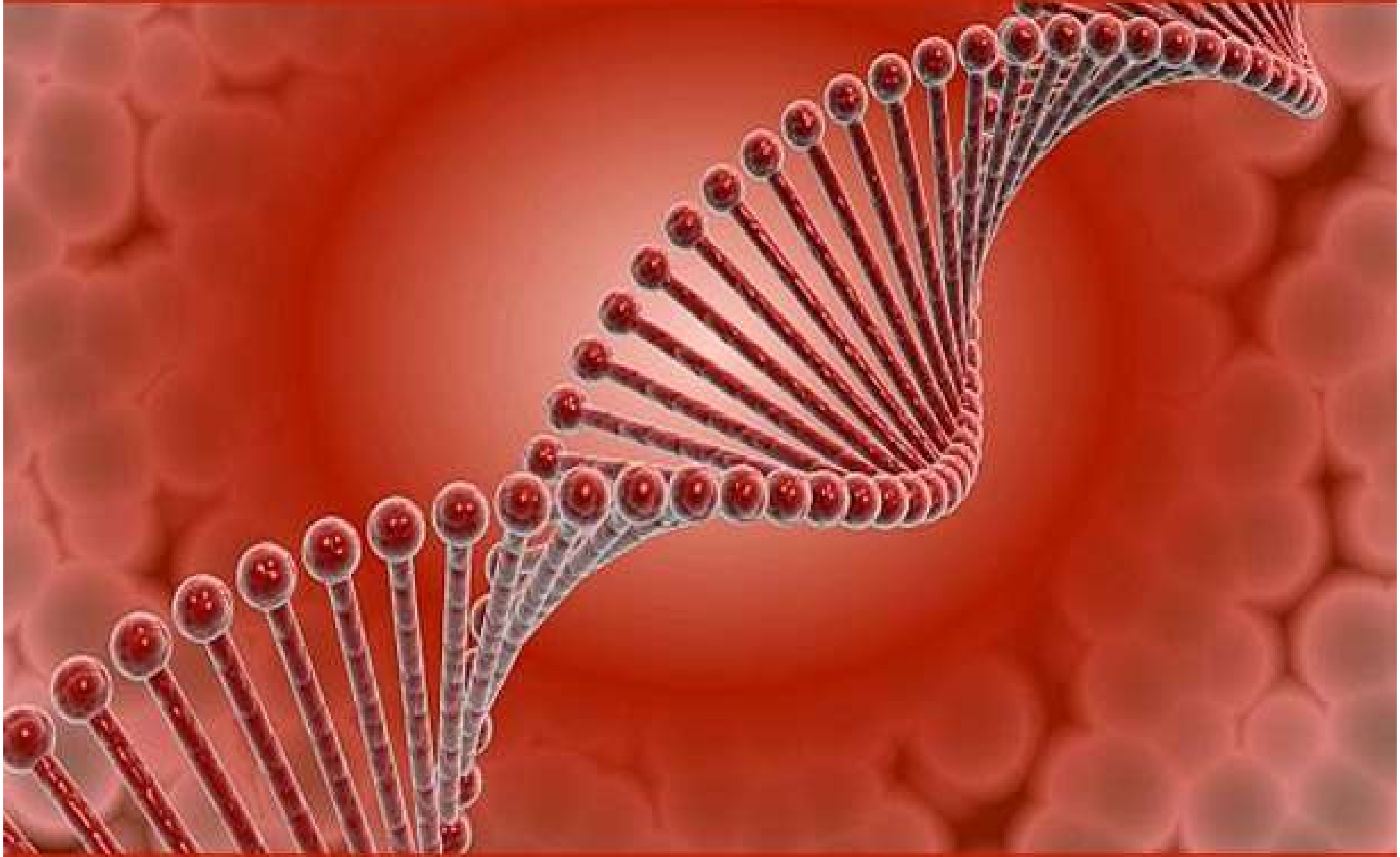
Gene Expression & DNA Sequencing



Lucy Balderas

Marquita Cornish

First Edition, 2012

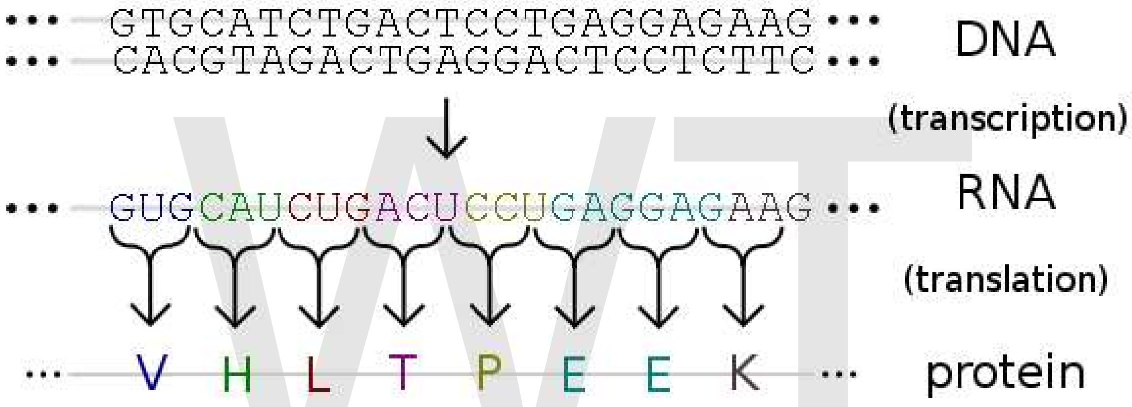
ISBN 978-81-323-0711-2

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

Table of Contents

Chapter- 1

Gene Expression


... GTGCATCTGACTCCTGAGGAGAAG ...
... CACGTAGACTGAGGACTCCTCTTC ...
DNA
(transcription)
... GUGCAUCUGACUCCUGAGGAGAAG ...
RNA
(translation)
... V H L T P E E K ...
protein

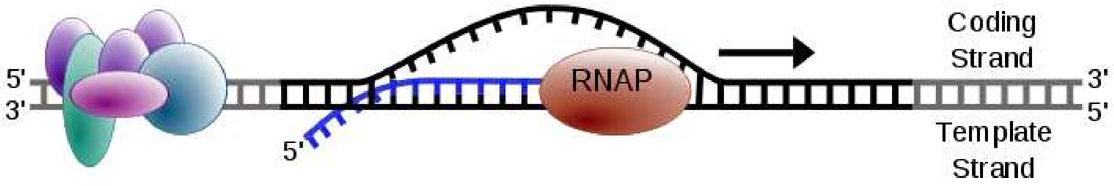
Genes are expressed by being transcribed into RNA, and this transcript may then be translated into protein.

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non-protein coding genes such as ribosomal RNA (rRNA) genes or transfer RNA (tRNA) genes, the product is a functional RNA. The process of gene expression is used by all known life - eukaryotes (including multicellular organisms), prokaryotes (bacteria and archaea) and viruses - to generate the macromolecular machinery for life. Several steps in the gene expression process may be modulated, including the transcription, RNA splicing, translation, and post-translational modification of a protein. Gene regulation gives the cell control over structure and function, and is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism. Gene regulation may also serve as a substrate for evolutionary change, since control of the timing, location, and amount of gene expression can have a profound effect on the functions (actions) of the gene in a cell or in a multicellular organism.

In genetics, gene expression is the most fundamental level at which genotype gives rise to the phenotype. The genetic code stored in DNA in form of nucleotide sequence is "interpreted" by gene expression, and the properties of the expression products give rise to the organism's phenotype.

Mechanism

Transcription



The process of transcription is carried out by RNA polymerase (RNAP), uses DNA (black) as a template and produces RNA (blue).

The gene itself is typically a long stretch of DNA which carries genetic information encoded by genetic code. Every molecule of DNA consists of two strands, each of them having 5' and 3' ends oriented in anti-parallel direction. The coding strand contains the genetic information while template strand (non-coding strand) serves as a blueprint for the production of RNA. The production of RNA copies of the DNA is called transcription, and is performed by RNA polymerase, which adds one RNA nucleotide at a time to a growing RNA strand. This RNA is complementary to the template 3' → 5' DNA strand, which is itself complementary to the coding 5' → 3' DNA strand. Therefore, the resulting 5' → 3' RNA strand is identical to the coding DNA strand with the exception that thymines (T) are replaced with uracils (U) in the RNA. A coding DNA strand reading "ATG" is transcribed as "AUG" in RNA.

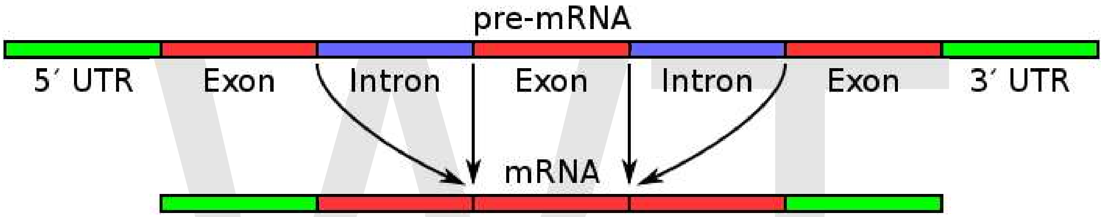
Transcription in prokaryotes is carried out by a single type of RNA polymerase, which needs DNA sequence called Pribnow box and sigma factor (σ factor) to start transcription. In eukaryotes, the transcription is done by three types of RNA polymerases, each of them needs special DNA sequence called promoter and a set of DNA-binding proteins - transcription factors to initiate the process. RNA polymerase I is responsible for transcription of rRNA genes, while RNA polymerase II transcribes all protein-coding genes but also some non-coding RNAs (e.g. snRNAs, snoRNAs or long non-coding RNAs) as well. It contains special part called C-terminal domain (CTD) that is rich of serines, which after being phosphorylated accumulate factors necessary for RNA modification and maturation. RNA polymerase III transcribes 5S rRNA and tRNA genes but also some small non-coding RNA genes (e.g. 7SK). Transcription ends on a special sequence called terminator.

RNA processing

While transcription of prokaryotic protein-coding genes creates messenger RNA (mRNA) which is ready for translation, transcription of eukaryotic genes leaves a primary transcript of RNA (pre-mRNA), which fist has to undergo series of modification to become a mature mRNA.

These include 5' *capping*, which is set of enzymatic reactions that add 7-methylguanosine (m^7G) to the 5' end of pre-mRNA and thus protect the RNA from degradation by exonucleases. The m^7G cap is then bound by cap binding complex heterodimer (CBC20/CBC80) which aids in mRNA export to cytoplasm and also protect the RNA from decapping.

Another modification is 3' *clevage and polyadenylation*. They occur if polyadenylation signal sequence (5'- AAUAAA-3') is present in pre-mRNA,which is usually between protein-coding sequence and terminator. The pre-mRNA is first cleaved and then a series of ~200 adenines (A) are added to form poly(A) tail which protects the RNA from degradation. Poly(A) tail is bound by multiple poly(A)-binding proteins (PABP) necessary for mRNA export and translation re-iniciation.



Simple illustration of exons and introns in pre-mRNA and the formation of mature mRNA by splicing. The UTRs are non-coding parts of exons at the ends of the mRNA.

Very important modification of eukaryotic pre-mRNA is *RNA splicing*. Majority of eukaryotic pre-mRNAs consist of alternating segments called exons and introns. During the process of splicing, RNA-protein catalytical complex known as spliceosome, catalyze two transesterification reactions, which remove intron and release it in form of lariat structure and then splice neighbouring exons together. In certain cases, some introns or exons can be either removed or retained in mature mRNA. This so-called alternative splicing creates series of different transcripts originating from a single gene. Because these transcripts can be potentially translated into different proteins, splicing extends the complexity of eukaryotic gene expression.

Extensive RNA processing may be an evolutionary advantage made possible by the nucleus of eukaryotes. In prokaryotes transcription and translation happen together whilst in eukaryotes the nuclear membrane separates the two processes giving time for RNA processing to occur.

non-coding RNA maturation

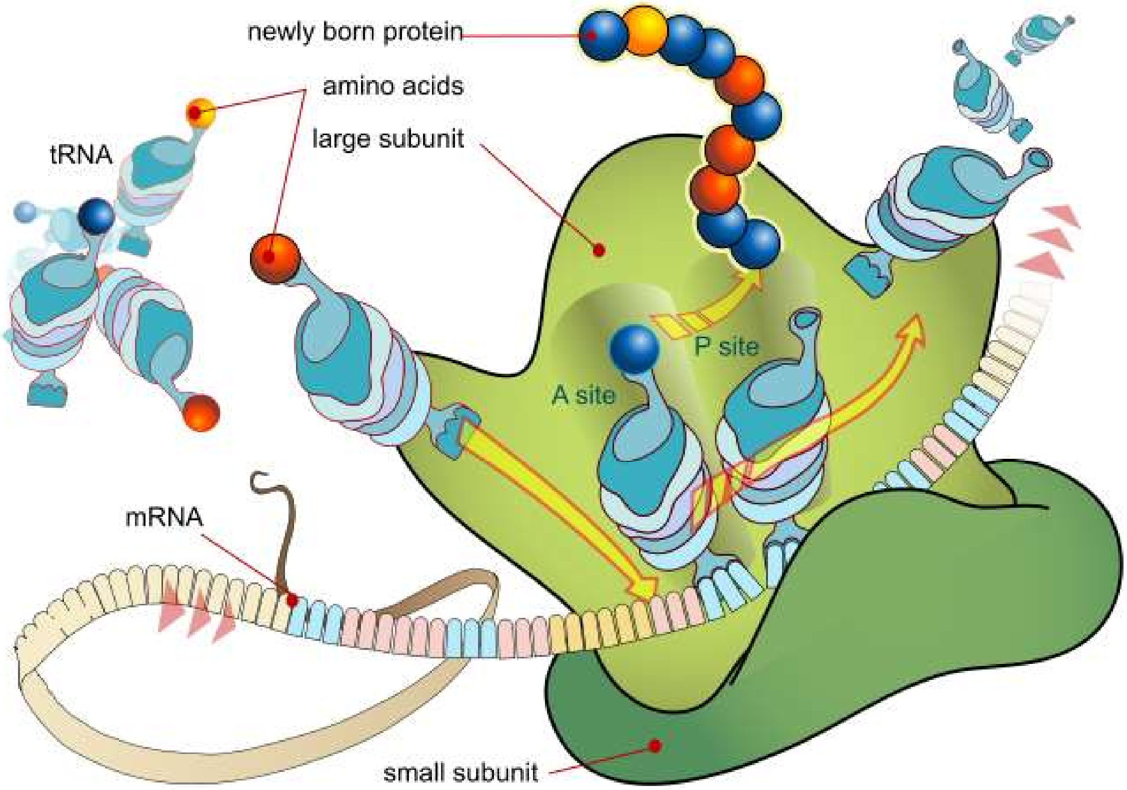
In most organisms non-coding genes (ncRNA) are transcribed as precursors which undergo further processing. In the case of ribosomal RNAs (rRNA), they are often transcribed as a pre-rRNA which contains one or more rRNAs, the pre-rRNA is cleaved and modified (2′-O-methylation and pseudouridine formation) at a specific sites by approximately 150 different small nucleolus-restricted RNA species, called snoRNAs. SnoRNAs associate with proteins, forming snoRNPs. While snoRNA part basepair with the target RNA and thus position the modification to precise site, the protein part performs the catalytical reaction. In eukaryotes, in particular a snoRNP, called RNase MRP cleaves the 45S pre-rRNA into the 28S, 5.8S, and 18S rRNAs. The rRNA and RNA processing factors form large aggregates called the nucleolus.

In the case of transfer RNA (tRNA), for example, the 5' sequence is removed by RNase P, whereas the 3' end is removed by the tRNase Z enzyme and the non-templated 3' CCA tail is added by a nucleotidyl transferase. In the case of micro RNA (miRNA), miRNAs are first transcribed as primary transcripts or pri-miRNA with a cap and poly-A tail and processed to short, 70-nucleotide stem-loop structures known as pre-miRNA in the cell nucleus by the enzymes Drosha and Pasha. After being exported, it is then processed to mature miRNAs in the cytoplasm by interaction with the endonuclease Dicer, which also initiates the formation of the RNA-induced silencing complex (RISC), composed of the Argonaute protein. Even snRNAs and snoRNAs themselves undergo series of modification before they become part of functional RNP complex. This is done either in the nucleoplasm or in the specialized compartments called Cajal bodies. Their bases are methylated or pseudouridinilated by a group of small Cajal body-specific RNAs (scaRNAs) which are structurally similar to snoRNAs.

RNA export

In eukaryotes most mature RNA must be exported to the cytoplasm from the nucleus. While some RNAs function in the nucleus, many RNAs are transported through the nuclear pores and into the cytosol. Notably this includes all RNA types involved in protein synthesis. In some cases RNAs are additionally transported to a specific part of the cytoplasm, such as a synapse; they are then towed by motor proteins that bind through linker proteins to specific sequences (called "zipcodes") on the RNA.

Translation



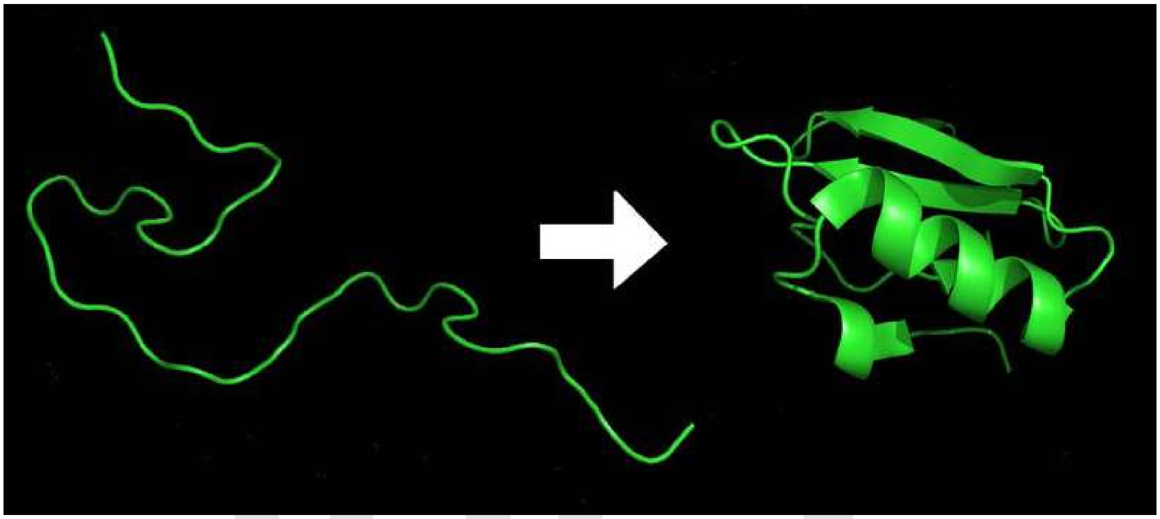
During the translation, tRNA charged with amino acid enters the ribosome and aligns with the correct mRNA triplet. Ribosome then adds amino acid to growing protein chain.

For some RNA (non-coding RNA) the mature RNA is the final gene product. In the case of messenger RNA (mRNA) the RNA is an information carrier coding for the synthesis of one or more proteins. mRNA carrying a single protein sequence (common in eukaryotes) is monocistronic whilst mRNA carrying multiple protein sequences (common in prokaryotes) is known as polycistronic.

Every mRNA consists of three parts - 5' untranslated region (5'UTR), protein-coding region or open reading frame (ORF) and 3' untranslated region (3'UTR). Coding region carries information for protein synthesis encoded by genetic code into form of triplets. Each triplet of nucleotides of the coding region is called codon and corresponds to a binding site complementary to an anticodon triplet in transfer RNA. Transfer RNAs with the same anticodon sequence always carry identical type of amino acid. Amino acids are then chained together by the ribosome according to order of triplets in the coding region. The ribosome helps transfer RNA to bind to messenger RNA and takes the amino acid from each transfer RNA and makes a structure-less protein out of it.

In prokaryotes translation generally occurs at the point of transcription (co-transcriptionally), often using a messenger RNA which is still in the process of being created. In eukaryotes translation can occur in a variety of regions of the cell depending on where the protein being written is supposed to be. Major locations are the cytoplasm for soluble cytoplasmic proteins and the membrane of endoplasmic reticulum for proteins which are for export from the cell or insertion into a cell membrane. Proteins which are supposed to be expressed at the endoplasmic reticulum are recognised part-way through the translation process. This is governed by the signal recognition particle - a protein which binds to the ribosome and directs it to the endoplasmic reticulum when it finds a signal sequence on the growing (nascent) amino acid chain.

Folding



Protein before (left) and after (right) folding

The polypeptide folds into its characteristic and functional three-dimensional structure from random coil. Each protein exists as an unfolded polypeptide or random coil when translated from a sequence of mRNA to a linear chain of amino acids. This polypeptide lacks any developed three-dimensional structure (the left hand side of the neighboring figure). Amino acids interact with each other to produce a well-defined three-dimensional structure, the folded protein (the right hand side of the figure), known as the native state. The resulting three-dimensional structure is determined by the amino acid sequence (Anfinsen's dogma).

The correct three-dimensional structure is essential to function, although some parts of functional proteins may remain unfolded Failure to fold into the intended shape usually produces inactive proteins with different properties including toxic prions. Several neurodegenerative and other diseases are believed to result from the accumulation of *misfolded* (incorrectly folded) proteins. Many allergies are caused by the folding of the proteins, for the immune system does not produce antibodies for certain protein structures.

Enzymes called chaperones assist the newly formed protein to attain (fold into) the 3-dimensional structure it needs to function. Similarly, RNA chaperones help RNAs attain their functional shapes. Assisting protein folding is one of the main roles of the endoplasmic reticulum in eukaryotes.

Protein transport

Many proteins are destined for other parts of the cell than the cytosol and a wide range of signalling sequences are used to direct proteins to where they are supposed to be. In prokaryotes this is normally a simple process due to limited compartmentalisation of the cell. However in eukaryotes there is a great variety of different targeting processes to ensure the protein arrives at the correct organelle.

Not all proteins remain within the cell and many are exported, for example digestive enzymes, hormones and extracellular matrix proteins. In eukaryotes the export pathway is well developed and the main mechanism for the export of these proteins is translocation to the endoplasmic reticulum, followed by transport via the Golgi apparatus.

Regulation of gene expression

The patchy colours of a tortoiseshell cat are the result of different levels of expression of pigmentation genes in different areas of the skin.

Regulation of gene expression refers to the control of the amount and timing of appearance of the functional product of a gene. Control of expression is vital to allow a cell to produce the gene products it needs when it needs them; in turn this gives cells the flexibility to adapt to a variable environment, external signals, damage to the cell, etc. Some simple examples of where gene expression is important are:

- Control of Insulin expression so it gives a signal for blood glucose regulation

- X chromosome inactivation in female mammals to prevent an "overdose" of the genes it contains.
- Cyclin expression levels control progression through the eukaryotic cell cycle

More generally gene regulation gives the cell control over all structure and function, and is the basis for cellular differentiation, morphogenesis and the versatility and adaptability of any organism.

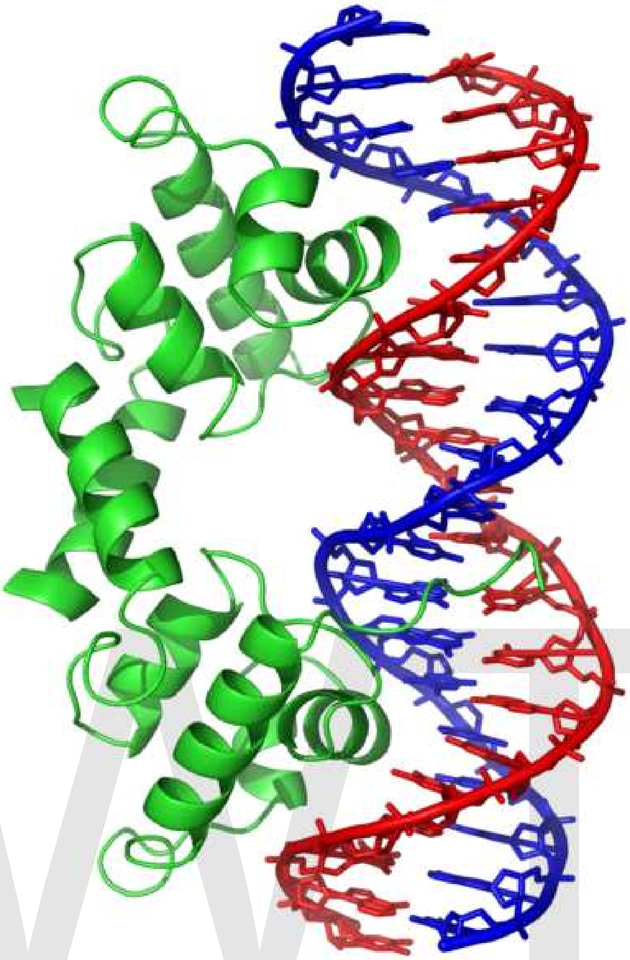
Any step of gene expression may be modulated, from the DNA-RNA transcription step to post-translational modification of a protein. The stability of the final gene product, whether it is RNA or protein, also contributes to the expression level of the gene - an unstable product results in a low expression level. In general gene expression is regulated through changes in the number and type of interactions between molecules that collectively influence transcription of DNA and translation of RNA.

Numerous terms are used to describe types of genes depending on how they are regulated, these include:

- A *constitutive gene* is a gene that is transcribed continually compared to a facultative gene which is only transcribed when needed.
- A *housekeeping gene* is typically a constitutive gene that is transcribed at a relatively constant level. The housekeeping gene's products are typically needed for maintenance of the cell. It is generally assumed that their expression is unaffected by experimental conditions. Examples include actin, GAPDH and ubiquitin.
- A *facultative gene* is a gene which is only transcribed when needed compared to a constitutive gene.
- An *inducible gene* is a gene whose expression is either responsive to environmental change or dependent on the position in the cell cycle.

Transcriptional regulation

Regulation of transcription can be broken down into three main routes of influence; genetic (direct interaction of a control factor with the gene), modulation (interaction of a control factor with the transcription machinery) and epigenetic (non-sequence changes in DNA structure which influence transcription).



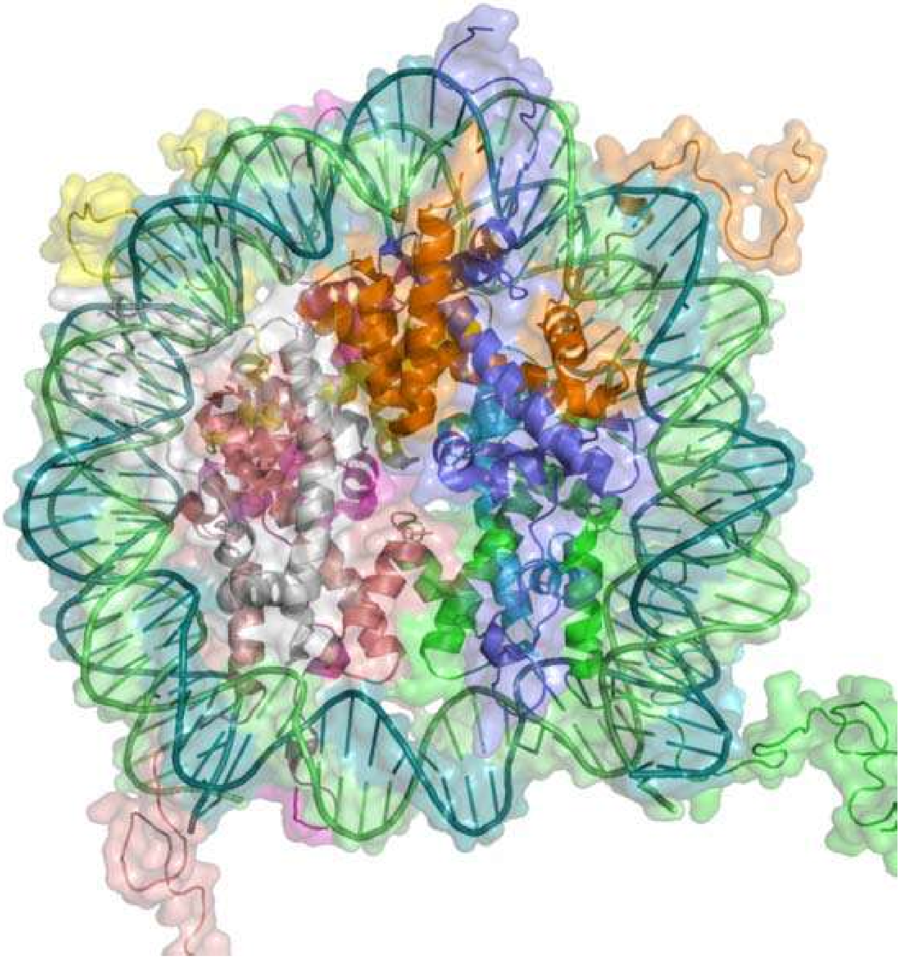
The lambda repressor transcription factor (green) binds as a dimer to major groove of DNA target (red and blue) and disables initiation of transcription.

Direct interaction with DNA is the simplest and the most direct method by which a protein can change transcription levels. Genes often have several protein binding sites around the coding region with the specific function of regulating transcription. There are many classes of regulatory DNA binding sites known as enhancers, insulators, repressors and silencers. The mechanisms for regulating transcription are very varied, from blocking key binding sites on the DNA for RNA polymerase to acting as an activator and promoting transcription by assisting RNA polymerase binding.

The activity of transcription factors is further modulated by intracellular signals causing protein post-translational modification including phosphorylated, acetylated, or glycosylated. These changes influence a transcription factor's ability to bind, directly or indirectly, to promoter DNA, to recruit RNA polymerase, or to favor elongation of a newly synthetized RNA molecule.

The nuclear membrane in eukaryotes allows further regulation of transcription factors by the duration of their presence in the nucleus which is regulated by reversible changes in their structure and by binding of other proteins. Environmental stimuli or endocrine signals may cause modification of regulatory proteins eliciting cascades of intracellular signals, which result in regulation of gene expression.

More recently it has become apparent that there is a huge influence of non-DNA-sequence specific effects on translation. These effects are referred to as epigenetic and involve the higher order structure of DNA, non-sequence specific DNA binding proteins and chemical modification of DNA. In general epigenetic effects alter the accessibility of DNA to proteins and so modulate transcription.



In eukaryotes, DNA is organized in form of nucleosomes. Note how the DNA (blue and green) is tightly wrapped around the protein core made of histone octamer (ribbon coils), restricting access to the DNA.

DNA methylation is a widespread mechanism for epigenetic influence on gene expression and is seen in bacteria and eukaryotes and has roles in heritable transcription silencing and transcription regulation. In eukaryotes the structure of chromatin, controlled by the histone code, regulates access to DNA with significant impacts on the expression of genes in euchromatin and heterochromatin areas.

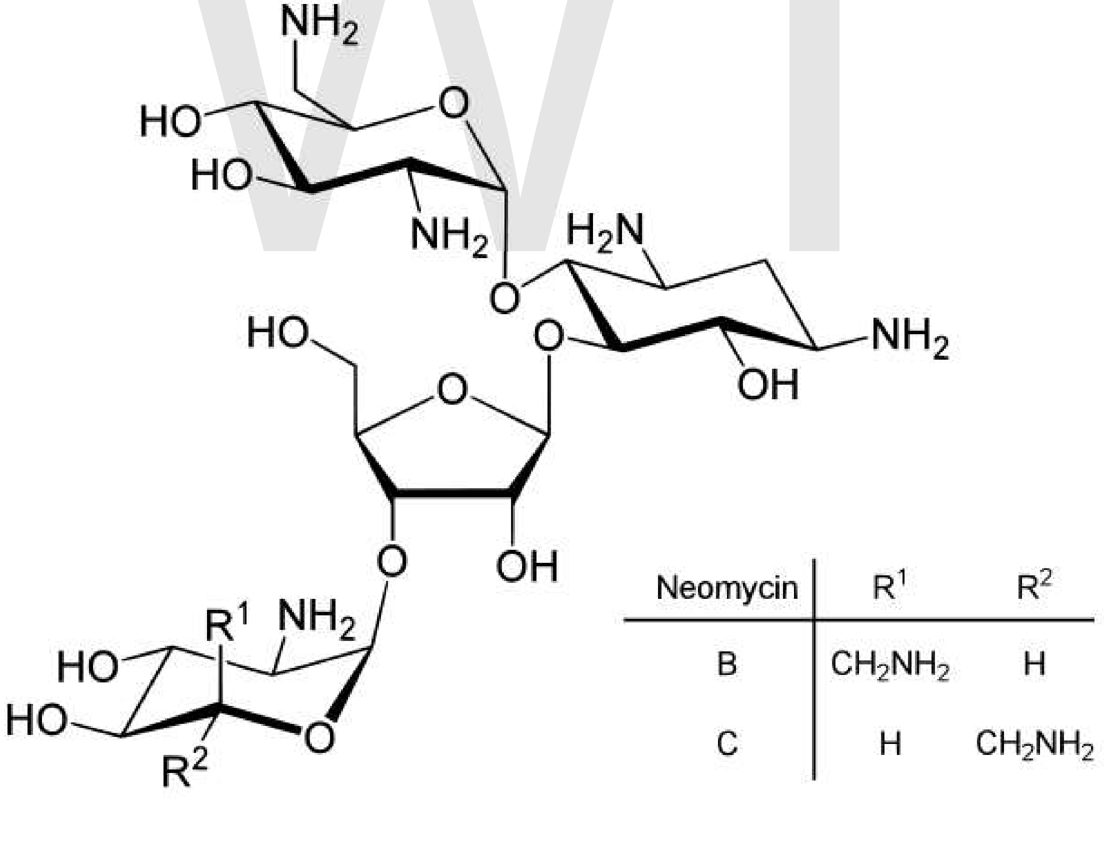
Post-transcriptional regulation

In eukaryotes, where export of RNA is required before translation is possible, nuclear export is thought to provide additional control over gene expression. All transport in and out of the nucleus is via the nuclear pore and transport is controlled by a wide range of importin and exportin proteins.

Expression of a gene coding for a protein is only possible if the messenger RNA carrying the code survives long enough to be translated. In a typical cell an RNA molecule is only stable if specifically protected from degradation. RNA degradation has particular importance in regulation of expression in eukaryotic cells where mRNA has to travel significant distances before being translated. In eukaryotes RNA is stabilised by certain post-transcriptional modifications, particularly the 5' cap and poly-adenylated tail.

Intentional degradation of mRNA is used not just as a defence mechanism from foreign RNA (normally from viruses) but also as a route of mRNA *destabilisation*. If an mRNA molecule has a complementary sequence to a small interfering RNA then it is targeted for destruction via the RNA interference pathway.

Translational regulation



Neomycin	R^1	R^2
B	CH_2NH_2	H
C	H	CH_2NH_2

Neomycin is an example of a small molecule which reduces expression of all protein genes inevitably leading to cell death, thus acts as an antibiotic.

Direct regulation of translation is less prevalent than control of transcription or mRNA stability but is occasionally used. Inhibition of protein translation is a major target for toxins and antibiotics in order to kill a cell by overriding its normal gene expression control. Protein synthesis inhibitors include the antibiotic neomycin and the toxin ricin.

Protein degradation

Once protein synthesis is complete the level of expression of that protein can be reduced by protein degradation. There are major protein degradation pathways in all prokaryotes and eukaryotes of which the proteasome is a common component. An unneeded or damaged protein is often labelled for degradation by addition of ubiquitin.

Measurement

Measuring gene expression is an important part of many life sciences - the ability to quantify the level at which a particular gene is expressed within a cell, tissue or organism can give a huge amount of information. For example measuring gene expression can:

- Identify viral infection of a cell (viral protein expression)
- Determine an individual's susceptibility to cancer (oncogene expression)
- Find if a bacterium is resistant to penicillin (beta-lactamase expression)

Similarly the analysis of the location of expression protein is a powerful tool and this can be done on an organism or cellular scale. Investigation of localisation is particularly important for study of development in multicellular organisms and as an indicator of protein function in single cells. Ideally measurement of expression is done by detecting the final gene product (for many genes this is the protein) however it is often easier to detect one of the precursors, typically mRNA, and infer gene expression level.

mRNA quantification

Levels of mRNA can be quantitatively measured by Northern blotting which gives size and sequence information about the mRNA molecules. A sample of RNA is separated on an agarose gel and hybridized to a radio-labeled RNA probe that is complementary to the target sequence. The radio-labeled RNA is then detected by an autoradiograph. The main problems with Northern blotting stem from the use of radioactive reagents (which make the procedure time consuming and potentially dangerous) and lower quality quantification than more modern methods (due to the fact that quantification is done by measuring band strength in an image of a gel). Northern blotting is, however, still widely used as the additional mRNA size information allows the discrimination of alternately spliced transcripts.

A more modern low-throughput approach for measuring mRNA abundance is reverse transcription quantitative polymerase chain reaction (RT-PCR followed with qPCR). RT-PCR first generates a DNA template from the mRNA by reverse transcription, which is called cDNA. This cDNA template is then used for qPCR where the change in fluorescence of a probe changes as the DNA amplification process progresses. With a carefully constructed standard curve qPCR can produce an absolute measurement such as number of copies of mRNA, typically in units of copies per nanolitre of homogenized tissue or copies per cell. qPCR is very sensitive (detection of a single mRNA molecule is possible), but can be expensive due to the fluorescent probes required.

Northern blots and RT-qPCR are good for detecting whether a single gene is being expressed, but it quickly becomes impractical if many genes within the sample are being studied. Using DNA microarrays transcript levels for many genes at once (expression profiling) can be measured. Recent advances in microarray technology allow for the quantification, on a single array, of transcript levels for every known gene in several organism's genomes, including humans.

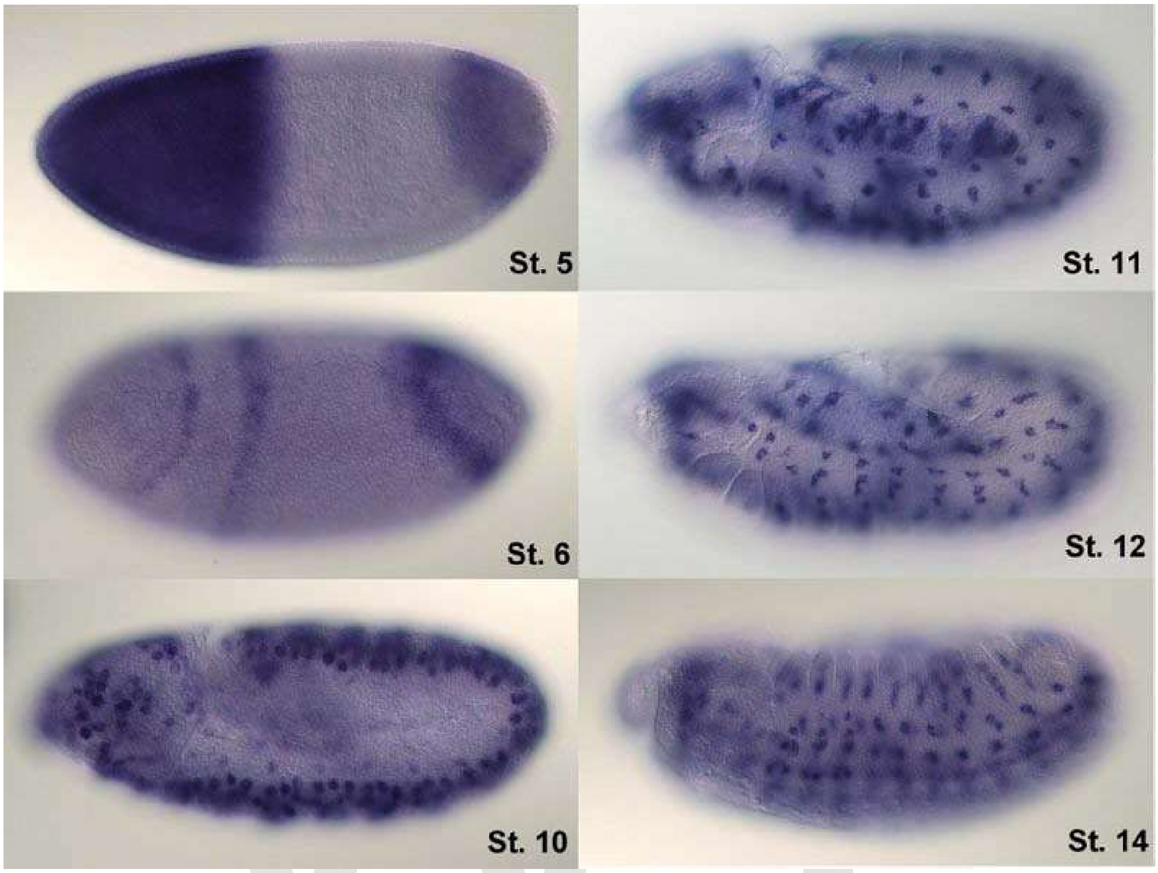
Alternatively "tag based" technologies like Serial analysis of gene expression (SAGE), which can provide a relative measure of the cellular concentration of different mRNAs, can be used. The great advantage of tag-based methods is the "open architecture", allowing for the exact measurement of any transcript, with a known or unknown sequence.

