4 can also be manipulated through demethylation using histone lysine demethylase (KDM). This recently identifited enzyme has a catalytically active site

called the Jumonji domain (JmjC). The demethylation occurs when JmjC utilizes multiple cofactors to hydroxylate the methyl group, thereby removing it. JmjC is capable of demethylating mono-, di-, and tri-methylated substrates.

Chromosomal regions can adopt stable and heritable alternative states resulting in bistable gene expression without changes to the DNA sequence. Epigenetic control is often associated with alternative covalent modifications of histones. The stability and heritability of states of larger chromosomal regions are often thought to involve positive feedback where modified nucleosomes recruit enzymes that similarly modify nearby nucleosomes. A simplified stochastic model for this type of epigenetics is found here.

Because DNA methylation and chromatin remodeling play such a central role in many types of epigenic inheritance, the word "epigenetics" is sometimes used as a synonym for these processes. However, this can be misleading. Chromatin remodeling is not always inherited, and not all epigenetic inheritance involves chromatin remodeling.

It has been suggested that the histone code could be mediated by the effect of small RNAs. The recent discovery and characterization of a vast array of small (21- to 26-nt), non-coding RNAs suggests that there is an RNA component, possibly involved in epigenetic gene regulation. Small interfering RNAs can modulate transcriptional gene expression via epigenetic modulation of targeted promoters.

RNA transcripts and their encoded proteins

Sometimes a gene, after being turned on, transcribes a product that (either directly or indirectly) maintains the activity of that gene. For example, Hnf4 and MyoD enhance the transcription of many liver- and muscle-specific genes, respectively, including their own, through the transcription factor activity of the proteins they encode. RNA signalling includes differential recruitment of a hierarchy of generic chromatin modifying complexes and DNA methyltransferases to specific loci by RNAs during differentiation and development. Other epigenetic changes are mediated by the production of different splice forms of RNA, or by formation of double-stranded RNA (RNAi). Descendants of the cell in which the gene was turned on will inherit this activity, even if the original stimulus for gene-activation is no longer present. These genes are most often turned on or off by signal transduction, although in some systems where syncytia or gap junctions are important, RNA may spread directly to other cells or nuclei by diffusion. A large amount of RNA and protein is contributed to the zygote by the mother during oogenesis or via nurse cells, resulting in maternal effect phenotypes. A smaller quantity of sperm RNA is transmitted from the father, but there is recent evidence that this epigenetic information can lead to visible changes in several generations of offspring.

Prions

Prions are infectious forms of proteins. Proteins generally fold into discrete units which perform distinct cellular functions, but some proteins are also capable of forming an infectious conformational state known as a prion. Although often viewed in the context of

infectious disease, prions are more loosely defined by their ability to catalytically convert other native state versions of the same protein to an infectious conformational state. It is in this latter sense that they can be viewed as epigenetic agents capable of inducing a phenotypic change without a modification of the genome.

Fungal prions are considered epigenetic because the infectious phenotype caused by the prion can be inherited without modification of the genome. PSI+ and URE3, discovered in yeast in 1965 and 1971, are the two best studied of this type of prion. Prions can have a phenotypic effect through the sequestration of protein in aggregates, thereby reducing that protein's activity. In PSI+ cells, the loss of the Sup35 protein (which is involved in termination of translation) causes ribosomes to have a higher rate of read-through of stop codons, an effect which results in suppression of nonsense mutations in other genes. The ability of Sup35 to form prions may be a conserved trait. It could confer an adaptive advantage by giving cells the ability to switch into a PSI+ state and express dormant genetic features normally terminated by premature stop codon mutations.

Structural inheritance systems

In ciliates such as *Tetrahymena* and *Paramecium*, genetically identical cells show heritable differences in the patterns of ciliary rows on their cell surface. Experimentally altered patterns can be transmitted to daughter cells. It seems existing structures act as templates for new structures. The mechanisms of such inheritance are unclear, but reasons exist to assume that multicellular organisms also use existing cell structures to assemble new ones.

Functions and consequences

Development

Somatic epigenetic inheritance, particularly through DNA methylation and chromatin remodeling, is very important in the development of multicellular eukaryotic organisms. The genome sequence is static (with some notable exceptions), but cells differentiate into many different types, which perform different functions, and respond differently to the environment and intercellular signalling. Thus, as individuals develop, morphogens activate or silence genes in an epigenetically heritable fashion, giving cells a "memory". In mammals, most cells terminally differentiate, with only stem cells retaining the ability to differentiate into several cell types ("totipotency" and "multipotency"). In mammals, some stem cells continue producing new differentiated cells throughout life, but mammals are not able to respond to loss of some tissues, for example, the inability to regenerate limbs, which some other animals are capable of. Unlike animals, plant cells do not terminally differentiate, remaining totipotent with the ability to give rise to a new individual plant. While plants do utilise many of the same epigenetic mechanisms as animals, such as chromatin remodeling, it has been hypothesised that plant cells do not have "memories", resetting their gene expression patterns at each cell division using positional information from the environment and surrounding cells to determine their fate.

Medicine

Epigenetics has many and varied potential medical applications. Congenital genetic disease is well understood, and it is also clear that epigenetics can play a role, for example, in the case of Angelman syndrome and Prader-Willi syndrome. These are normal genetic diseases caused by gene deletions or inactivation of the genes, but are unusually common because individuals are essentially hemizygous because of genomic imprinting, and therefore a single gene knock out is sufficient to cause the disease, where most cases would require both copies to be knocked out.

Evolution

Although epigenetics in multicellular organisms is generally thought to be a mechanism involved in differentiation, with epigenetic patterns "reset" when organisms reproduce, there have been some observations of transgenerational epigenetic inheritance (e.g., the phenomenon of paramutation observed in maize). Although most of these multigenerational epigenetic traits are gradually lost over several generations, the possibility remains that multigenerational epigenetics could be another aspect to evolution and adaptation. A sequestered germ line or Weismann barrier is specific to animals, and epigenetic inheritance is expected to be far more common in plants and microbes. These effects may require enhancements to the standard conceptual framework of the modern evolutionary synthesis.

Epigenetic features may play a role in short-term adaptation of species by allowing for reversible phenotype variability. The modification of epigenetic features associated with a region of DNA allows organisms, on a multigenerational time scale, to switch between phenotypes that express and repress that particular gene. When the DNA sequence of the region is not mutated, this change is reversible. It has also been speculated that organisms may take advantage of differential mutation rates associated with epigenetic features to control the mutation rates of particular genes. Interestingly, recent analysis have suggested that members of the APOBEC family of cytosine deaminases are capable of simultaneously mediating genetic and epigenetic inheritance using similar molecular mechanisms.

Epigenetic changes have also been observed to occur in response to environmental exposure—for example, mice given some dietary supplements have epigenetic changes affecting expression of the agouti gene, which affects their fur color, weight, and propensity to develop cancer.

More than 100 cases of transgenerational epigenetic inheritance phenomena have been reported in a wide range of organisms, including prokaryotes, plants, and animals.

Epigenetic effects in humans

Genomic imprinting and related disorders

Some human disorders are associated with genomic imprinting, a phenomenon in mammals where the father and mother contribute different epigenetic patterns for specific genomic loci in their germ cells. The best-known case of imprinting in human disorders is that of Angelman syndrome and Prader-Willi syndrome—both can be produced by the same genetic mutation, chromosome 15q partial deletion, and the particular syndrome that will develop depends on whether the mutation is inherited from the child's mother or from their father. This is due to the presence of genomic imprinting in the region. Beckwith-Wiedemann syndrome is also associated with genomic imprinting, often caused by abnormalities in maternal genomic imprinting of a region on chromosome 11.

Transgenerational epigenetic observations

Marcus Pembrey and colleagues also observed in the Överkalix study that the paternal (but not maternal) grandsons of Swedish boys who were exposed during preadolescence to famine in the 19th century were less likely to die of cardiovascular disease; if food was plentiful then diabetes mortality in the grandchildren increased, suggesting that this was a transgenerational epigenetic inheritance. The opposite effect was observed for females—the paternal (but not maternal) granddaughters of women who experienced famine while in the womb (and therefore while their eggs were being formed) lived shorter lives on average.

Cancer and developmental abnormalities

A variety of compounds are considered as epigenetic carcinogens—they result in an increased incidence of tumors, but they do not show mutagen activity (toxic compounds or pathogens that cause tumors incident to increased regeneration should also be excluded). Examples include diethylstilbestrol, arsenite, hexachlorobenzene, and nickel compounds.

Many teratogens exert specific effects on the fetus by epigenetic mechanisms. While epigenetic effects may preserve the effect of a teratogen such as diethylstilbestrol throughout the life of an affected child, the possibility of birth defects resulting from exposure of fathers or in second and succeeding generations of offspring has generally been rejected on theoretical grounds and for lack of evidence. However, a range of male-mediated abnormalities have been demonstrated, and more are likely to exist. FDA label information for Vidaza(tm), a formulation of 5-azacitidine (an unmethylatable analog of cytidine that causes hypomethylation when incorporated into DNA) states that "men should be advised not to father a child" while using the drug, citing evidence in treated male mice of reduced fertility, increased embryo loss, and abnormal embryo development. In rats, endocrine differences were observed in offspring of males exposed to morphine. In mice, second generation effects of diethylstilbesterol have been described occurring by epigenetic mechanisms.

Recent studies have shown that the Mixed Lineage Leukemia (MLL) gene causes leukemia by rearranging and fusing with other genes in different chromosomes, which is a process under epigenetic control.

Other investigations have concluded that alterations in histone acetylation and DNA methylation occur in various genes influencing prostate cancer.

In 2008, the National Institutes of Health announced that $190 million had been earmarked for epigenetics research over the next five years. In announcing the funding, government officials noted that epigenetics has the potential to explain mechanisms of aging, human development, and the origins of cancer, heart disease, mental illness, as well as several other conditions. Some investigators, like Randy Jirtle, PhD, of Duke University Medical Center, think epigenetics may ultimately turn out to have a greater role in disease than genetics.

DNA methylation in cancer

DNA methylation is an important regulator of gene transcription and a large body of evidence has demonstrated that aberrant DNA methylation is associated with unscheduled gene silencing, and the genes with high levels of 5-methylcytosine in their promoter region are transcriptionally silent. DNA methylation is essential during embryonic development, and in somatic cells, patterns of DNA methylation are generally transmitted to daughter cells with a high fidelity. Aberrant DNA methylation patterns have been associated with a large number of human malignancies and found in two distinct forms: hypermethylation and hypomethylation compared to normal tissue. Hypermethylation is one of the major epigenetic modifications that repress transcription via promoter region of tumour suppressor genes. Hypermethylation typically occurs at CpG islands in the promoter region and is associated with gene inactivation. Global hypomethylation has also been implicated in the development and progression of cancer through different mechanisms.

Variant histones H2A in cancer

The histone variants of the H2A family are highly conserved in mammals, playing critical roles in regulating many nuclear processes by altering chromatin structure. One of the key H2A variants, H2A.X, marks DNA damage, facilitating the recruitment of DNA repair proteins to restore genomic integrity. Another variant, H2A.Z, plays an important role in both gene activation and repression. A high level of H2A.Z expression is ubiquitously detected in many cancers and is significantly associated with cellular proliferation and genomic instability.

Cancer Treatment

Current research has shown that epigenetic pharmaceuticals could be a putative replacement or adjuvant therapy for currently accepted treatment methods such as radiation and chemotherapy, or could enhance the effects of these current treatments. It

has been shown that the epigenetic control of the proto-onco regions and the tumor suppressor sequences by conformational changes in histones directly affects the formation and progression of cancer Epigenetics also has the factor of reversibility, a characteristic that other cancer treatments do not offer.

Drug development has mainly focused on Histone Acetyltransferase (HAT) and Histone Deactylase (HDAC), including the introduction of the new pharmaceutical Vorinostat, a HDAC inhibitor, to the market. HDAC specifically has been shown to play an integral role in the progression of oral squamous cancer

Current front-runner candidates for new drug targets are Histone Lysine Methyltransferases (KMT) and Protein Arginine Methyltransferases (PRMT).

Twin studies

Recent studies involving both dizygotic and monozygotic twins have produced some evidence of epigenetic influence in humans.

Epigenetics in microorganisms

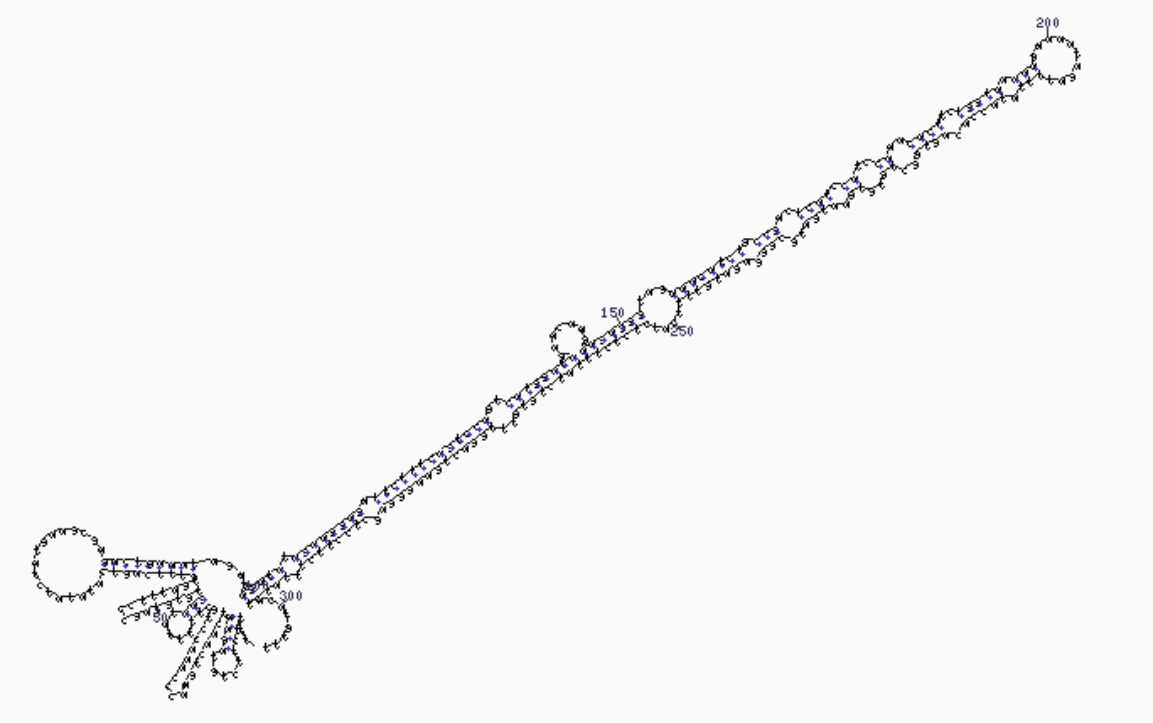
Bacteria make widespread use of postreplicative DNA methylation for the epigenetic control of DNA-protein interactions. Bacteria make use of DNA adenine methylation (rather than DNA cytosine methylation) as an epigenetic signal. DNA adenine methylation is important in bacteria virulence in organisms such as *Escherichia coli*, *Salmonella, Vibrio, Yersinia, Haemophilus*, and *Brucella*. In *Alphaproteobacteria*, methylation of adenine regulates the cell cycle and couples gene transcription to DNA replication. In *Gammaproteobacteria*, adenine methylation provides signals for DNA replication, chromosome segregation, mismatch repair, packaging of bacteriophage, transposase activity and regulation of gene expression.

The filamentous fungus *Neurospora crassa* is a prominent model system for understanding the control and function of cytosine methylation. In this organisms, DNA methylation is associated with relics of a genome defense system called RIP (repeat-induced point mutation) and silences gene expression by inhibiting transcription elongation.

The yeast prion PSI is generated by a conformational change of a translation termination factor, which is then inherited by daughter cells. This can provide a survival advantage under adverse conditions. This is an example of epigenetic regulation enabling unicellular organisms to respond rapidly to environmental stress. Prions can be viewed as epigenetic agents capable of inducing a phenotypic change without modification of the genome.

Chapter- 6

MicroRNA



The stem-loop secondary structure of a pre-microRNA from *Brassica oleracea*.

MicroRNAs (miRNAs) are short ribonucleic acid (RNA) molecules, on average only 22 nucleotides long and are found in all eukaryotic cells. miRNAs are post-transcriptional regulators that bind to complementary sequences on target messenger RNA transcripts (mRNAs), usually resulting in translational repression and gene silencing. The human genome may encode over 1000 miRNAs, which may target about 60% of mammalian genes and are abundant in many human cell types.

miRNAs show very different characteristics between plants and metazoans. In plants the miRNA complementarity to its mRNA target is nearly perfect, with no or few mismatched bases. In metazoans on the other hand miRNA complementarity is far from perfect and one miRNA can target many different sites on the same mRNA or on many different mRNAs. Another difference is the location of target sites on mRNAs. In metazoans the miRNA target sites are in the three prime untranslated regions (3'UTR) of

the mRNA. In plants targets can be located in the 3' UTR but are more often in the coding region itself. MiRNAs are well conserved in eukaryotic organism and are thought to be a vital and evolutionary ancient component of genetic regulation.

The first miRNAs were characterized in the early 1990s, but miRNAs were not recognized as a distinct class of biologic regulators with conserved functions until the early 2000s. Since then, miRNA research has revealed multiple roles in negative regulation (transcript degradation and sequestering, translational suppression) and possible involvement in positive regulation (transcriptional and translational activation). By affecting gene regulation, miRNAs are likely to be involved in most biologic processes. Different sets of expressed miRNAs are found in different cell types and tissues.

Aberrant expression of miRNAs has been implicated in numerous disease states, and miRNA-based therapies are under investigation.

History

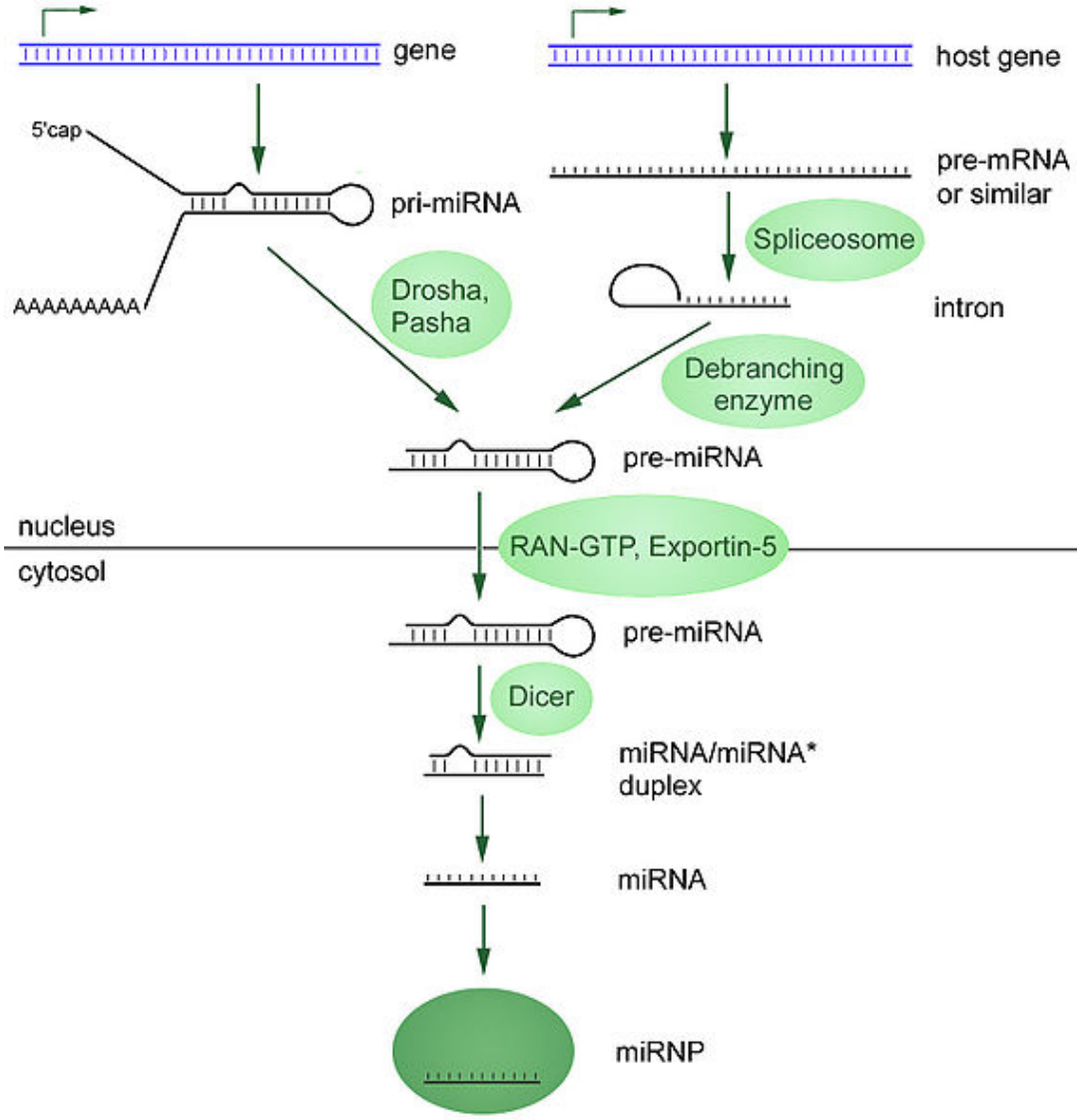
MicroRNAs were discovered in 1993 by Victor Ambros, Rosalind Lee and Rhonda Feinbaum during a study of the gene *lin-14* in *C. elegans* development. They found that LIN-14 protein abundance was regulated by a short RNA product encoded by the *lin-4* gene. A 61 nucleotide precursor from *lin-4* gene matured to a 22 nucleotide RNA containing sequences partially complementary to multiple sequences in the 3’ UTR of the *lin-14* mRNA. This complementarity was sufficient and necessary to inhibit the translation of *lin-14* mRNA into LIN-14 protein. Retrospectively, the *lin-4* small RNA was the first microRNA to be identified, though at the time, it was thought to be a nematode idiosyncrasy. Only in 2000 was a second RNA characterized: let-7, which repressed *lin-41*, *lin-14*, *lin-28*, *lin-42*, and *daf-12* expression during developmental stage transitions in *C. elegans*. let-7 was soon found to be conserved in many species, indicating the existence of a wider phenomenon.

Nomenclature

Under a standard nomenclature system, names are assigned to experimentally confirmed miRNAs before publication of their discovery. The prefix "mir" is followed by a dash and a number, the latter often indicating order of naming. For example, mir-123 was named and likely discovered prior to mir-456. The uncapitalized "mir-" refers to the pre-miRNA, while a capitalized "miR-" refers to the mature form. miRNAs with nearly identical sequences bar one or two nucleotides are annotated with an additional lower case letter. For example, miR-123a would be closely related to miR-123b. Pre-miRNAs that lead to 100% identical mature miRNAs but that are located at different places in the genome are indicated with an additional dash-number suffix. For example, the pre-miRNAs hsa-mir-194-1 and hsa-mir-194-2 lead to an identical mature miRNA (hsa-miR-194) but are located in different regions of the genome. Species of origin is designated with a three-letter prefix, e.g., hsa-miR-123 would be from human (*Homo sapiens*) and oar-miR-123 would be a sheep (*Ovis aries*) miRNA. Other common prefixes include 'v'

for viral (miRNA encoded by a viral genome) and 'd' for *Drosophila* miRNA (a fruit fly commonly studied in genetic research). When two mature microRNAs originate from opposite arms of the same pre-miRNA, they are denoted with a -3p or -5p suffix. (In the past, this distinction was also made with 's' (sense) and 'as' (antisense)). When relative expression levels are known, an asterisk following the name indicates an miRNA expressed at low levels relative to the miRNA in the opposite arm of a hairpin. For example, miR-123 and miR-123* would share a pre-miRNA hairpin, but more miR-123 would be found in the cell.

Biogenesis



MicroRNAs are produced from either their own genes or from introns

Most microRNA genes are found in intergenic regions or in anti-sense orientation to genes and contain their own miRNA gene promoter and regulatory units. As much as

40% of miRNA genes may lie in the introns of protein and non-protein coding genes or even in exons. These are usually, though not exclusively, found in a sense orientation. and thus usually are regulated together with their host genes. Other miRNA genes showing a common promoter include the 42-48% of all miRNAs originating from polycistronic units containing 2-7 discrete loops from which mature miRNAs are processed, although this does not necessarily mean the mature miRNAs of a family will be homologous in structure and function. The promoters mentioned have been shown to have some similarities in their motifs to promoters of other genes transcribed by RNA polymerase II such as protein coding genes. The DNA template is not the final word on mature miRNA production: 6% of human miRNAs show RNA editing, the site-specific modification of RNA sequences to yield products different from those encoded by their DNA. This increases the diversity and scope of miRNA action beyond that implicated from the genome alone.

Transcription

miRNA genes are usually transcribed by RNA polymerase II (Pol II). The polymerase often binds to a promoter found near the DNA sequence encoding what will become the hairpin loop of the pre-miRNA. The resulting transcript is capped with a specially-modified nucleotide at the 5’ end, polyadenylated with multiple adenosines (a poly(A) tail), and spliced. The product, called a primary miRNA (pri-miRNA), may be hundreds or thousands of nucleotides in length and contain one or more miRNA stem loops. When a stem loop precursor is found in the 3’ UTR, a transcript may serve as a pri-miRNA and a mRNA. RNA polymerase III (Pol III) transcribes some miRNAs, especially those with upstream Alu sequences, transfer RNAs (tRNAs), and mammalian wide interspersed repeat (MWIR) promoter units.

Nuclear processing

A single pri-miRNA may contain from one to six miRNA precursors. These hairpin loop structures are composed of about 70 nucleotides each. Each hairpin is flanked by sequences necessary for efficient processing. The double-stranded RNA structure of the hairpins in a pri-miRNA is recognized by a nuclear protein known as DiGeorge Syndrome Critical Region 8 (DGCR8 or "Pasha" in invertebrates), named for its association with DiGeorge Syndrome. DGCR8 associates with the enzyme Drosha, a protein that cuts RNA, to form the "Microprocessor" complex. In this complex, DGCR8 orients the catalytic RNase III domain of Drosha to liberate hairpins from pri-miRNAs by cleaving RNA about eleven nucleotides from the hairpin base (two helical RNA turns into the stem). The resulting hairpin, known as a pre-miRNA (precursor-miRNA), has a two-nucleotide overhang at its 3’ end; it has 3' hydroxyl and 5' phosphate groups.

pre-miRNAs that are spliced directly out of introns, bypassing the Microprocessor complex, are known as "mirtrons." Originally thought to exist only in *Drosophila* and *C. elegans*, mirtrons have now been found in mammals.

Perhaps as many as 16% of pri-miRNAs may be altered through nuclear RNA editing. Most commonly, enzymes known as adenosine deaminases acting on RNA (ADARs) catalyze adenosine to inosine (A to I) transitions. RNA editing can halt nuclear processing (for example, of pri-miR-142, leading to degradation by the ribonuclease Tudor-SN) and alter downstream processes including cytoplasmic miRNA processing and target specificity (e.g., by changing the seed region of miR-376 in the central nervous system).

Nuclear export

pre-miRNA hairpins are exported from the nucleus in a process involving the nucleocytoplasmic shuttle Exportin-5. This protein, a member of the *karyopherin* family, recognizes a two-nucleotide overhang left by the RNase III enzyme Drosha at the 3' end of the pre-miRNA hairpin. Exportin-5-mediated transport to the cytoplasm is energy-dependent, using GTP bound to the Ran protein.

Cytoplasmic processing

In the cytoplasm, the pre-miRNA hairpin is cleaved by the RNase III enzyme Dicer. This endoribonuclease interacts with the 3' end of the hairpin and cuts away the loop joining the 3' and 5' arms, yielding an imperfect miRNA:miRNA* duplex about 22 nucleotides in length. Overall hairpin length and loop size influence the efficiency of Dicer processing, and the imperfect nature of the miRNA:miRNA* pairing also affects cleavage. Although either strand of the duplex may potentially act as a functional miRNA, only one strand is usually incorporated into the RNA-induced silencing complex (RISC) where the miRNA and its mRNA target interact.