Protein quantification

For genes encoding proteins the expression level can be directly assessed by a number of means with some clear analogies to the techniques for mRNA quantification.

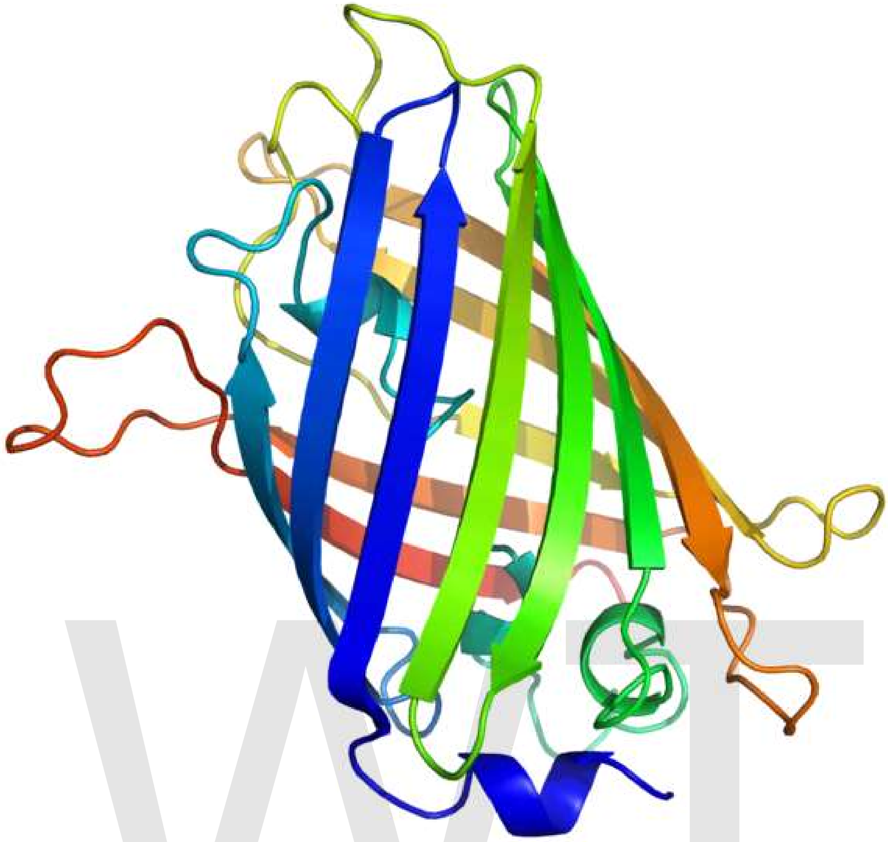
The most commonly used method is to perform a Western blot against the protein of interest - this gives information on the size of the protein in addition to its identity. A sample (often cellular lysate) is separated on a polyacrylamide gel, transferred to a membrane and then probed with an antibody to the protein of interest. The antibody can either be conjugated to a fluorophore or to horseradish peroxidase for imaging and/or quantification. The gel-based nature of this assay makes quantification less accurate but it has the advantage of being able to identify later modifications to the protein, for example proteolysis or ubiquitination, from changes in size.

Localisation



In situ-hybridization of Drosophila embryos at different developmental stages for the mRNA responsible for the expression of hunchback. High intensity of blue color marks places with high hunchback mRNA quantity.

Analysis of expression is not limited to only quantification; localisation can also be determined. mRNA can be detected with a suitably labelled complementary mRNA strand and protein can be detected via labelled antibodies. The probed sample is then observed by microscopy to identify where the mRNA or protein is.



The three-dimensional structure of green fluorescent protein. The residues in the centre of the "barrel" are responsible for production of green light after exposing to higher energetic blue light.

By replacing the gene with a new version fused a green fluorescent protein (or similar) marker expression may be directly quantified in live cells. This is done by imaging using a fluorescence microscope. It is very difficult to clone a GFP-fused protein into its native location in the genome without affecting expression levels so this method often cannot be used to measure endogenous gene expression. It is, however, widely used to measure the expression of a gene artificially introduced into the cell, for example via an expression vector. It is important to note that by fusing a target protein to a fluorescent reporter the protein's behavior, including its cellular localization and expression level, can be significantly changed.

The enzyme-linked immunosorbent assay works by using antibodies immobilised on a microtiter plate to capture proteins of interest from samples added to the well. Using a detection antibody conjugated to an enzyme or fluorophore the quantity of bound protein can be accurately measured by fluorometric or colourimetric detection. The detection process is very similar to that of a Western blot, but by avoiding the gel steps more accurate quantification can be achieved.

Expression system

An expression system is a system specifically designed for the production of a gene product of choice. This is normally a protein although may also be RNA, such as tRNA or a ribozyme. An expression system consists of a gene, normally encoded by DNA, and the molecular machinery required to transcribe the DNA into mRNA and translate the mRNA into protein using the reagents provided. In the broadest sense this includes every living cell but the term is more normally used to refer to expression as a laboratory tool. An expression system is therefore often artificial in some manner. Expression systems are, however, a fundamentally natural process. Viruses are an excellent example where they replicate by using the host cell as an expression system for the viral proteins and genome.

In nature

In addition to these biological tools, certain naturally observed configurations of DNA (genes, promoters, enhancers, repressors) and the associated machinery itself are referred to as an expression system. This term is normally used in the case where a gene or set of genes is switched on under well defined conditions. For example the simple repressor switch expression system in Lambda phage and the lac operator system in bacteria. Several natural expression systems are directly used or modified and used for artificial expression systems such as the Tet-on and Tet-off expression system.

Gene networks

Genes have sometimes been regarded as nodes in a network, with inputs being proteins such as transcription factors, and outputs being the level of gene expression. The node itself performs a function, and the operation of these functions have been interpreted as performing a kind of information processing within cell and determine cellular behavior.

Gene networks can also be constructed without formulating an explicit causal model. This is often the case when assembling networks from large expression data sets. Covariation and correlation of expression is computed across a large sample of cases and measurements (often transcriptome or proteome data). The source of variation can be either experimental or natural (observational). There are several ways to construct gene expression networks, but one common approach is to compute a matrix of all pair-wise correlations of expression across conditions, time points, or individuals and convert the matrix (after thresholding at some cut-off value) into a graphical representation in which nodes represent genes, transcripts, or proteins and edges connecting these nodes represent the strength of association.

Techniques and tools

The following experimental techniques are used to measure gene expression and are listed in roughly chronological order, starting with the older, more established technologies. They are divided into two groups based on their degree of multiplexity.

- Low-to-mid-plex techniques:
    - Reporter gene
    - Northern blot
    - Western blot
    - Fluorescent in situ hybridization
    - Reverse transcription PCR

- Higher-plex techniques:
    - SAGE
    - DNA microarray
    - Tiling array
    - RNA-Seq

Chapter- 2

Transcription

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

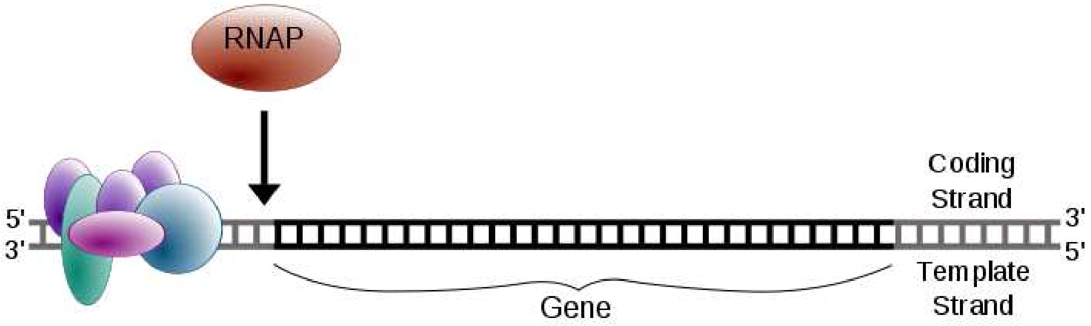
Pre-initiation

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.

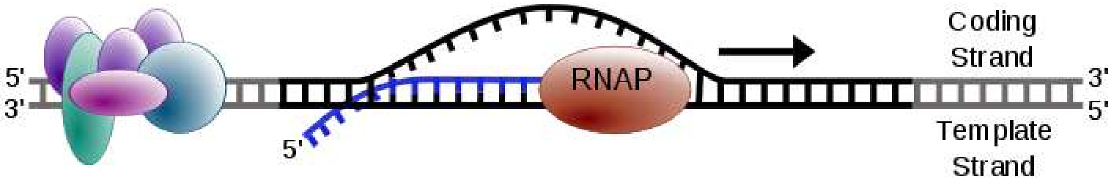
Transcription initiation is more complex in eukaryotes. Eukaryotic RNA polymerase does not directly recognize the core promoter sequences. Instead, a collection of proteins called transcription factors mediate the binding of RNA polymerase and the initiation of transcription. Only after certain transcription 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



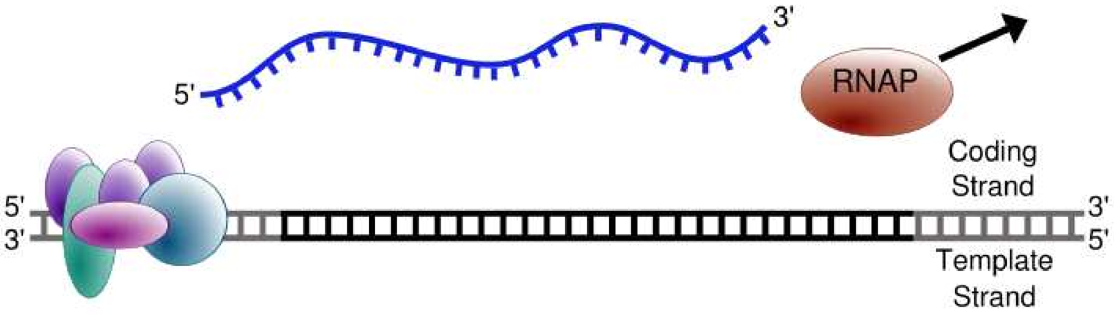
Simple diagram of transcription elongation

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

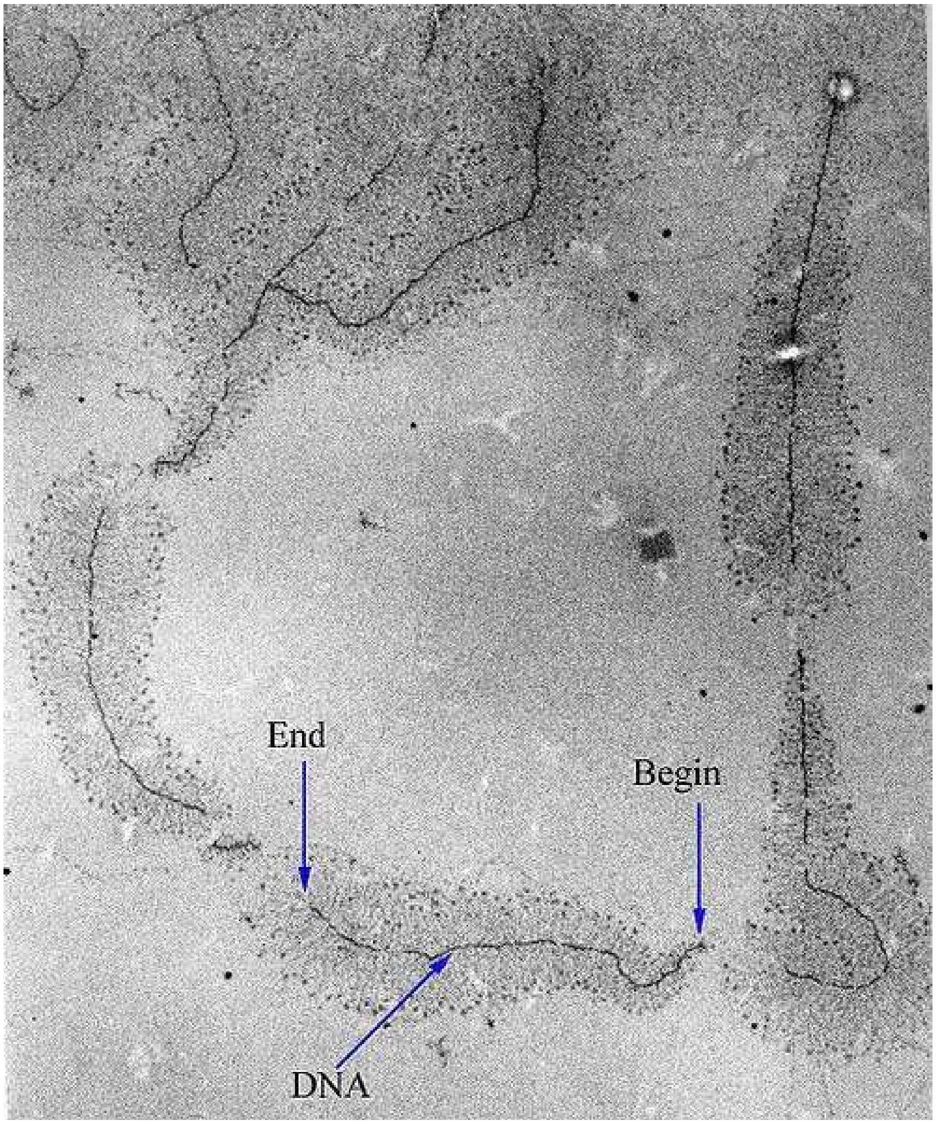


Simple diagram of transcription termination

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.

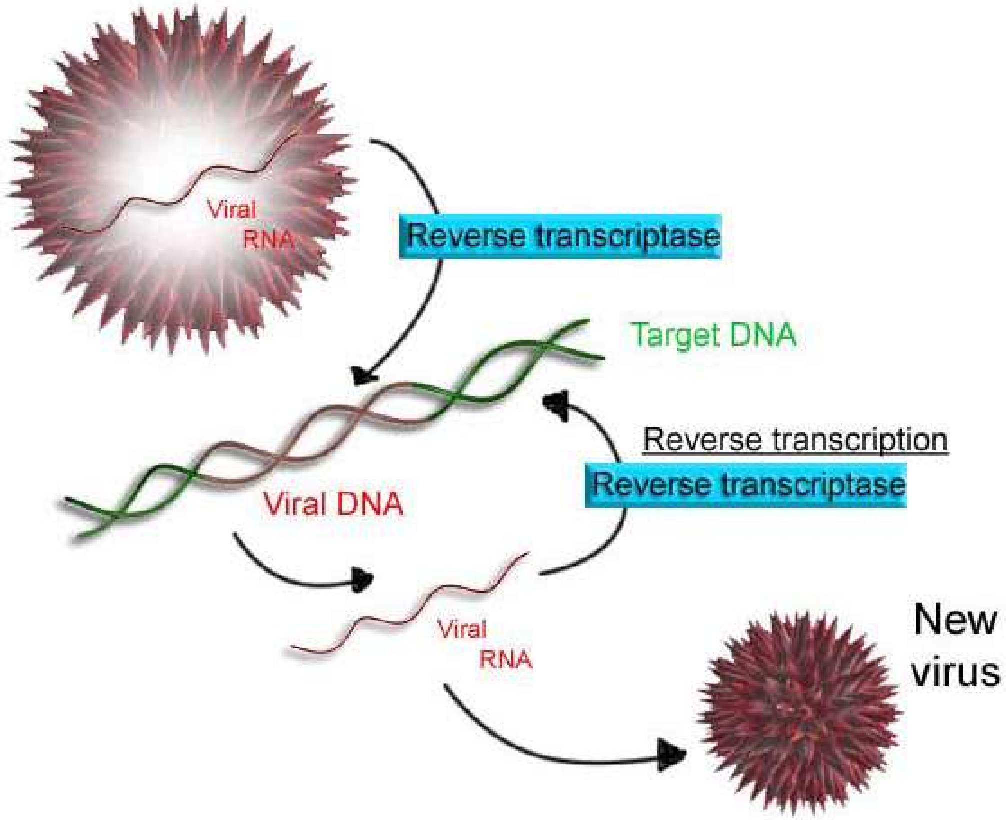
History

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.

Chapter- 3

Post-Transcriptional Modification and Transfer RNA

Post-transcriptional modification

Post-transcriptional modification is a process in cell biology by which, in eukaryotic cells, primary transcript RNA is converted into mature RNA. A notable example is the conversion of precursor messenger RNA into mature messenger RNA (mRNA), which includes splicing and occurs prior to protein synthesis. This process is vital for the correct translation of the genomes of eukaryotes as the human primary RNA transcript that is produced as a result of transcription contains both exons, which are coding sections of the primary RNA transcript and introns, which are the non coding sections of the primary RNA transcript.

mRNA processing

The pre-mRNA molecule undergoes three main modifications. These modifications are 5' capping, 3' polyadenylation, and RNA splicing, which occur in the cell nucleus before the RNA is translated.

5' Processing

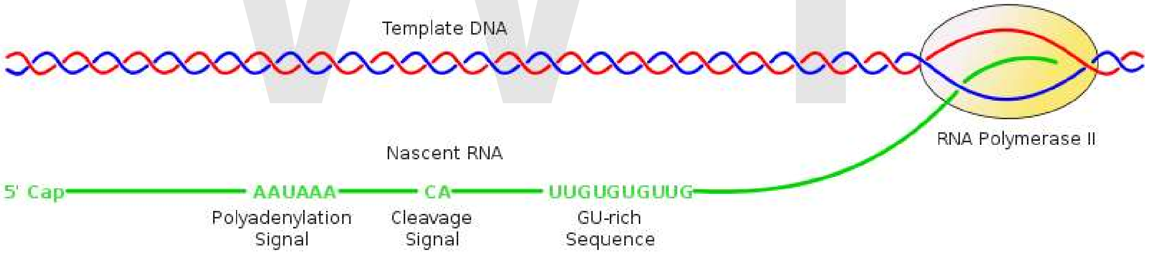
Capping

Capping of the pre-mRNA involves the addition of **7-methylguanosine (m^{7}G)** to the 5' end. To achieve this, the terminal 5' phosphate requires removal, which is done with the aid of a **phosphatase enzyme**. The enzyme **guanosyl transferase** then catalyses the reaction, which produces the diphosphate 5' end. The diphosphate 5' prime end then attacks the α phosphorus atom of a GTP molecule in order to add the guanine residue in a 5'5' triphosphate link. The enzyme **(guanine-*N*^{7}-)-methyltransferase** ("cap MTase") transfers a methyl group from S-adenosyl methionine to the guanine ring. This type of cap, with just the (m^{7}G) in position is called a **cap 0 structure**. The ribose of the adjacent nucleotide may also be methylated to give a **cap 1**. Methylation of nucleotides downstream of the RNA molecule produce **cap 2**, **cap 3** structures and so on. In these cases the methyl groups are added to the 2' OH groups of the ribose sugar. The cap protects the 5' end of the primary RNA transcript from attack by ribonucleases that have specificity to the 3'5' phosphodiester bonds.

3' Processing

Cleavage and Polyadenylation

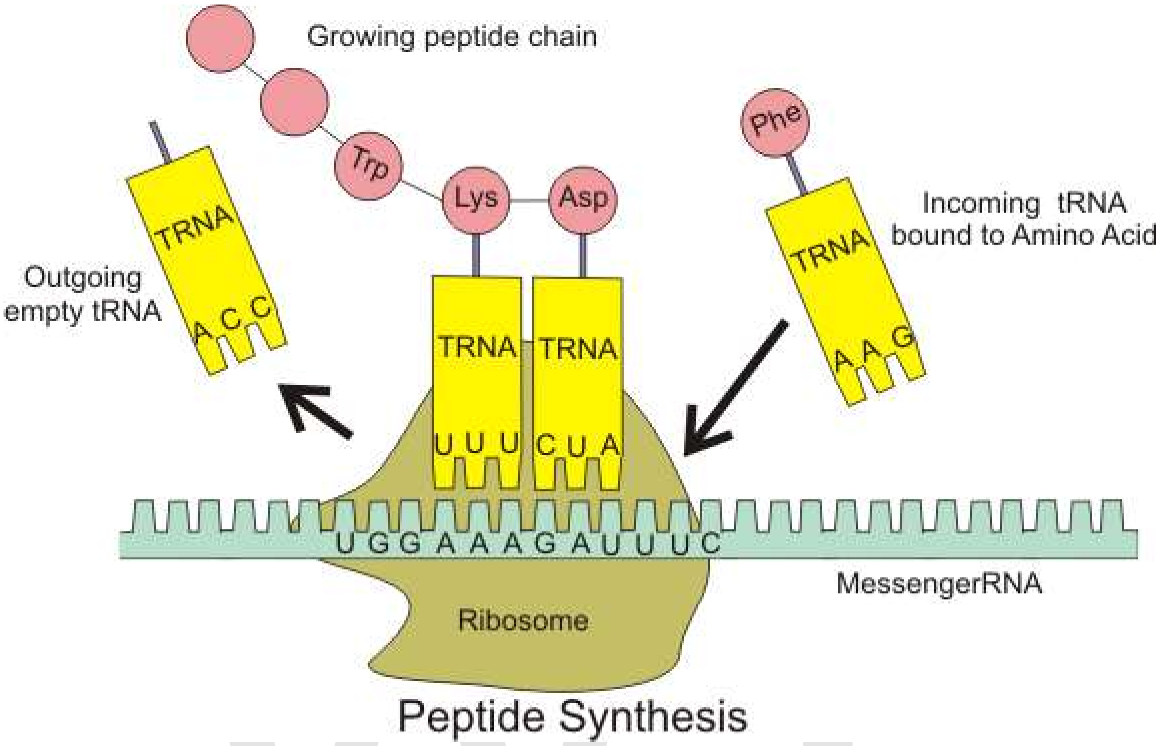
The pre-mRNA processing at the 3' end of the RNA molecule involves cleavage of its 3' end and then the addition of about 200 adenine residues to form a poly(A) tail. The cleavage and adenylation reactions occur if a polyadenylation signal sequence (5'-AAUAAA-3') is located near the 3' end of the pre-mRNA molecule, which is followed by another sequence, which is usually **(5'-CA-3')**. The second signal is the site of cleavage. A **GU-rich sequence** is also usually present further downstream on the pre-mRNA molecule. After the synthesis of the sequence elements, two multisubunit proteins called cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CStF) are transferred from RNA Polymerase II to the RNA molecule. The two factors bind to the sequence elements. A protein complex forms that contains additional cleavage factors and the enzyme Polyadenylate Polymerase (PAP). This complex cleaves the RNA between the polyadenylation sequence and the GU-rich sequence at the cleavage site marked by the (5'-CA-3') sequences. Poly(A) polymerase then adds about 200 adenine units to the new 3' end of the RNA molecule using ATP as a precursor. As the poly(A) tails is synthesised, it binds multiple copies of poly(A) binding protein, which protects the 3'end from ribonuclease digestion.



Splicing

RNA splicing is the process by which introns, regions of RNA that do not code for protein, are removed from the pre-mRNA and the remaining exons connected to re-form a single continuous molecule. Although most RNA splicing occurs after the complete synthesis and end-capping of the pre-mRNA, transcripts with many exons can be spliced co-transcriptionally. The splicing reaction is catalyzed by a large protein complex called the spliceosome assembled from proteins and small nuclear RNA molecules that recognize splice sites in the pre-mRNA sequence. Many pre-mRNAs, including those encoding antibodies, can be spliced in multiple ways to produce different mature mRNAs that encode different protein sequences. This process is known as alternative splicing, and allows production of a large variety of proteins from a limited amount of DNA.

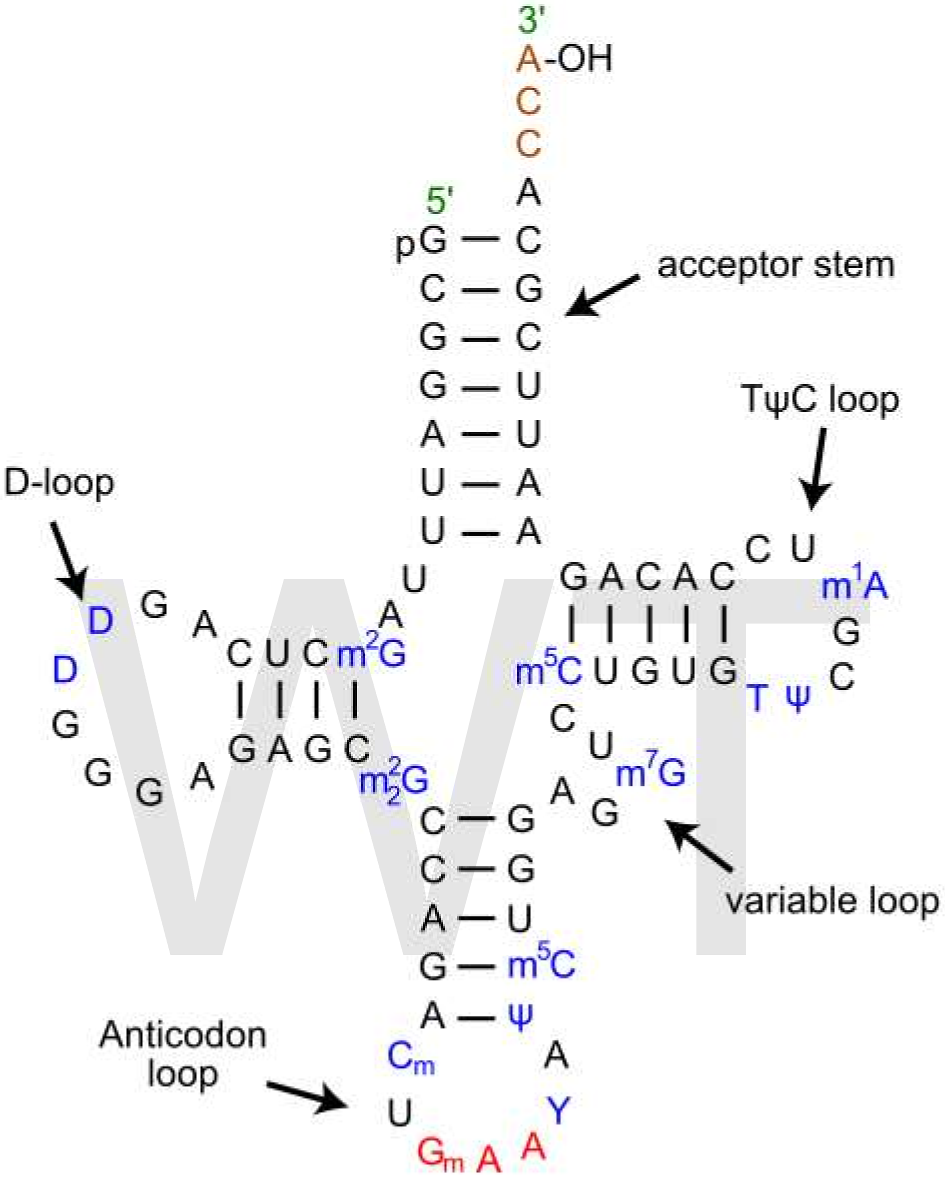
Transfer RNA



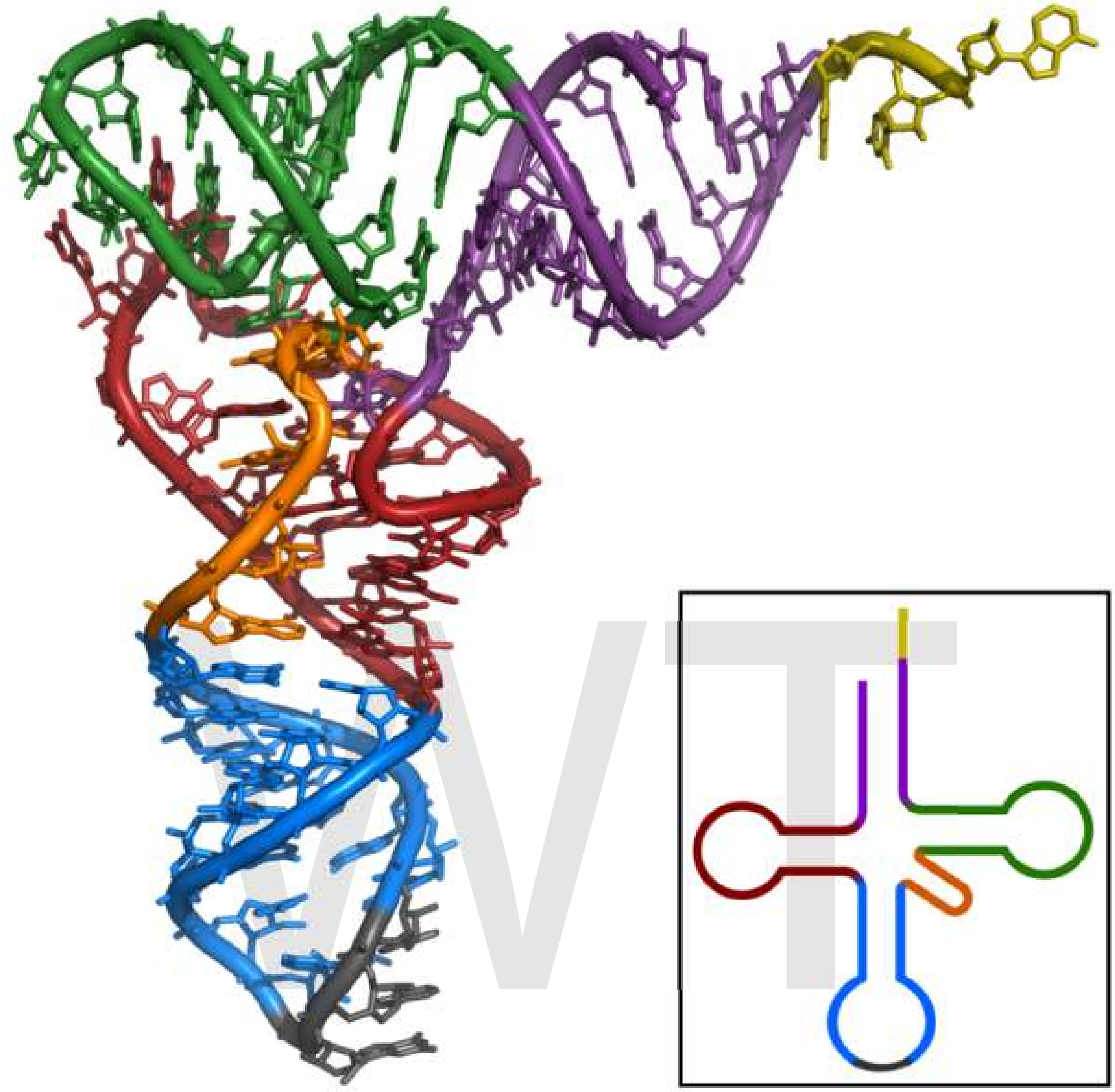
The interaction of tRNA and mRNA in protein synthesis

Transfer RNA (**tRNA**) is a small RNA molecule (usually about 73-95 nucleotides) that transfers a specific active amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation. It has a 3' terminal site for amino acid attachment. This covalent linkage is catalyzed by an aminoacyl tRNA synthetase. It also contains a three base region called the anticodon that can base pair to the corresponding three base codon region on mRNA. Each type of tRNA molecule can be attached to only one type of amino acid, but because the genetic code contains multiple codons that specify the same amino acid, tRNA molecules bearing different anticodons may also carry the same amino acid.

Structure



Secondary *cloverleaf structure* of tRNA^{Phe} from yeast



Tertiary structure of tRNA. *CCA tail* in orange, *Acceptor stem* in purple, *D arm* in red, *Anticodon arm* in blue with *Anticodon* in black, *T arm* in green.

The structure of tRNA can be decomposed into its primary structure, its secondary structure (usually visualized as the *cloverleaf structure*), and its tertiary structure (all tRNAs have a similar L-shaped 3D structure that allows them to fit into the P and A sites of the ribosome). The cloverleaf structure becomes the 3D L-shaped structure through coaxial stacking of the helices which is a common RNA Tertiary Structure motif.

1. The 5'-terminal phosphate group.
2. The acceptor stem is a 7-base pair (bp) stem made by the base pairing of the 5'-terminal nucleotide with the 3'-terminal nucleotide (which contains the CCA 3'-terminal group used to attach the amino acid). The acceptor stem may contain non-Watson-Crick base pairs.

3. The CCA tail is a cytosine-cytosine-adenine sequence at the 3' end of the tRNA molecule. This sequence is important for the recognition of tRNA by enzymes critical in translation. In prokaryotes, the CCA sequence is transcribed in some tRNA sequences. In most prokaryotic tRNAs and eukaryotic tRNAs, the CCA sequence is added during processing and therefore does not appear in the tRNA gene.
4. The D arm is a 4 bp stem ending in a loop that often contains dihydrouridine.
5. The anticodon arm is a 5-bp stem whose loop contains the anticodon.
6. The T arm is a 5 bp stem containing the sequence TΨC where Ψ is a pseudouridine.
7. Bases that have been modified, especially by methylation, occur in several positions throughout the tRNA. The first anticodon base, or wobble-position, is sometimes modified to inosine (derived from adenine), pseudouridine (derived from uracil) or lysidine (derived from cytosine).

Anticodon

An **anticodon** is a unit made up of three nucleotides that correspond to the three bases of the codon on the mRNA. Each tRNA contains a specific anticodon triplet sequence that can base-pair to one or more codons for an amino acid. For example, the codon for lysine is AAA; the anticodon of a lysine tRNA might be UUU. Some anticodons can pair with more than one codon due to a phenomenon known as wobble base pairing. Frequently, the first nucleotide of the anticodon is one of two not found on mRNA: inosine and pseudouridine, which can hydrogen bond to more than one base in the corresponding codon position. In the genetic code, it is common for a single amino acid to be specified by all four third-position possibilities, or at least by both Pyrimidines and Purines; for example, the amino acid glycine is coded for by the codon sequences GGU, GGC, GGA, and GGG.

To provide a one-to-one correspondence between tRNA molecules and codons that specify amino acids, 61 types of tRNA molecules would be required per cell. However, many cells contain fewer than 61 types of tRNAs because the wobble base is capable of binding to several, though not necessarily all, of the codons that specify a particular amino acid. A minimum of 31 tRNA are required to translate, unambiguously, all 61 sense codons of the standard genetic code.

Aminoacylation

Aminoacylation is the process of adding an aminoacyl group to a compound. It produces tRNA molecules with their CCA 3' ends covalently linked to an amino acid.

Each tRNA is aminoacylated (or *charged*) with a specific amino acid by an aminoacyl tRNA synthetase. There is normally a single aminoacyl tRNA synthetase for each amino acid, despite the fact that there can be more than one tRNA, and more than one anticodon, for an amino acid. Recognition of the appropriate tRNA by the synthetases is not mediated solely by the anticodon, and the acceptor stem often plays a prominent role.

Reaction:

1. amino acid + ATP → aminoacyl-AMP + PPi
2. aminoacyl-AMP + tRNA → aminoacyl-tRNA + AMP

Sometimes, certain organisms can have one or more aminoacyl tRNA synthetases missing. This leads to mischarging of the tRNA by a chemically related amino acid. The correct amino acid is made by enzymes that modify the mischarged amino acid to the correct one.

For example, *Helicobacter pylori* has glutaminyl tRNA synthetase missing. Thus, glutamate tRNA synthetase mischarges tRNA-glutamine(tRNA-Gln) with glutamate. An amidotransferase then converts the acid side chain of the glutamate to the amide, forming the correctly charged gln-tRNA-Gln.

Binding to ribosome

The ribosome has three binding sites for tRNA molecules: the A (aminoacyl), P (peptidyl), and E (exit) sites. During translation the A site binds an incoming aminoacyl-tRNA as directed by the codon currently occupying this site. This codon specifies the next amino acid to be added to the growing peptide chain. The A site only works after the first aminoacyl-tRNA has attached to the P site. The P-site codon is occupied by peptidyl-tRNA that is a tRNA with multiple amino acids attached as a long chain. The P site is actually the first to bind to aminoacyl tRNA. This tRNA in the P site carries the chain of amino acids that has already been synthesized. The E site is occupied by the empty tRNA as it's about to exit the ribosome.

tRNA genes

Organisms vary in the number of tRNA genes in their genome. The nematode worm *C. elegans*, a commonly used model organism in genetics studies, has 29,647 genes in its nuclear genome, of which 620 code for tRNA. The budding yeast *Saccharomyces cerevisiae* has 275 tRNA genes in its genome. In the human genome, which according to current estimates has about 27,161 genes in total, there are about 4,421 non-coding RNA genes, which include tRNA genes. There are 22 mitochondrial tRNA genes; 497 nuclear genes encoding cytoplasmic tRNA molecules and there are 324 tRNA-derived putative pseudogenes.

Cytoplasmic tRNA genes can be grouped into 49 families according to their anticodon features. These genes are found on all chromosomes, except 22 and Y chromosome. High clustering on 6p is observed (140 tRNA genes), as well on 1 chromosome.

tRNA biogenesis

In eukaryotic cells, tRNAs are transcribed by RNA polymerase III as pre-tRNAs in the nucleus. RNA polymerase III recognizes two internal promoter sequences (A-box B internal promoter) inside tRNA genes. The first promoter begins at nucleotide 8 of mature tRNAs and the second promoter is located 30-60 nucleotides downstream of the first promoter. The transcription terminates after a strech of four or more thymidines.

Pre-tRNAs undergo extensive modifications inside the nucleus. Some pre-tRNAs contain introns; in bacteria these self-splice, whereas in eukaryotes and archaea they are removed by tRNA splicing endonuclease.. The 5' sequence is removed by RNase P, whereas the 3' end is removed by the tRNase Z enzyme. A notable exception is in the archaeon *Nanoarchaeum equitans* which does not possess an RNase P enzyme and has a promoter placed such that transcription starts at the 5' end of the mature tRNA.. The non-templated 3' CCA tail is added by a nucleotidyl transferase. Before tRNAs are exported into the cytoplasm by Los1/Xpo-t, tRNAs are aminoacylated. The order of the processing events is not conserved. For example in yeast, the splicing is not carried out in the nucleus but at the cytoplasmic side of mitochondrial membranes.

History

The existence of tRNA was first hypothesized by Francis Crick, based on the assumption that there must exist an adapter molecule capable of mediating the translation of the RNA alphabet into the protein alphabet. Significant research on structure was conducted in the early 1960s by Alex Rich and Don Caspar, two researchers in Boston, the Jacques Fresco group in Princeton University and a United Kingdom group at King's College London. In 1965, a publication by Robert W. Holley reported the primary structure and suggested three secondary structures. The cloverleaf structure was ascertained by several other studies in the following years and was finally confirmed using X-ray crystallography studies in 1974. Two independent groups, Kim Sung-Hou working under Alexander Rich and a British group headed by Aaron Klug, published the same crystallography findings within a year.