Biogenesis in plants

miRNA biogenesis in plants differs from metazoan biogenesis mainly in the steps of nuclear processing and export. Instead of being cleaved by two different enzymes, once inside and once outside the nucleus, both cleavages of the plant miRNA is performed by a Dicer homolog, called Dicer-like1 (DL1). DL1 is only expressed in the nucleus of plant cells, which indicates that both reactions take place inside the nucleus. Before plant miRNA:miRNA* duplexes are transported out of the nucleus its 3' overhangs are methylated by a RNA methyltransferaseprotein called Hua-Enhancer1 (HEN1). The duplex is then transported out of the nucleus to the cytoplasm by a protein called Hasty (HST), an Exportin 5 homolog, where they disassemble and the mature miRNA is incorporated into the RISC.

The RNA-induced silencing complex

The mature miRNA is part of an active RNA-induced silencing complex (RISC) containing Dicer and many associated proteins. RISC is also known as a microRNA ribonucleoprotein complex (miRNP); RISC with incorporated miRNA is sometimes referred to as "miRISC."

Dicer processing of the pre-miRNA is thought to be coupled with unwinding of the duplex. Generally, only one strand is incorporated into the miRISC, selected on the basis of its thermodynamic instability and weaker base-pairing relative to the other strand. The position of the stem-loop may also influence strand choice. The other strand, called the passenger strand due to its lower levels in the steady state, is denoted with an asterisk (*) and is normally degraded. In some cases, both strands of the duplex are viable and become functional miRNA that target different mRNA populations.

Members of the argonaute (Ago) protein family are central to RISC function. Argonautes are needed for miRNA-induced silencing and contain two conserved RNA binding domains: a PAZ domain that can bind the single stranded 3’ end of the mature miRNA and a PIWI domain that structurally resembles ribonuclease-H and functions to interact with the 5’ end of the guide strand. They bind the mature miRNA and orient it for interaction with a target mRNA. Some argonautes, for example human Ago2, cleave target transcripts directly; argonautes may also recruit additional proteins to achieve translational repression. The human genome encodes eight argonaute proteins divided by sequence similarities into two families: AGO (with four members present in all mammalian cells and called E1F2C/hAgo in humans), and PIWI (found in the germ line and hematopoietic stem cells).

Additional RISC components include TRBP [human immunodeficiency virus (HIV) transactivating response RNA (TAR) binding protein], PACT (protein activator of the interferon induced protein kinase (PACT), the SMN complex, fragile X mental retardation protein (FMRP), and Tudor staphylococcal nuclease-domain-containing protein (Tudor-SN).

Mode of Silencing

Gene silencing may occur either via mRNA degradation or preventing mRNA from being translated. It has been demonstrated that if there is complete complementation between the miRNA and target mRNA sequence, Ago2 can cleave the mRNA and lead to direct mRNA degradation. Yet, if there isn't complete complementation the silencing is achieved by preventing translation.

miRNA turnover

Turnover of mature miRNA is needed for rapid changes in miRNA expression profiles. During miRNA maturation in the cytoplasm, uptake by the Argonaute protein is thought to stabilize the guide strand, while the opposite (* or "passenger") strand is preferentially destroyed. In what has been called a "Use it or lose it" strategy, Argonaute may preferentially retain miRNAs with many targets over miRNAs with few or no targets, leading to degradation of the non-targeting molecules.

Decay of mature miRNAs in animals is mediated by the 5′-to-3′ exoribonuclease XRN2, also known as Rat1p. In plants, SDN (small RNA degrading nuclease) family members

degrade miRNAs in the opposite (3'-to-5') direction. Similar enzymes are encoded in animal genomes, but their roles have not yet been described.

Several miRNA modifications affect miRNA stability. As indicated by work in the model organism *Arabidopsis thaliana* (thale cress), mature plant miRNAs appear to be stabilized by the addition of methyl moieties at the 3' end. The 2'-O-conjugated methyl groups block the addition of uracil (U) residues by uridyltransferase enzymes, a modification that may be associated with miRNA degradation. However, uridylation may also protect some miRNAs; the consequences of this modification are incompletely understood. Uridylation of some animal miRNAs has also been reported. Both plant and animal miRNAs may be altered by addition of adenine (A) residues to the 3' end of the miRNA. An extra A added to the end of mammalian miR-122, a liver-enriched miRNA important in Hepatitis C, stabilizes the molecule, and plant miRNAs ending with an adenine residue have slower decay rates.

Cellular functions

The function of miRNAs appears to be in gene regulation. For that purpose, a miRNA is complementary to a part of one or more messenger RNAs (mRNAs). Animal miRNAs are usually complementary to a site in the 3' UTR whereas plant miRNAs are usually complementary to coding regions of mRNAs. Perfect or near perfect base pairing with the target RNA promotes cleavage of the RNA. This is the primary mode of plant microRNAs. In animals, microRNAs more often only partially base pair and inhibit protein translation of the target mRNA (this exists in plants as well but is less common). MicroRNAs that are partially complementary to the target can also speed up deadenylation, causing mRNAs to be degraded sooner. For partially complementary microRNA to recognise their targets, the nucleotides 2–7 of the miRNA ('seed region') still have to be perfectly complementary. miRNAs occasionally also causes histone modification and DNA methylation of promoter sites and therefore affecting the expression of targeted genes.

Animal microRNAs target in particular developmental genes. In contrast, genes involved in functions common to all cells, such as gene expression, have very few microRNA target sites and seem to be under selection to avoid targeting by microRNAs.

dsRNA can also activate gene expression, a mechanism that has been termed "small RNA-induced gene activation" or RNAa. dsRNAs targeting gene promoters can induce potent transcriptional activation of associated genes. This was demonstrated in human cells using synthetic dsRNAs termed small activating RNAs (saRNAs), but has also been demonstrated for endogenous microRNA.

Evolution

MicroRNAs are significant phylogenetic markers because of their astonishingly low rate of evolution. Their origin may have permitted the development of morphological innovation, and by making gene expression more specific and 'fine-tunable', permitted the

genesis of complex organs and perhaps, ultimately, complex life. Indeed, rapid bursts of morphological innovation are generally associated with a high rate of microRNA accumulation.

MicroRNAs originate predominantly by the random formation of hairpins in "non-coding" sections of DNA (i.e. introns or intergene regions), but also by the duplication and modification of existing microRNAs. The rate of evolution (i.e. nucleotide substitution) in recently-originated microRNAs is comparable to that elsewhere in the non-coding DNA, implying evolution by neutral drift; however, older microRNAs have a much lower rate of change (often less than one substitution per hundred million years), suggesting that once a microRNA gains a function it undergoes extreme purifying selection. At this point, a microRNA is rarely lost from an animal's genome, although microRNAs which are more recently derived (and thus presumably non-functional) are frequently lost. This makes them a valuable phylogenetic marker, and they are being looked upon as a possible solution to such outstanding phylogenetic problems as the relationships of arthropods.

MicroRNAs feature in the genomes of most eukaryotic organisms, from the brown algae to the metazoa. Across all species, in excess of 5000 had been identified by March 2010. Whilst short RNA sequences (50 – hundreds of base pairs) of a broadly comparable function occur in bacteria, bacteria lack true microRNAs.

Experimental detection and manipulation of miRNA

MicroRNA expression can be quantified in a two-step polymerase chain reaction process of modified RT-PCR followed by quantitative real-time PCR. Variations of this method achieve absolute or relative quantification. miRNAs can also be hybridized to microarrays, slides or chips with probes to hundreds or thousands of miRNA targets, so that relative levels of miRNAs can be determined in different samples. MicroRNAs can be both discovered and profiled by high-throughput sequencing methods. The activity of an miRNA can be experimentally inhibited using a locked nucleic acid (LNA) oligo, a Morpholino oligo or a 2'-O-methyl RNA oligo. MicroRNA maturation can be inhibited at several points by steric-blocking oligos. The miRNA target site of an mRNA transcript can also be blocked by a steric-blocking oligo. Additionally, a specific miRNA can be silenced by a complementary antagomir. For the “in situ” detection of miRNA, the use of LNA is currently the only efficient method. The locked conformation of LNA results in enhanced hybridization properties and increases sensitivity and selectivity, making it ideal for detection of short miRNA.

miRNA and disease

Just as miRNA is involved in the normal functioning of eukaryotic cells, so has dysregulation of miRNA been associated with disease. A manually curated, publicly available database miR2Disease documents known relationships between miRNA dysregulation and human disease.

miRNA and cancer

Several miRNAs have been found to have links with some types of cancer.

A study of mice altered to produce excess c-Myc — a protein with mutated forms implicated in several cancers — shows that miRNA has an effect on the development of cancer. Mice that were engineered to produce a surplus of types of miRNA found in lymphoma cells developed the disease within 50 days and died two weeks later. In contrast, mice without the surplus miRNA lived over 100 days. Leukemia can be caused by the insertion of a viral genome next to the 17-92 array of microRNAs leading to increased expression of this microRNA.

Another study found that two types of miRNA inhibit the E2F1 protein, which regulates cell proliferation. miRNA appears to bind to messenger RNA before it can be translated to proteins that switch genes on and off.

By measuring activity among 217 genes encoding miRNA, patterns of gene activity that can distinguish types of cancers can be discerned. miRNA signatures may enable classification of cancer. This will allow doctors to determine the original tissue type which spawned a cancer and to be able to target a treatment course based on the original tissue type. miRNA profiling has already been able to determine whether patients with chronic lymphocytic leukemia had slow growing or aggressive forms of the cancer.

Transgenic mice that over-express or lack specific miRNAs have provided insight into the role of small RNAs in various malignancies.

A novel miRNA-profiling based screening assay for the detection of early-stage colorectal cancer has been developed and is currently in clinical trials. Early results showed that blood plasma samples collected from patients with early, resectable (Stage II) colorectal cancer could be distinguished from those of sex-and age-matched healthy volunteers. Sufficient selectivity and specificity could be achieved using small (less than 1 mL) samples of blood. The test has potential to be a cost-effective, non-invasive way to identify at-risk patients who should undergo colonoscopy.

miRNA and heart disease

The global role of miRNA function in the heart has been addressed by conditionally inhibiting miRNA maturation in the murine heart, and has revealed that miRNAs play an essential role during its development. miRNA expression profiling studies demonstrate that expression levels of specific miRNAs change in diseased human hearts, pointing to their involvement in cardiomyopathies. Furthermore, studies on specific miRNAs in animal models have identified distinct roles for miRNAs both during heart development and under pathological conditions, including the regulation of key factors important for cardiogenesis, the hypertrophic growth response, and cardiac conductance.

miRNA and the nervous system

miRNAs appear to regulate the nervous system. Neural miRNAs are involved at various stages of synaptic development, including dendritogenesis (involving miR-132, miR-134 and miR-124), synapse formation and synapse maturation (where miR-134 and miR-138 are thought to be involved). Some studies find altered miRNA expression in schizophrenia.

miRNA and non-coding RNAs

When the human genome project mapped its first chromosome in 1999, it was predicted the genome would contain over 100,000 protein coding genes. However, only around 20,000 were eventually identified (International Human Genome Sequencing Consortium, 2004). Since then, the advent of bioinformatics approaches combined with genome tiling studies examining the transcriptome, systematic sequencing of full length cDNA libraries, and experimental validation (including the creation of miRNA derived antisense oligonucleotides called antagomirs) have revealed that many transcripts are non protein-coding RNA, including several snoRNAs and miRNAs.

Chapter- 7

DNA Methylation

DNA methylation involves the addition of a methyl group to the 5 position of the cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring (cytosine and adenine are two of the four bases of DNA). This modification can be inherited through cell division. DNA methylation is typically removed during zygote formation and re-established through successive cell divisions during development. DNA methylation is a crucial part of normal organismal development and cellular differentiation in higher organisms. DNA methylation stably alters the gene expression pattern in cells such that cells can "remember where they have been" or decrease gene expression; for example, cells programmed to be pancreatic islets during embryonic development remain pancreatic islets throughout the life of the organism without continuing signals telling them that they need to remain islets. In addition, DNA methylation suppresses the expression of viral genes and other deleterious elements that have been incorporated into the genome of the host over time. DNA methylation also forms the basis of chromatin structure, which enables cells to form the myriad characteristics necessary for multicellular life from a single immutable sequence of DNA. DNA methylation also plays a crucial role in the development of nearly all types of cancer.

DNA methylation involves the addition of a methyl group to DNA — for example, to the number 5 carbon of the cytosine pyrimidine ring — in this case with the specific effect of reducing gene expression. DNA methylation at the 5 position of cytosine has been found in every vertebrate examined. In adult somatic tissues, DNA methylation typically occurs in a CpG dinucleotide context; non-CpG methylation is prevalent in embryonic stem cells.

In mammals

DNA methylation is essential for normal development and is associated with a number of key processes including genomic imprinting, X-chromosome inactivation, suppression of repetitive elements, and carcinogenesis.

Between 60% and 90% of all CpGs are methylated in mammals. Methylated C residues spontaneously deaminate to form T residues; hence CpG dinucleotides steadily mutate to TpG dinucleotides, which is evidenced by the under-representation of CpG dinucleotides in the human genome (they occur at only 21% of the expected frequency). (On the other

hand, spontaneous deamination of unmethylated C residues gives rise to U residues, a mutation that is quickly recognized and repaired by the cell.)

Unmethylated CpGs are often grouped in clusters called *CpG islands*, which are present in the 5' regulatory regions of many genes. In many disease processes, such as cancer, gene promoter CpG islands acquire abnormal hypermethylation, which results in transcriptional silencing that can be inherited by daughter cells following cell division. Alterations of DNA methylation have been recognized as an important component of cancer development. Hypomethylation, in general, arises earlier and is linked to chromosomal instability and loss of imprinting, whereas hypermethylation is associated with promoters and can arise secondary to gene (oncogene suppressor) silencing, but might be a target for epigenetic therapy.