Chapter- 4

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).

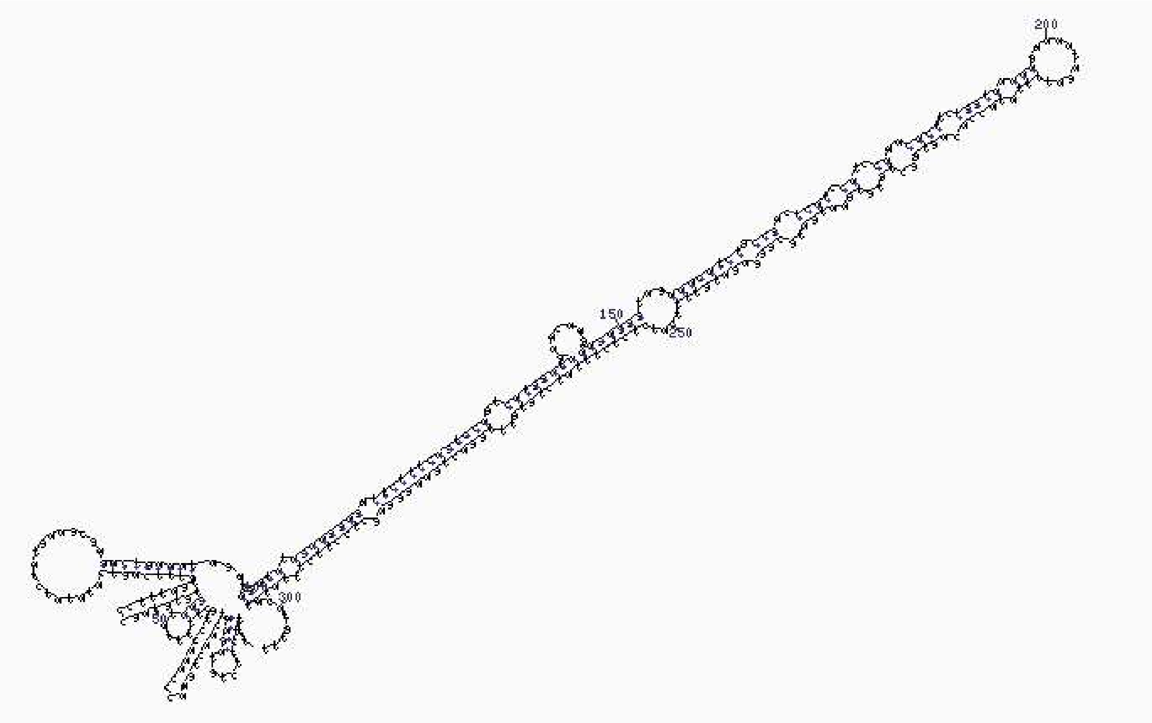
When studying gene expression, there are several methods to look at the various stages. In eukaryotes these include:

- 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- 5

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

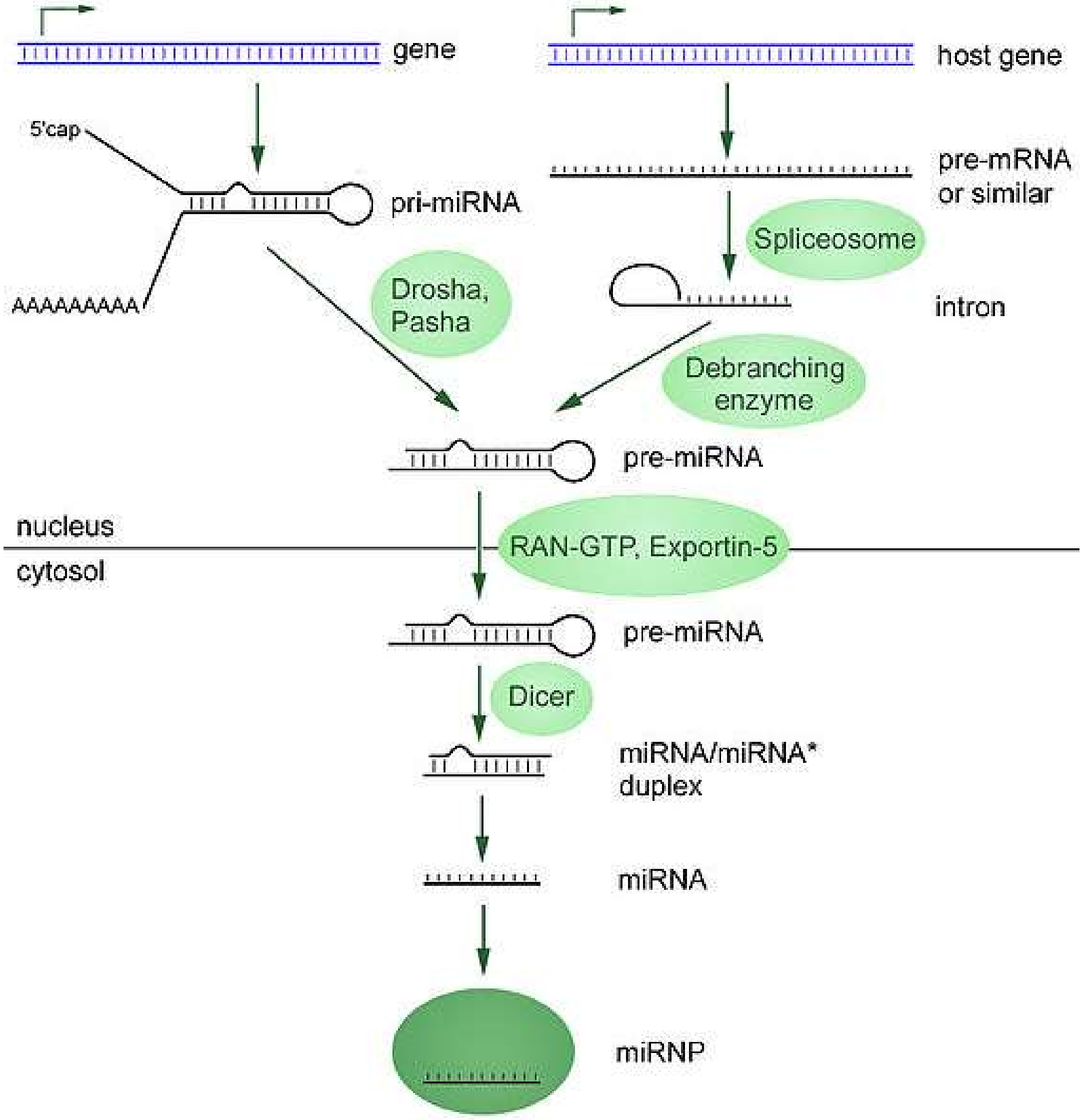
MicroRNAs were discovered in 1993 by Victor Ambros, Rosalind Lee and Rhonda Feinbaum during a study of the gene *lin-14* in *C. elegans* development. They found that LIN-14 protein abundance was regulated by a short RNA product encoded by the *lin-4* gene. A 61 nucleotide precursor from *lin-4* gene 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- 6

Translation (biology)

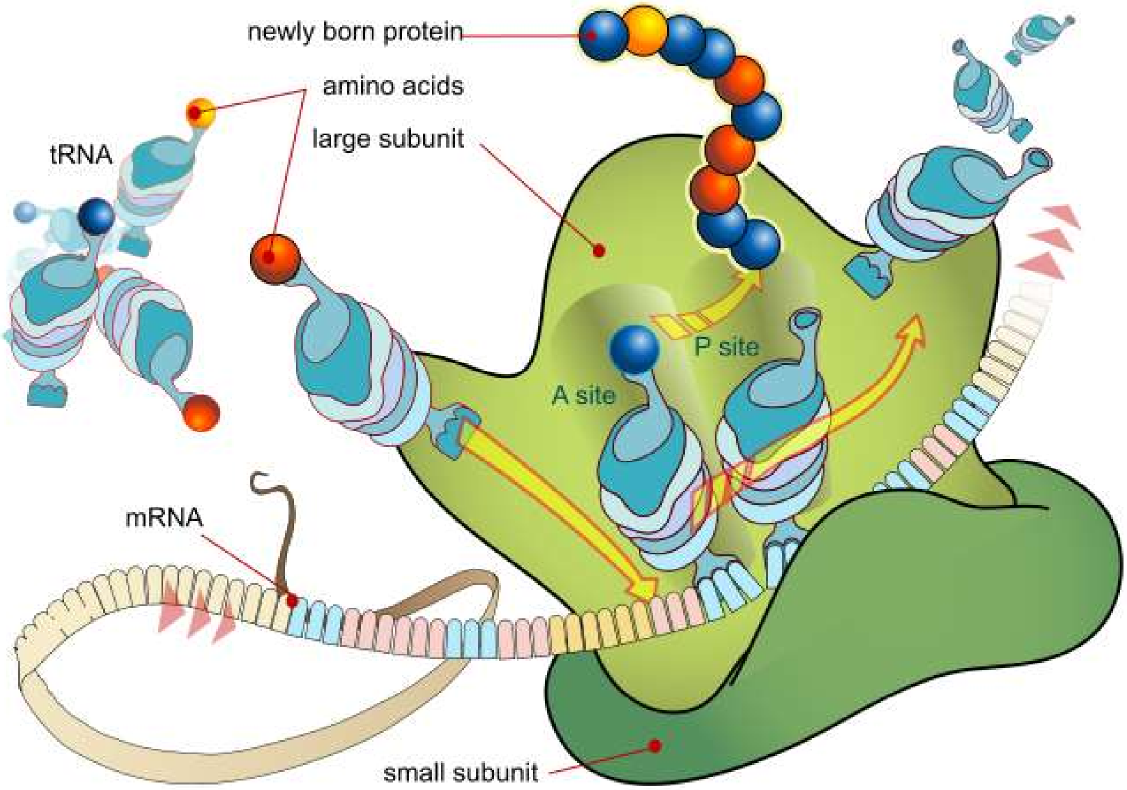


Diagram showing the translation of mRNA and the synthesis of proteins by a ribosome

In molecular biology and genetics, **translation** is the third stage of protein biosynthesis (part of the overall process of gene expression). In translation, messenger RNA (mRNA) produced by transcription is decoded by the ribosome to produce a specific amino acid chain, or polypeptide, that will later fold into an active protein. In Bacteria, translation occurs in the cell's cytoplasm, where the large and small subunits of the ribosome are located, and bind to the mRNA. In Eukaryotes, translation occurs across the membrane of the endoplasmic reticulum in a process called vectorial synthesis. The ribosome facilitates decoding by inducing the binding of tRNAs with complementary anticodon sequences to that of the mRNA. The tRNAs carry specific amino acids that are chained together into a polypeptide as the mRNA passes through and is "read" by the ribosome in a fashion reminiscent to that of a stock ticker and ticker tape.

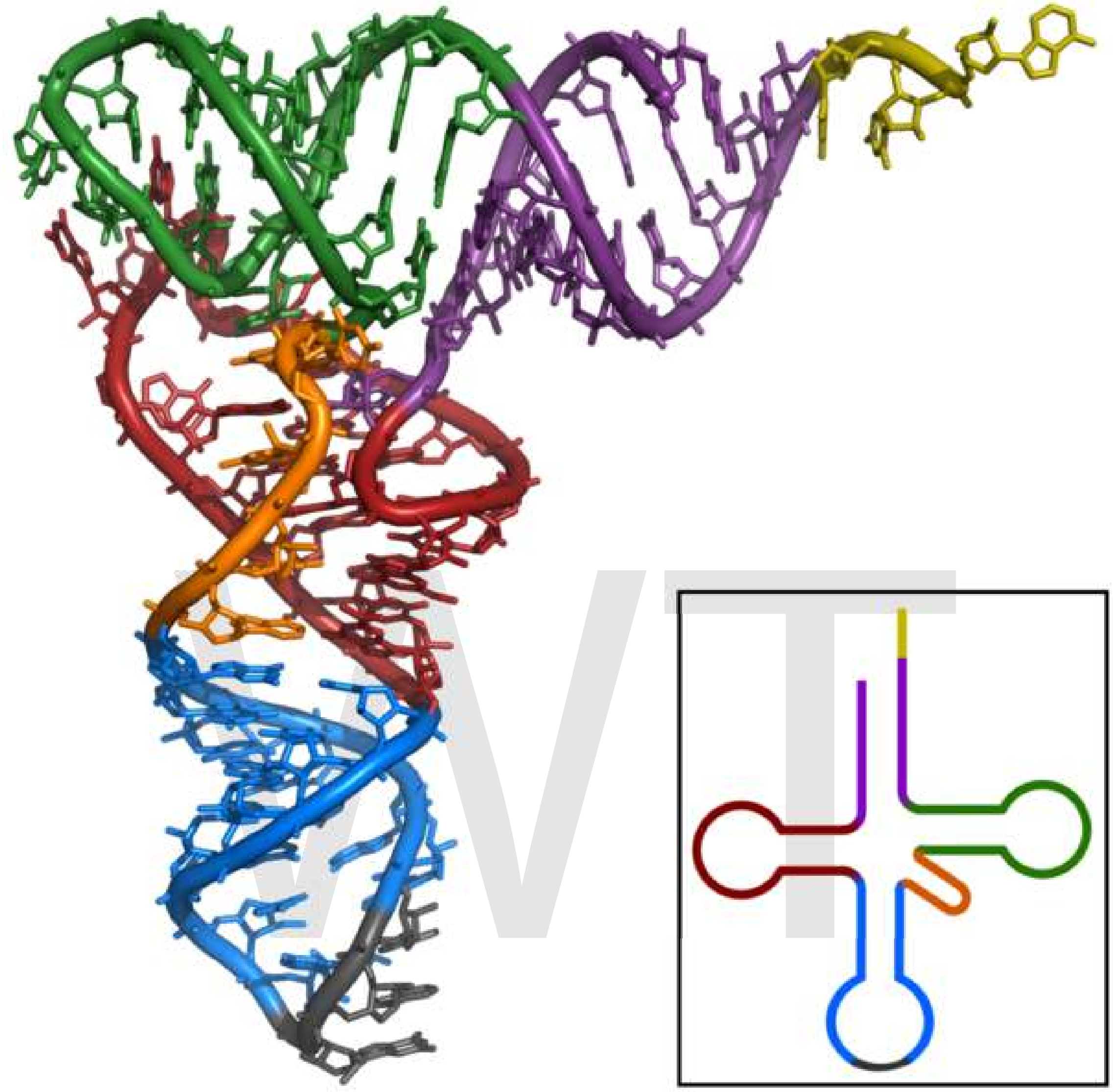
In many instances, the entire ribosome/mRNA complex will bind to the outer membrane of the rough endoplasmic reticulum and release the nascent protein polypeptide inside for later vesicle transport and secretion outside of the cell. Many types of transcribed RNA, such as transfer RNA, ribosomal RNA, and small nuclear RNA, do not undergo translation into proteins.

Translation proceeds in four phases: **activation**, **initiation**, **elongation** and **termination** (all describing the growth of the amino acid chain, or polypeptide that is the product of translation). Amino acids are brought to ribosomes and assembled into proteins.

In activation, the correct amino acid is covalently bonded to the correct transfer RNA (tRNA). The amino acid is joined by its carboxyl group to the 3' OH of the tRNA by a ester bond. When the tRNA has an amino acid linked to it, it is termed "charged". Initiation involves the small subunit of the ribosome binding to 5' end of mRNA with the help of initiation factors (IF). Termination of the polypeptide happens when the A site of the ribosome faces a stop codon (UAA, UAG, or UGA). No tRNA can recognize or bind to this codon. Instead, the stop codon induces the binding of a release factor protein that prompts the disassembly of the entire ribosome/mRNA complex.

A number of antibiotics act by inhibiting translation; these include anisomycin, cycloheximide, chloramphenicol, tetracycline, streptomycin, erythromycin, and puromycin, among others. Prokaryotic ribosomes have a different structure from that of eukaryotic ribosomes, and thus antibiotics can specifically target bacterial infections without any detriment to a eukaryotic host's cells.

Basic mechanisms



Tertiary structure of tRNA. *CCA tail* in orange, *Acceptor stem* in purple, *D arm* in red, *Anticodon arm* in blue with *Anticodon* in black, *T arm* in green.

The mRNA carries genetic information encoded as a ribonucleotide sequence from the chromosomes to the ribosomes. The ribonucleotides are "read" by translational machinery in a sequence of nucleotide triplets called codons. Each of those triplets codes for a specific amino acid.

The ribosome molecules translate this code to a specific sequence of amino acids. The ribosome is a multisubunit structure containing rRNA and proteins. It is the "factory" where amino acids are assembled into proteins. tRNAs are small noncoding RNA chains (74-93 nucleotides) that transport amino acids to the ribosome. tRNAs have a site for amino acid attachment, and a site called an anticodon. The anticodon is an RNA triplet complementary to the mRNA triplet that codes for their cargo amino acid.

Aminoacyl tRNA synthetase (an enzyme) catalyzes the bonding between specific tRNAs and the amino acids that their anticodon sequences call for. The product of this reaction is an aminoacyl-tRNA molecule. This aminoacyl-tRNA travels inside the ribosome, where mRNA codons are matched through complementary base pairing to specific tRNA anticodons. The amino acids that the tRNAs carry are then used to assemble a protein. After the new amino acid is added to the chain, the energy provided by the hydrolysis of a GTP bound to the translocase EF-G (in prokaryotes) and eEF-2 (in eukaryotes) moves the ribosome down one codon towards the 3' end. The energy required for translation of proteins is significant. For a protein containing *n* amino acids, the number of high-energy Phosphate bonds required to translate it is 4*n*+1. The rate of translation varies; it is significantly higher in prokaryotic cells (up to 17-21 amino acid residues per second) than in eukaryotic cells (up to 6-9 amino acid residues per second).

Genetic code

Whereas other aspects such as the 3D structure, called tertiary structure, of protein can only be predicted using sophisticated algorithms, the amino acid sequence, called primary structure, can be determined solely from the nucleic acid sequence with the aid of a translation table.

This approach may not give the correct amino acid composition of the protein, in particular if unconventional amino acids such as selenocysteine are incorporated into the protein, which is coded for by a conventional stop codon in combination with a downstream hairpin (SElenoCysteine Insertion Sequence, or SECIS).

There are many computer programs capable of translating a DNA/RNA sequence into a protein sequence. Normally this is performed using the Standard Genetic Code; many bioinformaticians have written at least one such program at some point in their education. However, few programs can handle all the "special" cases, such as the use of the alternative initiation codons. For instance, the rare alternative start codon CTG codes for Methionine when used as a start codon, and for Leucine in all other positions.

Example: Condensed translation table for the Standard Genetic Code (from the NCBI Taxonomy webpage).

```
   AAs  =
FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
 Starts = ---M---------------M---------------M-------------------------
---
 Base1  =
TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
 Base2  =
TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
 Base3  =
TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

Translation tables

Even when working with ordinary Eukaryotic sequences such as the Yeast genome, it is often desired to be able to use alternative translation tables—namely for translation of the mitochondrial genes. Currently the following translation tables are defined by the NCBI Taxonomy Group for the translation of the sequences in GenBank:

```
 1: The Standard
 2: The Vertebrate Mitochondrial Code
 3: The Yeast Mitochondrial Code
 4: The Mold, Protozoan, and Coelenterate Mitochondrial Code and the
    Mycoplasma/Spiroplasma Code
 5: The Invertebrate Mitochondrial Code
 6: The Ciliate, Dasycladacean and Hexamita Nuclear Code
 9: The Echinoderm and Flatworm Mitochondrial Code
10: The Euplotid Nuclear Codecbn dxh
11: The Bacterial and Plant Plastid Code
12: The Alternative Yeast Nuclear Code
13: The Ascidian Mitochondrial Code
14: The Alternative Flatworm Mitochondrial Code
15: Blepharisma Nuclear Code
16: Chlorophycean Mitochondrial Code
21: Trematode Mitochondrial Code
22: Scenedesmus obliquus mitochondrial Code
23: Thraustochytrium Mitochondrial Code
```

Software examples

- ApE (Mac, Windows, Unix)
- Serial Cloner (A DNA editing and manipulating software for MacOS and Windows)
- DNA Strider (Mac)
- ExPASy Translate Tool (webserver)
- Virtual Ribosome (webserver, cross-platform command-line)
- DNA to protein translation (webserver, 13 genomic codes or custom ones)

Example of computational translation - notice the indication of (alternative) start-codons:

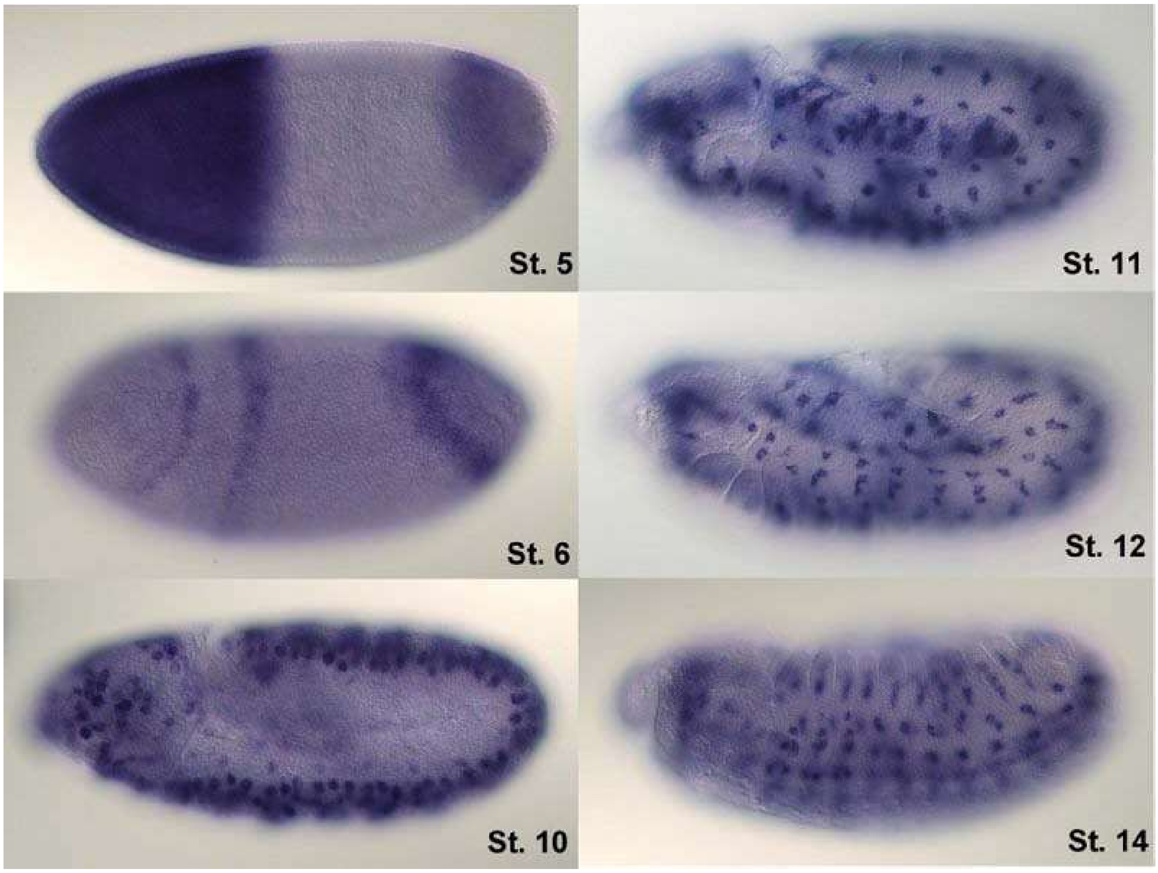
```
VIRTUAL RIBOSOME
----
Translation table: Standard SGC0

>Seq1
Reading frame: 1

   M  V  L  S  A  A  D  K  G  N  V  K  A  A  W  G  K  V  G  G  H  A  A
E  Y  G  A  E  A  L
5'
ATGGTGCTGTCTGCCGCCGACAAGGGCAATGTCAAGGCCGCCTGGGGCAAGGTTGGCGGCCACGCTGCAGA
GTATGGCGCAGAGGCCCTG 90
```

```
>>>...)))..............................................................
................)))

    E  R  M  F  L  S  F  P  T  T  K  T  Y  F  P  H  F  D  L  S  H  G  S
A  Q  V  K  G  H  G
5'
GAGAGGATGTTCCTGAGCTTCCCCACCACCAAGACCTACTTCCCCCACTTCGACCTGAGCCACGGCTCCGC
GCAGGTCAAGGGCCACGGC 180

......>>>...))).......................................)))..............
...................

    A  K  V  A  A  A  L  T  K  A  V  E  H  L  D  D  L  P  G  A  L  S  E
L  S  D  L  H  A  H
5'
GCGAAGGTGGCCGCCGCGCTGACCAAAGCGGTGGAACACCTGGACGACCTGCCCGGTGCCCTGTCTGAACT
GAGTGACCTGCACGCTCAC 270

..................)))..................))).......))).........))).......))
)......))).........

    K  L  R  V  D  P  V  N  F  K  L  L  S  H  S  L  L  V  T  L  A  S  H
L  P  S  D  F  T  P
5'
AAGCTGCGTGTGGACCCGGTCAACTTCAAGCTTCTGAGCCACTCCCTGCTGGTGACCCTGGCCTCCCACCT
CCCCAGTGATTTCACCCCC 360

...))).........................))).........))))))......))).............
...................

    A  V  H  A  S  L  D  K  F  L  A  N  V  S  T  V  L  T  S  K  Y  R  *
5'
GCGGTCCACGCCTCCCTGGACAAGTTCTTGGCCAACGTGAGCACCGTGCTGACCTCCAAATACCGTTAA
429

...............)))..........))).................. ))).................***

Annotation key:
>>> : START codon (strict)
))) : START codon (alternative)
*** : STOP
```

Chapter- 7

In Situ Hybridization and Protein Expression (Biotechnology)

In situ hybridization



In situ hybridization of wild type *Drosophila* embryos at different developmental stages for the RNA from a gene called hunchback.

In situ hybridization (ISH) is a type of hybridization that uses a labeled complementary DNA or RNA strand (i.e., probe) to localize a specific DNA or RNA sequence in a portion or section of tissue (*in situ*), or, if the tissue is small enough (e.g. plant seeds, *Drosophila* embryos), in the entire tissue (whole mount ISH). This is distinct from immunohistochemistry, which usually localizes proteins in tissue sections. DNA ISH can be used to determine the structure of chromosomes. Fluorescent DNA ISH (FISH) can, for example, be used in medical diagnostics to assess chromosomal integrity. RNA ISH (hybridization histochemistry) is used to measure and localize mRNAs and other transcripts within tissue sections or whole mounts.

Process

For hybridization histochemistry, sample cells and tissues are usually treated to fix the target transcripts in place and to increase access of the probe. As noted above, the probe is either a labeled complementary DNA or, now most commonly, a complementary RNA (riboprobe). The probe hybridizes to the target sequence at elevated temperature, and then the excess probe is washed away (after prior hydrolysis using RNase in the case of unhybridized, excess RNA probe). Solution parameters such as temperature, salt and/or detergent concentration can be manipulated to remove any non-identical interactions (i.e. only exact sequence matches will remain bound). Then, the probe that was labeled with either radio-, fluorescent- or antigen-labeled bases (e.g., digoxigenin) is localized and quantitated in the tissue using either autoradiography, fluorescence microscopy or immunohistochemistry, respectively. ISH can also use two or more probes, labeled with radioactivity or the other non-radioactive labels, to simultaneously detect two or more transcripts.

Basic Steps for Digoxigenin-labeled probes

1. permeabilisation of cells with proteinase K to open cell membranes (around 25 minutes, not needed for tissue sections or some early-stage embryos)
2. binding of mRNAs to marked RNA probe (usually overnight)
3. antibody-phosphatase binding to RNA-probe (some hours)
4. staining of antibody (e.g. with alkaline phosphatase)

The protocol takes around 2-3 days and takes some time to set up. Some companies sell robots to automate the process. As a result, large-scale screenings have been conducted in laboratories on thousands of genes.

Protein expression

Protein expression is a subcomponent of gene expression. It consists of the stages after DNA has been translated into polypeptide chains, which are ultimately folded into proteins. Protein expression is commonly used by proteomics researchers to denote the measurement of the presence and abundance of one or more proteins in a particular cell or tissue.

Protein expression systems are very widely used in the life sciences, biotechnology and medicine. Molecular biology research uses an enormous number of proteins and enzymes many of which are from expression systems; particularly DNA polymerase for PCR, reverse transcriptase for RNA analysis and restriction endonucleases for cloning. There are also significant medical applications for expression systems, notably the production of human insulin to treat diabetes.

Expression systems

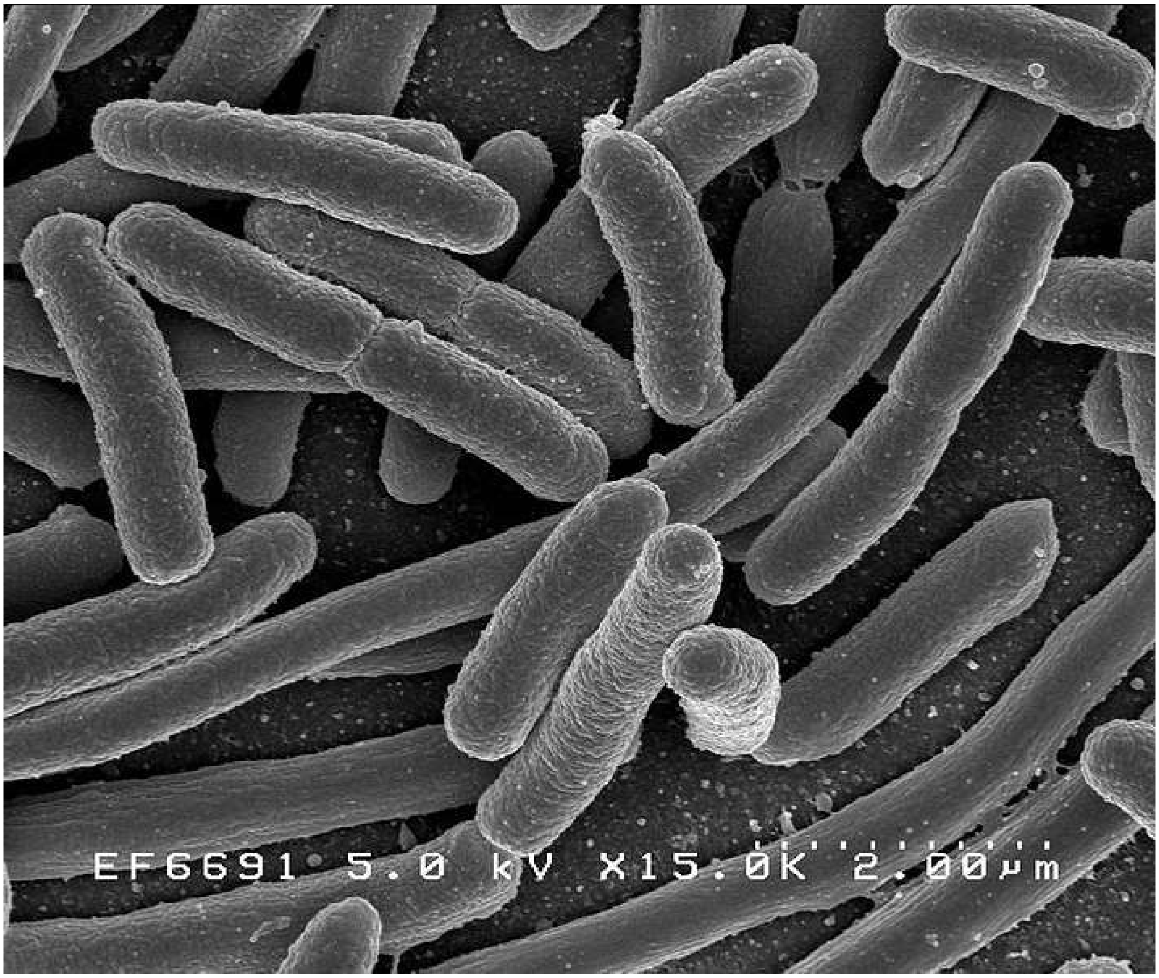
Commonly used protein expression systems include those derived from bacteria, yeast, baculovirus/insect, and mammalian cells.

Cell-based systems

The oldest and most widely used expression systems are cell-based and may be defined as the "*combination of an expression vector, its cloned DNA, and the host for the vector that provide a context to allow foreign gene function in a host cell, that is, produce proteins at a high level*". Expression is often done to a very high level and therefore referred to as overexpression.

There are many ways to introduce foreign DNA to a cell for expression, and there are many different host cells which may be used for expression - each expression system has distinct advantages and liabilities. Expression systems are normally referred to by the host and the DNA source or the delivery mechanism for the genetic material. For example, common hosts are bacteria (such as *E.coli*, *B. subtilis*), yeast (such as *S.cerevisiae*) or eukaryotic cell lines. Common DNA sources and delivery mechanisms are viruses (such as baculovirus, retrovirus, adenovirus), plasmids, artificial chromosomes and bacteriophage (such as lambda). The best expression system of choice depends on the gene involved, for example the *Saccharomyces cerevisiae* is often preferred for proteins that require significant posttranslational modification and Insect or mammal cell lines are used when human-like splicing of the mRNA is required. Nonetheless, bacterial expression has the advantage of easily producing large amounts of protein, which is required for X-ray crystallography or nuclear magnetic resonance experiments for structure determination.

Escherichia coli



E. coli, one of the most popular hosts for artificial gene expression

E. coli is one of the most widely used expression hosts, and DNA is normally introduced in a plasmid expression vector. The techniques for overexpression in *E. coli* are well developed and work by increasing the number of copies of the gene or increasing the binding strength of the promoter region so assisting trancription.

For example a DNA sequence for a protein of interest could be cloned or subcloned into a high copy-number plasmid containing the *lac* promoter, which is then transformed into the bacterium *Escherichia coli*. Addition of IPTG (a lactose analog) activates the lac promoter and causes the bacteria to express the protein of interest.

Corynebacterium

Non-pathogenic species of the gram-positive *Corynebacterium* are used for the commercial production of various amino acids. The *C. glutamicum* species is widely used for producing glutamate and lysine, components of human food, animal feed, and pharmaceutical products.

Expression of functionally active human epidermal growth factor has been done in *C. glutamicum*, thus demonstrating a potential for industrial-scale production of human proteins. Expressed proteins can be targeted for secretion through either the general secretory pathway (Sec) or the twin-arginine translocation pathway (Tat).

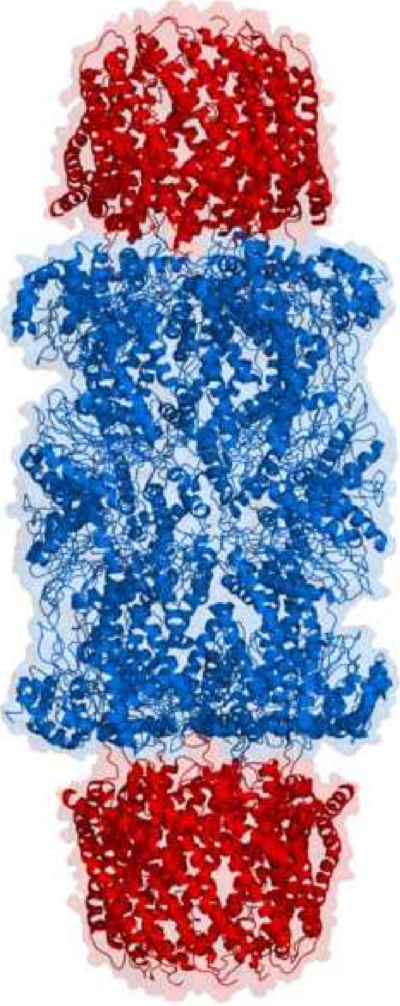
Unlike gram-negative bacteria, the gram-positive *Corynebacterium* lack lipopolysaccarides that function as antigenic endotoxins in humans.

Cell-free systems

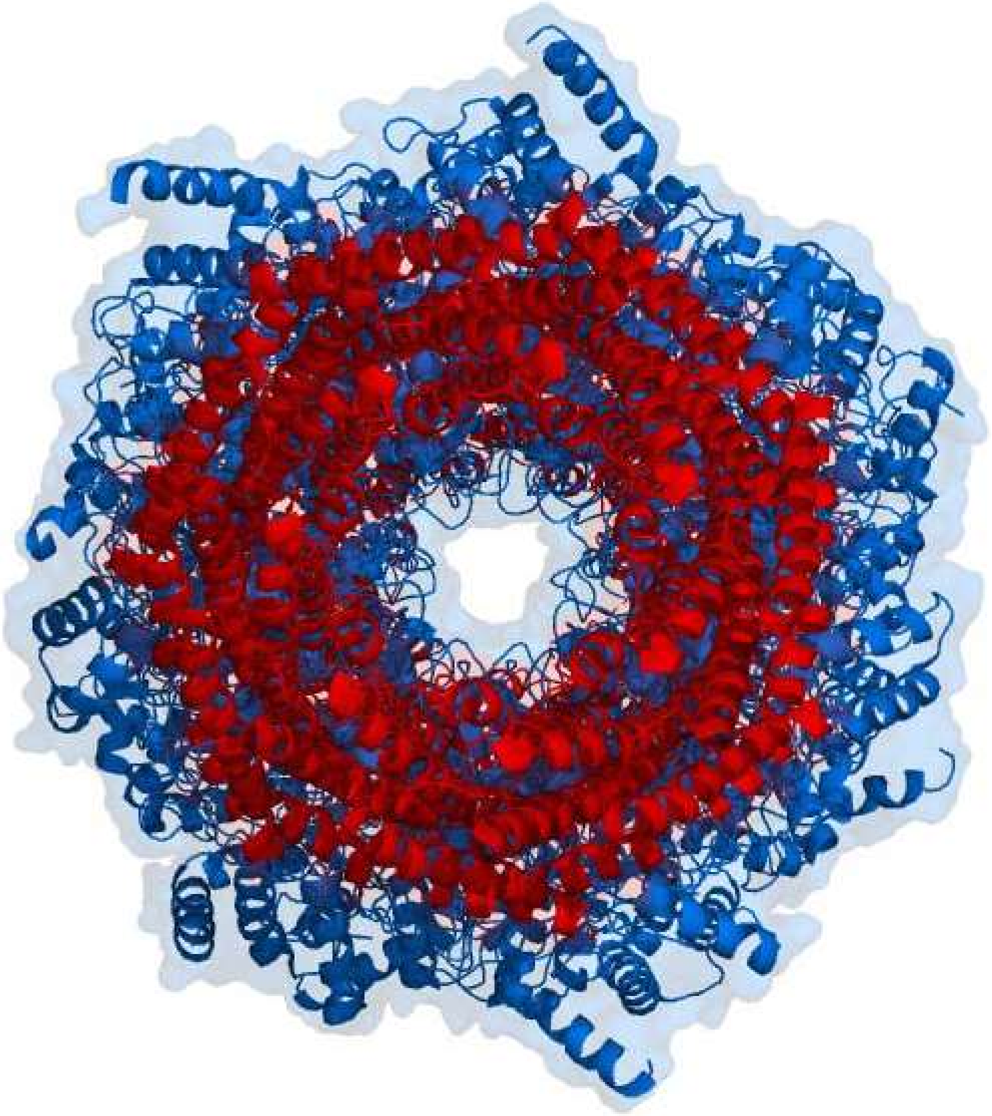
Cell-free expression of proteins is possible using purified RNA polymerase, ribosomes, tRNA and ribonucleotides. These reagents may be produced by extraction from cells or from a cell-based expression system. Due to the low expression levels and high cost of cell-free systems cell-based systems are more widely used.

Chapter- 8

Proteasome



Cartoon representation of a proteasome. Its active sites are sheltered inside the tube (blue). The caps (red; in this case, 11S regulatory particles) on the ends regulate entry into the destruction chamber, where the protein is degraded.



Top view of the proteasome above

Proteasomes are very large protein complexes inside all eukaryotes and archaea, and in some bacteria. In eukaryotes, they are located in the nucleus and the cytoplasm. The main function of the proteasome is to degrade unneeded or damaged proteins by proteolysis, a chemical reaction that breaks peptide bonds. Enzymes that carry out such reactions are called proteases. Proteasomes are part of a major mechanism by which cells regulate the concentration of particular proteins and degrade misfolded proteins. The degradation process yields peptides of about seven to eight amino acids long, which can then be further degraded into amino acids and used in synthesizing new proteins. Proteins are tagged for degradation with a small protein called ubiquitin. The tagging reaction is catalyzed by enzymes called ubiquitin ligases. Once a protein is tagged with a single ubiquitin molecule, this is a signal to other ligases to attach additional ubiquitin molecules. The result is a *polyubiquitin chain* that is bound by the proteasome, allowing it to degrade the tagged protein.

In structure, the proteasome is a cylindrical complex containing a "core" of four stacked rings around a central pore. Each ring is composed of seven individual proteins. The inner two rings are made of seven *β subunits* that contain the six protease active sites. These sites are located on the interior surface of the rings, so that the target protein must enter the central pore before it is degraded. The outer two rings each contain seven *α subunits* whose function is to maintain a "gate" through which proteins enter the barrel. These α subunits are controlled by binding to "cap" structures or *regulatory particles* that recognize polyubiquitin tags attached to protein substrates and initiate the degradation process. The overall system of ubiquitination and proteasomal degradation is known as the *ubiquitin-proteasome system*.

The proteasomal degradation pathway is essential for many cellular processes, including the cell cycle, the regulation of gene expression, and responses to oxidative stress. The importance of proteolytic degradation inside cells and the role of ubiquitin in proteolytic pathways was acknowledged in the award of the 2004 Nobel Prize in Chemistry to Aaron Ciechanover, Avram Hershko and Irwin Rose.

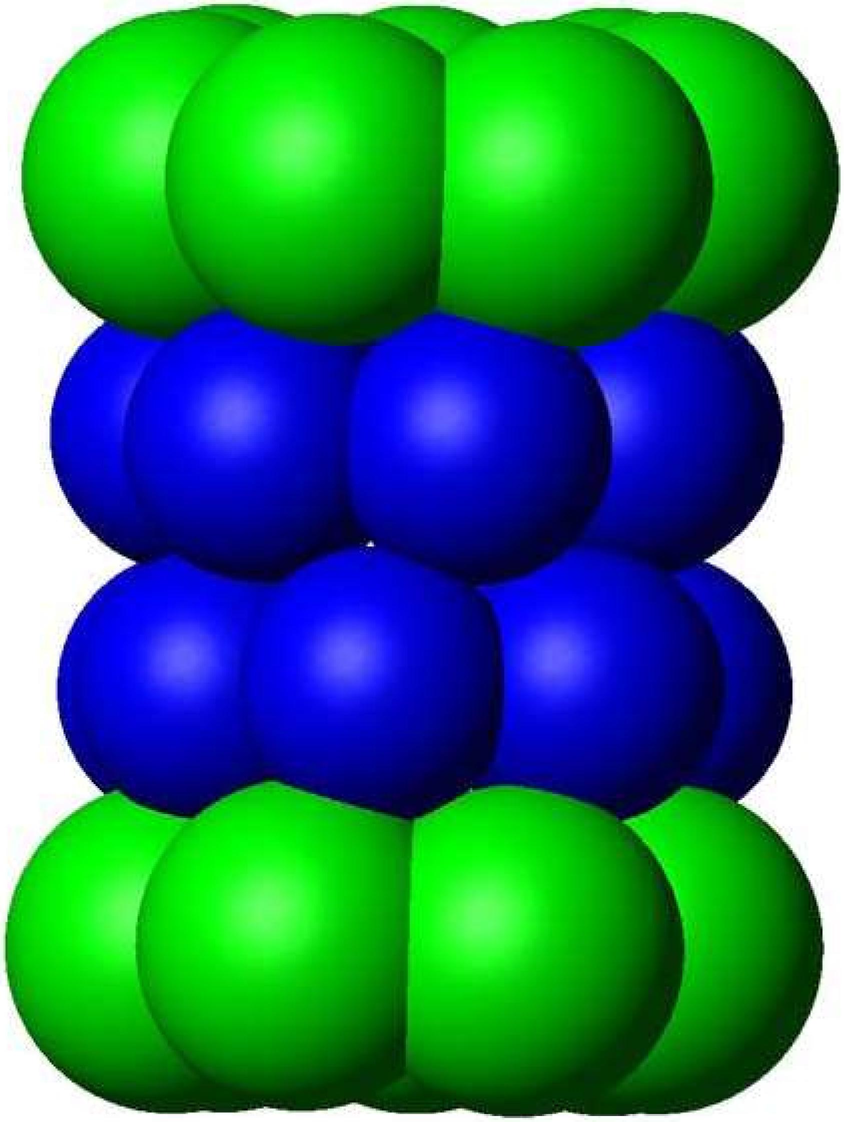
Discovery

Before the discovery of the ubiquitin proteasome system, protein degradation in cells was thought to rely mainly on lysosomes, membrane-bound organelles with acidic and protease-filled interiors that can degrade and then recycle exogenous proteins and aged or damaged organelles. However, work by Alfred Goldberg in 1977 on ATP-dependent protein degradation in reticulocytes, which lack lysosomes, suggested the presence of a second intracellular degradation mechanism. This was shown in 1978 to be composed of several distinct protein chains, a novelty among proteases at the time. Later work on modification of histones led to the identification of an unexpected covalent modification of the histone protein by a bond between a lysine side chain of the histone and the C-terminal glycine residue of ubiquitin, a protein which had no known function. It was then discovered that a previously identified protein associated with proteolytic degradation, known as ATP-dependent proteolysis factor 1 (APF-1), was the same protein as ubiquitin. Later, the ATP-dependent proteolytic complex that was responsible for ubiquitin-dependent protein degradation was discovered and was called the 26S proteasome.

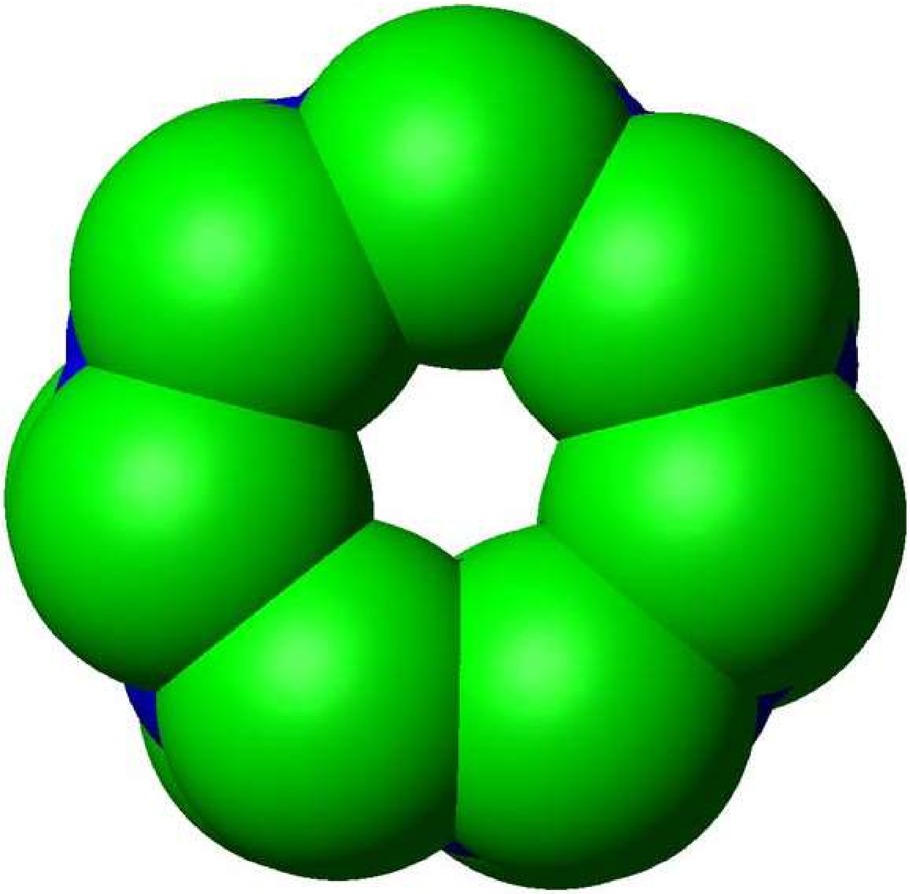
Much of the early work leading up to the discovery of the ubiquitin proteasome system occurred in the late 1970s and early 1980s at the Technion in the laboratory of Avram Hershko, where Aaron Ciechanover worked as a graduate student. Hershko's year-long sabbatical in the laboratory of Irwin Rose at the Fox Chase Cancer Center provided key conceptual insights, though Rose later downplayed his role in the discovery. The three shared the 2004 Nobel Prize in Chemistry for their work in discovering this system.

Although electron microscopy data revealing the stacked-ring structure of the proteasome became available in the mid-1980s, the first structure of the proteasome core particle was not solved by X-ray crystallography until 1994. As of 2006, no structure has been solved of the core particle in complex with the most common form of regulatory cap.

Structure and organization



A schematic diagram of the proteasome 20S core particle viewed from one side. The α subunits that make up the outer two rings are shown in green, and the β subunits that make up the inner two rings are shown in blue.



Top view of the same schematic, illustrating the seven-fold symmetry of the rings

The proteasome subcomponents are often referred to by their Svedberg sedimentation coefficient (denoted *S*). The most common form of the proteasome is known as the 26S proteasome, which is about 2000 kilodaltons (kDa) in molecular mass and contains one 20S core particle structure and two 19S regulatory caps. The core is hollow and provides an enclosed cavity in which proteins are degraded; openings at the two ends of the core allow the target protein to enter. Each end of the core particle associates with a 19S regulatory subunit that contains multiple ATPase active sites and ubiquitin binding sites; it is this structure that recognizes polyubiquitinated proteins and transfers them to the catalytic core. An alternative form of regulatory subunit called the 11S particle can associate with the core in essentially the same manner as the 19S particle; the 11S may play a role in degradation of foreign peptides such as those produced after infection by a virus.

20S core particle

The number and diversity of subunits contained in the 20S core particle depends on the organism; the number of distinct and specialized subunits is larger in multicellular than unicellular organisms and larger in eukaryotes than in prokaryotes. All 20S particles consist of four stacked heptameric ring structures that are themselves composed of two different types of subunits; α subunits are structural in nature, whereas β subunits are predominantly catalytic. The outer two rings in the stack consist of seven α subunits each, which serve as docking domains for the regulatory particles and the alpha subunits N-termini form a gate that blocks unregulated access of substrates to the interior cavity. The inner two rings each consist of seven β subunits and contain the protease active sites that perform the proteolysis reactions. The size of the proteasome is relatively conserved and is about 150 angstroms (Å) by 115 Å. The interior chamber is at most 53 Å wide, though the entrance can be as narrow as 13 Å, suggesting that substrate proteins must be at least partially unfolded to enter.

In archaea such as *Thermoplasma acidophilum*, all the α and all the β subunits are identical, while eukaryotic proteasomes such as those in yeast contain seven distinct types of each subunit. In mammals, the β1, β2, and β5 subunits are catalytic; although they share a common mechanism, they have three distinct substrate specificities considered chymotrypsin-like, trypsin-like, and peptidyl-glutamyl peptide-hydrolyzing (PHGH). Alternative β forms denoted β1i, β2i, and β5i can be expressed in hematopoietic cells in response to exposure to pro-inflammatory signals such as cytokines, in particular, interferon gamma. The proteasome assembled with these alternative subunits is known as the *immunoproteasome*, whose substrate specificity is altered relative to the normal proteasome.

19S regulatory particle

The 19S particle in eukaryotes consists of 19 individual proteins and is divisible into two subassemblies, a 10-protein base that binds directly to the α ring of the 20S core particle, and a 9-protein lid where polyubiquitin is bound. Six of the ten base proteins are ATPase subunits from the AAA Family, and an evolutionary homolog of these ATPases exists in archaea, called PAN (Proteasome Activating Nucleotidase). The association of the 19S and 20S particles requires the binding of ATP to the 19S ATPase subunits, and ATP hydrolysis is required for the assembled complex to degrade folded and ubiquitinated proteins. Interestingly, only the step of substrate unfolding requires energy from ATP hydrolysis, while ATP-binding alone can support all the other steps required for protein degradation (e.g. complex assembly, gate opening, translocation and proteolysis). In fact, ATP binding to the ATPases by itself supports the rapid degradation of unfolded proteins. However, while ATP hydrolysis is required for unfolding only it is not yet clear whether this energy may be used in the coupling of some of these steps. As of 2008, the atomic structure of the 26S proteasome has not been solved, despite massive efforts to do so. Nevertheless, it is understood generally how the 19S associates with and regulates the 20S core particle. In fact, the 19S and 11S particles bind to the same sites in the α rings of the 20S core particle although, they each induce gate opening by different mechanism.

Regulation of the 20S by the 19S

The 19S regulatory particle is responsible for stimulating the 20S to degrade proteins. A primary function of the 19S regulatory ATPases is to open the gate in the 20S that blocks the entry of substrates into the degradation chamber. The mechanism by which the proteasomal ATPase open this gate has been recently elucidated. 20S gate opening, and thus substrate degradation, requires the C-termini of the proteasomal ATPases, which contains a specific motif (i.e. HbYX motif). The ATPases C-termini bind into pockets in the top of the 20S, and tether the ATPase complex to the 20S proteolytic complex thus joining the substrate unfolding equipment with the 20S degradation machinery. Binding of these C-termini into these 20S pockets by themselves stimulates opening of the gate in the 20S much like a "key-in-a-lock" opens a door. The precise mechanism by which this "key-in-a-lock" mechanism functions has been structurally elucidated.

11S regulatory particle

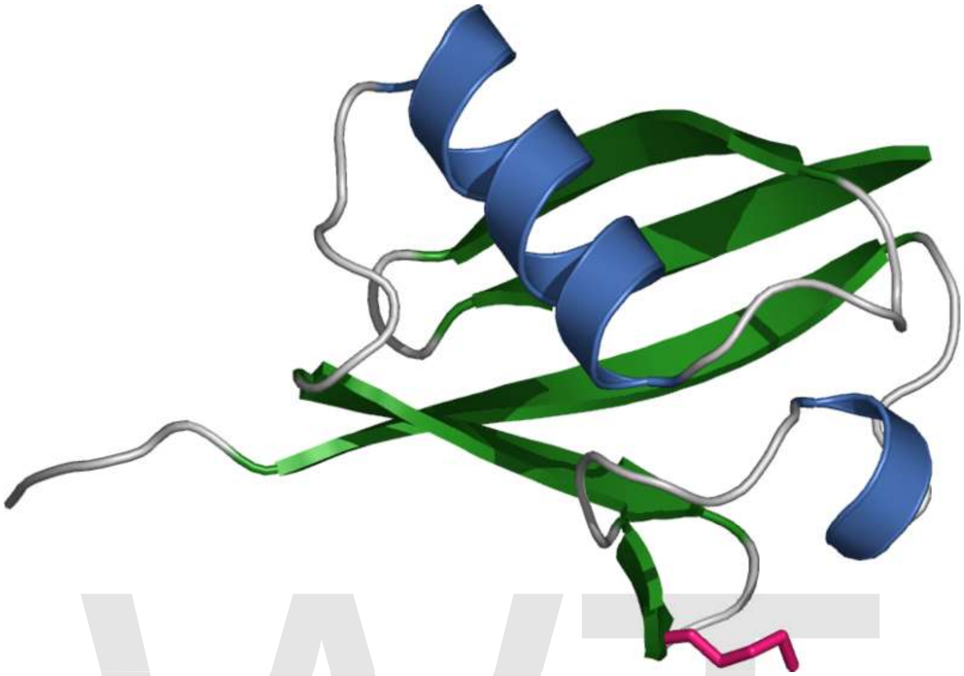
20S proteasomes can also associate with a second type of regulatory particle, the 11S regulatory particle, a heptameric structure that does not contain any ATPases and can promote the degradation of short peptides, but not of complete proteins. It is presumed that this is because the complex cannot unfold larger substrates. This structure is also known as PA28 or REG. The mechanisms by which it binds to the core particle through the C-terminal tails of its subunits and induces α-ring conformational changes to open the 20S gate suggest a similar mechanism for the 19S particle. The expression of the 11S particle is induced by interferon gamma and is responsible, in conjunction with the immunoproteasome β subunits, for the generation of peptides that bind to the major histocompatibility complex.

Assembly

The assembly of the proteasome is a complex process due to the number of subunits that must associate to form an active complex. The β subunits are synthesized with N-terminal "propeptides" that are post-translationally modified during the assembly of the 20S particle to expose the proteolytic active site. The 20S particle is assembled from two half-proteasomes, each of which consists of a seven-membered pro-β ring attached to a seven-membered α ring. The association of the β rings of the two half-proteasomes triggers threonine-dependent autolysis of the propeptides to expose the active site. These β interactions are mediated mainly by salt bridges and hydrophobic interactions between conserved alpha helices whose disruption by mutation damages the proteasome's ability to assemble. The assembly of the half-proteasomes, in turn, is initiated by the assembly of the α subunits into their heptameric ring, forming a template for the association of the corresponding pro-β ring. The assembly of α subunits has not been characterized.

In general, less is known about the assembly and maturation of the 19S regulatory particles. They are believed to assemble as two distinct subcomponents, the ATPase-containing base and the ubiquitin-recognizing lid. The six ATPases in the base may assemble in a pairwise manner mediated by coiled-coil interactions. The order in which the nineteen subunits of the regulatory particle are bound is a likely regulatory mechanism that prevents exposure of the active site before assembly is complete.

The protein degradation process



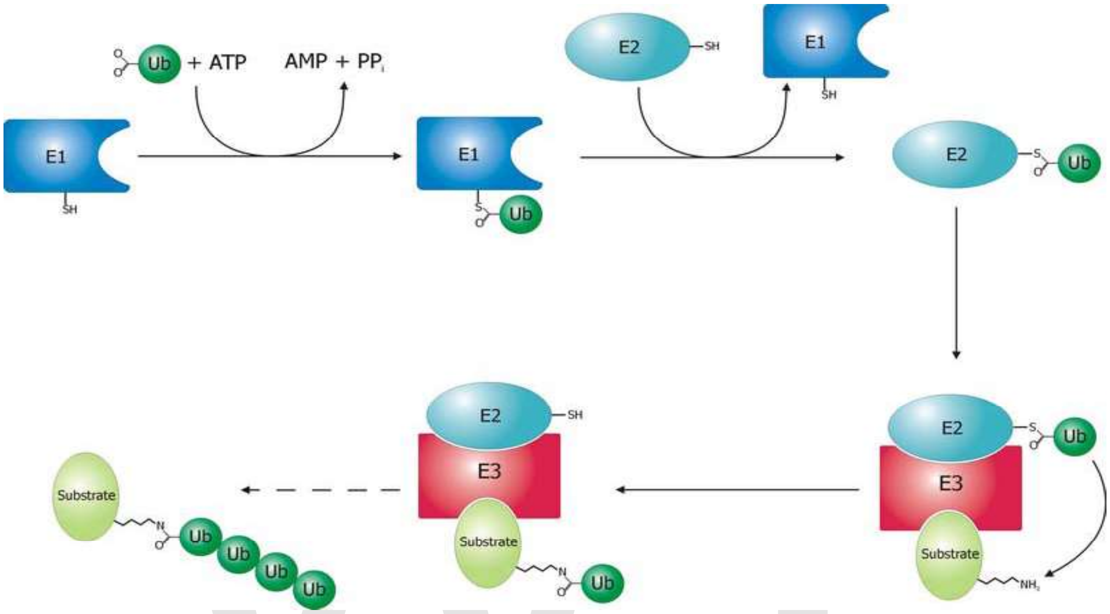
Ribbon diagram of ubiquitin, the highly conserved protein that serves as a molecular tag targeting proteins for degradation by the proteasome

Ubiquitination and targeting

Proteins are targeted for degradation by the proteasome by covalent modification of a lysine residue that requires the coordinated reactions of three enzymes. In the first step, a ubiquitin-activating enzyme (known as E1) hydrolyzes ATP and adenylates a ubiquitin molecule. This is then transferred to E1's active-site cysteine residue in concert with the adenylation of a second ubiquitin. This adenylated ubiquitin is then transferred to a cysteine of a second enzyme, ubiquitin-conjugating enzyme (E2). In the last step, a member of a highly diverse class of enzymes known as ubiquitin ligases (E3) recognizes the specific protein to be ubiquitinated and catalyzes the transfer of ubiquitin from E2 to this target protein. A target protein must be labeled with at least four ubiquitin monomers (in the form of a polyubiquitin chain) before it is recognized by the proteasome lid. It is therefore the E3 that confers substrate specificity to this system. The number of E1, E2, and E3 proteins expressed depends on the organism and cell type, but there are many different E3 enzymes present in humans, indicating that there is a huge number of targets for the ubiquitin proteasome system.

The mechanism by which a polyubiquitinated protein is targeted to the proteasome is not fully understood. Ubiquitin-receptor proteins have an N-terminal ubiquitin-like (UBL) domain and one or more ubiquitin-associated (UBA) domains. The UBL domains are recognized by the 19S proteasome caps and the UBA domains bind ubiquitin via three-helix bundles. These receptor proteins may escort polyubiquitinated proteins to the proteasome, though the specifics of this interaction and its regulation are unclear.

The ubiquitin protein itself is 76 amino acids long and was named due to its ubiquitous nature, as it has a highly conserved sequence and is found in all known eukaryotic organisms. The genes encoding ubiquitin in eukaryotes are arranged in tandem repeats, possibly due to the heavy transcription demands on these genes to produce enough ubiquitin for the cell. It has been proposed that ubiquitin is the slowest-evolving protein identified to date.



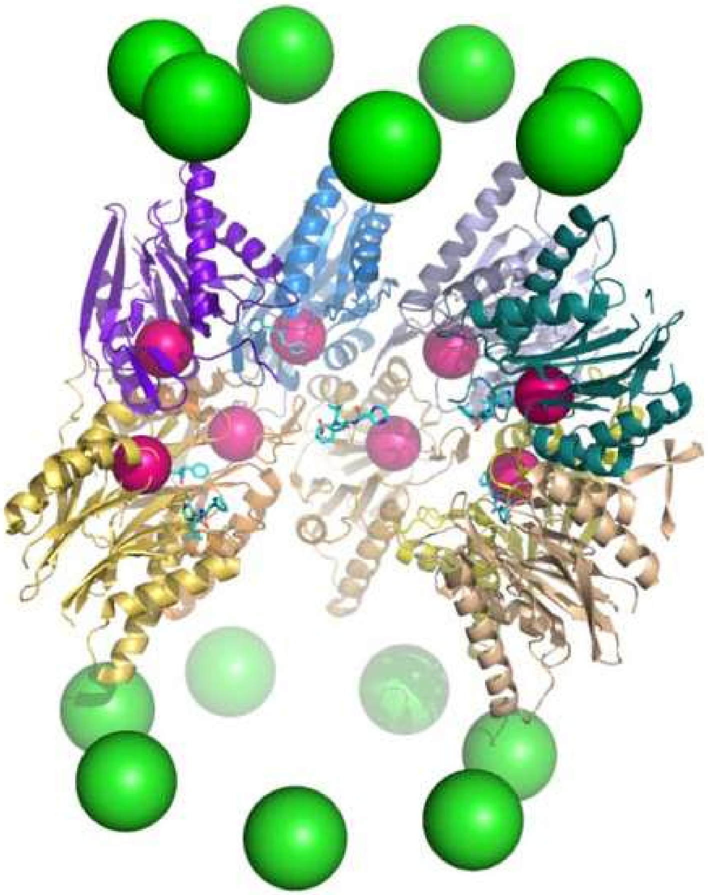
The ubiquitination pathway

Unfolding and translocation

After a protein has been ubiquitinated, it is recognized by the 19S regulatory particle in an ATP-dependent binding step. The substrate protein must then enter the interior of the 20S particle to come in contact with the proteolytic active sites. Because the 20S particle's central channel is narrow and gated by the N-terminal tails of the α ring subunits, the substrates must be at least partially unfolded before they enter the core. The passage of the unfolded substrate into the core is called *translocation* and necessarily occurs after deubiquitination. However, the order in which substrates are deubiquitinated and unfolded is not yet clear. Which of these processes is the rate-limiting step in the overall proteolysis reaction depends on the specific substrate; for some proteins, the unfolding process is rate-limiting, while deubiquitination is the slowest step for other proteins. The extent to which substrates must be unfolded before translocation is not known, but substantial tertiary structure, and in particular nonlocal interactions such as disulfide bonds, are sufficient to inhibit degradation.

The gate formed by the α subunits prevents peptides longer than about four residues from entering the interior of the 20S particle. The ATP molecules bound before the initial recognition step are hydrolyzed before translocation. While energy is needed for substrate unfolding it is not required for translocation. The assembled 26S proteasome can degrade unfolded proteins in the presence of a non-hydrolyzable ATP analog, but cannot degrade folded proteins, indicating that energy from ATP hydrolysis is used for substrate unfolding. Passage of the unfolded substrate through the opened gate occurs via facilitated diffusion if the 19S cap is in the ATP-bound state.

The mechanism for unfolding of globular proteins is necessarily general, but somewhat dependent on the amino acid sequence. Long sequences of alternating glycine and alanine have been shown to inhibit substrate unfolding decreasing the efficiency of proteasomal degradation; this results in the release of partially degraded byproducts, possibly due to the decoupling of the ATP hydrolysis and unfolding steps. Such glycine-alanine repeats are also found in nature, for example in silk fibroin; in particular, certain Epstein-Barr virus gene products bearing this sequence can stall the proteasome, helping the virus propagate by preventing antigen presentation on the major histocompatibility complex.



A cutaway view of the proteasome 20S core particle illustrating the locations of the active sites. The α subunits are represented as green spheres and the β subunits as protein backbones colored by individual polypeptide chain. The small pink spheres represent the location of the active-site threonine residue in each subunit. Light blue chemical structures are the inhibitor bortezomib bound to the active sites.

Proteolysis

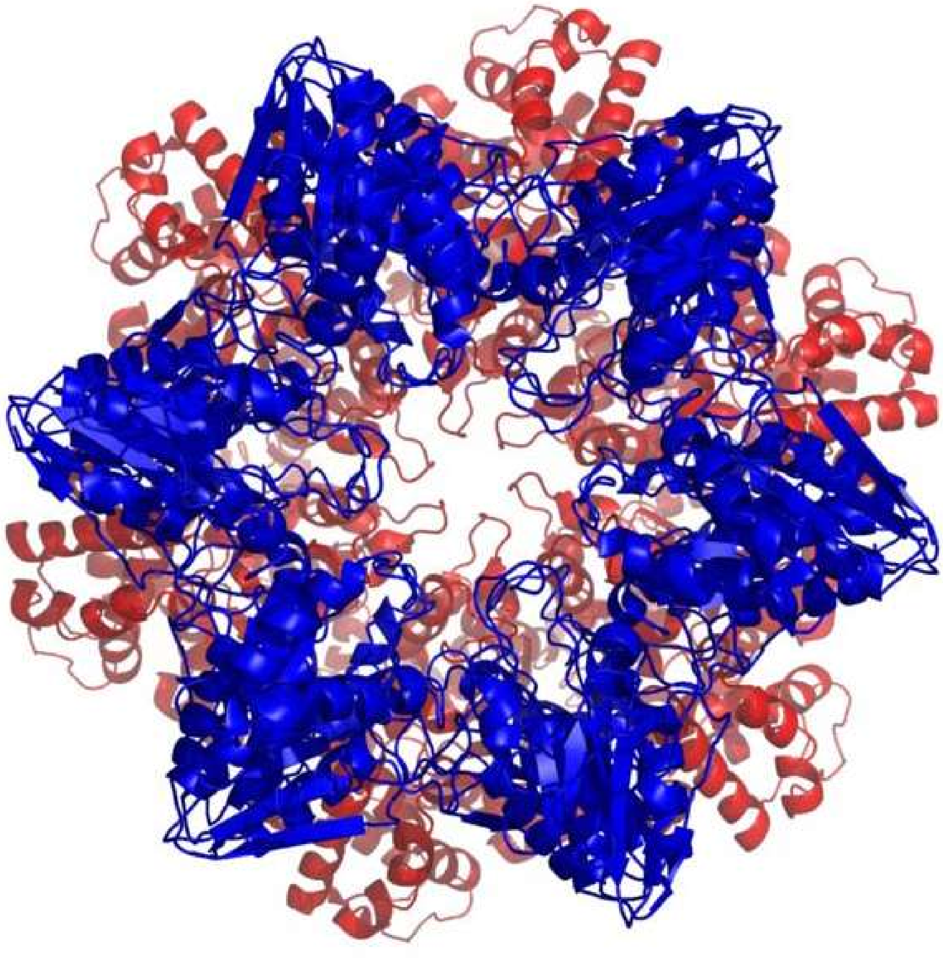
The mechanism of proteolysis by the β subunits of the 20S core particle is through a threonine-dependent nucleophilic attack. This mechanism may depend on an associated water molecule for deprotonation of the reactive threonine hydroxyl. Degradation occurs within the central chamber formed by the association of the two β rings and normally does not release partially degraded products, instead reducing the substrate to short polypeptides typically 7–9 residues long, though they can range from 4 to 25 residues depending on the organism and substrate. The biochemical mechanism that determines product length is not fully characterized. Although the three catalytic β subunits have a common mechanism, they have slightly different substrate specificities, which are considered chymotrypsin-like, trypsin-like, and peptidyl-glutamyl peptide-hydrolyzing (PHGH)-like. These variations in specificity are the result of interatomic contacts with local residues near the active sites of each subunit. Each catalytic β subunit also possesses a conserved lysine residue required for proteolysis.

Although the proteasome normally produces very short peptide fragments, in some cases these products are themselves biologically active and functional molecules. Certain transcription factors regulating the expression of specific genes, including one component of the mammalian complex NF-κB, are synthesized as inactive precursors whose ubiquitination and subsequent proteasomal degradation converts them to an active form. Such activity requires the proteasome to cleave the substrate protein internally: rather than processively degrading it from one terminus. It has been suggested that long loops on these proteins' surfaces serve as the proteasomal substrates and enter the central cavity, while the majority of the protein remains outside. Similar effects have been observed in yeast proteins; this mechanism of selective degradation is known as *regulated ubiquitin/proteasome dependent processing* (RUP).

Ubiquitin-independent degradation

Although most proteasomal substrates must be ubiquitinated before being degraded, there are some exceptions to this general rule, especially when the proteasome plays a normal role in the post-translational processing of the protein. The proteasomal activation of NF-κB by processing p105 into p50 via internal proteolysis is one major example. Some proteins that are hypothesized to be unstable due to intrinsically unstructured regions, are degraded in a ubiquitin-independent manner. The most well-known example of a ubiquitin-independent proteasome substrate is the enzyme ornithine decarboxylase. Ubiquitin-independent mechanisms targeting key cell cycle regulators such as p53 have also been reported, although p53 is also subject to ubiquitin-dependent degradation. Finally, structurally abnormal, misfolded, or highly oxidized proteins are also subject to ubiquitin-independent and 19S-independent degradation under conditions of cellular stress.

Evolution



The assembled complex of hslV (blue) and hslU (red) from *E. coli*. This complex of heat shock proteins is thought to resemble the ancestor of the modern proteasome.

The 20S proteasome is both ubiquitous and essential in eukaryotes. Some prokaryotes, including many archaea and the bacterial order Actinomycetales also share homologs of the 20S proteasome, whereas most bacteria possess heat shock genes hslV and hslU, whose gene products are a multimeric protease arranged in a two-layered ring and an ATPase. The hslV protein has been hypothesized to resemble the likely ancestor of the 20S proteasome. In general, HslV is not essential in bacteria, and not all bacteria possess it, whereas some protists possess both the 20S and the hslV systems.

Sequence analysis suggests that the catalytic β subunits diverged earlier in evolution than the predominantly structural α subunits. In bacteria that express a 20S proteasome, the β subunits have high sequence identity to archaeal and eukaryotic β subunits, whereas the α sequence identity is much lower. The presence of 20S proteasomes in bacteria may result from lateral gene transfer, while the diversification of subunits among eukaryotes is ascribed to multiple gene duplication events.

Cell cycle control

Cell cycle progression is controlled by ordered action of cyclin-dependent kinases (CDKs), activated by specific cyclins that demarcate phases of the cell cycle. Mitotic cyclins, which persist in the cell for only a few minutes, have one of the shortest life spans of all intracellular proteins. After a CDK-cyclin complex has performed its function, the associated cyclin is polyubiquitinated and destroyed by the proteasome, which provides directionality for the cell cycle. In particular, exit from mitosis requires the proteasome-dependent dissociation of the regulatory component cyclin B from the mitosis promoting factor complex. In vertebrate cells, "slippage" through the mitotic checkpoint leading to premature M phase exit can occur despite the delay of this exit by the spindle checkpoint.

Earlier cell cycle checkpoints such as post-restriction point check between G_1 phase and S phase similarly involve proteasomal degradation of cyclin A, whose ubiquitination is promoted by the anaphase promoting complex (APC), an E3 ubiquitin ligase. The APC and the Skp1/Cul1/F-box protein complex (SCF complex) are the two key regulators of cyclin degradation and checkpoint control; the SCF itself is regulated by the APC via ubiquitination of the adaptor protein, Skp2, which prevents SCF activity before the G1-S transition.

Individual components of the 19S particle have their own regulatory roles. Gankyrin, a recently identified oncoprotein, is one of the 19S subcomponents that also tightly binds the cyclin-dependent kinase CDK4 and plays a key role in recognizing ubiquitinated p53, via its affinity for the ubiquitin ligase MDM2. Gankyrin is anti-apoptotic and has been shown to be overexpressed in some tumor cell types such as hepatocellular carcinoma.

Regulation of plant growth