DNA methylation may affect the transcription of genes in two ways. First, the methylation of DNA may itself may physically impede the binding of transcriptional proteins to the gene, and second, and likely more important, methylated DNA may be bound by proteins known as methyl-CpG-binding domain proteins (MBDs). MBD proteins then recruit additional proteins to the locus, such as histone deacetylases and other chromatin remodeling proteins that can modify histones, thereby forming compact, inactive chromatin, termed silent chromatin. This link between DNA methylation and chromatin structure is very important. In particular, loss of methyl-CpG-binding protein 2 (MeCP2) has been implicated in Rett syndrome; and methyl-CpG-binding domain protein 2 (MBD2) mediates the transcriptional silencing of hypermethylated genes in cancer.

Research has suggested that long-term memory storage in humans may be regulated by DNA methylation.

DNA methylation in cancer

DNA methylation is an important regulator of gene transcription and a large body of evidence has demonstrated that aberrant DNA methylation is associated with unscheduled gene silencing, and the genes with high levels of 5-methylcytosine in their promoter region are transcriptionally silent. DNA methylation is essential during embryonic development, and in somatic cells, patterns of DNA methylation are generally transmitted to daughter cells with a high fidelity. Aberrant DNA methylation patterns have been associated with a large number of human malignancies and found in two distinct forms: hypermethylation and hypomethylation compared to normal tissue. Hypermethylation is one of the major epigenetic modifications that repress transcription via promoter region of tumour suppressor genes. Hypermethylation typically occurs at CpG islands in the promoter region and is associated with gene inactivation. Global hypomethylation has also been implicated in the development and progression of cancer through different mechanisms.

DNA methyltransferases

In mammalian cells, DNA methylation occurs mainly at the C5 position of CpG dinucleotides and is carried out by two general classes of enzymatic activities – maintenance methylation and *de novo* methylation.

Maintenance methylation activity is necessary to preserve DNA methylation after every cellular DNA replication cycle. Without the DNA methyltransferase (DNMT), the replication machinery itself would produce daughter strands that are unmethylated and, over time, would lead to passive demethylation. DNMT1 is the proposed maintenance methyltransferase that is responsible for copying DNA methylation patterns to the daughter strands during DNA replication. Mouse models with both copies of DNMT1 deleted are embryonic lethal at approximately day 9, due to the requirement of DNMT1 activity for development in mammalian cells.

It is thought that DNMT3a and DNMT3b are the *de novo* methyltransferases that set up DNA methylation patterns early in development. DNMT3L is a protein that is homologous to the other DNMT3s but has no catalytic activity. Instead, DNMT3L assists the *de novo* methyltransferases by increasing their ability to bind to DNA and stimulating their activity. Finally, DNMT2 (TRDMT1) has been identified as a DNA methyltransferase homolog, containing all 10 sequence motifs common to all DNA methyltransferases; however, DNMT2 (TRDMT1) does not methylate DNA but instead methylates cytosine-38 in the anticodon loop of aspartic acid transfer RNA.

Since many tumor suppressor genes are silenced by DNA methylation during carcinogenesis, there have been attempts to re-express these genes by inhibiting the DNMTs. 5-Aza-2'-deoxycytidine (decitabine) is a nucleoside analog that inhibits DNMTs by trapping them in a covalent complex on DNA by preventing the β-elimination step of catalysis, thus resulting in the enzymes' degradation. However, for decitabine to be active, it must be incorporated into the genome of the cell, which can cause mutations in the daughter cells if the cell does not die. In addition, decitabine is toxic to the bone marrow, which limits the size of its therapeutic window. These pitfalls have led to the development of antisense RNA therapies that target the DNMTs by degrading their mRNAs and preventing their translation. However, it is currently unclear whether targeting DNMT1 alone is sufficient to reactivate tumor suppressor genes silenced by DNA methylation.

In plants

Significant progress has been made in understanding DNA methylation in the model plant *Arabidopsis thaliana*. DNA methylation in plants differs from that of mammals: while DNA methylation in mammals mainly occurs on the cytosine nucleotide in a CpG site, in plants the cytosine can be methylated at CpG, CpHpG, and CpHpH sites, where H represents any nucleotide but guanine.

The principal *Arabidopsis* DNA methyltransferase enzymes, which transfer and covalently attach methyl groups onto DNA, are DRM2, MET1, and CMT3. Both the DRM2 and MET1 proteins share significant homology to the mammalian methyltransferases DNMT3 and DNMT1, respectively, whereas the CMT3 protein is unique to the plant kingdom. There are currently two classes of DNA methyltransferases: 1) the *de novo* class, or enzymes that create new methylation marks on the DNA; and 2) a maintenance class that recognizes the methylation marks on the parental strand of DNA and transfers new methylation to the daughters strands after DNA replication. DRM2 is the only enzyme that has been implicated as a *de novo* DNA methyltransferase. DRM2 has also been shown, along with MET1 and CMT3 to be involved in maintaining methylation marks through DNA replication. Other DNA methyltransferases are expressed in plants but have no known function.

It is not clear how the cell determines the locations of *de novo* DNA methylation, but evidence suggests that, for many (though not all) locations, RNA-directed DNA methylation (RdDM) is involved. In RdDM, specific RNA transcripts are produced from a genomic DNA template, and this RNA forms secondary structures called double-stranded RNA molecules. The double-stranded RNAs, through either the small interfering RNA (siRNA) or microRNA (miRNA) pathways direct de-novo DNA methylation of the original genomic location that produced the RNA. This sort of mechanism is thought to be important in cellular defense against RNA viruses and/or transposons, both of which often form a double-stranded RNA that can be mutagenic to the host genome. By methylating their genomic locations, through an as yet poorly-understood mechanism, they are shut off and are no longer active in the cell, protecting the genome from their mutagenic effect.

In fungi

It can be seen that many fungi have low levels (0.1 to 0.5%) of cytosine methylation, whereas other fungi have as much as 5% of the genome methylated.

This value seems to vary both among species and among isolates of the same species. There is also evidence that DNA methylation may be involved in state-specific control of gene expression in fungi.

Although brewers' yeast (*Saccharomyces*) and fission yeast (*Schizosaccharomyces*) have very little DNA methylation, the model filamentous fungus *Neurospora crassa* has a well-characterized methylation system. Several genes control methylation in *Neurospora* and mutation of the DNA methyl transferase, *dim-2*, eliminates all DNA methylation but does not affect growth or sexual reproduction. While the *Neurospora* genome has very little repeated DNA, half of the methylation occurs in repeated DNA including transposon relics and centromeric DNA. The ability to evaluate other important phenomena in a DNA methylase-deficient genetic background makes *Neurospora* an important system in which to study DNA methylation.

In bacteria

Adenine or cytosine methylation is part of the restriction modification system of many bacteria, in which specific DNA sequences are methylated periodically throughout the genome. A methylase is the enzyme that recognizes a specific sequence and methylates one of the bases in or near that sequence. Foreign DNAs (which are not methylated in this manner) that are introduced into the cell are degraded by sequence-specific restriction enzymes and cleaved. Bacterial genomic DNA is not recognized by these restriction enzymes. The methylation of native DNA acts as a sort of primitive immune system, allowing the bacteria to protect themselves from infection by bacteriophage.

E. coli DNA adenine methyltransferase (Dam) is an enzyme of ~32 kDa that does not belong to a restriction/modification system. The target recognition sequence for *E. coli* Dam is GATC, as the methylation occurs at the N6 position of the adenine in this sequence (G meATC). The three base pairs flanking each side of this site also influence DNA–Dam binding. Dam plays several key roles in bacterial processes, including mismatch repair, the timing of DNA replication, and gene expression. As a result of DNA replication, the status of GATC sites in the *E. coli* genome changes from fully methylated to hemimethylated. This is because adenine introduced into the new DNA strand is unmethylated. Re-methylation occurs within two to four seconds, during which time replication errors in the new strand are repaired. Methylation, or its absence, is the marker that allows the repair apparatus of the cell to differentiate between the template and nascent strands. It has been shown that altering Dam activity in bacteria results in increased spontaneous mutation rate. Bacterial viability is compromised in dam mutants that also lack certain other DNA repair enzymes, providing further evidence for the role of Dam in DNA repair.

One region of the DNA that keeps its hemimethylated status for longer is the origin of replication, which has an abundance of GATC sites. This is central to the bacterial mechanism for timing DNA replication. SeqA binds to the origin of replication, sequestering it and thus preventing methylation. Because hemimethylated origins of replication are inactive, this mechanism limits DNA replication to once per cell cycle.

Expression of certain genes, for example those coding for pilus expression in *E. coli*, is regulated by the methylation of GATC sites in the promoter region of the gene operon. The cells' environmental conditions just after DNA replication determine whether Dam is blocked from methylating a region proximal to or distal from the promoter region. Once the pattern of methylation has been created, the pilus gene transcription is locked in the on or off position until the DNA is again replicated. In *E. coli*, these pilus operons have important roles in virulence in urinary tract infections. It has been proposed that inhibitors of Dam may function as antibiotics.

On the other hand DNA cytosine methylase targets CCAGG and CCTGG sites to methylate cytosine at the C5 position (C meC(A/T)GG). The other methylase enzyme, EcoKI, causes mehtylation of adenine in the sequences AAC(N6A)GTGC and GCAC(N6A)GTT.

Most strains used by molecular biologists are derivatives of K-12, and possess both Dam and Dcm, but there are commercially available strains which possess dam-/dcm- activity. In fact, it is possible to unmethylate the DNA extracted from dam+/dcm+ strains by transforming into dam-/dcm- strains. This would help digest sequences that are not being recognized by methylation-sensitive restriction enzymes.

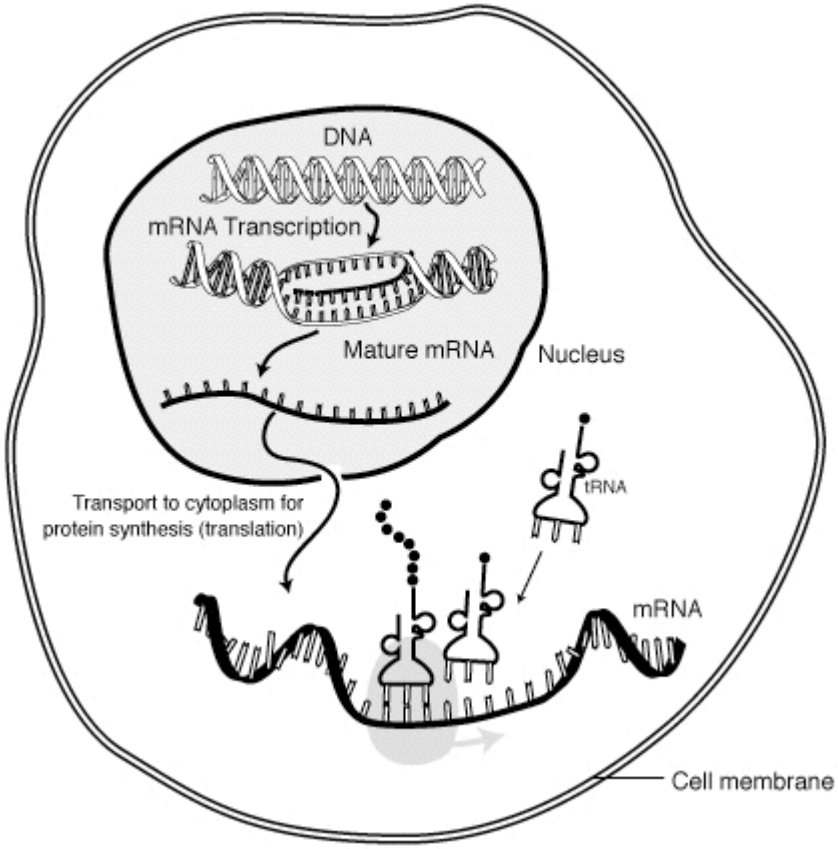
Detection

DNA methylation can be detected by the following assays currently used in scientific research:

- Methylation-Specific PCR (MSP), which is based on a chemical reaction of sodium bisulfite with DNA that converts unmethylated cytosines of CpG dinucleotides to uracil or UpG, followed by traditional PCR. However, methylated cytosines will not be converted in this process, and primers are designed to overlap the CpG site of interest, which allows one to determine methylation status as methylated or unmethylated.
- The HELP assay, which is based on restriction enzymes' differential ability to recognize and cleave methylated and unmethylated CpG DNA sites.
- ChIP-on-chip assays, which is based on the ability of commercially prepared antibodies to bind to DNA methylation-associated proteins like MCP2.
- Restriction landmark genomic scanning, a complicated and now rarely-used assay based upon restriction enzymes' differential recognition of methylated and unmethylated CpG sites; the assay is similar in concept to the HELP assay.
- Methylated DNA immunoprecipitation (MeDIP), analogous to chromatin immunoprecipitation, immunoprecipitation is used to isolate methylated DNA fragments for input into DNA detection methods such as DNA microarrays (MeDIP-chip) or DNA sequencing (MeDIP-seq).
- Molecular break light assay for DNA adenine methyltransferase activity – an assay that relies on the specificity of the restriction enzyme DpnI for fully methylated (adenine methylation) GATC sites in an oligonucleotide labeled with a fluorophore and quencher. The adenine methyltransferase methylates the oligonucleotide making it a substrate for DpnI. Cutting of the oligonucleotide by DpnI gives rise to a fluorescence increase.

Chapter- 8

Messenger RNA



The "life cycle" of an **mRNA** in a eukaryotic cell. RNA is transcribed in the nucleus; processed, it is transported to the cytoplasm and translated by the ribosome. At the end of its life, the mRNA is degraded.

Messenger RNA (**mRNA**) is a molecule of RNA encoding a chemical "blueprint" for a protein product. mRNA is transcribed from a DNA template, and carries coding information to the sites of protein synthesis: the ribosomes. Here, the nucleic acid

polymer is translated into a polymer of amino acids: a protein. In mRNA as in DNA, genetic information is encoded in the sequence of nucleotides arranged into codons consisting of three bases each. Each codon encodes for a specific amino acid, except the stop codons that terminate protein synthesis. This process requires two other types of RNA: transfer RNA (tRNA) mediates recognition of the codon and provides the corresponding amino acid, while ribosomal RNA (rRNA) is the central component of the ribosome's protein manufacturing machinery.

Synthesis, processing, and function

The brief existence of an mRNA molecule begins with transcription and ultimately ends in degradation. During its life, an mRNA molecule may also be processed, edited, and transported prior to translation. Eukaryotic mRNA molecules often require extensive processing and transport, while prokaryotic molecules do not.