In plants, signaling by auxins, or phytohormones that order the direction and tropism of plant growth, induces the targeting of a class of transcription factor repressors known as Aux/IAA proteins for proteasomal degradation. These proteins are ubiquitinated by SCFTIR1, or SCF in complex with the auxin receptor TIR1. Degradation of Aux/IAA proteins derepresses transcription factors in the auxin-response factor (ARF) family and induces ARF-directed gene expression. The cellular consequences of ARF activation depend on the plant type and developmental stage, but are involved in directing growth in roots and leaf veins. The specific response to ARF derepression is thought to be mediated by specificity in the pairing of individual ARF and Aux/IAA proteins.

Apoptosis

Both internal and external signals can lead to the induction of apoptosis, or programmed cell death. The resulting deconstruction of cellular components is primarily carried out by specialized proteases known as caspases, but the proteasome also plays important and diverse roles in the apoptotic process. The involvement of the proteasome in this process is indicated by both the increase in protein ubiquitination, and of E1, E2, and E3 enzymes that is observed well in advance of apoptosis, during apoptosis, proteasomes localized to the nucleus have also been observed to translocate to outer membrane blebs characteristic of apoptosis.

Proteasome inhibition has different effects on apoptosis induction in different cell types. In general, the proteasome is not required for apoptosis, although inhibiting it is pro-apoptotic in most cell types that have been studied. Apoptosis is mediated through disrupting the regulated degradation of pro-growth cell cycle proteins. However, some cell lines — in particular, primary cultures of quiescent and differentiated cells such as thymocytes and neurons — are prevented from undergoing apoptosis on exposure to proteasome inhibitors. The mechanism for this effect is not clear, but is hypothesized to be specific to cells in quiescent states, or to result from the differential activity of the pro-apoptotic kinase JNK. The ability of proteasome inhibitors to induce apoptosis in rapidly dividing cells has been exploited in several recently developed chemotherapy agents such as bortezomib and salinosporamide A.

Response to cellular stress

In response to cellular stresses – such as infection, heat shock, or oxidative damage – heat shock proteins that identify misfolded or unfolded proteins and target them for proteasomal degradation are expressed. Both Hsp27 and Hsp90—chaperone proteins have been implicated in increasing the activity of the ubiquitin-proteasome system, though they are not direct participants in the process. Hsp70, on the other hand, binds exposed hydrophobic patches on the surface of misfolded proteins and recruits E3 ubiquitin ligases such as CHIP to tag the proteins for proteasomal degradation. The CHIP protein (carboxyl terminus of Hsp70-interacting protein) is itself regulated via inhibition of interactions between the E3 enzyme CHIP and its E2 binding partner.

Similar mechanisms exist to promote the degradation of oxidatively damaged proteins via the proteasome system. In particular, proteasomes localized to the nucleus are regulated by PARP and actively degrade inappropriately oxidized histones. Oxidized proteins, which often form large amorphous aggregates in the cell, can be degraded directly by the 20S core particle without the 19S regulatory cap and do not require ATP hydrolysis or tagging with ubiquitin. However, high levels of oxidative damage increases the degree of cross-linking between protein fragments, rendering the aggregates resistant to proteolysis. Larger numbers and sizes of such highly oxidized aggregates are associated with aging.

Dysregulation of the ubiquitin proteasome system may contribute to several neural diseases. It may lead to brain tumors such as astrocytomas. In some of the late-onset neurodegenerative diseases that share aggregation of misfolded proteins as a common feature, such as Parkinson's disease and Alzheimer's disease, large insoluble aggregates of misfolded proteins can form and then result in neurotoxicity, through mechanisms that are not yet well understood. Decreased proteasome activity has been suggested as a cause of aggregation and Lewy body formation in Parkinson's. This hypothesis is supported by the observation that yeast models of Parkinson's are more susceptible to toxicity from α-synuclein, the major protein component of Lewy bodies, under conditions of low proteasome activity. Impaired proteasomal activity may underlie cognitive disorders such as the autism spectrum disorders, and muscle and nerve diseases such as inclusion body myopathy.

Role in the immune system

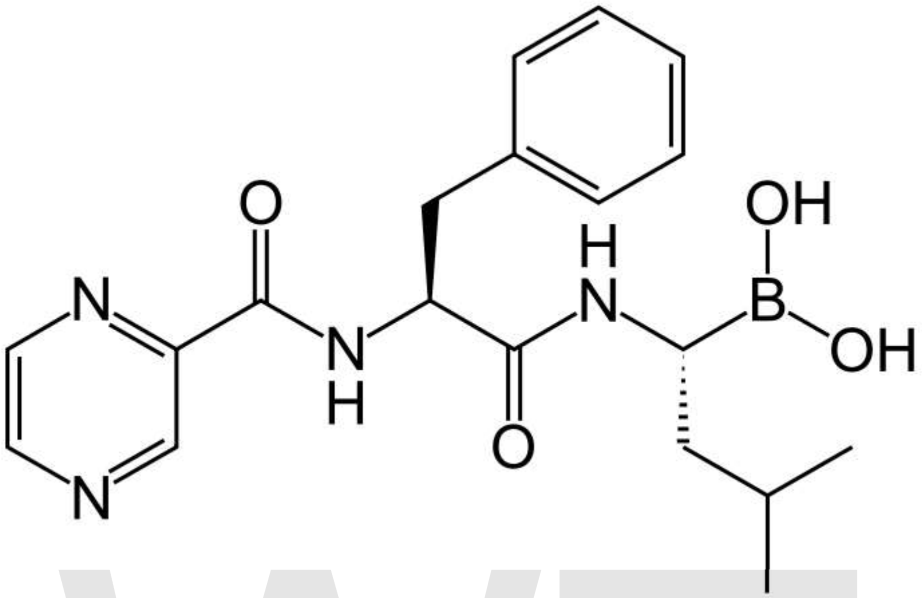
The proteasome plays a straightforward but critical role in the function of the adaptive immune system. Peptide antigens are displayed by the major histocompatibility complex class I (MHC) proteins on the surface of antigen-presenting cells. These peptides are products of proteasomal degradation of proteins originated by the invading pathogen. Although constitutively expressed proteasomes can participate in this process, a specialized complex composed of proteins whose expression is induced by interferon gamma produces peptides of the optimal size and composition for MHC binding. These proteins whose expression increases during the immune response include the 11S regulatory particle, whose main known biological role is regulating the production of MHC ligands, and specialized β subunits called β1i, β2i, and β5i with altered substrate specificity. The complex formed with the specialized β subunits is known as the *immunoproteasome*. Another β5i variant subunit, β5t, is expressed in the thymus, leading to a thymus-specific "thymoproteasome" whose function is as yet unclear.

The strength of MHC class I ligand binding is dependent on the composition of the ligand C-terminus, as peptides bind by hydrogen bonding and by close contacts with a region called the "B pocket" on the MHC surface. Many MHC class I alleles prefer hydrophobic C-terminal residues, and the immunoproteasome complex is more likely to generate hydrophobic C-termini.

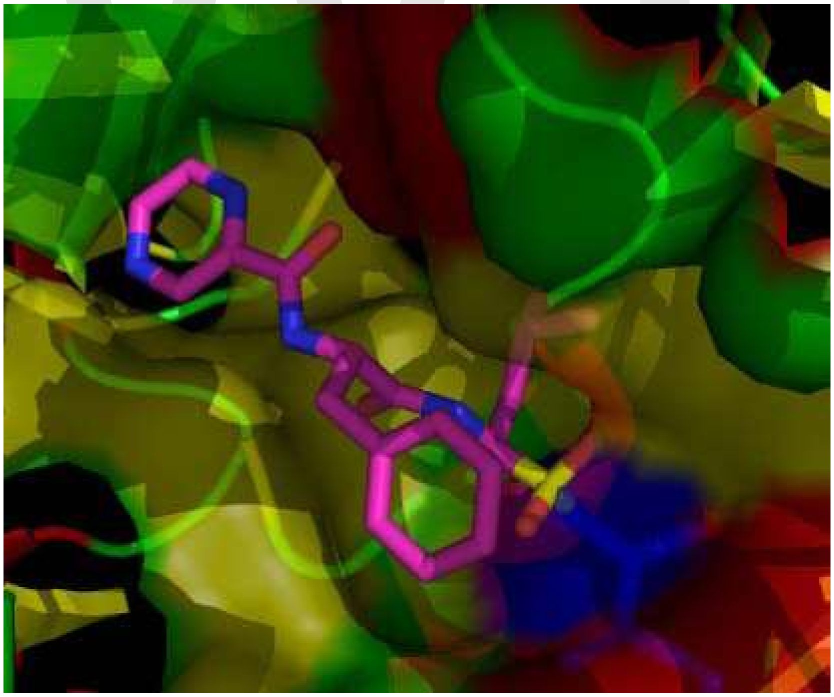
Due to its role in generating the activated form of NF-κB, an anti-apoptotic and pro-inflammatory regulator of cytokine expression, proteasomal activity has been linked to inflammatory and autoimmune diseases. Increased levels of proteasome activity correlate with disease activity and have been implicated in autoimmune diseases including systemic lupus erythematosus and rheumatoid arthritis.

The proteasome is also involved in Intracellular antibody-mediated proteolysis of antibody bound virions. In this neutralisation pathway, TRIM21 (a protein of the tripartite motif family) binds with immunoglobulin G to direct the virion to the proteasome where it is degraded.

Proteasome inhibitors



Chemical structure of bortezomib, a proteasome inhibitor used in chemotherapy that is particularly effective against multiple myeloma



Bortezomib bound to the core particle in a yeast proteasome. The bortezomib molecule is in the center colored by atom type (carbon = pink, nitrogen = blue, oxygen = red, boron = yellow), surrounded by the local protein surface. The blue patch is the catalytic threonine residue whose activity is blocked by the presence of bortezomib.

Proteasome inhibitors have effective anti-tumor activity in cell culture, inducing apoptosis by disrupting the regulated degradation of pro-growth cell cycle proteins. This approach of selectively inducing apoptosis in tumor cells has proven effective in animal models and human trials. Bortezomib, a molecule developed by Millennium Pharmaceuticals and marketed as Velcade, is the first proteasome inhibitor to reach clinical use as a chemotherapy agent. Bortezomib is used in the treatment of multiple myeloma. Notably, multiple myeloma has been observed to result in increased proteasome levels in blood serum that decrease to normal levels in response to successful chemotherapy. Studies in animals have indicated that bortezomib may also have clinically significant effects in pancreatic cancer. Preclinical and early clinical studies have been started to examine bortezomib's effectiveness in treating other B-cell-related cancers, particularly some types of non-Hodgkin's lymphoma.

The molecule ritonavir, marketed as Norvir, was developed as a protease inhibitor and used to target HIV infection. However, it has been shown to inhibit proteasomes as well as free proteases; to be specific, the chymotrypsin-like activity of the proteasome is inhibited by ritonavir, while the trypsin-like activity is somewhat enhanced. Studies in animal models suggest that ritonavir may have inhibitory effects on the growth of glioma cells.

Proteasome inhibitors have also shown promise in treating autoimmune diseases in animal models. For example, studies in mice bearing human skin grafts found a reduction in the size of lesions from psoriasis after treatment with a proteasome inhibitor. Inhibitors also show positive effects in rodent models of asthma.

Labeling and inhibition of the proteasome is also of interest in laboratory settings for both *in vitro* and *in vivo* study of proteasomal activity in cells. The most commonly used laboratory inhibitors are lactacystin, a natural product synthesized by *Streptomyces* bacteria, and peptide MG132. Fluorescent inhibitors have also been developed to specifically label the active sites of the assembled proteasome.

Chapter- 9

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.

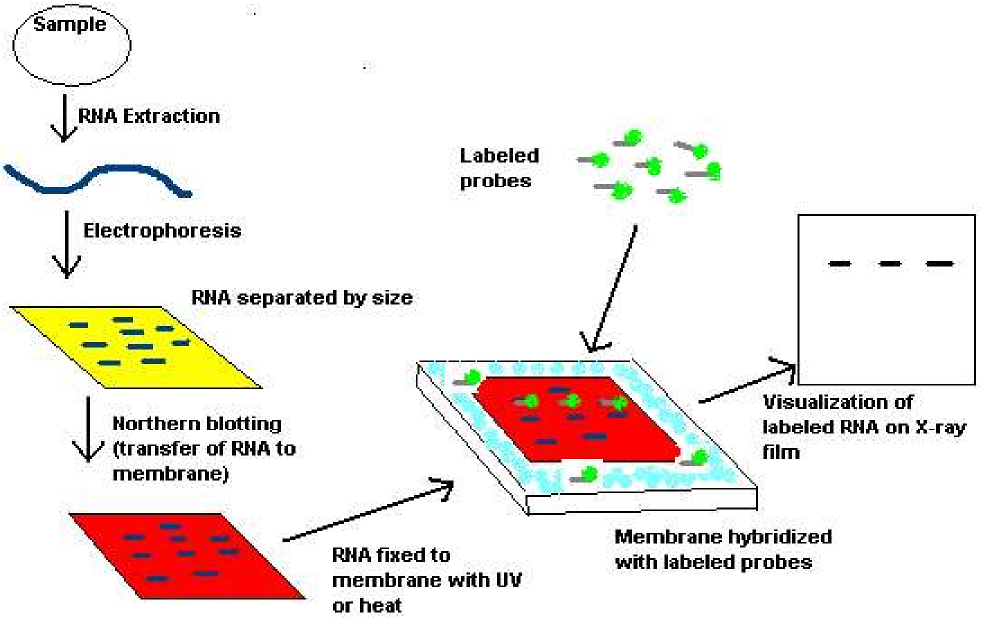
Network connectivity

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- 10

Northern Blot

The **northern blot** is a technique used in molecular biology research to study gene expression by detection of RNA (or isolated mRNA) in a sample.

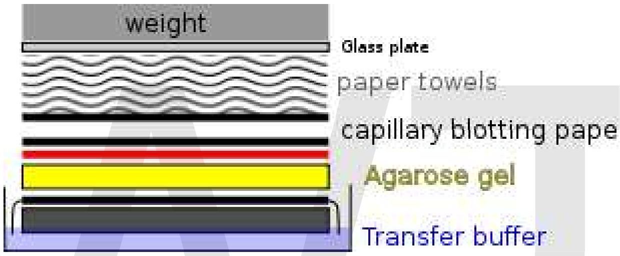

Sample
RNA Extraction
Electrophoresis
RNA separated by size
Northern blotting (transfer of RNA to membrane)
RNA fixed to membrane with UV or heat
Labeled probes
Membrane hybridized with labeled probes
Visualization of labeled RNA on X-ray film

Flow diagram outlining the general procedure for RNA detection by northern blotting

With northern blotting it is possible to observe cellular control over structure and function by determining the particular gene expression levels during differentiation, morphogenesis, as well as abnormal or diseased conditions. Northern blotting involves the use of electrophoresis to separate RNA samples by size, and detection with a hybridization probe complementary to part of or the entire target sequence. The term 'northern blot' actually refers specifically to the capillary transfer of RNA from the electrophoresis gel to the blotting membrane, however the entire process is commonly referred to as northern blotting. The northern blot technique was developed in 1977 by James Alwine, David Kemp, and George Stark at Stanford University. Northern blotting takes its name from its similarity to the first blotting technique, the Southern blot, named for biologist Edwin Southern. The major difference is that RNA, rather than DNA, is analyzed in the Northern blot.

Procedure

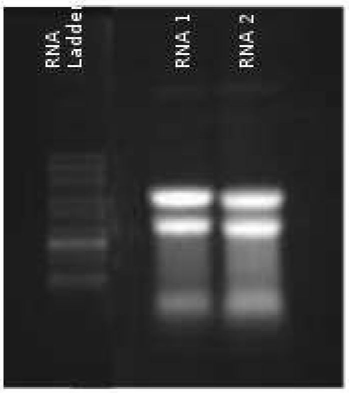
A general blotting procedure starts with extraction of total RNA from a homogenized tissue sample. The mRNA can then be isolated through the use of oligo (dT) cellulose chromatography to maintain only those RNAs with a poly(A) tail. RNA samples are then separated by gel electrophoresis. Since the gels are fragile and the probes are unable to enter the matrix, the RNA samples, now separated by size, are transferred to a nylon membrane through a capillary or vacuum blotting system.



Capillary blotting system setup for the transfer of RNA from an electrophoresis gel to a blotting membrane.

A nylon membrane with a positive charge is the most effective for use in northern blotting since the negatively charged nucleic acids have a high affinity for them. The transfer buffer used for the blotting usually contains formamide because it lowers the annealing temperature of the probe-RNA interaction preventing RNA degradation by high temperatures. Once the RNA has been transferred to the membrane it is immobilized through covalent linkage to the membrane by UV light or heat. After a probe has been labeled, it is hybridized to the RNA on the membrane. Experimental conditions that can affect the efficiency and specificity of hybridization include ionic strength, viscosity, duplex length, mismatched base pairs, and base composition. The membrane is washed to ensure that the probe has bound specifically and to avoid background signals from arising. The hybrid signals are then detected by X-ray film and can be quantified by densitometry. To create controls for comparison in a northern blot, samples not displaying the gene product of interest can be used after determination by microarrays or RT-PCR.

Gels



Formaldehyde gel (1%) with RNA samples run at 100V for 1 hour in 1x MOPS buffe

RNA run on a formaldehyde agarose gel to highlight the 28S (top band) and 18S (lower band) ribosomal subunits.

The RNA samples are most commonly separated on agarose gels containing formaldehyde as a denaturing agent for the RNA to limit secondary structure. The gels can be stained with ethidium bromide (EtBr) and viewed under UV light to observe the quality and quantity of RNA before blotting. Polyacrylamide gel electrophoeresis with urea can also be used in RNA separation but it is most commonly used for fragmented RNA or microRNAs. An RNA ladder is often run alongside the samples on an electrophoresis gel to observe the size of fragments obtained but in total RNA samples the ribosomal subunits can act as size markers. Since the large ribosomal subunit is 28S (approximately 5kb) and the small ribosomal subunit is 18S (approximately 2kb) two prominent bands will appear on the gel, the larger at close to twice the intensity of the smaller.

Probes

Probes for northern blotting are composed of nucleic acids with a complementary sequence to all or part of the RNA of interest, they can be DNA, RNA, or oligonucleotides with a minimum of 25 complementary bases to the target sequence. RNA probes (riboprobes) that are transcribed in vitro are able to withstand more rigorous washing steps preventing some of the background noise. Commonly cDNA is created with labelled primers for the RNA sequence of interest to act as the probe in the northern blot. The probes need to be labelled either with radioactive isotopes (^{32}P) or with chemiluminescence in which alkaline phosphatase or horseradish peroxidase breakdown chemiluminescent substrates producing a detectable emission of light. The chemiluminescent labelling can occur in two ways: either the probe is attached to the enzyme, or the probe is labelled with a ligand (e.g. biotin) for which the antibody (e.g. avidin or streptavidin) is attached to the enzyme. X-ray film can detect both the radioactive and chemiluminescent signals and many researchers prefer the chemiluminescent signals because they are faster, more sensitive, and reduce the health hazards that go along with radioactive labels. The same membrane can be probed up to five times without a significant loss of the target RNA.

Applications

Northern blotting allows one to observe a particular gene's expression pattern between tissues, organs, developmental stages, environmental stress levels, pathogen infection, and over the course of treatment. The technique has been used to show overexpression of oncogenes and downregulation of tumor-suppressor genes in cancerous cells when compared to 'normal' tissue, as well as the gene expression in the rejection of transplanted organs. If an upregulated gene is observed by an abundance of mRNA on the northern blot the sample can then be sequenced to determine if the gene is known to researchers or if it is a novel finding. The expression patterns obtained under given conditions can provide insight into the function of that gene. Since the RNA is first separated by size, if only one probe type is used variance in the level of each band on the membrane can provide insight into the size of the product, suggesting alternative splice products of the same gene or repetitive sequence motifs. The variance in size of a gene product can also indicate deletions or errors in transcript processing, by altering the probe target used along the known sequence it is possible to determine which region of the RNA is missing.

BlotBase is an online database publishing northern blots. BlotBase has over 700 published northern blots of human and mouse samples, in over 650 genes across more than 25 different tissue types. Northern blots can be searched by a blot ID, paper reference, gene identifier, or by tissue. The results of a search provide the blot ID, species, tissue, gene, expression level, blot image (if available), and links to the publication that the work originated from. This new database provides sharing of information between members of the science community that was not previously seen in northern blotting as it was in sequence analysis, genome determination, protein structure, etc.

Disadvantages and Advantages

Analysis of gene expression can be done by several different methods including RT-PCR, RNase protection assays, microarrays, serial analysis of gene expression (SAGE), as well as northern blotting. Microarrays are quite commonly used and are usually consistent with data obtained from northern blots; however, at times northern blotting is able to detect small changes in gene expression that microarrays cannot. The advantage that microarrays have over northern blots is that thousands of genes can be visualized at a time, while northern blotting is usually looking at one or a small number of genes.

A problem in northern blotting is often sample degradation by RNases (both endogenous to the sample and through environmental contamination), which can be avoided by proper sterilization of glassware and the use of RNase inhibitors such as DEPC (diethylpyrocarbonate). The chemicals used in most northern blots can be a risk to the researcher, since formaldehyde, radioactive material, ethidium bromide, DEPC, and UV light are all harmful under certain exposures. Compared to RT-PCR, northern blotting has a low sensitivity, but it also has a high specificity which is important to reduce false positive results.

The advantages of using northern blotting include the detection of RNA size, the observation of alternate splice products, the use of probes with partial homology, the quality and quantity of RNA can be measured on the gel prior to blotting, and the membranes can be stored and reprobed for years after blotting.

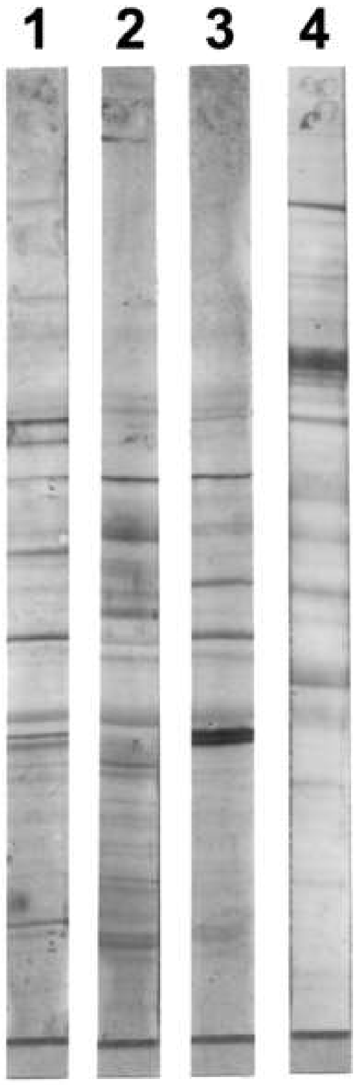
Reverse northern blot

A variant of the procedure known as the reverse northern blot is occasionally used. In this procedure, the substrate nucleic acid (that is affixed to the membrane) is a collection of isolated DNA fragments, and the probe is RNA extracted from a tissue and radioactively labelled.

The use of DNA microarrays that have come into widespread use in the late 1990s and early 2000s is more akin to the reverse procedure, in that they involve the use of isolated DNA fragments affixed to a substrate, and hybridization with a probe made from cellular RNA. Thus the reverse procedure, though originally uncommon, enabled northern analysis to evolve into gene expression profiling, in which many (possibly all) of the genes in an organism may have their expression monitored.

Chapter- 11

Western Blot



Western blot analysis of proteins separated by SDS-PAGE

The **Western blot** (alternatively, **protein immunoblot**) is an extremely useful analytical technique used to detect specific proteins in the given sample of tissue homogenate or extract. It uses gel electrophoresis to separate native or denatured proteins by the length of the polypeptide (denaturing conditions) or by the 3-D structure of the protein (native/ non-denaturing conditions). The proteins are then transferred to a membrane (typically nitrocellulose or PVDF), where they are probed (detected) using antibodies specific to the target protein.

There are now many reagent companies that specialize in providing antibodies (both monoclonal and polyclonal antibodies) against tens of thousands of different proteins. Commercial antibodies can be expensive, although the unbound antibody can be reused between experiments. This method is used in the fields of molecular biology, biochemistry, immunogenetics and other molecular biology disciplines.

Other related techniques include using antibodies to detect proteins in tissues and cells by immunostaining and enzyme-linked immunosorbent assay (ELISA).

The method originated from the laboratory of George Stark at Stanford. The name *Western blot* was given to the technique by W. Neal Burnette and is a play on the name Southern blot, a technique for DNA detection developed earlier by Edwin Southern. Detection of RNA is termed northern blotting and the detection of post-translational modification of protein is termed eastern blotting.

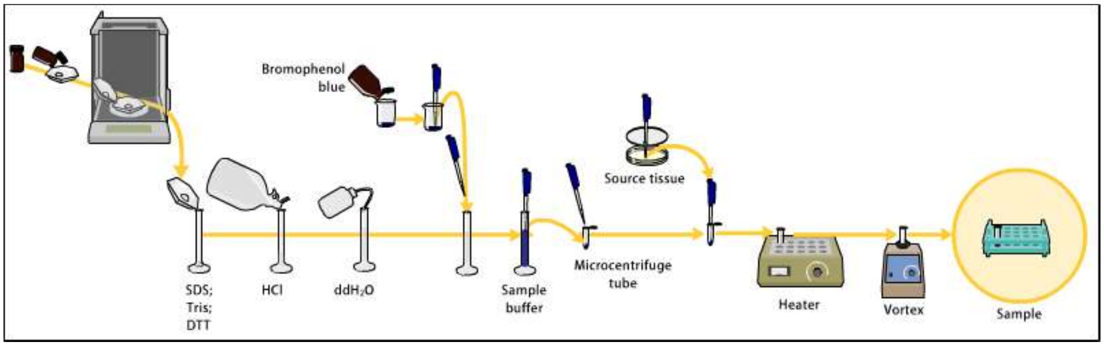
Steps in a Western blot

Tissue preparation

Samples may be taken from whole tissue or from cell culture. In most cases, solid tissues are first broken down mechanically using a blender (for larger sample volumes), using a homogenizer (smaller volumes), or by sonication. Cells may also be broken open by one of the above mechanical methods. However, bacteria, virus or environmental samples can be the source of protein and thus Western blotting is not restricted to cellular studies only.

Assorted detergents, salts, and buffers may be employed to encourage lysis of cells and to solubilize proteins. Protease and phosphatase inhibitors are often added to prevent the digestion of the sample by its own enzymes. Tissue preparation is often done at cold temperatures to avoid protein denaturing and degradation.

A combination of biochemical and mechanical techniques – including various types of filtration and centrifugation – can be used to separate different cell compartments and organelles.

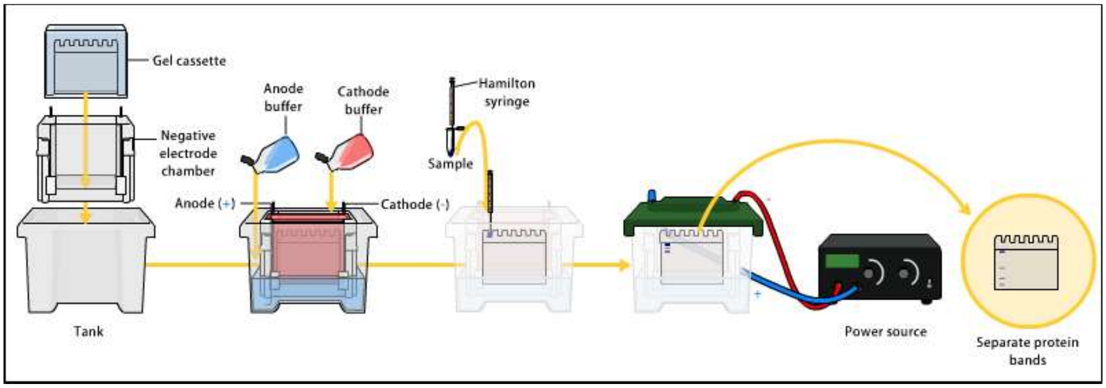


Gel electrophoresis

The proteins of the sample are separated using gel electrophoresis. Separation of proteins may be by isoelectric point (pI), molecular weight, electric charge, or a combination of these factors. The nature of the separation depends on the treatment of the sample and the nature of the gel. This is a very useful way to determine a protein.

By far the most common type of gel electrophoresis employs polyacrylamide gels and buffers loaded with sodium dodecyl sulfate (SDS). SDS-PAGE (SDS polyacrylamide gel electrophoresis) maintains polypeptides in a denatured state once they have been treated with strong reducing agents to remove secondary and tertiary structure (e.g. disulfide bonds [S-S] to sulfhydryl groups [SH and SH]) and thus allows separation of proteins by their molecular weight. Sampled proteins become covered in the negatively charged SDS and move to the positively charged electrode through the acrylamide mesh of the gel. Smaller proteins migrate faster through this mesh and the proteins are thus separated according to size (usually measured in kilodaltons, kDa). The concentration of acrylamide determines the resolution of the gel - the greater the acrylamide concentration the better the resolution of lower molecular weight proteins. The lower the acrylamide concentration the better the resolution of higher molecular weight proteins. Proteins travel only in one dimension along the gel for most blots.

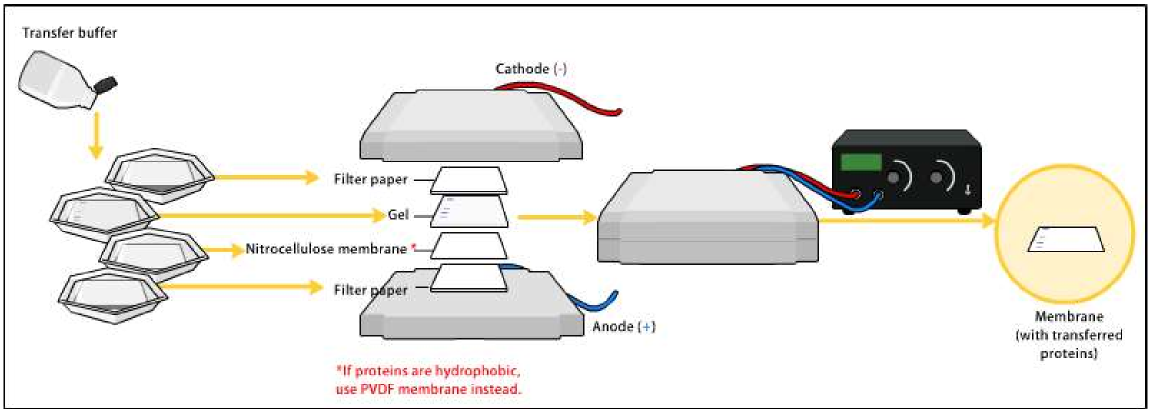
Samples are loaded into *wells* in the gel. One lane is usually reserved for a *marker* or *ladder*, a commercially available mixture of proteins having defined molecular weights, typically stained so as to form visible, coloured bands. When voltage is applied along the gel, proteins migrate into it at different speeds. These different rates of advancement (different *electrophoretic mobilities*) separate into *bands* within each *lane*.



It is also possible to use a two-dimensional (2-D) gel which spreads the proteins from a single sample out in two dimensions. Proteins are separated according to isoelectric point (pH at which they have neutral net charge) in the first dimension, and according to their molecular weight in the second dimension.

Transfer

In order to make the proteins accessible to antibody detection, they are moved from within the gel onto a membrane made of *nitrocellulose or polyvinylidene difluoride (PVDF*). The membrane is placed on top of the gel, and a stack of filter papers placed on top of that. The entire stack is placed in a buffer solution which moves up the paper by capillary action, bringing the proteins with it. Another method for transferring the proteins is called electroblotting and uses an electric current to pull proteins from the gel into the PVDF or nitrocellulose membrane. The protein move from within the gel onto the membrane while maintaining the organization they had within the gel. As a result of this "blotting" process, the proteins are exposed on a thin surface layer for detection (see below). Both varieties of membrane are chosen for their non-specific protein binding properties (i.e. binds all proteins equally well). Protein binding is based upon hydrophobic interactions, as well as charged interactions between the membrane and protein. Nitrocellulose membranes are cheaper than PVDF, but are far more fragile and do not stand up well to repeated probings.


Transfer buffer
Cathode (-)
Filter paper
Gel
Nitrocellulose membrane *
Filter paper
Anode (+)
Membrane (with transferred proteins)
*If proteins are hydrophobic, use PVDF membrane instead.

The uniformity and overall effectiveness of transfer of protein from the gel to the membrane can be checked by staining the membrane with Coomassie Brilliant Blue or Ponceau S dyes. Ponceau S is the more common of the two, due to Ponceau S's higher sensitivity and its water solubility makes it easier to subsequently destain and probe the membrane as described below.

Blocking

Since the membrane has been chosen for its ability to bind protein and as both antibodies and the target are proteins, steps must be taken to prevent interactions between the membrane and the antibody used for detection of the target protein. Blocking of non-specific binding is achieved by placing the membrane in a dilute solution of protein - typically 3-5% Bovine serum albumin (BSA) or non-fat dry milk (both are inexpensive) in Tris-Buffered Saline (TBS), with a minute percentage of detergent such as Tween 20 or Triton X-100. The protein in the dilute solution attaches to the membrane in all places where the target proteins have not attached. Thus, when the antibody is added, there is no room on the membrane for it to attach other than on the binding sites of the specific target protein. This reduces "noise" in the final product of the Western blot, leading to clearer results, and eliminates false positives.

Detection

During the detection process the membrane is "probed" for the protein of interest with a modified antibody which is linked to a reporter enzyme, which when exposed to an appropriate substrate drives a colourimetric reaction and produces a colour. For a variety of reasons, this traditionally takes place in a two-step process, although there are now one-step detection methods available for certain applications.

Two steps

- Primary antibody

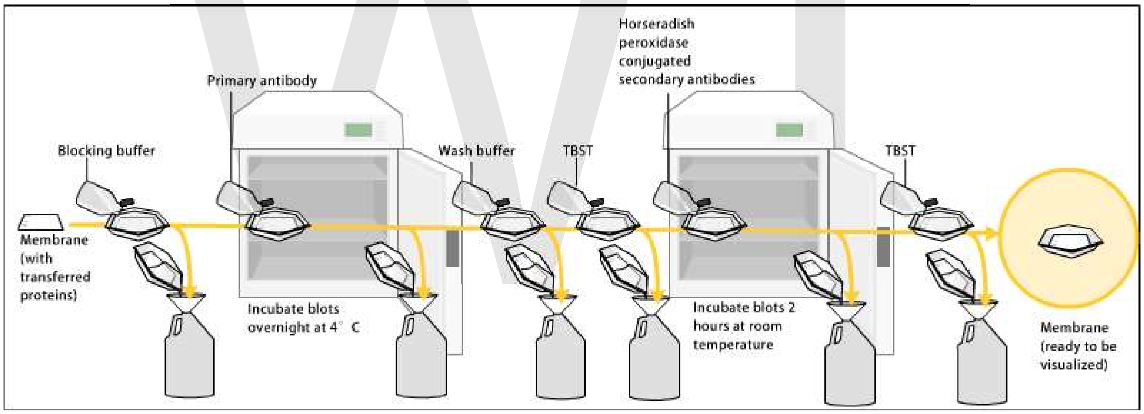
Antibodies are generated when a host species or immune cell culture is exposed to the protein of interest (or a part thereof). Normally, this is part of the immune response, whereas here they are harvested and used as sensitive and specific detection tools that bind the protein directly.

After blocking, a dilute solution of primary antibody (generally between 0.5 and 5 micrograms/mL) is incubated with the membrane under gentle agitation. Typically, the solution is composed of buffered saline solution with a small percentage of detergent, and sometimes with powdered milk or BSA. The antibody solution and the membrane can be sealed and incubated together for anywhere from 30 minutes to overnight. It can also be incubated at different temperatures, with warmer temperatures being associated with more binding, both specific (to the target protein, the "signal") and non-specific ("noise").

- Secondary antibody

After rinsing the membrane to remove unbound primary antibody, the membrane is exposed to another antibody, directed at a species-specific portion of the primary antibody. Antibodies come from animal sources (or animal sourced hybridoma cultures); an anti-mouse secondary will bind to almost any mouse-sourced primary antibody, which allows some cost savings by allowing an entire lab to share a single source of mass-produced antibody, and provides far more consistent results. This is known as a secondary antibody, and due to its targeting properties, tends to be referred to as "anti-mouse," "anti-goat," etc. The secondary antibody is usually linked to biotin or to a reporter enzyme such as alkaline phosphatase or horseradish peroxidase. This means that several secondary antibodies will bind to one primary antibody and enhance the signal.

Most commonly, a horseradish peroxidase-linked secondary is used to cleave a chemiluminescent agent, and the reaction product produces luminescence in proportion to the amount of protein. A sensitive sheet of photographic film is placed against the membrane, and exposure to the light from the reaction creates an image of the antibodies bound to the blot. A cheaper but less sensitive approach utilizes a 4-chloronaphthol stain with 1% hydrogen peroxide; reaction of peroxide radicals with 4-chloronaphthol produces a dark brown stain that can be photographed without using specialized photographic film.


Horseradish peroxidase conjugated secondary antibodies
Primary antibody
Blocking buffer
Wash buffer
TBST
TBST
Membrane (with transferred proteins)
Incubate blots overnight at 4° C
Incubate blots 2 hours at room temperature
Membrane (ready to be visualized)

As with the ELISPOT and ELISA procedures, the enzyme can be provided with a substrate molecule that will be converted by the enzyme to a colored reaction product that will be visible on the membrane.

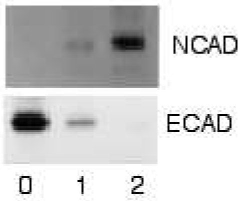
Another method of secondary antibody detection utilizes a near-infrared (NIR) fluorophore-linked antibody. Light produced from the excitation of a fluorescent dye is static, making fluorescent detection a more precise and accurate measure of the difference in signal produced by labeled antibodies bound to proteins on a Western blot. Proteins can be accurately quantified because the signal generated by the different amounts of proteins on the membranes is measured in a static state, as compared to chemiluminescence, in which light is measured in a dynamic state.

A third alternative is to use a radioactive label rather than an enzyme coupled to the secondary antibody, such as labeling an antibody-binding protein like *Staphylococcus*

Protein A or Streptavidin with a radioactive isotope of iodine. Since other methods are safer, quicker, and cheaper, this method is now rarely used; however, an advantage of this approach is the sensitivity of auto-radiography based imaging, which enables highly accurate protein quantification when combined with optical software (e.g. Optiquant).

One step