Transcription

During transcription, RNA polymerase makes a copy of a gene from the DNA to mRNA as needed. This process is similar in eukaryotes and prokaryotes. One notable difference, however, is that prokaryotic RNA polymerase associates with mRNA processing enzymes during transcription so that processing can proceed quickly after the start of transcription. The short-lived, unprocessed or partially processed, product is termed *pre-mRNA*; once completely processed, it is termed *mature mRNA*.

Eukaryotic pre-mRNA processing

Processing of mRNA differs greatly among eukaryotes, bacteria and archea. Non-eukaryotic mRNA is essentially mature upon transcription and requires no processing, except in rare cases. Eukaryotic pre-mRNA, however, requires extensive processing.

5' cap addition

A *5′ cap* (also termed an RNA cap, an RNA 7-methylguanosine cap or an RNA m^7G cap) is a modified guanine nucleotide that has been added to the "front" or 5' end of a eukaryotic messenger RNA shortly after the start of transcription. The 5' cap consists of a terminal 7-methylguanosine residue which is linked through a 5'-5'-triphosphate bond to the first transcribed nucleotide. Its presence is critical for recognition by the ribosome and protection from RNases.

Cap addition is coupled to transcription, and occurs co-transcriptionally, such that each influences the other. Shortly after the start of transcription, the 5' end of the mRNA being synthesized is bound by a cap-synthesizing complex associated with RNA polymerase. This enzymatic complex catalyzes the chemical reactions that are required for mRNA capping. Synthesis proceeds as a multi-step biochemical reaction.

Splicing

Splicing is the process by which pre-mRNA is modified to remove certain stretches of non-coding sequences called introns; the stretches that remain include protein-coding sequences and are called exons. Sometimes pre-mRNA messages may be spliced in several different ways, allowing a single gene to encode multiple proteins. This process is called alternative splicing. Splicing is usually performed by an RNA-protein complex called the spliceosome, but some RNA molecules are also capable of catalyzing their own splicing.

Editing

In some instances, an mRNA will be edited, changing the nucleotide composition of that mRNA. An example in humans is the apolipoprotein B mRNA, which is edited in some tissues, but not others. The editing creates an early stop codon, which upon translation, produces a shorter protein.

Polyadenylation

Polyadenylation is the covalent linkage of a polyadenylyl moiety to a messenger RNA molecule. In eukaryotic organisms, most messenger RNA (mRNA) molecules are polyadenylated at the 3' end. The poly(A) tail and the protein bound to it aid in protecting mRNA from degradation by exonucleases. Polyadenylation is also important for transcription termination, export of the mRNA from the nucleus, and translation. mRNA can also be polyadenylated in prokaryotic organisms, where poly(A) tails act to facilitate, rather than impede, exonucleolytic degradation.

Polyadenylation occurs during and immediately after transcription of DNA into RNA. After transcription has been terminated, the mRNA chain is cleaved through the action of an endonuclease complex associated with RNA polymerase. After the mRNA has been cleaved, around 250 adenosine residues are added to the free 3' end at the cleavage site. This reaction is catalyzed by polyadenylate polymerase. Just as in alternative splicing, there can be more than one polyadenylation variant of a mRNA.

Transport

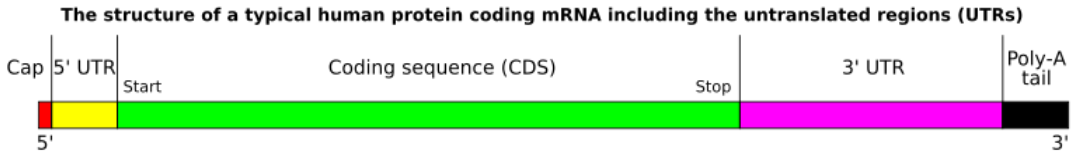
Another difference between eukaryotes and prokaryotes is mRNA transport. Because eukaryotic transcription and translation is compartmentally separated, eukaryotic mRNAs must be exported from the nucleus to the cytoplasm. Mature mRNAs are recognized by their processed modifications and then exported through the nuclear pore. In neurons mRNA must be transported from the soma to the dendrites where local translation occurs in response to external stimuli. Many messages are marked with so-called "zip codes" which targets their transport to a specific location.

Translation

Because prokaryotic mRNA does not need to be processed or transported, translation by the ribosome can begin immediately after the end of transcription. Therefore, it can be said that prokaryotic translation is *coupled* to transcription and occurs *co-transcriptionally*.

Eukaryotic mRNA that has been processed and transported to the cytoplasm (i.e. mature mRNA) can then be translated by the ribosome. Translation may occur at ribosomes free-floating in the cytoplasm, or directed to the endoplasmic reticulum by the signal recognition particle. Therefore, unlike in prokaryotes, eukaryotic translation *is not* directly coupled to transcription.

Structure



The structure of a mature eukaryotic mRNA. A fully processed mRNA includes a 5' cap, 5' UTR, coding region, 3' UTR, and poly(A) tail.

5' cap

The *5′ cap* is a modified guanine nucleotide added to the "front" (5' end) of the pre-mRNA using a 5'-5'-triphosphate linkage. This modification is critical for recognition and proper attachment of mRNA to the ribosome, as well as protection from 5' exonucleases. It may also be important for other essential processes, such as splicing and transport.

Coding regions

Coding regions are composed of codons, which are decoded and translated (in eukaryotes usually into one and in prokaryotes usually into several) proteins by the ribosome. Coding regions begin with the start codon and end with a stop codon. Generally, the start codon is an AUG triplet and the stop codon is UAA, UAG, or UGA. The coding regions tend to be stabilised by internal base pairs, this impedes degradation. In addition to being protein-coding, portions of coding regions may serve as regulatory sequences in the pre-mRNA as exonic splicing enhancers or exonic splicing silencers.

Untranslated regions

Untranslated regions (UTRs) are sections of the mRNA before the start codon and after the stop codon that are not translated, termed the five prime untranslated region (5' UTR) and three prime untranslated region (3' UTR), respectively. These regions are transcribed

with the coding region and thus are exonic as they are present in the mature mRNA. Several roles in gene expression have been attributed to the untranslated regions, including mRNA stability, mRNA localization, and translational efficiency. The ability of a UTR to perform these functions depends on the sequence of the UTR and can differ between mRNAs.

The stability of mRNAs may be controlled by the 5' UTR and/or 3' UTR due to varying affinity for RNA degrading enzymes called ribonucleases and for ancillary proteins that can promote or inhibit RNA degradation.

Translational efficiency, including sometimes the complete inhibition of translation, can be controlled by UTRs. Proteins that bind to either the 3' or 5' UTR may affect translation by influencing the ribosome's ability to bind to the mRNA. MicroRNAs bound to the 3' UTR also may affect translational efficiency or mRNA stability.

Cytoplasmic localization of mRNA is thought to be a function of the 3' UTR. Proteins that are needed in a particular region of the cell can actually be translated there; in such a case, the 3' UTR may contain sequences that allow the transcript to be localized to this region for translation.

Some of the elements contained in untranslated regions form a characteristic secondary structure when transcribed into RNA. These structural mRNA elements are involved in regulating the mRNA. Some, such as the SECIS element, are targets for proteins to bind. One class of mRNA element, the riboswitches, directly bind small molecules, changing their fold to modify levels of transcription or translation. In these cases, the mRNA regulates itself.

Poly(A) tail

The 3' poly(A) tail is a long sequence of adenine nucleotides (often several hundred) added to the 3' end of the pre-mRNA. This tail promotes export from the nucleus and translation, and protects the mRNA from degradation.

Monocistronic versus polycistronic mRNA

An mRNA molecule is said to be monocistronic when it contains the genetic information to translate only a single protein. This is the case for most of the eukaryotic mRNAs. On the other hand, polycistronic mRNA carries the information of several genes, which are translated into several proteins. These proteins usually have a related function and are grouped and regulated together in an operon. Most of the mRNA found in bacteria and archea are polycistronic. Dicistronic or bicistronic is the term used to describe an mRNA that encodes only two proteins.

mRNA circularization

In eukaryotes it is thought that mRNA molecules form circular structures due to an interaction between the cap binding complex and poly(A)-binding protein. Circularization is thought to promote recycling of ribosomes on the same message leading to efficient translation.

Degradation

Different mRNAs within the same cell have distinct lifetimes (stabilities). In bacterial cells, individual mRNAs can survive from seconds to more than an hour; in mammalian cells, mRNA lifetimes range from several minutes to days. The greater the stability of an mRNA, the more protein may be produced from that mRNA. The limited lifetime of mRNA enables a cell to alter protein synthesis rapidly in response to its changing needs. There are many mechanisms that lead to the destruction of a mRNA, some of which are described below.

Prokaryotic mRNA degradation

In prokaryotes the lifetime of mRNA is generally much shorter than in eukaryotes. Prokaryotes degrade messages by using a combination of ribonucleases, including endonucleases, 3' exonucleases, and 5' exonucleases. In some instances, small RNA molecules (sRNA) tens to hundreds of nucleotides long can stimulate the degradation of specific mRNAs by base pairing with complementary sequences and facilitating ribonuclease cleavage. It was recently shown that bacteria also have a sort of 5' cap consisting of a triphosphate on the 5' end. Removal of two of the phosphates leaves a 5' monophosphate, causing the message to be destroyed by the endonuclease RNase E.

Eukaryotic mRNA turnover

Inside eukaryotic cells there is a balance between the processes of translation and mRNA decay. Messages that are being actively translated are bound by ribosomes, the eukaryotic initiation factors eIF-4E and eIF-4G, and poly(A)-binding protein. eIF-4E and eIF-4G block the decapping enzyme (DCP2), and poly(A)-binding protein blocks the exosome complex, protecting the ends of the message. The balance between translation and decay is reflected in the size and abundance of cytoplasmic structures known as P-bodies The poly(A) tail of the mRNA is shortened by specialized exonucleases that are targeted to specific messenger RNAs by a combination of cis-regulatory sequences on the RNA and trans-acting RNA-binding proteins. Poly(A) tail removal is thought to disrupt the circular structure of the message and destabilize the cap binding complex. The message is then subject to degradation by either the exosome complex or the decapping complex. In this way, translationally inactive messages can be destroyed quickly, while active messages remain intact. The mechanism by which translation stops and the message is handed-off to decay complexes is not understood in detail.

AU-rich element decay

The presence of AU-rich elements in some mammalian mRNAs tends to destabilize those transcripts through the action of cellular proteins that bind these sequences and stimulate poly(A) tail removal. Loss of the poly(A) tail is thought to promote mRNA degradation by facilitating attack by both the exosome complex and the decapping complex. Rapid mRNA degradation via AU-rich elements is a critical mechanism for preventing the overproduction of potent cytokines such as tumor necrosis factor (TNF) and granulocyte-macrophage colony stimulating factor (GM-CSF). AU-rich elements also regulate the biosynthesis of proto-oncogenic transcription factors like c-Jun and c-Fos.

Nonsense mediated decay

Eukaryotic messages are subject to surveillance by nonsense mediated decay (NMD), which checks for the presence of premature stop codons (nonsense codons) in the message. These can arise via incomplete splicing, V(D)J recombination in the adaptive immune system, mutations in DNA, transcription errors, leaky scanning by the ribosome causing a frame shift, and other causes. Detection of a premature stop codon triggers mRNA degradation by 5' decapping, 3' poly(A) tail removal, or endonucleolytic cleavage.

Small interfering RNA (siRNA)

In metazoans, small interfering RNAs (siRNAs) processed by Dicer are incorporated into a complex known as the RNA-induced silencing complex or RISC. This complex contains an endonuclease that cleaves perfectly complementary messages to which the siRNA binds. The resulting mRNA fragments are then destroyed by exonucleases. siRNA is commonly used in laboratories to block the function of genes in cell culture. It is thought to be part of the innate immune system as a defense against double-stranded RNA viruses.

MicroRNA (miRNA)

MicroRNAs (miRNAs) are small RNAs that typically are partially complementary to sequences in metazoan messenger RNAs. Binding of a miRNA to a message can repress translation of that message and accelerate poly(A) tail removal, thereby hastening mRNA degradation. The mechanism of action of miRNAs is the subject of active research.

Other decay mechanisms

There are other ways in which messages can be degraded, including non-stop decay, silencing by Piwi-interacting RNA (piRNA), and surely other means.

Chapter- 9

Cis-Regulatory Module

***Cis*-regulatory module** (CRM) is a stretch of DNA, usually 100-1000 DNA base pairs in length, where a number of transcription factors can bind and regulate expression of nearby genes. One *cis*-regulatory element can regulate several genes, and conversely, one gene can have several *cis*-regulatory modules. *Cis*-regulatory modules are one of several types of functional regulatory elements. Regulatory elements are binding sites for transcription factors, which are involved in gene regulation. *Cis*-regulatory modules perform a large amount of developmental information processing. *Cis*-regulatory modules are non-random clusters at their specified target site that contain transcription factor binding sites . They are labeled as *cis* because they are typically located on the same DNA as the genes they control as opposed to *trans*, which refers to effects on genes not located on the same strand or farther away, such as transcription factors. The original definition presented cis-regulatory modules as enhancers of cis-acting DNA, which increased the rate of transcription from a linked promoter. However, this definition has changed to define *cis*-regulatory modules as a DNA sequence with transcription factor binding sites which are clustered into modular structures, including -but not limited to- locus control regions, promoters, enhancers, silencers, boundary control elements and other modulators . *Cis*-regulatory modules can be divided into classes. Enhancers regulate gene expression positively. Insulators work indirectly by interacting with other nearby *cis*-regulatory modules. Silencers act by silencing genes. *Cis*-regulatory modules carry out their function by integrating the active transcription factors and the associated co-factors at a specific time and place in the cell, this information is read and an output is given.

Gene-regulation function of a cis*-regulatory module*

The design of *cis*-regulatory modules is such that transcription factors and epigenetic modifications serve as inputs, and the output of the module is the command given to the transcription machinery, which in turn determines the rate of gene transcription or whether it is turned on or off. There are two types of transcription factor inputs: those that determine when the target gene is to be expressed and those that serve as functional *drivers*, which come into play only during specific situations during development. These inputs can come from different time points, can represent different signal ligands, or can come from different domains or lineages of cells. However, a lot still remains unknown.