Historically, the probing process was performed in two steps because of the relative ease of producing primary and secondary antibodies in separate processes. This gives researchers and corporations huge advantages in terms of flexibility, and adds an amplification step to the detection process. Given the advent of high-throughput protein analysis and lower limits of detection, however, there has been interest in developing one-step probing systems that would allow the process to occur faster and with less consumables. This requires a probe antibody which both recognizes the protein of interest and contains a detectable label, probes which are often available for known protein tags. The primary probe is incubated with the membrane in a manner similar to that for the primary antibody in a two-step process, and then is ready for direct detection after a series of wash steps.



Western blot using radioactive detection system

Analysis

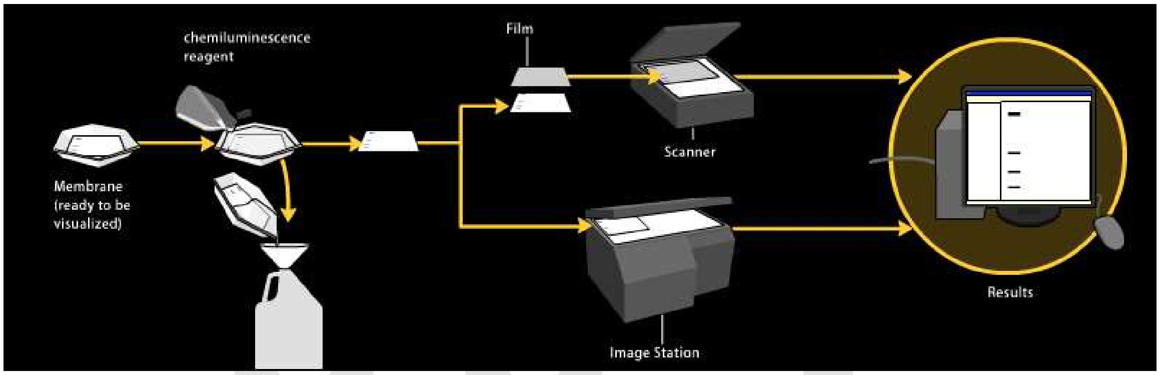
After the unbound probes are washed away, the Western blot is ready for detection of the probes that are labeled and bound to the protein of interest. In practical terms, not all Westerns reveal protein only at one band in a membrane. Size approximations are taken by comparing the stained bands to that of the marker or ladder loaded during electrophoresis. The process is repeated for a structural protein, such as actin or tubulin, that should not change between samples. The amount of target protein is indexed to the structural protein to control between groups. This practice ensures correction for the amount of total protein on the membrane in case of errors or incomplete transfers.

Colorimetric detection

The colorimetric detection method depends on incubation of the Western blot with a substrate that reacts with the reporter enzyme (such as peroxidase) that is bound to the secondary antibody. This converts the soluble dye into an insoluble form of a different color that precipitates next to the enzyme and thereby stains the membrane. Development of the blot is then stopped by washing away the soluble dye. Protein levels are evaluated through densitometry (how intense the stain is) or spectrophotometry.

Chemiluminescent detection

Chemiluminescent detection methods depend on incubation of the Western blot with a substrate that will luminesce when exposed to the reporter on the secondary antibody. The light is then detected by photographic film, and more recently by CCD cameras which capture a digital image of the Western blot. The image is analysed by densitometry, which evaluates the relative amount of protein staining and quantifies the results in terms of optical density. Newer software allows further data analysis such as molecular weight analysis if appropriate standards are used.



Radioactive detection

Radioactive labels do not require enzyme substrates, but rather allow the placement of medical X-ray film directly against the Western blot which develops as it is exposed to the label and creates dark regions which correspond to the protein bands of interest. The importance of radioactive detections methods is declining, because it is very expensive, health and safety risks are high, and ECL (enhanced chemiluminescence) provides a useful alternative.

Fluorescent detection

The fluorescently labeled probe is excited by light and the emission of the excitation is then detected by a photosensor such as CCD camera equipped with appropriate emission filters which captures a digital image of the Western blot and allows further data analysis such as molecular weight analysis and a quantitative Western blot analysis. Fluorescence is considered to be among the most sensitive detection methods for blotting analysis.

Secondary probing

One major difference between nitrocellulose and PVDF membranes relates to the ability of each to support "stripping" antibodies off and reusing the membrane for subsequent antibody probes. While there are well-established protocols available for stripping nitrocellulose membranes, the sturdier PVDF allows for easier stripping, and for more reuse before background noise limits experiments. Another difference is that, unlike nitrocellulose, PVDF must be soaked in 95% ethanol, isopropanol or methanol before use. PVDF membranes also tend to be thicker and more resistant to damage during use.

2-D gel electrophoresis

2-dimensional SDS-PAGE uses the principles and techniques outlined above. 2-D SDS-PAGE, as the name suggests, involves the migration of polypeptides in 2 dimensions. For example, in the first dimension polypeptides are separated according to isoelectric point, while in the second dimension polypeptides are separated according to their molecular weight. The isoelectric point of a given protein is determined by the relative number of positively (e.g. lysine and arginine) and negatively (e.g. glutamate and aspartate) charged amino acids, with negatively charged amino acids contributing to a high isoelectric point and positively charged amino acids contributing to a low isoelectric point. Samples could also be separated first under nonreducing conditions using SDS-PAGE and under reducing conditions in the second dimension, which breaks apart disulfide bonds that hold subunits together. SDS-PAGE might also be coupled with urea-PAGE for a 2-dimensional gel.

In principle, this method allows for the separation of all cellular proteins on a single large gel. A major advantage of this method is that it often distinguishes between different isoforms of a particular protein - e.g. a protein that has been phosphorylated (by addition of a negatively charged group). Proteins that have been separated can be cut out of the gel and then analysed by mass spectrometry, which identifies the protein.

Medical diagnostic applications

- The confirmatory HIV test employs a Western blot to detect anti-HIV antibody in a human serum sample. Proteins from known HIV-infected cells are separated and blotted on a membrane as above. Then, the serum to be tested is applied in the primary antibody incubation step; free antibody is washed away, and a secondary anti-human antibody linked to an enzyme signal is added. The stained bands then indicate the proteins to which the patient's serum contains antibody.
- A Western blot is also used as the definitive test for Bovine spongiform encephalopathy (BSE, commonly referred to as 'mad cow disease').
- Some forms of Lyme disease testing employ Western blotting.
- Western blot can also be used as a confirmatory test for Hepatitis B infection.
- In veterinary medicine, Western blot is sometimes used to confirm FIV+ status in cats

Chapter- 12

DNA Sequencing

The term **DNA sequencing** refers to sequencing methods for determining the order of the nucleotide bases—adenine, guanine, cytosine, and thymine—in a molecule of DNA.

Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA sequencing, and in numerous applied fields such as diagnostic, biotechnology, forensic biology and biological systematics. The advent of DNA sequencing has significantly accelerated biological research and discovery. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of the human genome, in the Human Genome Project. Related projects, often by scientific collaboration across continents, have generated the complete DNA sequences of many animal, plant, and microbial genomes.

The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of dye-based sequencing methods with automated analysis, DNA sequencing has become easier and orders of magnitude faster.

History

RNA sequencing was one of the earliest forms of nucleotide sequencing. The major landmark of RNA sequencing is the sequence of the first complete gene and the complete genome of Bacteriophage MS2, identified and published by Walter Fiers and his coworkers at the University of Ghent (Ghent, Belgium), between 1972 and 1976.

Prior to the development of rapid DNA sequencing methods in the early 1970s by Frederick Sanger at the University of Cambridge, in England and Walter Gilbert and Allan Maxam at Harvard, a number of laborious methods were used. For instance, in 1973, Gilbert and Maxam reported the sequence of 24 basepairs using a method known as wandering-spot analysis.

The chain-termination method developed by Sanger and coworkers in 1975 soon became the method of choice, owing to its relative ease and reliability.

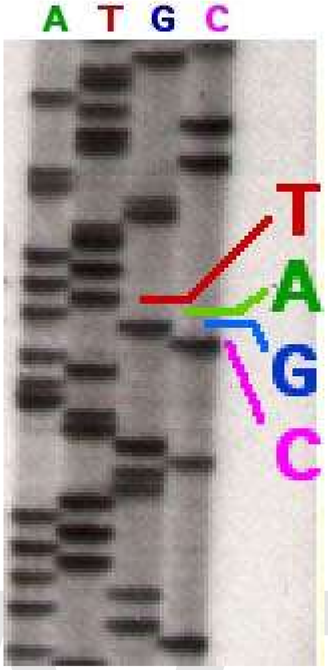
Maxam–Gilbert sequencing

In 1976–1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases. Although Maxam and Gilbert published their chemical sequencing method two years after the ground-breaking paper of Sanger and Coulson on plus-minus sequencing, Maxam–Gilbert sequencing rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each read start be cloned for production of single-stranded DNA. However, with the improvement of the chain-termination method (see below), Maxam-Gilbert sequencing has fallen out of favour due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties with scale-up.

The method requires radioactive labeling at one 5' end of the DNA (typically by a kinase reaction using gamma-^{32}P ATP) and purification of the DNA fragment to be sequenced. Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). For example, the purines (A+G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C+T) are methylated using hydrazine. The addition of salt (sodium chloride) to the hydrazine reaction inhibits the methylation of thymine for the C-only reaction. The modified DNAs are then cleaved by hot piperidine at the position of the modified base. The concentration of the modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule. The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabeled DNA fragment, from which the sequence may be inferred.

Also sometimes known as "chemical sequencing", this method originated in the study of DNA-protein interactions (DNase I footprinting) and nucleic acid structure, and within these it still has important applications.

Chain-termination methods



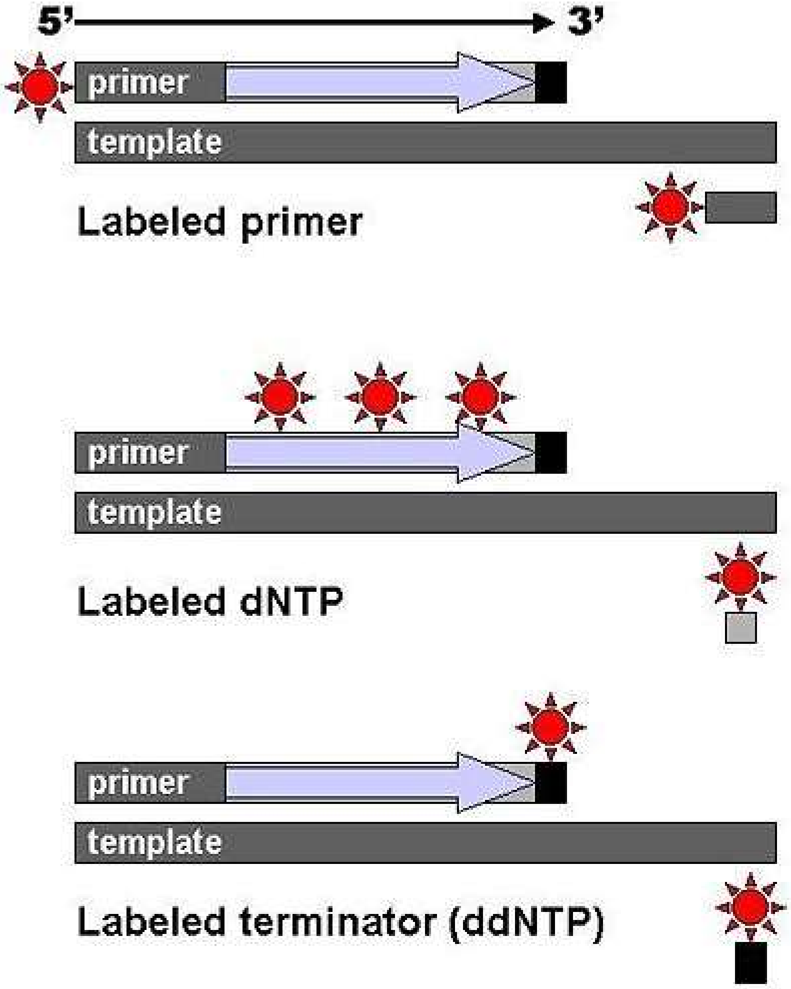
Part of a radioactively labelled sequencing gel

Because the chain-terminator method (or Sanger method after its developer Frederick Sanger) is more efficient and uses fewer toxic chemicals and lower amounts of radioactivity than the method of Maxam and Gilbert, it rapidly became the method of choice. The key principle of the Sanger method was the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.

The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotidephosphates (dNTPs), and modified nucleotides (dideoxyNTPs) that terminate DNA strand elongation. These ddNTPs will also be radioactively or fluorescently labelled for detection in automated sequencing machines. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) which are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length.

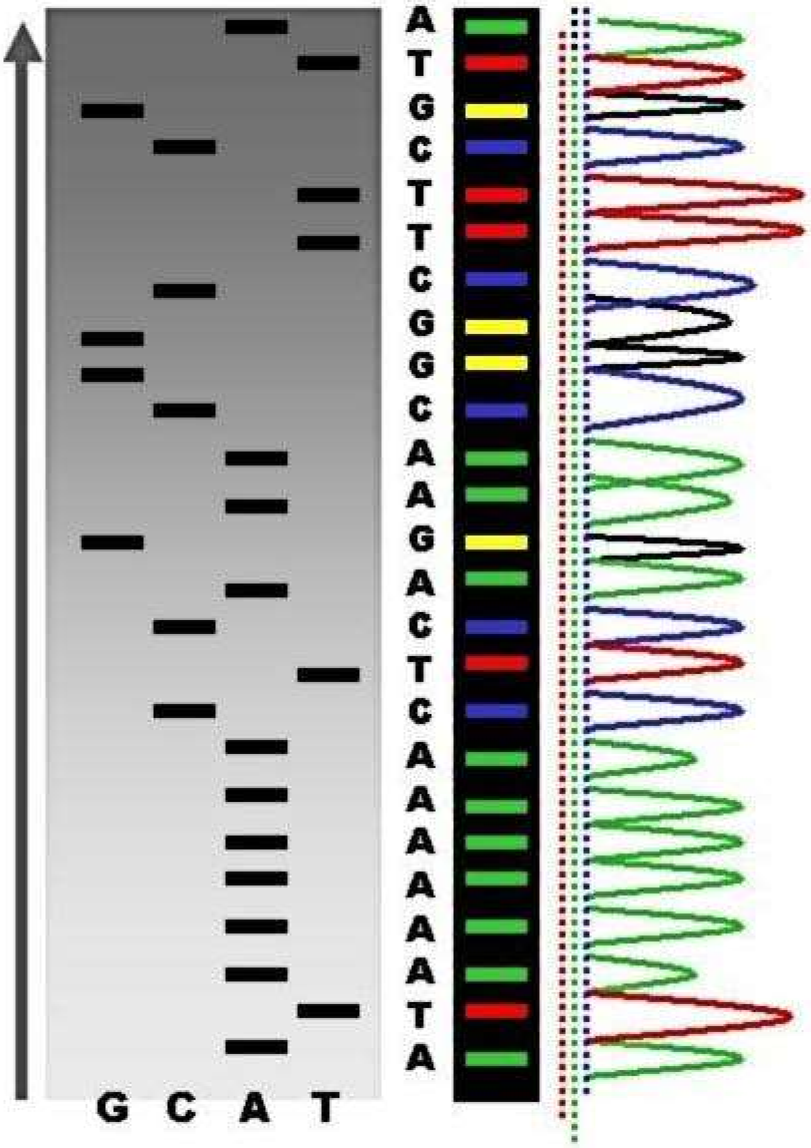
The newly synthesized and labelled DNA fragments are heat denatured, and separated by size (with a resolution of just one nucleotide) by gel electrophoresis on a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C); the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image. In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to

DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.



DNA fragments are labelled with a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.

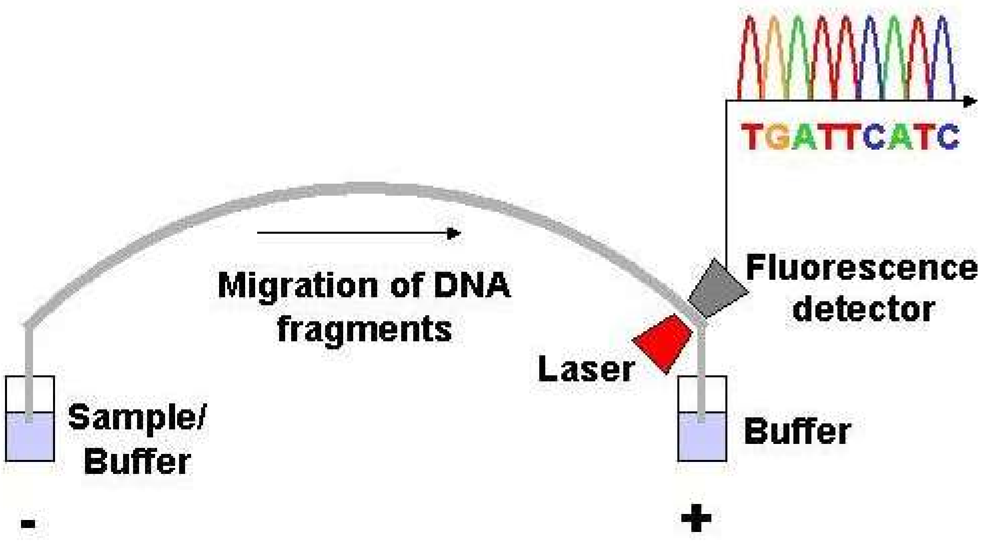
Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radiolabelling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation. The later development by Leroy Hood and coworkers of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.



Sequence ladder by radioactive sequencing compared to fluorescent peaks

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for sequencing, pre-aliquoted and ready to use. Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.

Dye-terminator sequencing



Capillary electrophoresis

Dye-terminator sequencing utilizes labelling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labelled with fluorescent dyes, each of which emit light at different wavelengths.

Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Its limitations include dye effects due to differences in the incorporation of the dye-labelled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis.

This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs". The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.

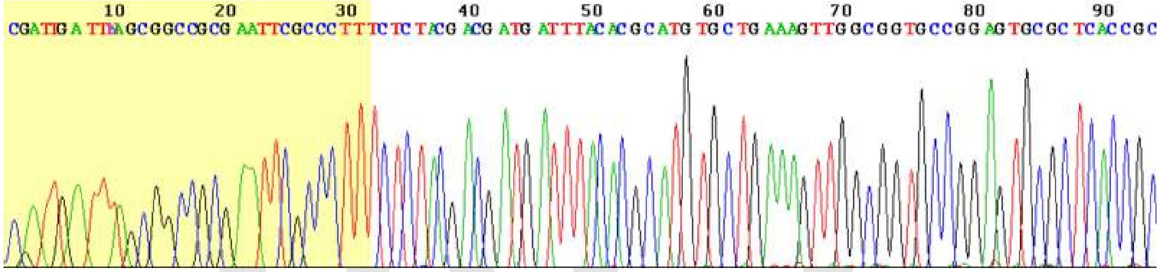
Challenges

Common challenges of DNA sequencing include poor quality in the first 15–40 bases of the sequence and deteriorating quality of sequencing traces after 700–900 bases. Base calling software typically gives an estimate of quality to aid in quality trimming.

In cases where DNA fragments are cloned before sequencing, the resulting sequence may contain parts of the cloning vector. In contrast, PCR-based cloning and emerging sequencing technologies based on pyrosequencing often avoid using cloning vectors. Recently, one-step Sanger sequencing (combined amplification and sequencing) methods such as Ampliseq and SeqSharp have been developed that allow rapid sequencing of target genes without cloning or prior amplification.

Current methods can directly sequence only relatively short (300–1000 nucleotides long) DNA fragments in a single reaction. The main obstacle to sequencing DNA fragments above this size limit is insufficient power of separation for resolving large DNA fragments that differ in length by only one nucleotide. In all cases the use of a primer with a free 5' end is essential.

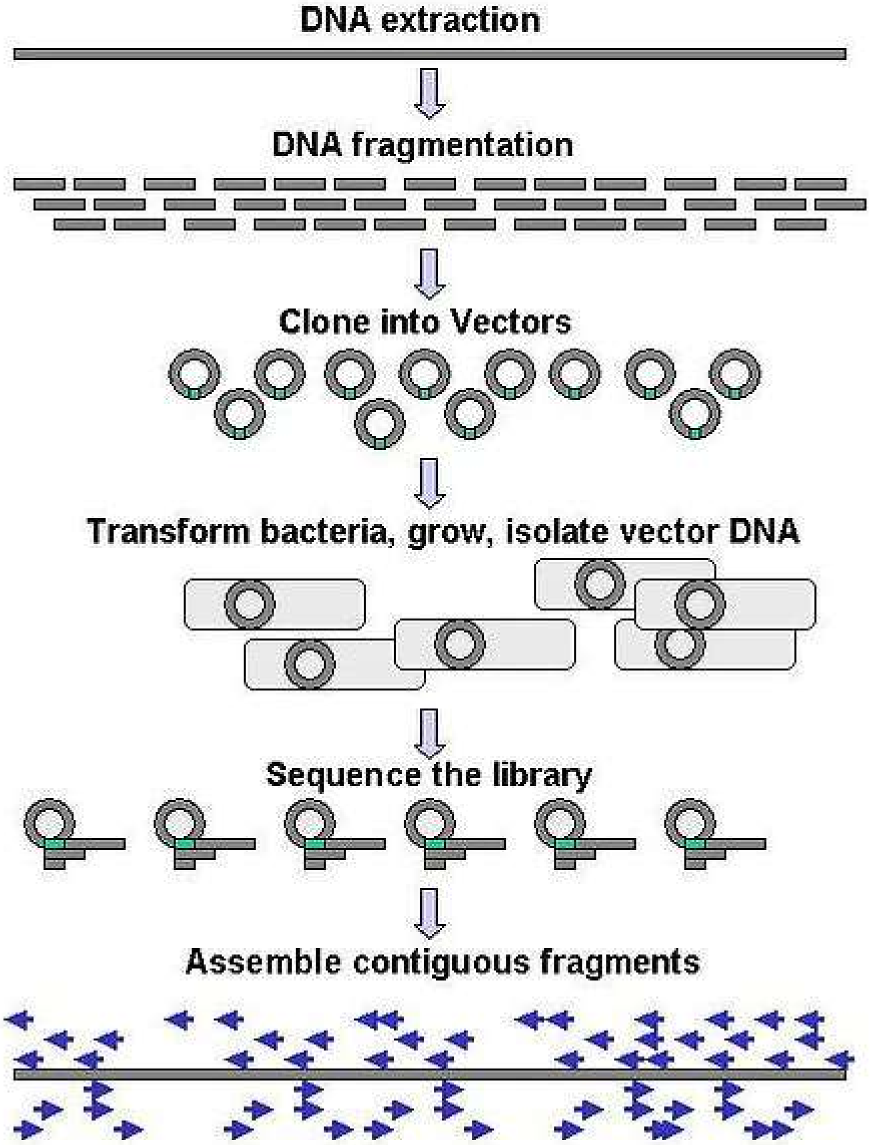
Automation and sample preparation



View of the start of an example dye-terminator read

Automated DNA-sequencing instruments (DNA sequencers) can sequence up to 384 DNA samples in a single batch (run) in up to 24 runs a day. DNA sequencers carry out capillary electrophoresis for size separation, detection and recording of dye fluorescence, and data output as fluorescent peak trace chromatograms. Sequencing reactions by thermocycling, cleanup and re-suspension in a buffer solution before loading onto the sequencer are performed separately. A number of commercial and non-commercial software packages can trim low-quality DNA traces automatically. These programs score the quality of each peak and remove low-quality base peaks (generally located at the ends of the sequence). The accuracy of such algorithms is below visual examination by a human operator, but sufficient for automated processing of large sequence data sets.

Amplification and clonal selection


DNA extraction
DNA fragmentation
Clone into Vectors
Transform bacteria, grow, isolate vector DNA
Sequence the library
Assemble contiguous fragments

Genomic DNA is fragmented into random pieces and cloned as a bacterial library. DNA from individual bacterial clones is sequenced and the sequence is assembled by using overlapping DNA regions.

Large-scale sequencing aims at sequencing very long DNA pieces, such as whole chromosomes. Common approaches consist of cutting (with restriction enzymes) or shearing (with mechanical forces) large DNA fragments into shorter DNA fragments. The fragmented DNA is cloned into a DNA vector, and amplified in *Escherichia coli*. Short DNA fragments purified from individual bacterial colonies are individually sequenced and assembled electronically into one long, contiguous sequence.

This method does not require any pre-existing information about the sequence of the DNA and is referred to as *de novo* sequencing. Gaps in the assembled sequence may be filled by primer walking. The different strategies have different tradeoffs in speed and accuracy; *shotgun methods* are often used for sequencing large genomes, but its assembly is complex and difficult, particularly with sequence repeats often causing gaps in genome assembly.

Most sequencing approaches use an *in vitro* cloning step to amplify individual DNA molecules, because their molecular detection methods are not sensitive enough for single molecule sequencing. Emulsion PCR isolates individual DNA molecules along with primer-coated beads in aqueous droplets within an oil phase. Polymerase chain reaction (PCR) then coats each bead with clonal copies of the DNA molecule followed by immobilization for later sequencing. Emulsion PCR is used in the methods by Marguilis et al. (commercialized by 454 Life Sciences), Shendure and Porreca et al. (also known as "Polony sequencing") and SOLiD sequencing, (developed by Agencourt, now Applied Biosystems).

Another method for *in vitro* clonal amplification is *bridge PCR*, where fragments are amplified upon primers attached to a solid surface, used in the Illumina Genome Analyzer. The single-molecule method developed by Stephen Quake's laboratory (later commercialized by Helicos) is an exception: it uses bright fluorophores and laser excitation to detect pyrosequencing events from individual DNA molecules fixed to a surface, eliminating the need for molecular amplification.

High-throughput sequencing

The high demand for low-cost sequencing has driven the development of high-throughput sequencing technologies that parallelize the sequencing process, producing thousands or millions of sequences at once. High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods.

Lynx Therapeutics' Massively Parallel Signature Sequencing (MPSS)

The first of the "next-generation" sequencing technologies, MPSS was developed in 1990s at Lynx Therapeutics, a company founded in 1992 by Sidney Brenner and Sam Eletr. MPSS was a bead-based method that used a complex approach of adapter ligation followed by adapter decoding, reading the sequence in increments of four nucleotides; this method made it susceptible to sequence-specific bias or loss of specific sequences. Because the technology was so complex, MPSS was only performed 'in-house' by Lynx Therapeutics and no machines were sold; when the merger with Solexa later lead to the development of sequencing-by-synthesis, a more simple approach with numerous advantages, MPSS became obsolete. However, the essential properties of the MPSS output were typical of later "next-gen" data types, including hundreds of thousands of short DNA sequences. In the case of MPSS, these were typically used for sequencing cDNA for measurements of gene expression levels. Lynx Therapeutics merged with Solexa in 2004, and this company was later purchased by Illumina.

Polony Sequencing

Polony sequencing, developed in George Church's lab at Harvard, was among the first next-generation sequencing systems used to sequence a full genome in 2005. It combined an in vitro paired-tag library with emulsion PCR, an automated microscope, and ligation-based sequencing chemistry to sequence an E. coli genome at an accuracy of > 99.9999% and a cost approximately 1/10th that of Sanger sequencing. The technology was licensed to Agencourt Biosciences, subsequently spun out into Agencourt Personal Genomics, and ultimately incorporated into the Applied Biosystems SOLiD platform.

454 pyrosequencing

A parallelized version of pyrosequencing was developed by 454 Life Sciences. The method amplifies DNA inside water droplets in an oil solution (emulsion PCR), with each droplet containing a single DNA template attached to a single primer-coated bead that then forms a clonal colony. The sequencing machine contains many picolitre-volume wells each containing a single bead and sequencing enzymes. Pyrosequencing uses luciferase to generate light for detection of the individual nucleotides added to the nascent DNA, and the combined data are used to generate sequence read-outs. This technology provides intermediate read length and price per base compared to Sanger sequencing on one end and Solexa and SOLiD on the other. 454 Life Sciences has since been acquired by Roche Diagnostics.

Illumina (Solexa) sequencing

Solexa, now part of Illumina developed a sequencing technology based on reversible dye-terminators. DNA molecules are first attached to primers on a slide and amplified so that local clonal colonies are formed (bridge amplification). Four types of ddNTPs are added, and non-incorporated nucleotides are washed away. Unlike pyrosequencing, the DNA can only be extended one nucleotide at a time. A camera takes images of the fluorescently labeled nucleotides then the dye along with the terminal 3' blocker is chemically removed from the DNA, allowing a next cycle.

SOLiD sequencing

Applied Biosystems' SOLiD technology employs sequencing by ligation. Here, a pool of all possible oligonucleotides of a fixed length are labeled according to the sequenced position. Oligonucleotides are annealed and ligated; the preferential ligation by DNA ligase for matching sequences results in a signal informative of the nucleotide at that position. Before sequencing, the DNA is amplified by emulsion PCR. The resulting bead, each containing only copies of the same DNA molecule, are deposited on a glass slide. The result is sequences of quantities and lengths comparable to Illumina sequencing.

Future methods

Sequencing by hybridization is a non-enzymatic method that uses a DNA microarray. A single pool of DNA whose sequence is to be determined is fluorescently labeled and hybridized to an array containing known sequences. Strong hybridization signals from a given spot on the array identifies its sequence in the DNA being sequenced. Mass spectrometry may be used to determine mass differences between DNA fragments produced in chain-termination reactions.

DNA sequencing methods currently under development include labeling the DNA polymerase, reading the sequence as a DNA strand transits through nanopores, and microscopy-based techniques, such as AFM or electron microscopy that are used to identify the positions of individual nucleotides within long DNA fragments (>5,000 bp) by nucleotide labeling with heavier elements (e.g., halogens) for visual detection and recording.

In microfluidic Sanger sequencing the entire thermocycling amplification of DNA fragments as well as their separation by electrophoresis is done on a single glass wafer (approximately 10 cm in diameter) thus reducing the reagent usage as well as cost. In some instances researchers have shown that they can increase the throughput of conventional sequencing through the use of microchips. Research will still need to be done in order to make this use of technology effective.

In October 2006, the X Prize Foundation established an initiative to promote the development of full genome sequencing technologies, called the Archon X Prize, intending to award $10 million to "the first Team that can build a device and use it to sequence 100 human genomes within 10 days or less, with an accuracy of no more than one error in every 100,000 bases sequenced, with sequences accurately covering at least 98% of the genome, and at a recurring cost of no more than $10,000 (US) per genome."

Each year NHGRI promotes grants for new research and developments in genomics. 2010 grants and 2011 candidates include continuing work in microfluidic, polony and base-heavy sequencing methodologies

Major landmarks in DNA sequencing

- 1953 Discovery of the structure of the DNA double helix.
- 1972 Development of recombinant DNA technology, which permits isolation of defined fragments of DNA; prior to this, the only accessible samples for sequencing were from bacteriophage or virus DNA.
- 1977 The first complete DNA genome to be sequenced is that of bacteriophage φX174.

- 1977 Allan Maxam and Walter Gilbert publish "DNA sequencing by chemical degradation". Frederick Sanger, independently, publishes "DNA sequencing with chain-terminating inhibitors".

- 1984 Medical Research Council scientists decipher the complete DNA sequence of the Epstein-Barr virus, 170 kb.

- 1986 Leroy E. Hood's laboratory at the California Institute of Technology and Smith announce the first semi-automated DNA sequencing machine.

- 1987 Applied Biosystems markets first automated sequencing machine, the model ABI 370.

- 1990 The U.S. National Institutes of Health (NIH) begins large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* (at US$0.75/base).

- 1991 Sequencing of human expressed sequence tags begins in Craig Venter's lab, an attempt to capture the coding fraction of the human genome.

- 1995 Craig Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) publish the first complete genome of a free-living organism, the bacterium Haemophilus influenzae. The circular chromosome contains 1,830,137 bases and its publication in the journal Science marks the first use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.

- 1996 Pål Nyrén and his student Mostafa Ronaghi at the Royal Institute of Technology in Stockholm publish their method of pyrosequencing

- 1998 Phil Green and Brent Ewing of the University of Washington publish "`phred`" for sequencer data analysis.

- 2000 Lynx Therapeutics publishes and markets "MPSS" - a parallelized, adapter/ligation-mediated, bead-based sequencing technology, launching "next-generation" sequencing.

- 2001 A draft sequence of the human genome is published.

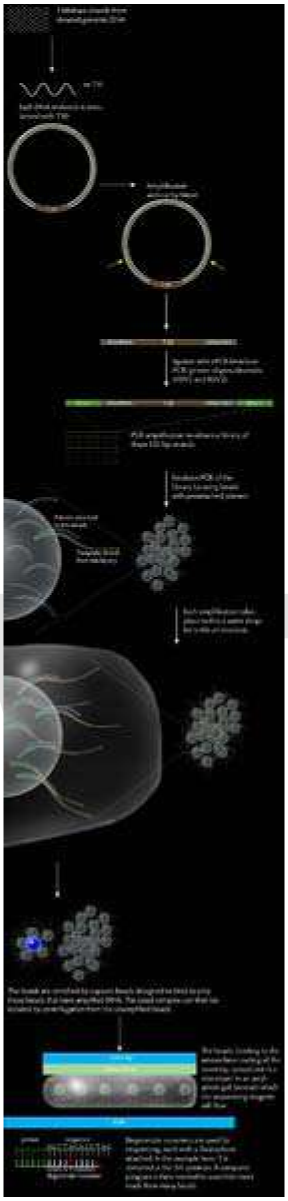
- 2004 454 Life Sciences markets a parallelized version of pyrosequencing. The first version of their machine reduced sequencing costs 6-fold compared to automated Sanger sequencing, and was the second of a new generation of sequencing technologies, after MPSS.

Chapter- 13

Polony Sequencing

Polony Sequencing is an inexpensive but highly accurate multiplex sequencing technique that can be used to “read” millions of immobilized DNA sequences in parallel. This technique was first developed by Dr. George Church group in Harvard Medical School. Unlike other sequencing technique, **Polony sequencing** technology is an open platform with freely downloadable, open source software and protocols. Also, the hardware of this technique can be easily set up with a commonly available epifluorescence microscope and a computer-controlled flowcell/ fluidics system. Polony sequencing is generally performed on paired-end Tags library that each molecule of DNA template is of 135bp in length with two 17-18bp paired genomic tags separated and flanked by common sequences. The current read length of this technique is 26 bases per amplicon and 13 bases per tag, leaving a 4-5 bases gap in each tag.

Workflow



An illustrated procedure for Polony sequencing.

The protocol of Polony sequencing can be break into three main parts which are the paired end-tag library construction, template amplification and DNA sequencing.

Paired end-tag library construction

This protocol begins by randomly shearing the tested genomic DNA into a tight size distribution. The sheared DNA molecules are then subjected for the end repair and A-tailed treatment. The end repair treatment converts any damaged or incompatible protruding ends of DNA to 5’-phosphorylated and blunt-ended DNA, enabling immediate blunt-end ligation. While the A-tailing treatment adds an A to the 3’ end of the sheared DNA. DNA molecules with a length of 1kb are selected by loading on the 6% TBE

PAGE gel. Next step, the DNA molecules are circularized with T-tailed 30bp long synthetic oligonucleotides (T30), which contains two outward-facing MmeI recognition sites, and the resulting circularized DNA undergoes rolling circle replication. The amplified circularized DNA molecules are then digested with MmeI (type IIs restriction endonuclease) which will cuts at a distance from its recognition site, releasing the T30 fragment flanked by 17-18 bp tags (~70 bp in length). The paired-tag molecules need to be end-repaired prior to the ligation of ePCR (emulsion PCR) primer oligonucleotides (FDV2 and RDV2) to their both ends. The resulting 135 bp library molecules are size-selected and nick translated. Lastly, amplify the 135bp paired end-tag library molecules with PCR to increase the amount of library material and eliminate extraneous ligation products in a single step. The resulted DNA template consists a 44bp FDV sequence, a 17-18 bp proximal tag, the T30 sequence, a 17-18 bp distal tag, and a 25 bp RDV sequence.

Template amplification

- **Emulsion PCR** The Monosized, paramagnetic streptavidin –coated beads are pre-loaded with dual biotin forward primer. Streptavidin has a very strong affinity for biotin, thus the forward primer will bind firmly on the surface of the beads. Next, an aqueous phase is prepared with the pre-loaded beads, PCR mixture, forward and reverse primers, and the paired end-tag library. This is mixed and vortexed with an oil phase to create the emulsion. Ideally, each droplet of water in the oil emulsion has one bead and one molecule of template DNA, permitting millions of non-interacting amplification within a milliliter-scale volume by performing PCR.
- **Emulsion breaking** After amplification, the emulsion from preceding step is broken using isopropanol and detergent buffer (10mM Tris pH 7.5, 1mM EDTA pH 8.0, 100mM NaCl, 1% (v/v) Triton X‐100, 1% (w/v) SDS), following with a series of vortexing, centrifuging, and magnetic separation. The resulted solution is a suspension of empty, clonal and non-clonal beads, which arise from emulsion droplets that initially have zero, one or multiple DNA template molecules, respectively. The amplified bead could be enriched in the following step.
- **Bead enrichment** The enrichment of amplified beads is achieved through hybridization to a larger, low density, non-magnetic polystyrene beads that pre-loaded with a biotinylated capture oligonucleotides (DNA sequence that complementary to ePCR amplicon sequence). The mixture is then centrifuged to separate the amplified and capture beads complex from the unamplified beads. The amplified, capture bead complex has a lower density and thus will remain in the supernatant while the unamplified beads form a pellet. The supernatant is recovered and treated with NaOH which will break the complex. The paramagnetic amplified beads are separated from the non-magnetic capture beads by magnetic separation. This enrichment protocol is capable in enriching five times of amplified beads.
- **Bead capping** The purpose of bead capping is to attach a "capping" oligonucleotide to the 3' end of both unextended forward ePCR primers and the RDV segment of template DNA. The cap that being use is an amino group that prevents fluorescent probes from ligating to these ends and at the same time, helping the subsequent coupling of template DNA to the aminosilanated flow cell coverslip.
- **Coverslip arraying** First, the coverslips are washed and aminosilane-treated, enabling

the subsequent covalent coupling of template DNA on it and eliminating any fluorescent contamination. The amplified, enriched beads are mixed with acrylamide and poured into a shallow mold formed by a Teflon-masked microscope slide. Immediately, place the aminosilane-treated coverslip on top of the acrylamide gel and allow to polymerize for 45 minutes. Next, invert the slide/coverslip stack and remove the microscope slide from gel. The silane treated coverslips will bind covalently to the gel while the Teflon on the surface of microscope slide will enable the better removal of slide from the acrylamide gel. The coverslips then bonded to the flow cell body and any unattached beads will be removed.

DNA sequencing

The biochemistry of Polony sequencing mainly rely on the discriminatory capacities of polymerases and ligases. First, a series of anchor primers are flowed through the cells and hybridize to the synthetic oligonucleotide sequences at the immediate 3’ or 5’ end of the 17-18bp proximal or distal genomic DNA tags. Next, an enzymatic ligation reaction of the anchor primer to a population of degenerate nonamers that are labeled with fluorescent dyes is performed.
Differentially labeled nonamers:

```
5' Cy5 - NNNNNNNNT
5' Cy3 - NNNNNNNNA
5' TexasRed - NNNNNNNNC
5' 6FAM - NNNNNNNNG
```

The fluorophore-tagged nonamers selectively ligate onto the anchor primer, providing a fluorescent signal that indicates whether there is an A, C, G, or T at the query position on the genomic DNA tag. After four colour imaging, the anchor primer/nonamer complexes are stripped off and a new cycle is begun by replacing the anchor primer. A new mixture of the fluorescently tagged nonamers is introduced, for which the query position is shifted one base further into the genomic DNA tag.

```
5' Cy5 - NNNNNNNTN
5' Cy3 - NNNNNNNAN
5' TexasRed - NNNNNNNCN
5' 6FAM - NNNNNNNGN
```

Seven bases from the 5’ to 3’ direction and six bases from the 3’ end could be queried in this fashion. The ultimate result is a read length of 26 bases per run (13 bases from each of the paired tags) with a 4 to 5 bases gap in the middle of each tag.

Analysis and software

The polony sequencing generates millions of 26 reads per run and this information needed to be normalized and converted to sequence. These can be done by the software that has been developed by Church Lab. All of the software is free and could be downloaded from the website.

Strength and Weaknesses

Polony sequencing allows for a high throughput and high consensus accuracies of DNA sequencing based on a commonly available, inexpensive instrument. Also, it is a very flexible technique that enables variable application including BAC (bacterial artificial chromosome) and bacterial genome resequencing, as well as SAGE (serial analysis of gene expression) tag and barcode sequencing. Furthermore, the polony sequencing technique is emphasized as an open system that shares everything including the software that have been developed, protocol and reagents.

However, although the raw data acquisition could be achieved as high as 786 gigabits but only 1 bit of information out of 10,000 bits collected is useful. Another challenge of this technique is the uniformity of the relative amplification of individual targets. The non-uniform amplification could lower the efficiency of sequencing and posted as the biggest obstacle in this technique.

Cost

The sequencing instrument used in this technique could be set up by the commonly available fluorescence microscope and a computer controlled flowcell. According to the calculation in the year of 2005, the set up of a complete set of instrument will cost around US$ 130,000. However, the cost could be further lower to US$ 100,000 in the near future. A biotech company, Dover, is in collaboration with the Church Laboratory of Harvard Medical School to produce an automated sequencing machine, Polonator G.007, based on polony sequencing technique. The current selling price of this machine is around US$ 170,000. According to the calculation in year 2005, every kilobase of generated raw sequence was estimated to $0.11, while omitting the paired-end tag library construction cost, the cost of every kilobase of sequence could drops to $0.08.

History

The polony sequencing is a “distant relative” of the classical polony technology which mainly developed by Dr. Rob Mitra. Together with a MD PhD student, Jay Shendure, Dr. Rob Mitra worked out ways to sequence in situ polonies using single-base extension which can achieved 5-6 bases reads. However, the existing polony sequencing technology was mainly developed by Jay Shendure and Greg Porreca. They have changed almost everything that was there in order to make this multiplex sequencing technology work.

Also, the highly parallel sequencing-by-ligation method of polony sequencing has contributed in forming the basis for ABI Solid Sequencing and others.

Chapter- 14

DNA Sequencing Theory

DNA sequencing theory is the broad body of work that attempts to lay analytical foundations for DNA sequencing. The practical aspects revolve around designing and optimizing sequencing projects, predicting project performance, troubleshooting experimental results, characterizing factors such as sequence bias and the effects of software processing algorithms, and comparing various sequencing methods to one another. In this sense, it could be considered a branch of systems engineering or operations research. The permanent archive of work is primarily mathematical, although numerical calculations are often conducted for particular problems too. DNA sequencing theory addresses *physical processes* related to sequencing DNA and should not be confused with theories of analyzing resultant DNA sequences, e.g. sequence alignment. Publications sometimes do not make a careful distinction, but the latter are primarily concerned with algorithmic issues.

Sequencing as a covering problem

All mainstream methods of DNA sequencing rely on reading small fragments of DNA and subsequently reconstructing these data to infer the original DNA target, either via assembly or alignment to a reference. The abstraction common to these methods is that of a mathematical covering problem. For example, one can imagine a line segment representing the target and a subsequent process where smaller segments are "dropped" onto random locations of the target. The target is considered "sequenced" when adequate coverage accumulates, for example when no gaps remain.

The abstract properties of covering have been studied by mathematicians for over a century. However, direct application of these results has not generally been possible. Closed-form mathematical solutions, especially for probability distributions, often cannot be readily evaluated. That is, they involve inordinately large amounts of computer time for parameters characteristic of DNA sequencing. Stevens' configuration is one such example. Results obtained from the perspective of pure mathematics also do not account for factors that are actually important in sequencing, for instance detectable overlap in sequencing fragments, double-stranding, edge-effects, and target multiplicity. Consequently, development of sequencing theory has proceeded more according to the philosophy of applied mathematics. In particular, it has been problem-focused and makes expedient use of approximations, simulations, etc.

Early uses derived From elementary probability theory

The earliest result was actually borrowed directly from elementary probability theory. If we model the above process and take *L* and *G* as the fragment length and target length, respectively, then the probability of "covering" any given location on the target *with one particular fragment* is *L* / *G*. Note that this presumes $L \ll G$, which is valid for many, though not all sequencing scenarios. Utilizing concepts from the binomial distribution, it can then be shown that the probability that the location is covered by at least one of *N* fragments is

$$P = 1 - \left[1 - \frac{L}{G}\right]^N .$$

This equation was first used to characterize plasmid libraries, but is often more useful in a modified form. For most projects $N \gg 1$, so that, to a good degree of approximation

$$\left[1 - \frac{L}{G}\right]^N \sim \exp(-NL/G),$$

where *R* = *NL* / *G* is called the *redundancy*. Note the significance of redundancy as representing the average number of times a position is covered with fragments. Note also that in considering the covering process over all positions in the target, this probability is identical to the expected value of the random variable *C*, which represents the fraction of the target coverage. The final result,

$$E\langle C\rangle = 1 - e^{-R},$$

remains in widespread use as a "back of the envelope" estimator and predicts that coverage for all projects evolves along a universal curve that is a function only of the redundancy.

Lander-Waterman theory

In 1988, Eric Lander and Michael Waterman published an important paper examining the covering problem from the standpoint of gaps. Although they focused on the so-called mapping problem, the abstraction to sequencing is much the same. They furnished a number of useful results that were adopted as the standard theory from the earliest days of "large-scale" genome sequencing. Their model was also used in designing the Human Genome Project and continues to play an important role in DNA sequencing.

Ultimately, the main goal of a sequencing project is to close all gaps, so the "gap perspective" was a logical basis of developing a sequencing model. One of the more frequently used results from this model is the expected number of contigs, given the number of fragments sequenced. If one neglects the amount of sequence that is essentially "wasted" by having to detect overlaps, their theory yields

$$E\langle contigs\rangle = Ne^{-R}.$$

In 1995, Roach proposed a model that appeared to be essentially different and asserted that Lander-Waterman theory gave contradictory results for large values of *R*. Wendl and Waterston later showed, based on Stevens' method, that subtle differences in interpretation explained the anomalies and that both models were indeed essentially identical and consistent.

The basic ideas of Lander-Waterman theory led to a number of additional results for particular variations in mapping techniques. However, technological advancements have rendered mapping and its more esoteric theories largely obsolete.

Recent advancements

The physical processes and protocols of DNA sequencing have continued to evolve, largely driven by advancements in bio-chemical methods, hardware, and automation techniques. There is now a wide range of problems that DNA sequencing has made in-roads into, including metagenomics and medical (cancer) sequencing. There are important factors in these scenarios that classical theory does not account for. Recent work has begun to focus on resolving the effects of some of these issues. The level of mathematics becomes commensurately more sophisticated.

Multiplicity

Biologists have developed methods to filter highly-repetitive, essentially un-sequenceable regions of genomes. These procedures are important for organisms whose genomes consist mostly of such DNA, for example corn. They yield multitudes of small islands of sequenceable DNA products. Wendl and Barbazuk proposed an extension to Lander-Waterman Theory to account for "gaps" in the target due to filtering and the so-called "edge-effect". The latter is a position-specific sampling bias, for example the terminal base position has only a 1 / *G* chance of being covered, as opposed to *L* / *G* for interior positions. For $R < 1$, classical Lander-Waterman Theory still gives good predictions, but dynamics change for higher redundancies.

Paired-end sequencing

Modern sequencing methods usually sequence both ends of a larger fragment, which provides linking information for *de novo* assembly and improved probabilities for alignment to reference sequence. Researchers generally believe that longer lengths of data (read lengths) enhance performance for very large DNA targets, an idea consistent with predictions from distribution models. However, Wendl showed that smaller fragments provide better coverage on small, linear targets because they reduce the edge effect in linear molecules. These findings have implications for sequencing the products of DNA filtering procedures. Read-pairing and fragment size evidently have negligible influence for large, whole-genome class targets.

Diploid sequencing

Sequencing is emerging as an important tool in medicine, for example in cancer research. Here, the ability to detect heterozygous mutations is important and this can only be done if the sequence of the diploid genome is obtained. In the pioneering efforts to sequence individuals, Levy *et al.* and Wheeler *et al.*, who sequenced Craig Venter and Jim Watson, respectively, outlined models for covering both alleles in a genome. Wendl and Wilson followed with a more general theory that allowed for an arbitrary number of coverings of each allele and arbitrary ploidy. These results point to the general conclusion that the amount of data needed for such projects is significantly higher than for traditional haploid projects.

Limitations

DNA sequencing theories often invoke the assumption that certain random variables in a model are independently and identically distributed. For example, in Lander-Waterman Theory, a sequenced fragment is presumed to have the same probability of covering each region of a genome and all fragments are assumed to be independent of one another. In actuality, sequencing projects are subject to various types of bias, including differences of how well regions can be cloned, sequencing anomalies, biases in the target sequence (which is *not* random), and software-dependent errors and biases. In general, theory will agree well with observation up to the point that enough data have been generated to expose latent biases. The kinds of biases related to the underlying target sequence are particularly difficult to model, since the sequence itself may not be known *a priori*. This presents a type of "chicken and egg" closure problem.

Academic status

Sequencing theory is based on elements of mathematics, biology, and systems engineering, so it is highly interdisciplinary. Although many universities now have programs in computational biology, there does not yet seem to be a strong focus at the graduate level on this topic. Academic contributions have mainly been limited to a small number of PhD dissertations.

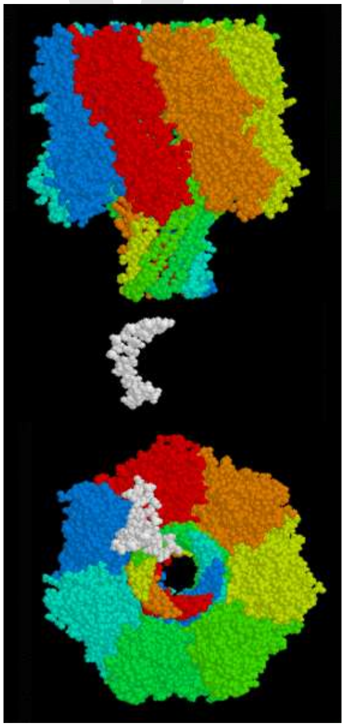
Chapter- 15

Nanopore Sequencing and Sequencing by Ligation

Nanopore sequencing

Nanopore sequencing is a method under development since 1995 for determining the order in which nucleotides occur on a strand of DNA.

A nanopore is simply a small hole, of the order of 1 nanometer in internal diameter. Certain transmembrane cellular proteins act as nanopores, and nanopores have also been made by etching a somewhat larger hole (several tens of nanometers) in a piece of silicon, and then gradually filling it in using ion-beam sculpting methods which results in a much smaller diameter hole: the nanopore.



alpha-hemolysin pore (made up of 7 identical subunits in 7 colors) and 12-mer single-stranded DNA (in white) on the same scale to illustrate DNA effects on conductance when moving through a nanopore. Below is an orthogonal view of the same molecules.

The theory behind nanopore sequencing has to do with what occurs when the nanopore is immersed in a conducting fluid and a potential (voltage) is applied across it: under these conditions a slight electric current due to conduction of ions through the nanopore can be observed, and the amount of current is very sensitive to the size and shape of the nanopore. If single nucleotides (bases) or strands of DNA pass through the nanopore, this can create a characteristic change in the magnitude of the current through the nanopore.

DNA could be passed through the nanopore for various reasons. For example, electrophoresis might attract the DNA towards the nanopore, and it might eventually pass through it. Or, enzymes attached to the nanopore might guide DNA towards the nanopore. The scale of the nanopore means that the DNA may be forced through the hole as a long string, one base at a time, rather like thread through the eye of a needle. As it does so, each nucleotide on the DNA molecule may obstruct the nanopore to a different, characteristic degree. The amount of current which can pass through the nanopore at any given moment therefore varies depending on whether the nanopore is blocked by an A, a C, a G or a T. The change in the current through the nanopore as the DNA molecule passes through the nanopore represents a direct reading of the DNA sequence. Alternatively, a nanopore might be used to identify individual DNA bases as they pass through the nanopore in the correct order - this approach has been shown by Oxford Nanopore Technologies and Professor Hagan Bayley.

The potential is that a single molecule of DNA can be sequenced directly using a nanopore, without the need for an intervening PCR amplification step or a chemical labelling step or the need for optical instrumentation to identify the chemical label. As of July 2010, information available to the public indicates that nanopore sequencing is still in the development stage, with some laboratory-based data to back up the different components of the sequencing method, but not yet commercially available, parallelized, routineized, nor cost-effective enough yet to compete with out "next generation sequencing" methods. Nanopore-based DNA analysis techniques are being industrially developed by Oxford Nanopore Technologies (developing direct exonuclease sequencing and strand sequencing using protein nanopores, and solid-state sequencing through internal R&D and collaborations with academic institutions), NabSys (using a library of DNA probes and using nanopores to detect where these probes have hybridized to single stranded DNA) and NobleGen(using nanopores in combination with fluorescent labels). IBM has noted research projects on computer simulations of translocation of a DNA strand through a solid-state nanopore, but not projects on identifying the DNA bases on that strand.

One challenge for the 'strand sequencing' method is in refining the method to improve its resolution to be able to detect single bases. In the early papers methods, a nucleotide needed to be repeated in a sequence about 100 times successively in order to produce a measurable characteristic change. This low resolution is due to the fact that the DNA strand moves rapidly at the rate of 1 to 5μs per base through the nanopore. This makes recording difficult and prone to background noise, failing in obtaining single-nucleotide resolution. The problem is being tackled by either improving the recording technology or by controlling the speed of DNA strand by various protein engineering strategies. More recently effects of single bases due to secondary structure or released mononucleotides have been shown.. Professor Hagan Bayley, founder of Oxford Nanopore, recently proposed that creating two recognition sites within an alpha hemolysin pore may confer advantages in base recognition.

One challenge for the 'exonuclease approach', where a processive enzyme feeds individual bases, in the correct order, into the nanopore, is to integrate the exonuclease and the nanopore detection systems. In particular, the problem is that when an exonuclease hydrolyzes the phosphodieseter bonds between nucleotides in DNA, the subsequentially released nucleotide is not necessarily guaranteed to directly move in to, say, a nearby alpha-hemolysin nanopore. One idea is to attach the exonuclease to the nanopore, perhaps through biotinylation to the beta barrel hemolsyin. The central pore of the protein may be lined with charged residues arranged so that the positive and negative charges appear on opposite sides of the pore. However, this mechanism is primarily discriminatory and does not constitute a mechanism to guide nucleotides down some particular path.

Commercialization

Agilent Laboratories was the first to license and develop nanopores but does not have any current disclosed research in the area.

The company Oxford Nanopore Technologies in 2008 licensed technology from Harvard, UCSC and other universities and is developing protein and solid state nanopore technology with the aim of sequencing DNA and identifying biomarkers, drugs of abuse and a range of other molecules.

Sequenom licensed nanopore technology from Harvard in 2007 using an approach that combines nanopores and fluorescent labels. This technology was subsequently licensed to Noblegen.

NABsys was spun out of Brown University and is researching nanopores as a method of identifying areas of single stranded DNA that have been hybridized with specific DNA probes.

Sequencing by ligation

Sequencing by ligation is a DNA sequencing method that uses the enzyme DNA ligase to identify the nucleotide present at a given position in a DNA sequence. Unlike most currently popular DNA sequencing methods, this method does not use a DNA polymerase to create a second strand. Instead, the mismatch sensitivity of a DNA ligase enzyme is used to determine the underlying sequence of the target DNA molecule.

Process

DNA ligase is an enzyme that joins together ends of DNA molecules. Although commonly represented as joining two pairs of ends at once, as in the ligation of restriction enzyme fragments, ligase can also join the ends on only one of the two strands (for example, when the other strand is already continuous or lacks a terminal phosphate necessary for ligation). DNA ligase is sensitive to the structure of DNA and has very low efficiency when there are mismatches between the bases of the two strands.

Sequencing by ligation relies upon the sensitivity of DNA ligase for base-pairing mismatches. The target molecule to be sequenced is a single strand of unknown DNA sequence, flanked on at least one end by a known sequence. A short "anchor" strand is brought in to bind the known sequence.

A mixed pool of probe oligonucleotides is then brought in (eight or nine bases long), labeled (typically with fluorescent dyes) according to the position that will be sequenced. These molecules hybridize to the target DNA sequence, next to the anchor sequence, and DNA ligase preferentially joins the molecule to the anchor when its bases match the unknown DNA sequence. Based on the fluorescence produced by the molecule, one can infer the identity of the nucleotide at this position in the unknown sequence.

The oligonucleotide probes may also be constructed with cleavable linkages which can be cleaved after identifying the label. This will both remove the label and regenerate a 5' phosphate on the end of the ligated probe, preparing the system for another round of ligation. This cycle can be repeated several times to read longer sequences. This sequences every Nth base, where N is the length of the probe left behind after cleavage. To sequence the skipped positions, the anchor and ligated oligonucleotides may be stripped off the target DNA sequence, and another round of sequencing by ligation started with an anchor one or more bases shorter.

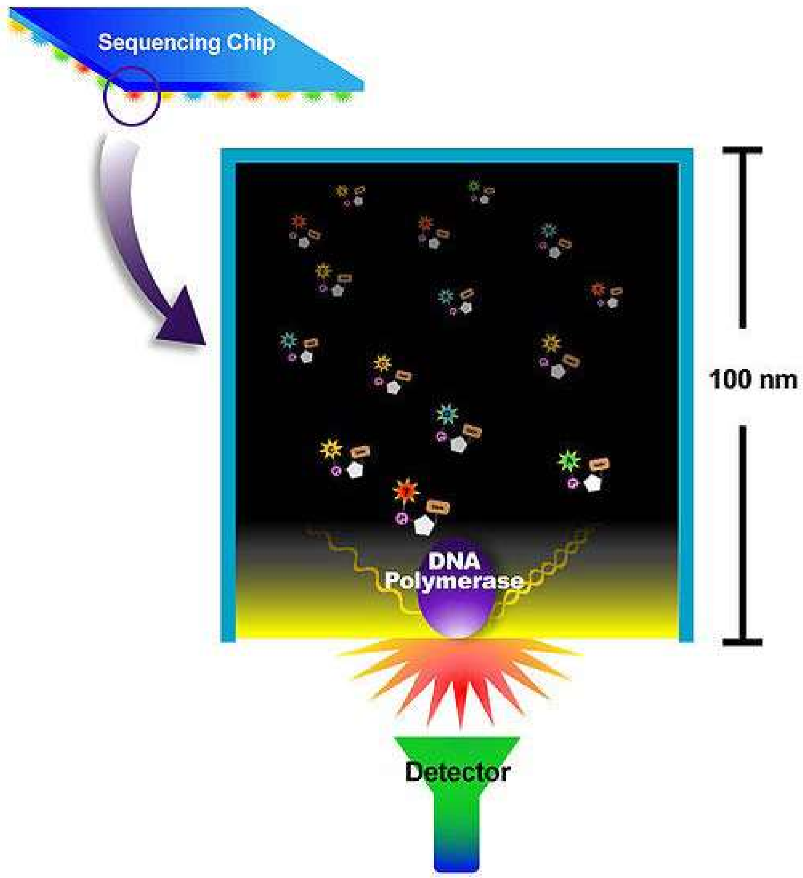
A simpler, albeit more limited, technique is to do repeated rounds of a single ligation where the label corresponds to different position in the probe, followed by stripping the anchor and ligated probe.

Sequencing by ligation can proceed in either direction (either 5'-3' or 3'-5') depending on which end of the probe oligonucleotides are blocked by the label. The 3'-5' direction is more efficient for doing multiple cycles of ligation. Note that this is the opposite direction to polymerase based sequencing methods.

Chapter- 16

Single Molecule Real Time Sequencing and Pyrosequencing

Single molecule real time sequencing



Overview of Single Molecule Real Time Sequencing

Single molecule real time sequencing (also known as SMRT) is a parallelized single molecule DNA sequencing by synthesis technology developed by Pacific Biosciences. Single molecule real time sequencing utilizes the zero-mode waveguide (ZMW), developed in the laboratories of Harold G. Craighead and Watt W. Webb at Cornell University. A single DNA polymerase enzyme is affixed at the bottom of a ZMW with a single molecule of DNA as a template. The ZMW is a structure that creates an illuminated observation volume that is small enough to observe only a single nucleotide of DNA (also known as a base) being incorporated by DNA polymerase. Each of the four DNA bases is attached to one of four different fluorescent dyes. When a nucleotide is incorporated by the DNA polymerase, the fluorescent tag is cleaved off and diffuses out of the observation area of the ZMW where its fluorescence is no longer observable. A detector detects the fluorescent signal of the nucleotide incorporation, and the base call is made according to the corresponding fluorescence of the dye.

Technology

The DNA sequencing is done on a chip that contains many ZMWs. Inside each ZMW, a single active DNA polymerase with a single molecule of single stranded DNA template is immobilized to the bottom through which light can penetrate and create a visualization chamber that allows monitoring of the activity of the DNA polymerase at a single molecule level. The signal from a phospho-linked nucleotide incorporated by the DNA polymerase is detected as the DNA synthesis proceeds which results in the DNA sequencing in real time.

Phospholinked nucleotide

For each of the nucleotide bases, there are four corresponding fluorescent dye molecules that enable the detector to identify the base being incorporated by the DNA polymerase as it performs the DNA synthesis. The fluorescent dye molecule is attached to the phosphate chain of the nucleotide. When the nucleotide is incorporated by the DNA polymerase, the fluorescent dye is cleaved off with the phosphate chain as a part of a natural DNA synthesis process during which a phosphodiester bond is created to elongate the DNA chain. The cleaved fluorescent dye molecule then diffuses out of the detection volume so that the fluorescent signal is no longer detected.

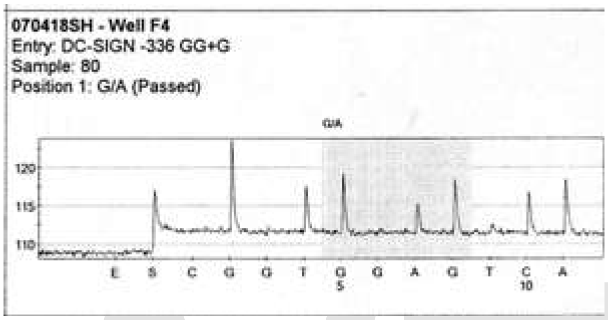
Zero-mode waveguide

The zero-mode waveguide (ZMW) is a nanophotonic confinement structure that consists of a circular hole in an aluminum cladding film deposited on a clear silica substrate. The ZMW holes are ~70 nm in diameter and ~100 nm in depth. Due to the behavior of light when it travels through a small aperture, the optical field decays exponentially inside the chamber. The observation volume within an illuminated ZMW is ~20 zeptoliters (20 X 10^{-21} liters). Within this volume, the activity of DNA polymerase incorporating a single nucleotide can be readily detected.

Sequencing performance

Pacific Biosciences expects to commercialize SMRT sequencing in 2010 or 2011. The prototype of the SMRT chip contains ~3000 ZMW holes that allow parallelized DNA sequencing. Each of the ZMW holes produces approximately 1,500 bp (base pair) read lengths at a speed of 10 bp per second.

Pyrosequencing



Example of a pyrogram showing the nucleotide sequence in a specific section of DNA. The tops represent light emission and nucleotide binding.

Pyrosequencing is a method of DNA sequencing (determining the order of nucleotides in DNA) based on the "sequencing by synthesis" principle. It differs from Sanger sequencing, in that it relies on the detection of pyrophosphate release on nucleotide incorporation, rather than chain termination with dideoxynucleotides. The technique was developed by Pål Nyrén and Mostafa Ronaghi at the Royal Institute of Technology in Stockholm in 1996.

Procedure

"Sequencing by synthesis" involves taking a single strand of the DNA to be sequenced and then synthesizing its complementary strand enzymatically. The Pyrosequencing method is based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemiluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step. The template DNA is immobile, and solutions of A, C, G, and T nucleotides are sequentially added and removed from the reaction. Light is produced only when the nucleotide solution complements the first unpaired base of the template. The sequence of solutions which produce chemiluminescent signals allows the determination of the sequence of the template.

ssDNA template is hybridized to a sequencing primer and incubated with the enzymes DNA polymerase, ATP sulfurylase, luciferase and apyrase, and with the substrates adenosine 5´ phosphosulfate (APS) and luciferin.

1. The addition of one of the four deoxynucleotide triphosphates (dNTPs)(in the case of dATP we add dATPαS which is not a substrate for a luciferase) initiates the second step. DNA polymerase incorporates the correct, complementary dNTPs onto the template. This incorporation releases pyrophosphate (PPi) stoichiometrically.
2. ATP sulfurylase quantitatively converts PPi to ATP in the presence of adenosine 5´ phosphosulfate. This ATP acts as fuel to the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are proportional to the amount of ATP. The light produced in the luciferase-catalyzed reaction is detected by a camera and analyzed in a program.
3. Unincorporated nucleotides and ATP are degraded by the apyrase, and the reaction can restart with another nucleotide.

Currently, a limitation of the method is that the lengths of individual reads of DNA sequence are in the neighborhood of 300-500 nucleotides, shorter than the 800-1000 obtainable with chain termination methods (e.g. Sanger sequencing). This can make the process of genome assembly more difficult, particularly for sequences containing a large amount of repetitive DNA. As of 2007, pyrosequencing is most commonly used for resequencing or sequencing of genomes for which the sequence of a close relative is already available.

The templates for pyrosequencing can be made both by solid phase template preparation (streptavidin-coated magnetic beads) and enzymatic template preparation (apyrase+exonuclease).

Commercialization

The company **Pyrosequencing AB** in Uppsala, Sweden commercialized machinery and reagents for sequencing short stretches of DNA using the pyrosequencing technique. **Pyrosequencing AB** was renamed to **Biotage** in 2003 which was acquired by Qiagen in 2008. Pyrosequencing technology was further licensed to 454 Life Sciences. 454 developed an array-based pyrosequencing technology which has emerged as a platform for large-scale DNA sequencing. Most notable are the applications for genome sequencing and metagenomics. *GS FLX*, the latest pyrosequencing platform by 454 Life Sciences (now owned by Roche Diagnostics), can generate 400 million nucleotide data in a 10 hour run with a single machine. Each run would cost about 5,000-7,000 USD, with multiple fold coverage required for accuracy this pushes de novo sequencing of mammalian genomes into the million dollar range.

Chapter- 17

Nucleotide

Nucleotides are molecules that, when joined together, make up the structural units of RNA and DNA. In addition, nucleotides play central roles in metabolism. In that capacity, they serve as sources of chemical energy (adenosine triphosphate and guanosine triphosphate), participate in cellular signaling (cyclic guanosine monophosphate and cyclic adenosine monophosphate), and are incorporated into important cofactors of enzymatic reactions (coenzyme A, flavin adenine dinucleotide, flavin mononucleotide, and nicotinamide adenine dinucleotide phosphate).

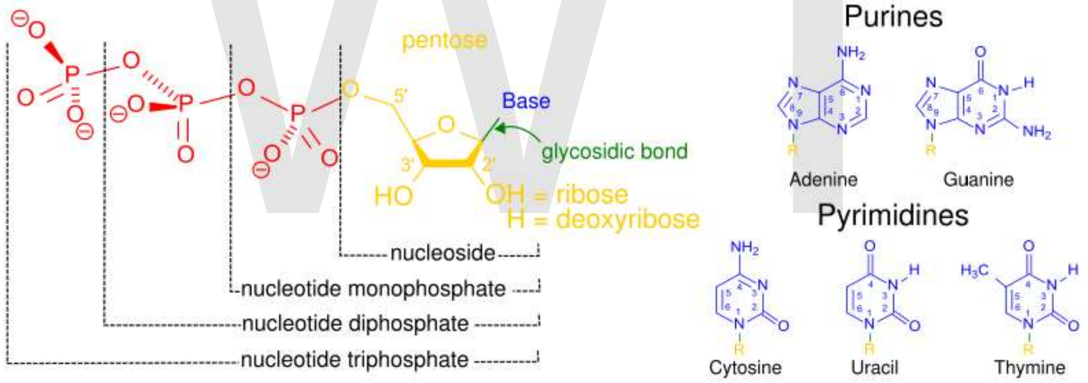


Figure 1: Structural elements of the most common nucleotides

Nucleotide structure

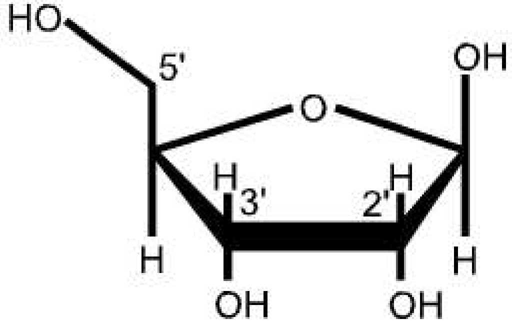


Figure 2: Ribose structure indicating numbering of carbon atoms

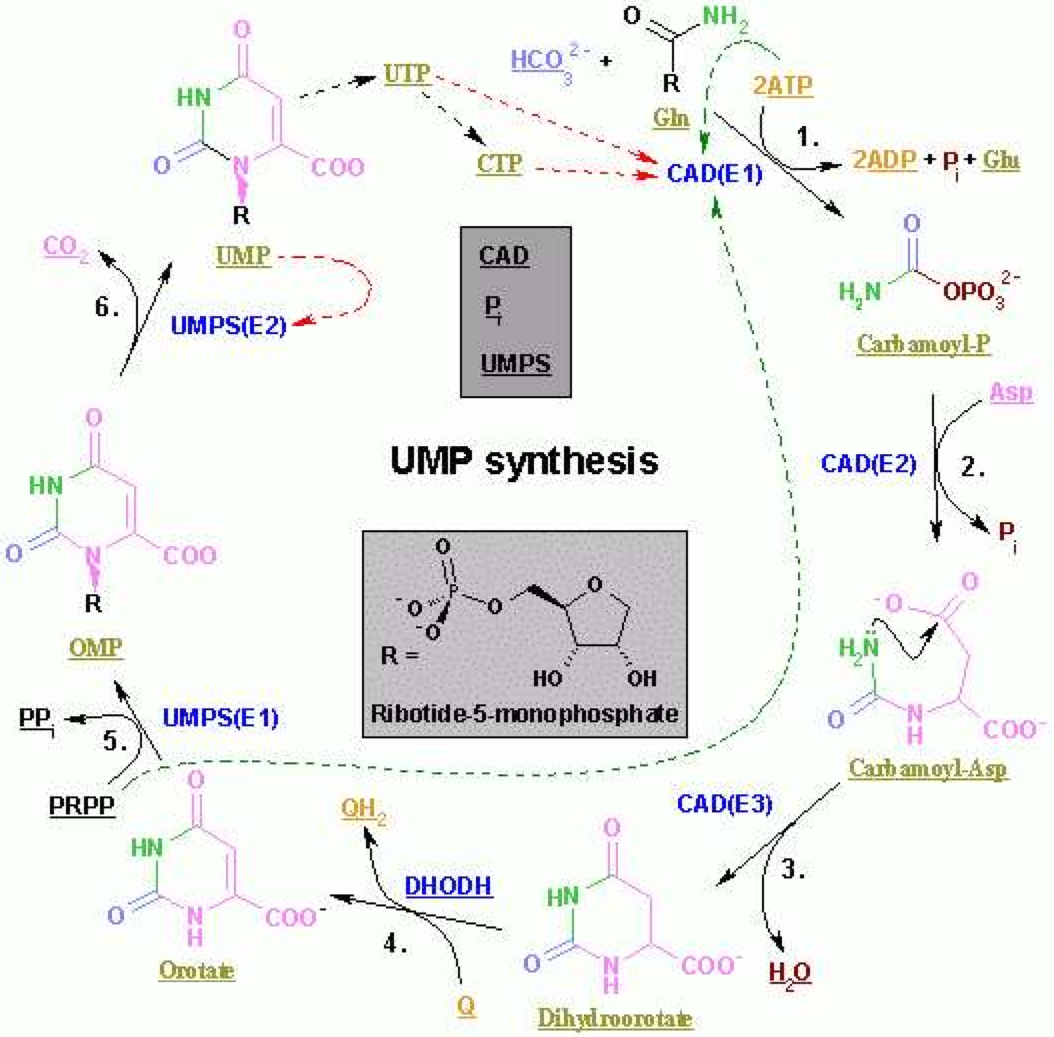
A nucleotide is composed of a nucleobase (nitrogenous base), a five-carbon sugar (either ribose or 2'-deoxyribose), and one to three phosphate groups. Together, the nucleobase and sugar comprise a nucleoside. The phosphate groups form bonds with either the 2, 3, or 5-carbon of the sugar, with the 5-carbon site most common. Cyclic nucleotides form when the phosphate group is bound to two of the sugar's hydroxyl groups. Ribonucleotides are nucleotides where the sugar is ribose, and deoxyribonucleotides contain the sugar deoxyribose. Nucleotides can contain either a purine or a pyrimidine base.

Nucleic acids are polymeric macromolecules made from nucleotide monomers. In DNA, the purine bases are adenine and guanine, while the pyrimidines are thymine and cytosine. RNA uses uracil in place of thymine. Adenine always pairs with thymine by 2 hydrogen bonds, while guanine pairs with cytosine through 3 hydrogen bonds, each due to their unique structures.

Synthesis

Nucleotides can be synthesized by a variety of means both in vitro and in vivo. In vivo, nucleotides can be synthesised de novo or recycled through salvage pathways. Nucleotides undergo breakdown such that useful parts can be reused in synthesis reactions to create new nucleotides. In vitro, protecting groups may be used during laboratory production of nucleotides. A purified nucleoside is protected to create a phosphoramidite, which can then be used to obtain analogues not found in nature and/or to synthesize an oligonucleotide.

Pyrimidine ribonucleotides

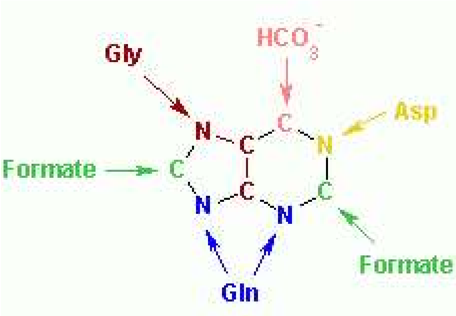


The synthesis of UMP.
The color scheme is as follows: **enzymes, coenzymes, substrate names, inorganic molecules**

Pyrimidine nucleotide synthesis starts with the formation of carbamoyl phosphate from glutamine and CO_2. The cyclisation reaction between carbamoyl phosphate reacts with aspartate, yielding orotate in subsequent steps. Orotate reacts with 5-phosphoribosyl α-diphosphate (PRPP), yielding orotidine monophosphate (OMP), which is decarboxylated to form uridine monophosphate (UMP). It is from UMP that other pyrimidine nucleotides are derived. UMP is phosphorylated to uridine triphosphate (UTP) via two sequential reactions with ATP. Cytidine monophosphate (CMP) is derived from conversion of UTP to cytidine triphosphate (CTP) with subsequent loss of two phosphates.

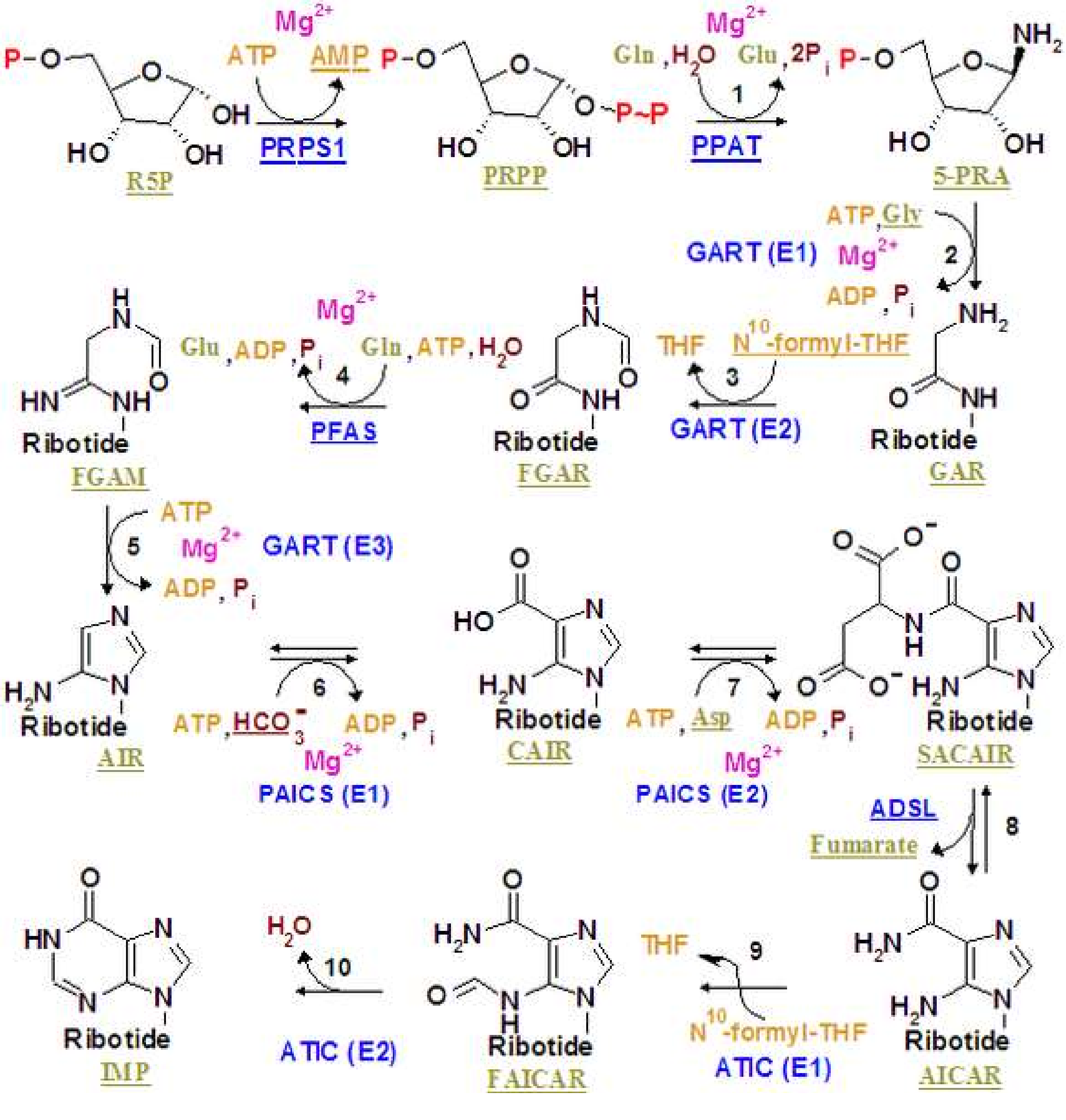
Purine ribonucleotides

The atoms which are used to build the purine nucleotides come from a variety of sources:



The biosynthetic origins of purine ring atoms

N1 arises from the amine group of Asp
C2 and C8 originate from formate
N3 and N9 are contributed by the amide group of Gln
C4, C5 and N7 are derived from Gly
C6 comes from HCO_3^- (CO_2)



The synthesis of IMP. The color scheme is as follows: **enzymes, coenzymes, substrate names, metal ions, inorganic molecules**

The de novo synthesis of purine nucleotides by which these precursors are incorporated into the purine ring proceeds by a 10-step pathway to the branch-point intermediate IMP, the nucleotide of the base hypoxanthine. AMP and GMP are subsequently synthesized from this intermediate via separate, two-step pathways. Thus, purine moieties are initially formed as part of the ribonucleotides rather than as free bases.

Six enzymes take part in IMP synthesis. Three of them are multifunctional:

- GART (reactions 2, 3, and 5)
- PAICS (reactions 6, and 7)
- ATIC (reactions 9, and 10)

Reaction 1. The pathway starts with the formation of PRPP. PRPS1 is the enzyme that activates R5P, which is formed primarily by the pentose phosphate pathway, to PRPP by reacting it with ATP. The reaction is unusual in that a pyrophosphoryl group is directly transferred from ATP to C1 of R5P and that the product has the **α** configuration about C1. This reaction is also shared with the pathways for the synthesis of the pyrimidine nucleotides, Trp, and His. As a result of being on (a) such (a) major metabolic crossroad and the use of energy, this reaction is highly regulated.

Reaction 2. In the first reaction unique to purine nucleotide biosynthesis, PPAT catalyzes the displacement of PRPP's pyrophosphate group (PP_i) by Gln's amide nitrogen. The reaction occurs with the inversion of configuration about ribose C1, thereby forming **β**-5-phosphorybosylamine (5-PRA) and establishing the anomeric form of the future nucleotide. This reaction, which is driven to completion by the subsequent hydrolysis of the released PP_i, is the pathway's flux-generating step and is therefore regulated, too.

Length unit

Nucleotide (abbreviated nt) is a common length unit for single-stranded RNA, similar to how base pair is a length unit for double-stranded DNA.

Abbreviation codes for degenerate bases

The IUPAC has designated the symbols for nucleotides. Apart from the five (A, G, C, T/U) bases, often degenerate bases are used especially for designing PCR primers. These nucleotide codes are listed here.

IUPAC nucleotide code	**Base**
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G

Y	C or T (U)
S	G or C
W	A or T (U)
K	G or T (U)
M	A or C
B	C or G or T (U)
D	A or G or T (U)
H	A or C or T (U)
V	A or C or G
N	any base
. or -	gap

Chapter- 18

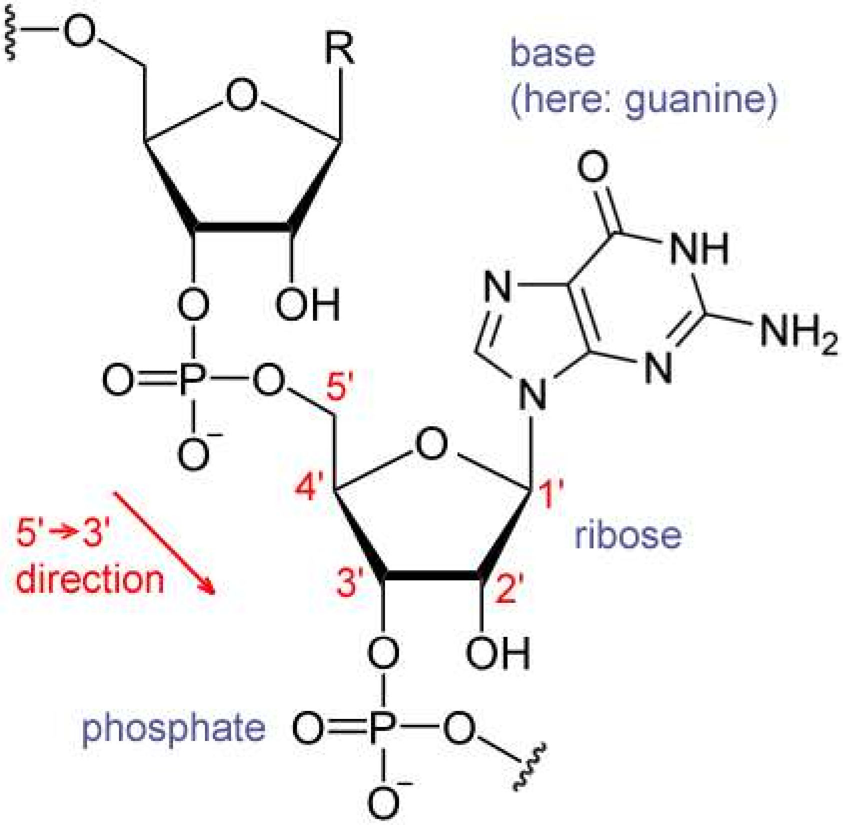
Nucleic Acid Sequence

The **sequence** or **primary structure of a nucleic acid** is the exact specification of its atomic composition and the chemical bonds connecting those atoms. As nucleic acids, e.g. DNA and RNA, are unbranched polymers, this is equivalent to specifying exact sequence of nucleotides that comprise the whole molecule. This sequence is written as a succession of letters representing a real or hypothetical DNA molecule or strand. By convention, the primary structure of a DNA or RNA molecule is reported from the 5' end to the 3' end.

The sequence has capacity to carry information. When used in reference to biological DNA, which carries the information which directs the functions of living beings, the term **genetic sequence** is often used. Sequences can be read from the biological raw material through DNA sequencing methods.

Primary structure is sometimes mistakenly termed *primary sequence*, but there is no such term, as well as no parallel concept of secondary or tertiary sequence.

Nucleotides



Chemical structure of RNA

Nucleic acids consist of a chain of linked units called nucleotides. Each nucleotide consists of three subunits: a phosphate group and a sugar (ribose in the case of RNA, deoxyribose in DNA) make up the backbone of the nucleic acid strand, and attached to the sugar is one of a set of nucleobases. The nucleobases are important in base pairing of strands to form higher-level secondary and tertiary structure such as the famed double helix.

The possible letters are *A*, *C*, *G*, and *T*, representing the four nucleotide bases of a DNA strand — adenine, cytosine, guanine, thymine — covalently linked to a phosphodiester backbone. In the typical case, the sequences are printed abutting one another without gaps, as in the sequence AAAGTCTGAC, read left to right in the 5' to 3' direction. With regards to transcription, a sequence is on the coding strand if it has the same order as the transcribed RNA.

One sequence can be complementary to another sequence, meaning that they have the base on each position is the complementary (i.e. A to T, C to G) and in the reverse order. For example, the complementary sequence to TTAC is GTAA. If one strand of the double-stranded DNA is considered the sense strand, then the other strand, considered the antisense strand, will have the complementary sequence to the sense strand.

Notation

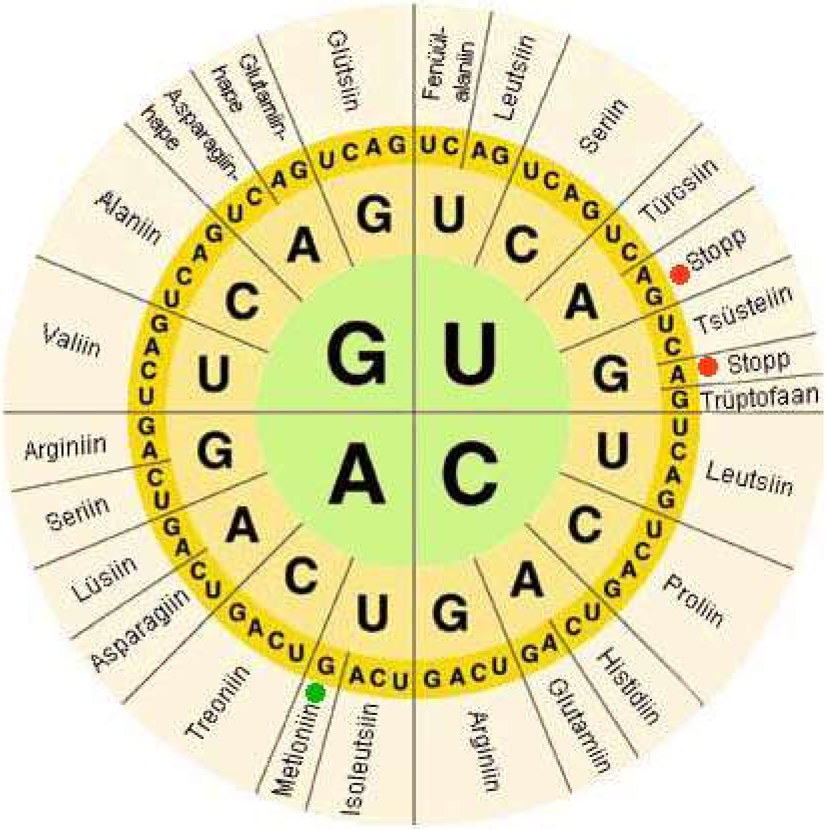
While A, T, C, and G represent a particular nucleotide at a position, there are also letters that represent ambiguity. Of all the molecules sampled, there is more than one kind of nucleotide at that position. The rules of the International Union of Pure and Applied Chemistry (IUPAC) are as follows:

- **A** = adenine
- **C** = cytosine
- **G** = guanine
- **T** = thymine
- **R** = G A (purine)
- **Y** = T C (pyrimidine)
- **K** = G T (keto)
- **M** = A C (amino)
- **S** = G C (strong bonds)
- **W** = A T (weak bonds)
- **B** = G T C (all but A)
- **D** = G A T (all but C)
- **H** = A C T (all but G)
- **V** = G C A (all but T)
- **N** = A G C T (any)

These symbols are also valid for RNA, except with U (uracil) replacing T (thymine).

Apart from adenine (A), cytosine (C), guanine (G), thymine (T) and uracil (U), DNA and RNA also contain bases that have been modified after the nucleic acid chain has been formed. In DNA, the most common modified base is 5-methylcytidine (m5C). In RNA, there are many modified bases, including pseudouridine (Ψ), dihydrouridine (D), inosine (I), ribothymidine (rT) and 7-methylguanosine (m7G). Hypoxanthine and xanthine are two of the many bases created through mutagen presence, both of them through deamination (replacement of the amine-group with a carbonyl-group). Hypoxanthine is produced from adenine, xanthine from guanine. Similarly, deamination of cytosine results in uracil.

Biological significance

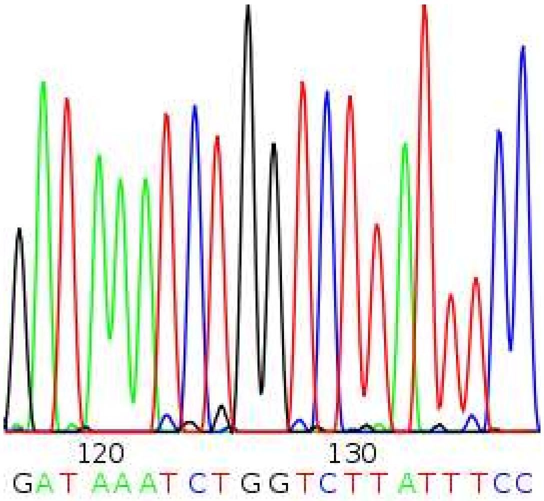


A depiction of the genetic code, by which the information contained in nucleic acids are translated into amino acid sequences in proteins.

In biological systems, nucleic acids contain information which is used by a living cell to construct specific proteins. The sequence of nucleobases on a nucleic acid strand is translated by cell machinery into a sequence of amino acids making up a protein strand. Each group of three bases, called a codon, corresponds to a single amino acid, and there is a specific genetic code by which each possible combination of three bases corresponds to a specific amino acid.

The central dogma of molecular biology outlines the mechanism by which proteins are constructed using information contained in nucleic acids. DNA is transcribed into mRNA molecules, which travels to the ribosome where the mRNA is used as a template for the construction of the protein strand. Since nucleic acids can bind to molecules with complementary sequences, there is a distinction between "sense" sequences which code for proteins, and the complementary "antisense" sequence which is by itself nonfunctional, but can bind to the sense strand.

Sequence determination



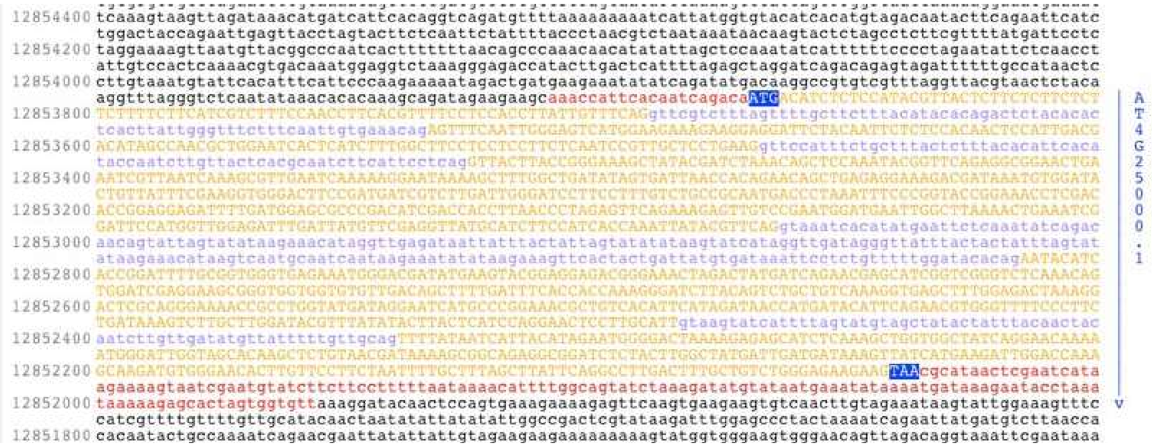
Electropherogram printout from automated sequencer for determining part of a DNA sequence

DNA sequencing is the process of determining the nucleotide sequence of a given DNA fragment. The sequence of DNA encodes the necessary information for living things to survive and reproduce. Determining the sequence is therefore useful in fundamental research into why and how organisms live, as well as in applied subjects. Because of the key nature of DNA to living things, knowledge of DNA sequence may come in useful in practically any biological research. For example, in medicine it can be used to identify, diagnose and potentially develop treatments for genetic diseases. Similarly, research into pathogens may lead to treatments for contagious diseases. Biotechnology is a burgeoning discipline, with the potential for many useful products and services.

RNA is not sequenced directly. Instead, it is copied to a DNA by reverse transcriptase, and this DNA is then sequenced.

Current sequencing methods rely on the discriminatory ability of DNA polymerases, and can therefore only distinguish four bases. An inosine (created from adenosine during RNA editing) will be read as a G, and 5-methyl-cytosine (created from cytosine by DNA methylation) will be read as a C. It is also currently difficult to sequence small amounts of DNA, as the signal will be too weak to measure. This is overcome by PCR amplification.

Digital format