Additionally, the regulation of chromatin structure and nuclear organization also play a role in determining and controlling the function of cis-regulatory modules. Thus gene-regulation functions (GRF) provide a unique characteristic of a cis-regulatory module (CRM), relating the concentrations of transcription factors (input) to the promoter activities (output). The challenge is to predict GRFs. This challenge still remains unsolved. In general, gene-regulation functions do not use Boolean logic, although in some cases the approximation of the Boolean logic is still very useful.

The Boolean logic assumption

Within the assumption of the Boolean logic, four principles guiding the operation of these cis-regulatory modules are : 1) The design of the *cis*-regulatory module determines the regulatory function. 2) In relation to development, these modules can generate both positive and negative outputs. 3) The output of each module is a product of the various operations performed on it. Common operations: “OR” logic gate- This design indicates that in an output will be given when either input is given. “AND” logic gate– In this design two different regulatory factors are necessary to make sure that a positive output results. “Toggle Switches”- This design occurs when the signal ligand is absent while the transcription factor is present; this transcription factor ends up acting as a dominant repressor. However, once the signal ligand is present the transcription factor’s role as repressor is eliminated and transcription can occur . Other Boolean logic operations can occur as well, such as sequence specific transcriptional repressors, which when they bind to the *cis*-regulatory module lead to an output of zero. Additionally, besides influence from the different logic operations, the output of a "cis"-regulatory module will also be influenced by prior events. 4) *Cis*-regulatory modules must interact with other regulatory elements. For the most part, even with the presence of functional overlap between *cis*-regulatory modules of a gene, the modules’ inputs and outputs tend to not be the same .

Identification and computational prediction of CRMs

Besides experimentally determining CRMs, there are various bioinformatics algorithms for predicting them. Most algorithms try to search for significant combinations of transcription factor binding sites (DNA binding sites) in promoter sequences of co-expressed genes. More advanced methods combine the search for significant motifs with correlation in gene expression datasets between transcription factors and target genes. Both methods have been implemented, for example, in the ModuleMaster. Other programs created for the identification and prediction if "cis-*regulatory modules include:*

Stubb uses hidden Markov models to identify statistically significant clusters of transcription factor combinations. It also uses a second related genome to improve the prediction accuracy of the model.

Bayesian Networks use an algorithm that combines site predictions and tissue-specific expression data for transcription factors and target genes of interest. This model also uses regression trees to depict the relationship between the identified *cis*-regulatory module and the possible binding set of transcription factors.

CRÈME examine clusters of target sites for transcription factors of interest. This program uses a database of confirmed transcription factor binding sites that were annotated across the human genome. A search algorithm is applied to the data set to identify possible combinations of transcription factors, which have binding sites that are close to the promoter of the gene set of interest. The possible cis-regulatory modules are then statistically analyzed and the significant combinations are graphically represented

Active *cis*-regulatory modules in a genomic sequence have been difficult to identify. Problems in identification arise due to the fact that often scientists find themselves with a small set of known transcription factors, so it makes it harder to identify statistically significant clusters of transcription factor binding sites. Additionally, high costs limit the use of large whole genome tiling arrays.

Classification of cis*-regulatory modules*

Cis-regulatory modules can be characterized by the information processing that they encode and the organization of their transcription factor binding sites. Additionally, *cis*-regulatory modules are also characterized by the way they affect the probability, proportion, and rate of transcription. Highly cooperative and coordinated *cis*-regulatory modules are classified as enhanceosomes . The architecture and the arrangement of the transcription factor binding sites are critical because disruption of the arrangement could cancel out the function . Functional flexible *cis*-regulatory modules are called billboards. Their transcriptional output is the summation effect of the bound transcription factors. Enhancers affect the probability of a gene being activated, but have little or no effect on rate. Binary response model acts like an on/off switch for transcription. This model will increase or decrease the amount of cells that transcribe a gene, however it does not affect the rate of transcription. Rheostatic response model describes cis-regulatory modules as regulators of the initiation rate of transcription of its associated gene.

Mode of action

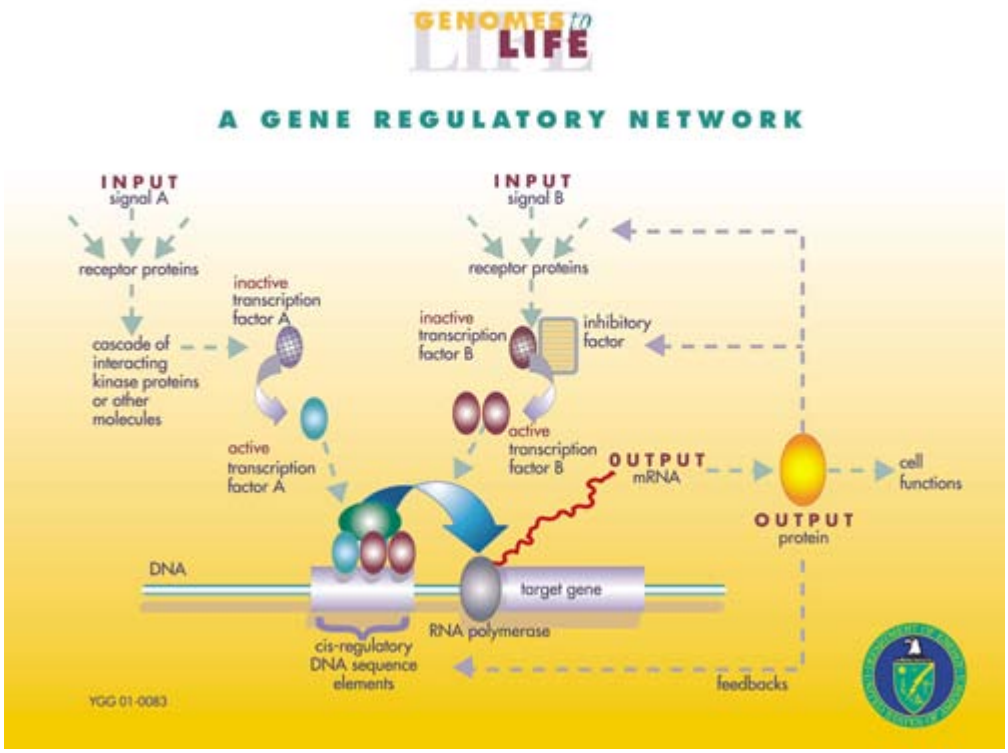
Cis-regulatory modules can regulate their target genes over large distances, several models have been proposed to describe the way that these modules may communicate with their target gene promoter. DNA Scanning Model: In this model the transcription factor and cofactor complex form at the *cis*-regulatory module and then continues to move along the DNA sequence until it finds the target gene promoter. Looping Model: In this model the transcription factor binds to the *cis*-regulatory module, which then causes the *looping* of the DNA sequence and allows for the interaction with the target gene promoter. Facilitated Tracking Model: This model combines parts of the two previous models. The transcription factor-*cis*-regulatory module complex causes the looping of the DNA sequence slowly towards the target promoter and forms a stable looped configuration.

Cis-*regulatory module in gene regulatory network*

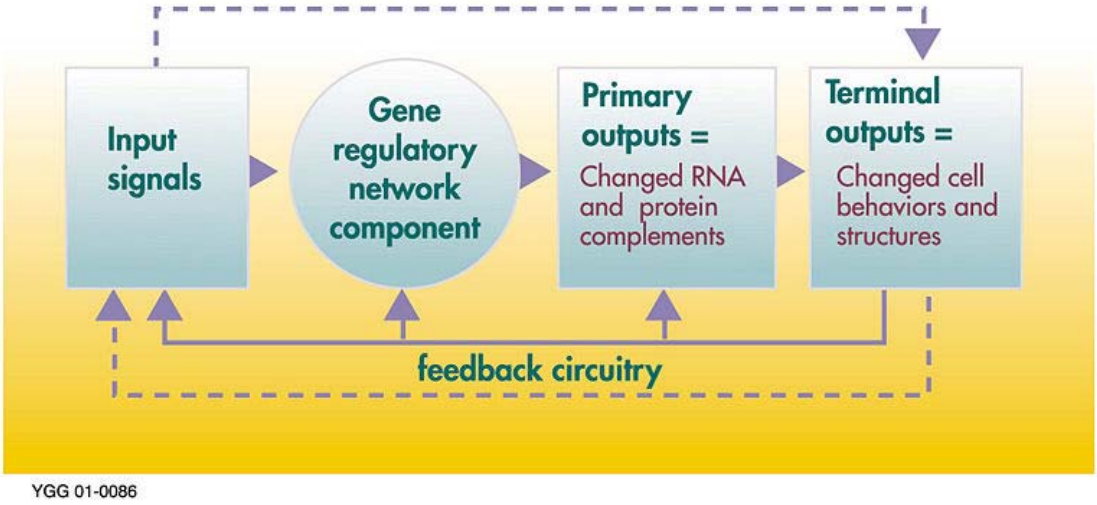
The function of a gene regulatory network depends on the architecture of the nodes, whose function is dependent on the multiple *cis*-regulatory modules. The layout of *cis*-regulatory modules can provide enough information to generate spatial and temporal patterns of gene expression. During development each domain, where each domain represents a different spatial regions of the embryo, of gene expression will be under the control of different *cis*-regulatory module(s). The design of regulatory modules help in producing feedback, feed forward, and cross-regulatory loops.

Chapter- 10

Gene Regulatory Network



Structure of a Gene Regulatory Network



Control process of a Gene Regulatory Network

A **gene regulatory network** or **genetic regulatory network (GRN)** is a collection of DNA segments in a cell which interact with each other (indirectly through their RNA and protein expression products) and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed into mRNA. In general, each mRNA molecule goes on to make a specific protein (or set of proteins). In some cases this protein will be structural, and will accumulate at the cell-wall or within the cell to give it particular structural properties. In other cases the protein will be an enzyme; a micro-machine that catalyses a certain reaction, such as the breakdown of a food source or toxin. Some proteins though serve only to activate other genes, and these are the transcription factors that are the main players in regulatory networks or cascades. By binding to the promoter region at the start of other genes they turn them on, initiating the production of another protein, and so on. Some transcription factors are inhibitory.

In single-celled organisms regulatory networks respond to the external environment, optimising the cell at a given time for survival in this environment. Thus a yeast cell, finding itself in a sugar solution, will turn on genes to make enzymes that process the sugar to alcohol. This process, which we associate with wine-making, is how the yeast cell makes its living, gaining energy to multiply, which under normal circumstances would enhance its survival prospects.

In multicellular animals the same principle has been put in the service of gene cascades that control body-shape. Each time a cell divides, two cells result which, although they contain the same genome in full, can differ in which genes are turned on and making proteins. Sometimes a 'self-sustaining feedback loop' ensures that a cell maintains its identity and passes it on. Less understood is the mechanism of epigenetics by which chromatin modification may provide cellular memory by blocking or allowing transcription. A major feature of multicellular animals is the use of morphogen gradients,

which in effect provide a positioning system that tells a cell where in the body it is, and hence what sort of cell to become. A gene that is turned on in one cell may make a product that leaves the cell and diffuses through adjacent cells, entering them and turning on genes only when it is present above a certain threshold level. These cells are thus induced into a new fate, and may even generate other morphogens that signal back to the original cell. Over longer distances morphogens may use the active process of signal transduction. Such signalling controls embryogenesis, the building of a body plan from scratch through a series of sequential steps. They also control maintain adult bodies through feedback processes, and the loss of such feedback because of a mutation can be responsible for the cell proliferation that is seen in cancer. In parallel with this process of building structure, the gene cascade turns on genes that make structural proteins that give each cell the physical properties it needs. It has been suggested that, because biological molecular interactions are intrinsically stochastic, gene networks are the result of cellular processes and not their cause. (i.e. Cellular Darwinism) However, recent experimental evidence has favored the attractor view of cell fates.

Overview

At one level, biological cells can be thought of as "partially-mixed bags" of biological chemicals – in the discussion of gene regulatory networks, these chemicals are mostly the mRNAs and proteins that arise from gene expression. These mRNA and proteins interact with each other with various degrees of specificity. Some diffuse around the cell. Others are bound to cell membranes, interacting with molecules in the environment. Still others pass through cell membranes and mediate long range signals to other cells in a multi-cellular organism. These molecules and their interactions comprise a *gene regulatory network*. A typical gene regulatory network looks something like this:

The nodes of this network are proteins, their corresponding mRNAs, and protein/protein complexes. Nodes that are depicted as lying along vertical lines are associated with the cell/environment interfaces, while the others are free-floating and diffusible. Implied are genes, the DNA sequences which are transcribed into the mRNAs that translate into proteins. Edges between nodes represent individual molecular reactions, the protein/protein and protein/mRNA interactions through which the products of one gene affect those of another, though the lack of experimentally obtained information often implies that some reactions are not modeled at such a fine level of detail. These interactions can be inductive (the arrowheads), with an increase in the concentration of one leading to an increase in the other, or inhibitory (the filled circles), with an increase in one leading to a decrease in the other. A series of edges indicates a chain of such dependences, with cycles corresponding to feedback loops. The network structure is an abstraction of the system's chemical dynamics, describing the manifold ways in which one substance affects all the others to which it is connected. In practice, such GRNs are inferred from the biological literature on a given system and represent a distillation of the collective knowledge about a set of related biochemical reactions.

Genes can be viewed as nodes in the network, with input being proteins such as transcription factors, and outputs being the level of gene expression. The node itself can

also be viewed as a function which can be obtained by combining basic functions upon the inputs (in the Boolean network described below these are Boolean functions, typically AND, OR, and NOT). These functions have been interpreted as performing a kind of information processing within the cell, which determines cellular behavior. The basic drivers within cells are concentrations of some proteins, which determine both spatial (location within the cell or tissue) and temporal (cell cycle or developmental stage) coordinates of the cell, as a kind of "cellular memory". The gene networks are only beginning to be understood, and it is a next step for biology to attempt to deduce the functions for each gene "node", to help understand the behavior of the system in increasing levels of complexity, from gene to signaling pathway, cell or tissue level.

Mathematical models of GRNs have been developed to capture the behavior of the system being modeled, and in some cases generate predictions corresponding with experimental observations. In some other cases, models have proven to make accurate novel predictions, which can be tested experimentally, thus suggesting new approaches to explore in an experiment that sometimes wouldn't be considered in the design of the protocol of an experimental laboratory. The most common modeling technique involves the use of coupled ordinary differential equations (ODEs). Several other promising modeling techniques have been used, including Boolean networks, Petri nets, Bayesian networks, graphical Gaussian models, Stochastic, and Process Calculi. Conversely, techniques have been proposed for generating models of GRNs that best explain a set of time series observations.

Modelling

Coupled ODEs

It is common to model such a network with a set of coupled ordinary differential equations (ODEs) or stochastic ODEs, describing the reaction kinetics of the constituent parts. Suppose that our regulatory network has N nodes, and let $S_1(t), S_2(t), \ldots, S_N(t)$ represent the concentrations of the N corresponding substances at time t. Then the temporal evolution of the system can be described approximately by

$$\frac{dS_j}{dt} = f_j\left(S_1, S_2, \ldots, S_N\right)$$

where the functions f_j express the dependence of S_j on the concentrations of other substances present in the cell. The functions f_j are ultimately derived from basic principles of chemical kinetics or simple expressions derived from these e.g. Michaelis-Menten enzymatic kinetics. Hence, the functional forms of the f_j are usually chosen as low-order polynomials or Hill functions that serve as an ansatz for the real molecular dynamics. Such models are then studied using the mathematics of nonlinear dynamics. System-specific information, like reaction rate constants and sensitivities, are encoded as constant parameters.

By solving for the fixed point of the system:

$$\frac{dS_j}{dt} = 0$$

for all *j*, one obtains (possibly several) concentration profiles of proteins and mRNAs that are theoretically sustainable (though not necessarily stable). Steady states of kinetic equations thus correspond to potential cell types, and oscillatory solutions to the above equation to naturally cyclic cell types. Mathematical stability of these attractors can usually be characterized by the sign of higher derivatives at critical points, and then correspond to biochemical stability of the concentration profile. Critical points and bifurcations in the equations correspond to critical cell states in which small state or parameter perturbations could switch the system between one of several stable differentiation fates. Trajectories correspond to the unfolding of biological pathways and transients of the equations to short-term biological events.

Boolean network

The following example illustrates how a Boolean network can model a GRN together with its gene products (the outputs) and the substances from the environment that affect it (the inputs). Stuart Kauffman was amongst the first biologists to use the metaphor of Boolean networks to model genetic regulatory networks.

1. Each gene, each input, and each output is represented by a node in a directed graph in which there is an arrow from one node to another if and only if there is a causal link between the two nodes.
2. Each node in the graph can be in one of two states: on or off.
3. For a gene, "on" corresponds to the gene being expressed; for inputs and outputs, "on" corresponds to the substance being present.
4. Time is viewed as proceeding in discrete steps. At each step, the new state of a node is a Boolean function of the prior states of the nodes with arrows pointing towards it.

The validity of the model can be tested by comparing simulation results with time series observations.

Continuous networks

Continuous network models of GRNs are an extension of the boolean networks described above. Nodes still represent genes and connections between them regulatory influences on gene expression. Genes in biological systems display a continuous range of activity levels and it has been argued that using a continuous representation captures several properties of gene regulatory networks not present in the Boolean model. Formally most of these approaches are similar to an artificial neural network, as inputs to a node are summed up and the result serves as input to a sigmoid function, e.g., but proteins do often control gene expression in a synergistic, i.e. non-linear, way. However there is now a

continuous network model that allows grouping of inputs to a node thus realizing another level of regulation. This model is formally closer to a higher order recurrent neural network. The same model has also been used to mimic the evolution of cellular differentiation and even multicellular morphogenesis.

Stochastic gene networks

Recent experimental results have demonstrated that gene expression is a stochastic process. Thus, many authors are now using the stochastic formalism, after the work by. Works on single gene expression and small synthetic genetic networks, such as the genetic toggle switch of Tim Gardner and Jim Collins, provided additional experimental data on the phenotypic variability and the stochastic nature of gene expression. The first versions of stochastic models of gene expression involved only instantaneous reactions and were driven by the Gillespie algorithm.

Since some processes, such as gene transcription, involve many reactions and could not be correctly modeled as an instantaneous reaction in a single step, it was proposed to model these reactions as single step multiple delayed reactions in order to account for the time it takes for the entire process to be complete.

From here, a set of reactions were proposed that allow generating GRNs. These are then simulated using a modified version of the Gillespie algorithm, that can simulate multiple time delayed reactions (chemical reactions where each of the products is provided a time delay that determines when will it be released in the system as a "finished product").

For example, basic transcription of a gene can be represented by the following single-step reaction (RNAP is the RNA polymerase, RBS is the RNA ribosome binding site, and Pro_i is the promoter region of gene *i*):

$$\mathrm{RNAP} + \mathrm{Pro}_i \xrightarrow{k_{i,bas}} \mathrm{Pro}_i(\tau_i^1) + \mathrm{RBS}_i(\tau_i^1) + \mathrm{RNAP}(\tau_i^2)$$

A recent work proposed a simulator (SGNSim, *Stochastic Gene Networks Simulator*), that can model GRNs where transcription and translation are modeled as multiple time delayed events and its dynamics is driven by a stochastic simulation algorithm (SSA) able to deal with multiple time delayed events. The time delays can be drawn from several distributions and the reaction rates from complex functions or from physical parameters. SGNSim can generate ensembles of GRNs within a set of user-defined parameters, such as topology. It can also be used to model specific GRNs and systems of chemical reactions. Genetic perturbations such as gene deletions, gene over-expression, insertions, frame shift mutations can also be modeled as well.

The GRN is created from a graph with the desired topology, imposing in-degree and out-degree distributions. Gene promoter activities are affected by other genes expression products that act as inputs, in the form of monomers or combined into multimers and set as direct or indirect. Next, each direct input is assigned to an operator site and different transcription factors can be allowed, or not, to compete for the same operator site, while

indirect inputs are given a target. Finally, a function is assigned to each gene, defining the gene's response to a combination of transcription factors (promoter state). The transfer functions (that is, how genes respond to a combination of inputs) can be assigned to each combination of promoter states as desired.

In other recent work, multiscale models of gene regulatory networks have been developed that focus on synthetic biology applications. Simulations have been used that model all biomolecular interactions in transcription, translation, regulation, and induction of gene regulatory networks, guiding the design of synthetic systems.

Network connectivity

Empirical data indicate that biological gene networks are sparsely connected, and that the average number of upstream-regulators per gene is less than two. Theoretical results show that selection for robust gene networks will favor minimally complex, more sparsely connected, networks. These results suggest that a sparse, minimally connected, genetic architecture may be a fundamental design constraint shaping the evolution of gene network complexity.

Chapter- 11

Transcription (Genetics)

Transcription is the process of creating a complementary RNA copy of a sequence of DNA. Both RNA and DNA are nucleic acids, which use base pairs of nucleotides as a complementary language that can be converted back and forth from DNA to RNA by the action of the correct enzymes. During transcription, a DNA sequence is read by RNA polymerase, which produces a complementary, antiparallel RNA strand. As opposed to DNA replication, transcription results in an RNA complement that includes uracil (U) in all instances where thymine (T) would have occurred in a DNA complement.

Transcription can be explained easily in 4 or 5 simple steps, each moving like a wave along the DNA.

1. DNA unwinds/"unzips" as the Hydrogen Bonds Break.
2. The free nucleotides of the RNA, pair with complementary DNA bases.
3. RNA sugar-phosphate backbone forms. (Aided by RNA Polymerase.)
4. Hydrogen bonds of the untwisted RNA+DNA "ladder" break, freeing the new RNA.
5. If the cell has a nucleus, the RNA is further processed and then moves through the small nuclear pores to the cytoplasm.

Transcription is the first step leading to gene expression. The stretch of DNA transcribed into an RNA molecule is called a *transcription unit* and encodes at least one gene. If the gene transcribed encodes a protein, the result of transcription is messenger RNA (mRNA), which will then be used to create that protein via the process of translation. Alternatively, the transcribed gene may encode for either ribosomal RNA (rRNA) or transfer RNA (tRNA), other components of the protein-assembly process, or other ribozymes.

A DNA transcription unit encoding for a protein contains not only the sequence that will eventually be directly translated into the protein (the *coding sequence*) but also *regulatory sequences* that direct and regulate the synthesis of that protein. The regulatory sequence before (upstream from) the coding sequence is called the five prime untranslated region (5'UTR), and the sequence following (downstream from) the coding sequence is called the three prime untranslated region (3'UTR).

Transcription has some proofreading mechanisms, but they are fewer and less effective than the controls for copying DNA; therefore, transcription has a lower copying fidelity than DNA replication.

As in DNA replication, DNA is read from 3' → 5' during transcription. Meanwhile, the complementary RNA is created from the 5' → 3' direction. This means its 5' end is created first in base pairing. Although DNA is arranged as two antiparallel strands in a double helix, only one of the two DNA strands, called the template strand, is used for transcription. This is because RNA is only single-stranded, as opposed to double-stranded DNA. The other DNA strand is called the coding strand, because its sequence is the same as the newly created RNA transcript (except for the substitution of uracil for thymine). The use of only the 3' → 5' strand eliminates the need for the Okazaki fragments seen in DNA replication.

Transcription is divided into 5 stages: *pre-initiation*, *initiation*, *promoter clearance*, *elongation* and *termination*.

Major steps

Pre-initiation

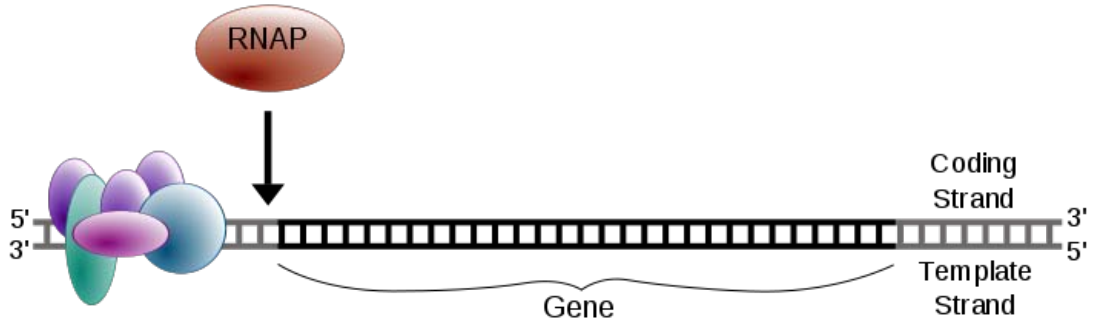
In eukaryotes, RNA polymerase, and therefore the initiation of transcription, requires the presence of a core promoter sequence in the DNA. Promoters are regions of DNA that promote transcription and, in eukaryotes, are found at -30, -75, and -90 base pairs upstream from the start site of transcription. Core promoters are sequences within the promoter that are essential for transcription initiation. RNA polymerase is able to bind to core promoters in the presence of various specific transcription factors.

The most common type of core promoter in eukaryotes is a short DNA sequence known as a TATA box, found 25-30 base pairs upstream from the start site of transcription. The TATA box, as a core promoter, is the binding site for a transcription factor known as TATA-binding protein (TBP), which is itself a subunit of another transcription factor, called Transcription Factor II D (TFIID). After TFIID binds to the TATA box via the TBP, five more transcription factors and RNA polymerase combine around the TATA box in a series of stages to form a preinitiation complex. One transcription factor, DNA helicase, has helicase activity and so is involved in the separating of opposing strands of double-stranded DNA to provide access to a single-stranded DNA template. However, only a low, or basal, rate of transcription is driven by the preinitiation complex alone. Other proteins known as activators and repressors, along with any associated coactivators or corepressors, are responsible for modulating transcription rate.

Thus, preinitiation complex contains: 1. Core Promoter Sequence 2. Transcription Factors 3. DNA Helicase 4. RNA Polymerase 5. Activators and Repressors The transcription preinitiation in archaea is, in essence, homologous to that of eukaryotes, but is much less complex. The archaeal preinitiation complex assembles at a TATA-box

binding site; however, in archaea, this complex is composed of only RNA polymerase II, TBP, and TFB (the archaeal homologue of eukaryotic transcription factor II B (TFIIB)).

Initiation



Simple diagram of transcription initiation. RNAP = RNA polymerase

In bacteria, transcription begins with the binding of RNA polymerase to the promoter in DNA. RNA polymerase is a core enzyme consisting of five subunits: 2 α subunits, 1 β subunit, 1 β' subunit, and 1 ω subunit. At the start of initiation, the core enzyme is associated with a sigma factor that aids in finding the appropriate -35 and -10 base pairs downstream of promoter sequences.

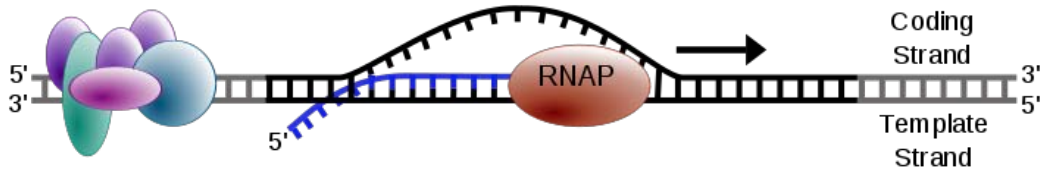
Transcription initiation is more complex in eukaryotes. Eukaryotic RNA polymerase does not directly recognize the core promoter sequences. Instead, a collection of proteins called transcription factors mediate the binding of RNA polymerase and the initiation of transcription. Only after certain transcription factors are attached to the promoter does the RNA polymerase bind to it. The completed assembly of transcription factors and RNA polymerase bind to the promoter, forming a transcription initiation complex. Transcription in the archaea domain is similar to transcription in eukaryotes.

Promoter clearance

After the first bond is synthesized, the RNA polymerase must clear the promoter. During this time there is a tendency to release the RNA transcript and produce truncated transcripts. This is called *abortive initiation* and is common for both eukaryotes and prokaryotes. Abortive initiation continues to occur until the σ factor rearranges, resulting in the transcription elongation complex (which gives a 35 bp moving footprint). The σ factor is released before 80 nucleotides of mRNA are synthesized. Once the transcript reaches approximately 23 nucleotides, it no longer slips and elongation can occur. This, like most of the remainder of transcription, is an energy-dependent process, consuming adenosine triphosphate (ATP).

Promoter clearance coincides with phosphorylation of serine 5 on the carboxy terminal domain of RNA Pol in eukaryotes, which is phosphorylated by TFIIH.

Elongation



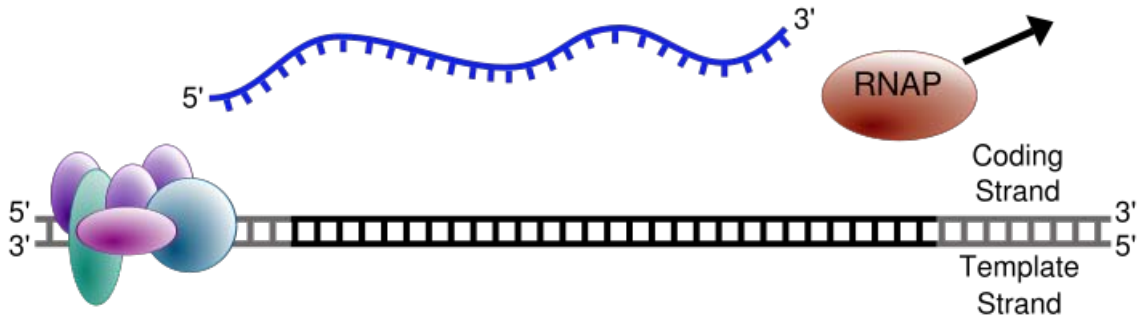
Simple diagram of transcription elongation

One strand of the DNA, the *template strand* (or noncoding strand), is used as a template for RNA synthesis. As transcription proceeds, RNA polymerase traverses the template strand and uses base pairing complementarity with the DNA template to create an RNA copy. Although RNA polymerase traverses the template strand from 3' → 5', the coding (non-template) strand and newly-formed RNA can also be used as reference points, so transcription can be described as occurring 5' → 3'. This produces an RNA molecule from 5' → 3', an exact copy of the coding strand (except that thymines are replaced with uracils, and the nucleotides are composed of a ribose (5-carbon) sugar where DNA has deoxyribose (one less oxygen atom) in its sugar-phosphate backbone).

Unlike DNA replication, mRNA transcription can involve multiple RNA polymerases on a single DNA template and multiple rounds of transcription (amplification of particular mRNA), so many mRNA molecules can be rapidly produced from a single copy of a gene.

Elongation also involves a proofreading mechanism that can replace incorrectly incorporated bases. In eukaryotes, this may correspond with short pauses during transcription that allow appropriate RNA editing factors to bind. These pauses may be intrinsic to the RNA polymerase or due to chromatin structure.

Termination



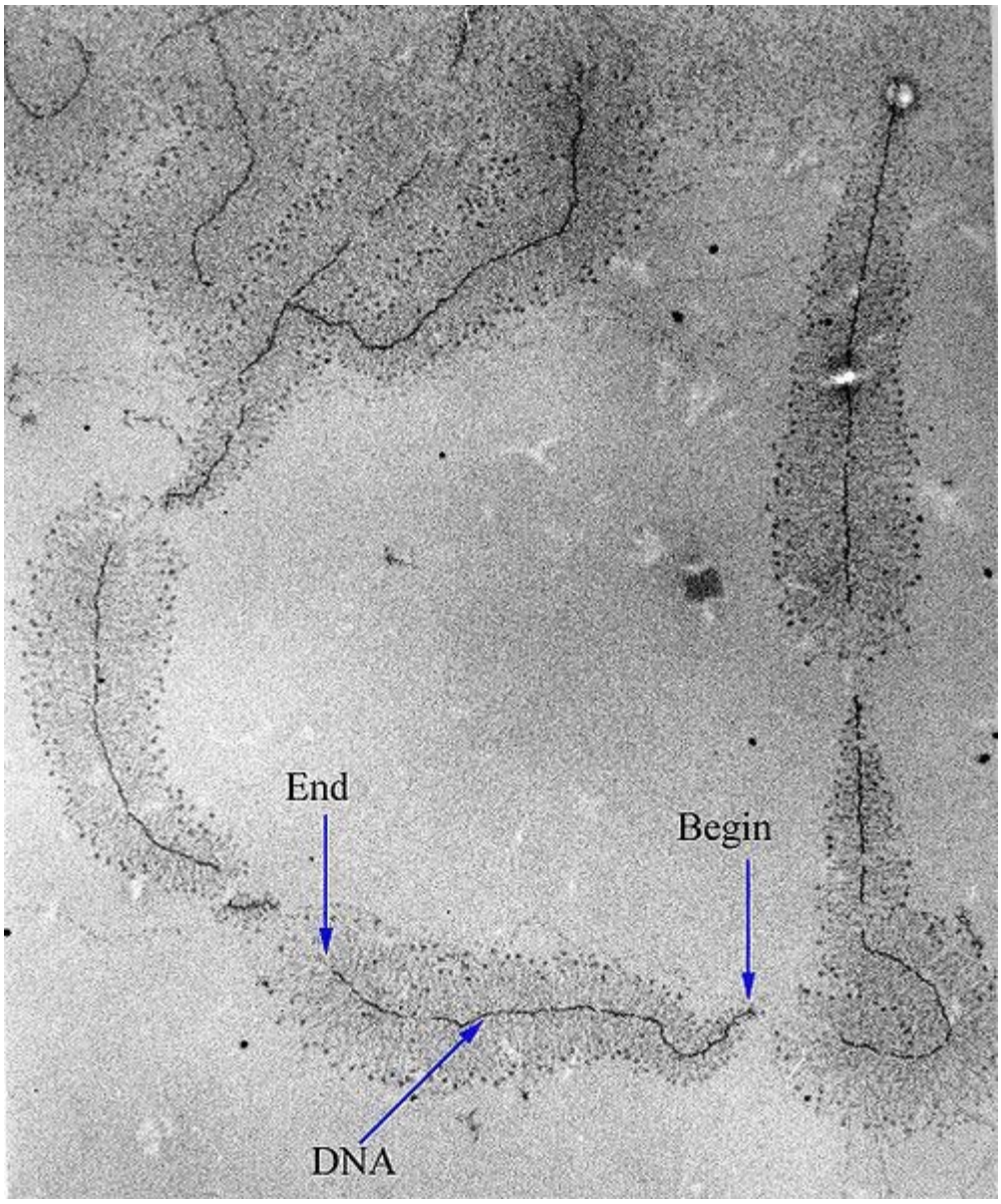
Simple diagram of transcription termination

Bacteria use two different strategies for transcription termination. In Rho-independent transcription termination, RNA transcription stops when the newly synthesized RNA molecule forms a G-C-rich hairpin loop followed by a run of Us. When the hairpin forms, the mechanical stress breaks the weak rU-dA bonds, now filling the DNA-RNA hybrid.

This pulls the poly-U transcript out of the active site of the RNA polymerase, in effect, terminating transcription. In the "Rho-dependent" type of termination, a protein factor called "Rho" destabilizes the interaction between the template and the mRNA, thus releasing the newly synthesized mRNA from the elongation complex.

Transcription termination in eukaryotes is less understood but involves cleavage of the new transcript followed by template-independent addition of *A*s at its new 3' end, in a process called polyadenylation.

Measuring and detecting transcription



Electron micrograph of the ribosomal transcription process. The forming mRNA strands are visible as branches from the main DNA strand.

Transcription can be measured and detected in a variety of ways:

- Nuclear Run-on assay: measures the relative abundance of newly formed transcripts
- RNase protection assay and ChIP-Chip of RNAP: detect active transcription sites
- RT-PCR: measures the absolute abundance of total or nuclear RNA levels, which may however differ from transcription rates
- DNA microarrays: measures the relative abundance of the global total or nuclear RNA levels; however, these may differ from transcription rates
- In situ hybridization: detects the presence of a transcript
- MS2 tagging: by incorporating RNA stem loops, such as MS2, into a gene, these become incorporated into newly synthesized RNA. The stem loops can then be detected using a fusion of GFP and the MS2 coat protein, which has a high affinity, sequence-specific interaction with the MS2 stem loops. The recruitment of GFP to the site of transcription is visualised as a single fluorescent spot. This remarkable new approach has revealed that transcription occurs in discontinuous bursts, or pulses. With the notable exception of in situ techniques, most other methods provide cell population averages, and are not capable of detecting this fundamental property of genes.
- Northern blot: the traditional method, and until the advent of RNA-Seq, the most quantitative
- RNA-Seq: applies next-generation sequencing techniques to sequence whole transcriptomes, which allows the measurement of relative abundance of RNA, as well as the detection of additional variations such as fusion genes, post-translational edits and novel splice sites

Transcription factories

Active transcription units are clustered in the nucleus, in discrete sites called transcription factories or euchromatin. Such sites can be visualized by allowing engaged polymerases to extend their transcripts in tagged precursors (Br-UTP or Br-U) and immuno-labeling the tagged nascent RNA. Transcription factories can also be localized using fluorescence in situ hybridization or marked by antibodies directed against polymerases. There are ~10,000 factories in the nucleoplasm of a HeLa cell, among which are ~8,000 polymerase II factories and ~2,000 polymerase III factories. Each polymerase II factory contains ~8 polymerases. As most active transcription units are associated with only one polymerase, each factory usually contains ~8 different transcription units. These units might be associated through promoters and/or enhancers, with loops forming a ‘cloud’ around the factor.

History

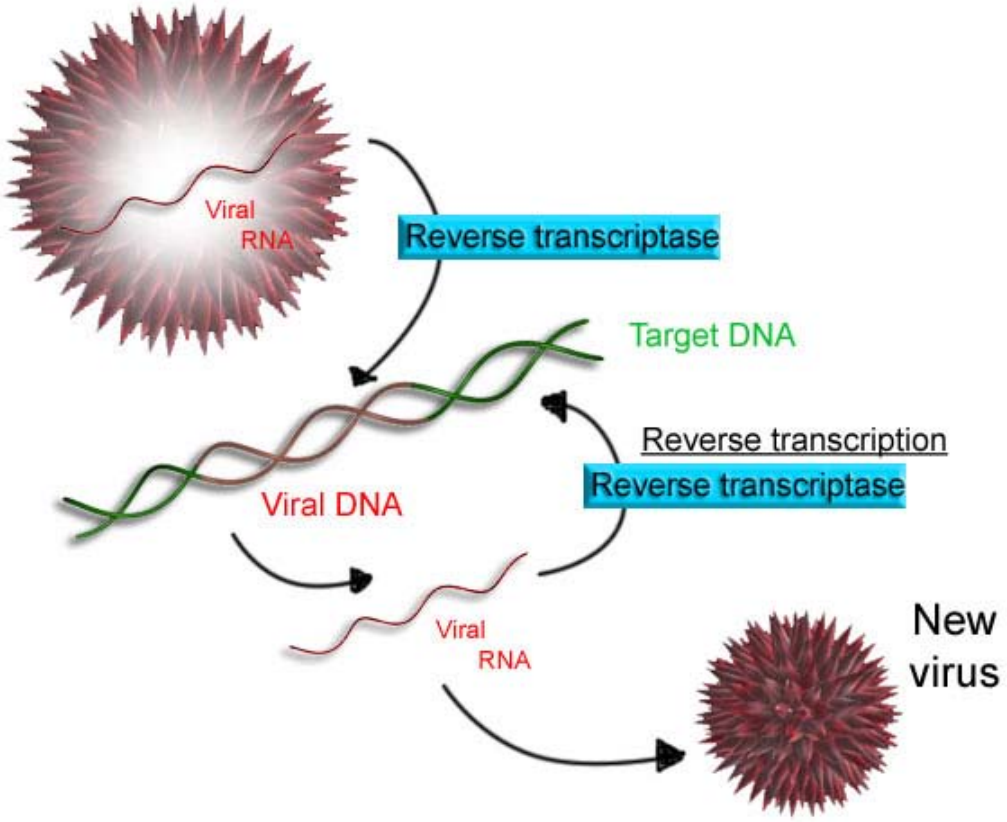
A molecule that allows the genetic material to be realized as a protein was first hypothesized by François Jacob and Jacques Monod. RNA synthesis by RNA polymerase

was established *in vitro* by several laboratories by 1965; however, the RNA synthesized by these enzymes had properties that suggested the existence of an additional factor needed to terminate transcription correctly.

In 1972, Walter Fiers became the first person to actually prove the existence of the terminating enzyme.

Roger D. Kornberg won the 2006 Nobel Prize in Chemistry "for his studies of the molecular basis of eukaryotic transcription".

Reverse transcription



Scheme of reverse transcription

Some viruses (such as HIV, the cause of AIDS), have the ability to transcribe RNA into DNA. HIV has an RNA genome that is duplicated into DNA. The resulting DNA can be merged with the DNA genome of the host cell. The main enzyme responsible for synthesis of DNA from an RNA template is called reverse transcriptase. In the case of HIV, reverse transcriptase is responsible for synthesizing a complementary DNA strand (cDNA) to the viral RNA genome. An associated enzyme, ribonuclease H, digests the RNA strand, and reverse transcriptase synthesises a complementary strand of DNA to form a double helix DNA structure. This cDNA is integrated into the host cell's genome via another enzyme (integrase) causing the host cell to generate viral proteins that

reassemble into new viral particles. Subsequent to this, the host cell undergoes programmed cell death, apoptosis.

Some eukaryotic cells contain an enzyme with reverse transcription activity called telomerase. Telomerase is a reverse transcriptase that lengthens the ends of linear chromosomes. Telomerase carries an RNA template from which it synthesizes DNA repeating sequence, or "junk" DNA. This repeated sequence of DNA is important because, every time a linear chromosome is duplicated, it is shortened in length. With "junk" DNA at the ends of chromosomes, the shortening eliminates some of the non-essential, repeated sequence rather than the protein-encoding DNA sequence farther away from the chromosome end. Telomerase is often activated in cancer cells to enable cancer cells to duplicate their genomes indefinitely without losing important protein-coding DNA sequence. Activation of telomerase could be part of the process that allows cancer cells to become *immortal*. However, the true *in vivo* significance of telomerase has still not been empirically proven.