```
12854400 tcaaagtaagttagataaacatgatcattcacaggtcagatgttttaaaaaaaaatcattatggtgtacatcacatgtagacaatacttcagaattcatc
         tggactaccagaattgagttacctagtacttctcaattctattttaccctaacgtctaataaataacaagtactctagcctcttcgttttatgattcctc
12854200 taggaaaagttaatgttacggcccaatcactttttttaacagcccaaacaacatatattagctccaaatatcatttttcccctagaatattctcaacct
         attgtccactcaaaacgtgacaaatggaggtctaaagggagaccatacttgactcatttagagctaggatcagacagagtagatttttgccataactc
12854000 cttgtaaatgtattcacatttcattcccaagaaaaatagactgatgaagaaatatatcagatatgacaaggccgtgtcgtttaggttacgtaactctaca
         aggtttagggtctcaatataaacacacaaagcagatagaagaagcaaaccattcacaatcagacaATGACATCTCTCCATACGTTACTCTTCTCTTCTCT
12853800 TCTTTTCTTCATCGTCTTTCCAACCTTCACGTTTTCCTCCACCTTATTGTTTCAGgttcgtctttagttttgcttctttacatacacagactctacacac
         tcacttattgggtttctttcaattgtgaaacagAGTTTCAATTGGAGTCATGGAAGAAAGAAGGAGGATTCTACAATTCTCTCCACAACTCCATTGACG
12853600 ACATAGCCAACGCTGGAATCACTCATCTTTGGCTTCCTCCTCCTTCTCAATCCGTTGCTCCTGAAGgttccatttctgctttactctttacacattcaca
         taccaatcttgttactcacgcaatcttcattcctcagGTTACTTACCGGGAAAGCTATACGATCTAAACAGCTCCAAATACGGTTCAGAGGCGGAACTGA
12853400 AATCGTTAATCAAAGCGTTGAATCAAAAAGGAATAAAAGCTTTGGCTGATATAGTGATTAACCACAGAACAGCTGAGAGGAAAGACGATAAATGTGGATA
         CTGTTATTTCGAAGGTGGGACTTCCGATGATCGTCTTGATTGGGATCCTTCCTTTGTCTGCCGCAATGACCCTAAATTTCCCGGTACCGGAAACCTCGAC
12853200 ACCGGAGGAGATTTTGATGGAGCGCCCGACATCGACCACCTTAACCCTAGAGTTCAGAAAGAGTTGTCCGAATGGATGAATTGGCTTAAAACTGAAATCG
         GATTCCATGGTTGGAGATTTGATTATGTTCGAGGTTATGCATCTTCCATCACCAAATTATACGTTCAGgtaaatcacatatgaattctcaaatatcagac
12853000 aacagtattagtatataagaaacataggttgagataattatttactattagtatatataagtatcataggttgatagggttatttactactatttagtat
         ataagaaacataagtcaatgcaatcaataagaaatatataagaaagttcactactgattatgtgataaattcctctgtttttggatacacagAATACATC
12852800 ACCGGATTTTGCGGTGGGTGAGAAATGGACGATATGAAGTACGGAGGAGACGGGAAACTAGACTATGATCAGAACGAGCATCGGTCGGGTCTCAAACAG
         TGGATCGAGGAAGCGGGTGGTGGTGTGTTGACAGCTTTTGATTTCACCACCAAAGGGATCTTACAGTCTGCTGTCAAAGGTGAGCTTTGGAGACTAAAGG
12852600 ACTCGCAGGGAAAACCGCCTGGTATGATAGGAATCATGCCCGGAAACGCTGTCACATTCATAGATAACCATGATACATTCAGAACGTGGGTTTTCCCTTC
         TGATAAAGTCTTGCTTGGATACGTTTATATACTTACTCATCCAGGAACTCCTTGCATTgtaagtatcattttagtatgtagctatactatttacaactac
12852400 aatcttgttgatatgttatttttgttgcagTTTTATAATCATTACATAGAATGGGGACTAAAAGAGAGCATCTCAAAGCTGGTGGCTATCAGGAACAAAA
         ATGGGATTGGTAGCACAAGCTCTGTAACGATAAAAGCGGCAGAGGCGGATCTCTACTTGGCTATGATTGATGATAAAGTTATCATGAAGATTGGACCAAA
12852200 GCAAGATGTGGGAACACTTGTTCCTTCTAATTTTGCTTTAGCTTATTCAGGCCTTGACTTTGCTGTCTGGGAGAAGAAGTAAcgcataactcgaatcata
         agaaaagtaatcgaatgtatcttcttcctttttaataaaacatttggcagtatctaaagatatgtataatgaaatataaaatgataaagaatacctaaa
12852000 taaaaagagcactagtggtgttaaaggatacaactccagtgaaagaaaagagttcaagtgaagaagtgtcaacttgtagaaataagtattggaaagtttc
         catcgttttgttttgttgcatacaactaatatattatatattggccgactcgtataagatttggagccctactaaaatcagaattatgatgtcttaacca
12851800 cacaatactgccaaaatcagaacgaattatattattgtagaagaagaaaaaaaaagtatggtgggaagtgggaacagttagacaggtaaattcgaataaa
```

Genetic sequence in digital format

Once a nucleic acid sequence has been obtained from an organism, it is stored *in silico* in digital format. Digital genetic sequences may be stored in sequence databases, be analyzed, be digitally altered and/or be used as templates for creating new actual DNA using artificial gene synthesis.

Sequence analysis

Digital genetic sequences may be analyzed using the tools of bioinformatics to attempt to determine its function.

Genetic testing

The DNA in an organism's genome can be analyzed to diagnose vulnerabilities to inherited diseases, and can also be used to determine a child's paternity (genetic father) or a person's ancestry. Normally, every person carries two copies of every gene, one inherited from their mother, one inherited from their father. The human genome is believed to contain around 20,000 - 25,000 genes. In addition to studying chromosomes to the level of individual genes, genetic testing in a broader sense includes biochemical tests for the possible presence of genetic diseases, or mutant forms of genes associated with increased risk of developing genetic disorders.

Genetic testing identifies changes in chromosomes, genes, or proteins. Most of the time, testing is used to find changes that are associated with inherited disorders. The results of a genetic test can confirm or rule out a suspected genetic condition or help determine a person's chance of developing or passing on a genetic disorder. Several hundred genetic tests are currently in use, and more are being developed.

Sequence alignment

In bioinformatics, a **sequence alignment** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels (that is, insertion or deletion mutations) introduced in one or both lineages in the time since they diverged from one another. In sequence alignments of proteins, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineages. The absence of substitutions, or the presence of only very conservative substitutions (that is, the substitution of amino acids whose side chains have similar biochemical properties) in a particular region of the sequence, suggest that this region has structural or functional importance. Although DNA and RNA nucleotide bases are more similar to each other than are amino acids, the conservation of base pairs can indicate a similar functional or structural role.

Computational phylogenetics makes extensive use of sequence alignments in the construction and interpretation of phylogenetic trees, which are used to classify the evolutionary relationships between homologous genes represented in the genomes of divergent species. The degree to which sequences in a query set differ is qualitatively related to the sequences' evolutionary distance from one another. Roughly speaking, high sequence identity suggests that the sequences in question have a comparatively young most recent common ancestor, while low identity suggests that the divergence is more ancient. This approximation, which reflects the "molecular clock" hypothesis that a roughly constant rate of evolutionary change can be used to extrapolate the elapsed time since two genes first diverged (that is, the coalescence time), assumes that the effects of mutation and selection are constant across sequence lineages. Therefore it does not account for possible difference among organisms or species in the rates of DNA repair or the possible functional conservation of specific regions in a sequence. (In the case of nucleotide sequences, the molecular clock hypothesis in its most basic form also discounts the difference in acceptance rates between silent mutations that do not alter the meaning of a given codon and other mutations that result in a different amino acid being incorporated into the protein.) More statistically accurate methods allow the evolutionary rate on each branch of the phylogenetic tree to vary, thus producing better estimates of coalescence times for genes.

Sequence motifs

Frequently the primary structure encodes motifs that are of functional importance. Some examples of sequence motifs are: the C/D and H/ACA boxes of snoRNAs, Sm binding site found in spliceosomal RNAs such as U1, U2, U4, U5, U6, U12 and U3, the Shine-Dalgarno sequence, the Kozak consensus sequence and the RNA polymerase III terminator.

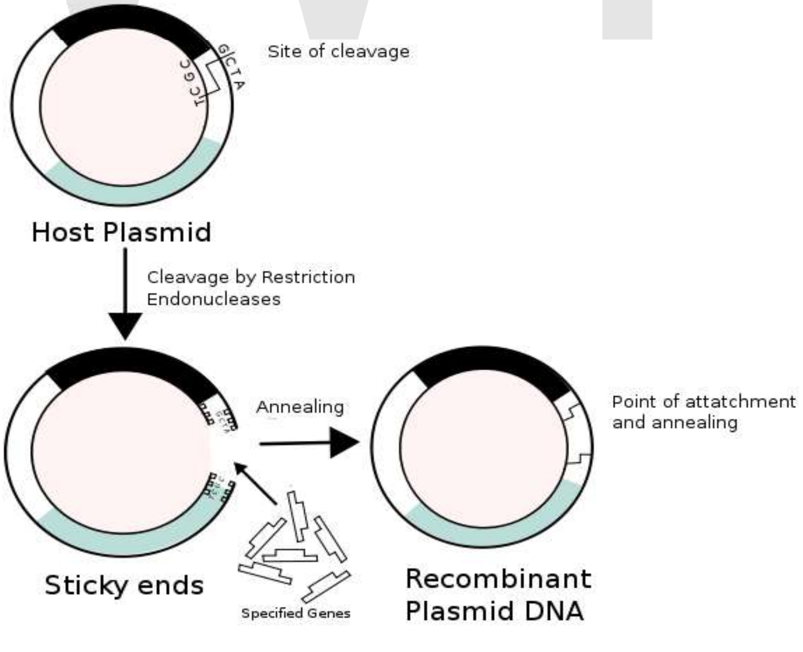
Chapter- 19

Recombinant DNA

Recombinant DNA (rDNA) is a form of artificial DNA that is created by combining two or more sequences that would not normally occur together through the process of gene splicing. In terms of genetic modification, it is created through the introduction of relevant DNA into an existing organismal DNA, such as the plasmids of bacteria, to code for or alter different traits for a specific purpose, such as antibiotic resistance. It differs from genetic recombination in that it does not occur through natural processes within the cell, but is engineered. A **recombinant protein** is a protein that is derived from recombinant DNA.

Methods

Cloning and relation to plasmids



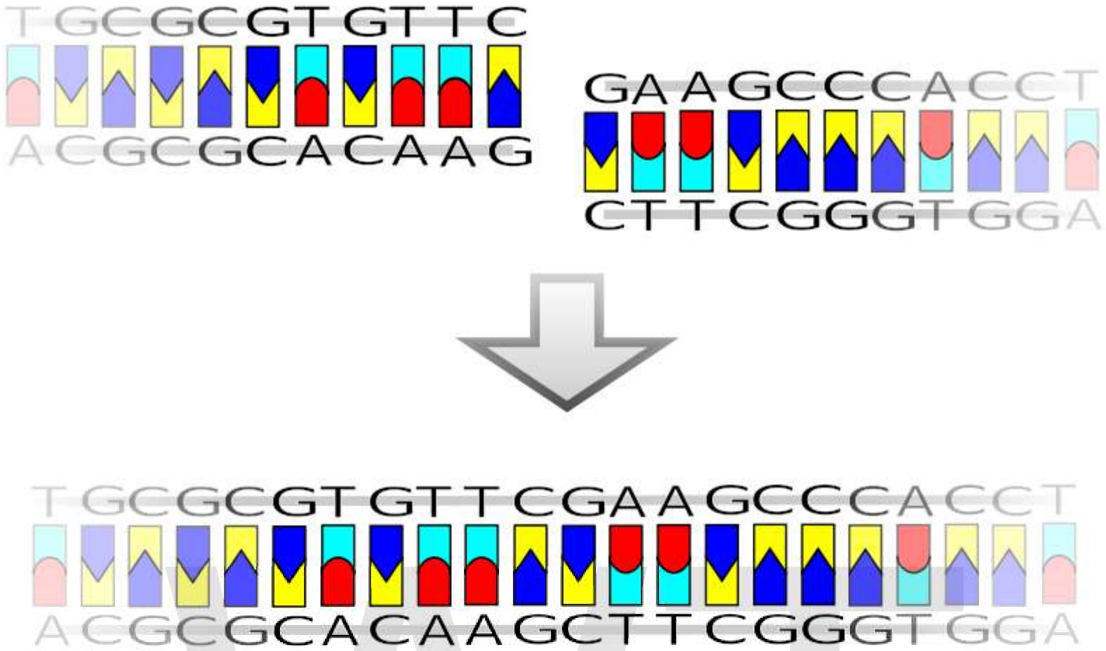
A simple example of how a desired gene is inserted into a plasmid. In this example, the gene specified in the white color becomes useless as the new gene is added.

The use of cloning is interrelated with recombinant DNA in classical biology, as the term "clone" refers to a cell or organism derived from a parental organism, with modern biology referring to the term as a collection of cells derived from the same cell that remain identical. In the classical instance, the use of recombinant DNA provides the initial cell from which the host organism is then expected to recapitulate when it undergoes further cell division, with bacteria remaining a prime example due to the use of viral vectors in medicine that contain recombinant DNA inserted into a structure known as a plasmid.

Plasmids are extrachromosomal self-replicating circular forms of DNA present in most bacteria, such as *Escherichia coli* (E. Coli), containing genes related to catabolism and metabolic activity, and allowing the carrier bacterium to survive and reproduce in conditions present within other species and environments. These genes represent characteristics of resistance to bacteriophages and antibiotics and some heavy metals, but can also be fairly easily removed or separated from the plasmid by restriction endonucleases, which regularly produce "sticky ends" and allow the attachment of a selected segment of DNA, which codes for more "reparative" substances, such as peptide hormone medications including insulin, growth hormone, and oxytocin. In the introduction of useful genes into the plasmid, the bacteria are then used as a viral vector, which are encouraged to reproduce so as to recapitulate the altered DNA within other cells it infects, and increase the amount of cells with the recombinant DNA present within them.

The use of plasmids is also key within gene therapy, where their related viruses are used as **cloning vectors** or carriers, which are means of transporting and passing on genes in recombinant DNA through viral reproduction throughout an organism. Plasmids contain three common features—a **replicator**, **selectable marker** and a **cloning site**. The replicator or "ori" refers to the origin of replication with regard to location and bacteria where replication begins. The marker refers to a particular gene that usually contains resistance to an antibiotic, but may also refer to a gene that is attached alongside the desired one, such as that which confers luminescence to allow identification of successfully recombined DNA. The cloning site is a sequence of nucleotides representing one or more positions where cleavage by restriction endonucleases occurs. Most eukaryotes do not maintain canonical plasmids; yeast is a notable exception. In addition, the Ti plasmid of the bacterium *Agrobacterium tumefaciens* can be used to integrate foreign DNA into the genomes of many plants. Other methods of introducing or creating recombinant DNA in eukaryotes include homologous recombination and transfection with modified viruses.

Chimeric plasmids



An example of chimeric plasmid formation from two "blunt ends" via the enzyme, T4 Ligase.

When recombinant DNA is then further altered or changed to host additional strands of DNA, the molecule formed is referred to as "chimeric" DNA molecule, with reference to the mythological chimera, which consisted as a composite of several animals. The presence of chimeric plasmid molecules is somewhat regular in occurrence, as, throughout the lifetime of an organism, the propagation by vectors ensures the presence of hundreds of thousands of organismal and bacterial cells that all contain copies of the original chimeric DNA.

In the production of chimeric(from chimera) plasmids, the processes involved can be somewhat uncertain, as the intended outcome of the addition of foreign DNA may not always be achieved and may result in the formation of unusable plasmids. Initially, the plasmid structure is linearised to allow the addition by bonding of complementary foreign DNA strands to single-stranded "overhangs" or "sticky ends" present at the ends of the DNA molecule from staggered, or "S-shaped" cleavages produced by restriction endonucleases.

A common vector used for the donation of plasmids originally was the bacterium Escherichia coli and, later, the EcoRI derivative, which was used for its versatility with addition of new DNA by "relaxed" replication when inhibited by chloramphenicol and spectinomycin, later being replaced by the pBR322 plasmid. In the case of EcoRI, the plasmid can anneal with the presence of foreign DNA via the route of sticky-end ligation, or with "blunt ends" via blunt-end ligation, in the presence of the phage T_4 ligase, which forms covalent links between 3-carbon OH and 5-carbon PO_4 groups present on blunt ends. Both sticky-end, or overhang ligation and blunt-end ligation can occur between foreign DNA segments, and cleaved ends of the original plasmid depending upon the restriction endonuclease used for cleavage.

Applications

There are multitudinous proteins that are created from recombinant DNA and used as medications. Some can alternatively be produced from animal extracts or harvested from humans, such as human growth hormone (rHGH), human insulin, follicle-stimulating hormone (FSH) and factor VIII. Other proteins, when used as medication, only has recombinant DNA as a source, such as with erythropoietin.

History

The recombinant DNA technique was first proposed by Peter Lobban, a graduate student, with A. Dale Kaiser at the Stanford University Department of Biochemistry. The technique was then realized by Lobban and Kaiser; Jackson, Symons and Berg; and Stanley Norman Cohen, Chang, Herbert Boyer and Helling, in 1972–74. They published their findings in papers including the 1972 paper "*Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of Escherichia coli*", the 1973 paper "*Enzymatic end-to-end joining of DNA molecules*" and the 1973 paper *"Construction of Biologically Functional Bacterial Plasmids* in vitro*"*, all of which described techniques to isolate and amplify genes or DNA segments and insert them into another cell with precision, creating a transgenic bacterium.

Exploitation of recombinant DNA technology was facilitated by the discovery, isolation and application of restriction endonucleases by Werner Arber, Daniel Nathans, and Hamilton Smith, for which they received the 1978 Nobel Prize in Medicine. Cohen and Boyer applied for a patent on the Process for producing biologically functional molecular chimeras which could not exist in nature in 1974. The patent was granted in 1980.

A breakthrough in the application of recombinant DNA technology occurred in 1977 when Herbert Boyer produced biosynthetic "human" insulin in the lab. The specific gene sequence, or polynucleotide, that codes for insulin production in humans was introduced to a sample colony of the *E. coli* bacteria. It was the first medicine made via recombinant DNA technology to be approved by the FDA and commercially available under the brand name Humulin. The vast majority of insulin currently used worldwide is now biosynthetic recombinant "human" insulin or its analogs.

Chapter- 20

Polymerase Chain Reaction

A strip of eight PCR tubes, each containing a 100 µl reaction mixture

The **polymerase chain reaction** (**PCR**) is a scientific technique in molecular biology to amplify a single or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

Developed in 1983 by Kary Mullis, PCR is now a common and often indispensable technique used in medical and biological research labs for a variety of applications. These include DNA cloning for sequencing, DNA-based phylogeny, or functional analysis of genes; the diagnosis of hereditary diseases; the identification of genetic fingerprints (used in forensic sciences and paternity testing); and the detection and diagnosis of infectious diseases. In 1993, Mullis was awarded the Nobel Prize in Chemistry for his work on PCR.

The method relies on thermal cycling, consisting of cycles of repeated heating and cooling of the reaction for DNA melting and enzymatic replication of the DNA. Primers (short DNA fragments) containing sequences complementary to the target region along with a DNA polymerase (after which the method is named) are key components to enable selective and repeated amplification. As PCR progresses, the DNA generated is itself used as a template for replication, setting in motion a chain reaction in which the DNA template is exponentially amplified. PCR can be extensively modified to perform a wide array of genetic manipulations.

Almost all PCR applications employ a heat-stable DNA polymerase, such as Taq polymerase, an enzyme originally isolated from the bacterium *Thermus aquaticus*. This DNA polymerase enzymatically assembles a new DNA strand from DNA building blocks, the nucleotides, by using single-stranded DNA as a template and DNA oligonucleotides (also called DNA primers), which are required for initiation of DNA synthesis. The vast majority of PCR methods use thermal cycling, i.e., alternately heating and cooling the PCR sample to a defined series of temperature steps. These thermal cycling steps are necessary first to physically separate the two strands in a DNA double helix at a high temperature in a process called DNA melting. At a lower temperature, each strand is then used as the template in DNA synthesis by the DNA polymerase to selectively amplify the target DNA. The selectivity of PCR results from the use of primers that are complementary to the DNA region targeted for amplification under specific thermal cycling conditions.

PCR principles and procedure

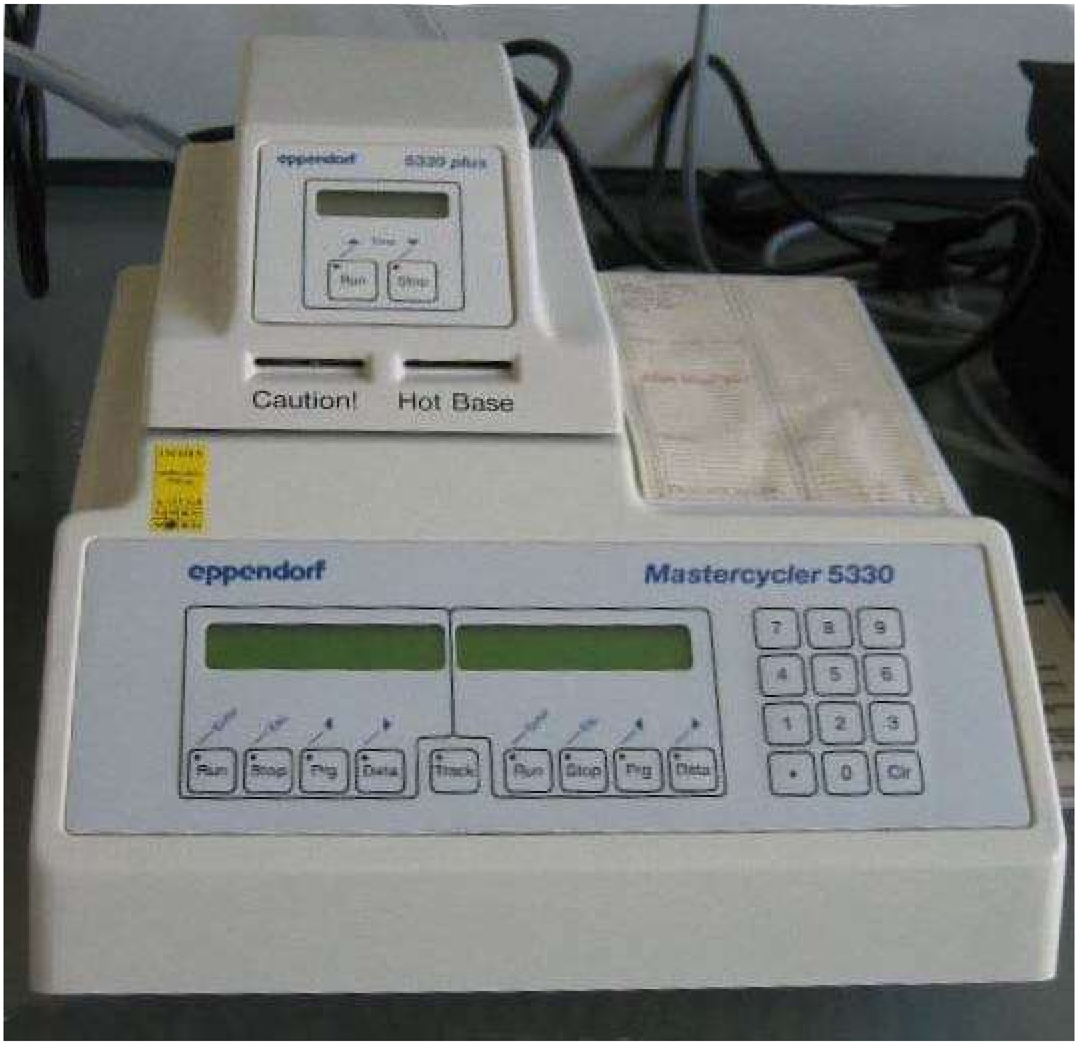


Figure 1a: A thermal cycler for PCR

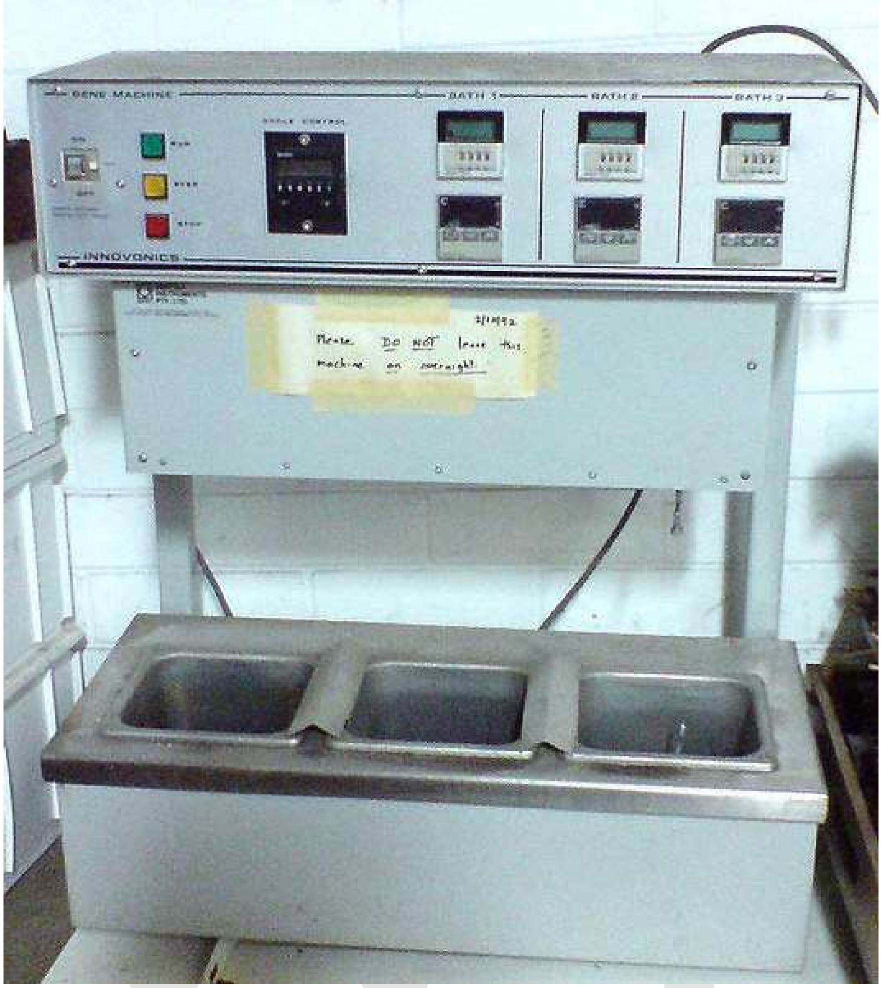


Figure 1b: An older model three-temperature thermal cycler for PCR

PCR is used to amplify a specific region of a DNA strand (the DNA target). Most PCR methods typically amplify DNA fragments of up to ~10 kilo base pairs (kb), although some techniques allow for amplification of fragments up to 40 kb in size.

A basic PCR set up requires several components and reagents. These components include:

- *DNA template* that contains the DNA region (target) to be amplified.
- Two *primers* that are complementary to the 3' (three prime) ends of each of the sense and anti-sense strand of the DNA target.
- *Taq polymerase* or another DNA polymerase with a temperature optimum at around 70 °C.
- *Deoxynucleotide triphosphates* (dNTPs), the building blocks from which the DNA polymerases synthesizes a new DNA strand.
- *Buffer solution,* providing a suitable chemical environment for optimum activity and stability of the DNA polymerase.

- *Divalent cations,* magnesium or manganese ions; generally Mg^{2+} is used, but Mn^{2+} can be utilized for PCR-mediated DNA mutagenesis, as higher Mn^{2+} concentration increases the error rate during DNA synthesis
- *Monovalent cation* potassium ions.

The PCR is commonly carried out in a reaction volume of 10–200 μl in small reaction tubes (0.2–0.5 ml volumes) in a thermal cycler. The thermal cycler heats and cools the reaction tubes to achieve the temperatures required at each step of the reaction (see below). Many modern thermal cyclers make use of the Peltier effect which permits both heating and cooling of the block holding the PCR tubes simply by reversing the electric current. Thin-walled reaction tubes permit favorable thermal conductivity to allow for rapid thermal equilibration. Most thermal cyclers have heated lids to prevent condensation at the top of the reaction tube. Older thermocyclers lacking a heated lid require a layer of oil on top of the reaction mixture or a ball of wax inside the tube.

Procedure

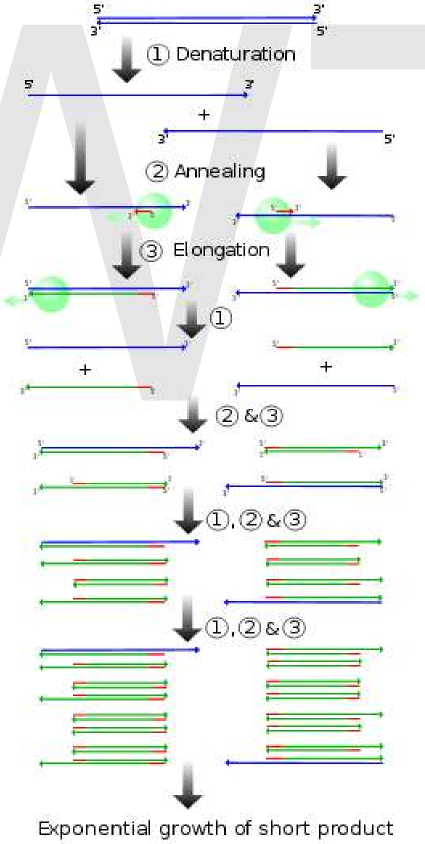


Figure 2: Schematic drawing of the PCR cycle. **(1) Denaturing at 94–96 °C. (2) Annealing at ~65 °C (3) Elongation at 72 °C**. Four cycles are shown here. The blue lines represent the DNA template to which primers (red arrows) anneal that are extended by the DNA polymerase (light green circles), to give shorter DNA products (green lines), which themselves are used as templates as PCR progresses.

Typically, PCR consists of a series of 20-40 repeated temperature changes, called cycles, with each cycle commonly consisting of 2-3 discrete temperature steps, usually three (Fig. 2). The cycling is often preceded by a single temperature step (called *hold*) at a high temperature (>90°C), and followed by one hold at the end for final product extension or brief storage. The temperatures used and the length of time they are applied in each cycle depend on a variety of parameters. These include the enzyme used for DNA synthesis, the concentration of divalent ions and dNTPs in the reaction, and the melting temperature (Tm) of the primers.

- *Initialization step*: This step consists of heating the reaction to a temperature of 94–96 °C (or 98 °C if extremely thermostable polymerases are used), which is held for 1–9 minutes. It is only required for DNA polymerases that require heat activation by hot-start PCR.

- *Denaturation step*: This step is the first regular cycling event and consists of heating the reaction to 94–98 °C for 20–30 seconds. It causes DNA melting of the DNA template by disrupting the hydrogen bonds between complementary bases, yielding single-stranded DNA molecules.

- *Annealing step*: The reaction temperature is lowered to 50–65 °C for 20–40 seconds allowing annealing of the primers to the single-stranded DNA template. Typically the annealing temperature is about 3-5 degrees Celsius below the Tm of the primers used. Stable DNA-DNA hydrogen bonds are only formed when the primer sequence very closely matches the template sequence. The polymerase binds to the primer-template hybrid and begins DNA synthesis.

- *Extension/elongation step*: The temperature at this step depends on the DNA polymerase used; Taq polymerase has its optimum activity temperature at 75–80 °C, and commonly a temperature of 72 °C is used with this enzyme. At this step the DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template in 5' to 3' direction, condensing the 5'-phosphate group of the dNTPs with the 3'-hydroxyl group at the end of the nascent (extending) DNA strand. The extension time depends both on the DNA polymerase used and on the length of the DNA fragment to be amplified. As a rule-of-thumb, at its optimum temperature, the DNA polymerase will polymerize a thousand bases per minute. Under optimum conditions, i.e., if there are no limitations due to limiting substrates or reagents, at each extension step, the amount of DNA target is doubled, leading to exponential (geometric) amplification of the specific DNA fragment.

- *Final elongation*: This single step is occasionally performed at a temperature of 70–74 °C for 5–15 minutes after the last PCR cycle to ensure that any remaining single-stranded DNA is fully extended.

- *Final hold*: This step at 4–15 °C for an indefinite time may be employed for short-term storage of the reaction.

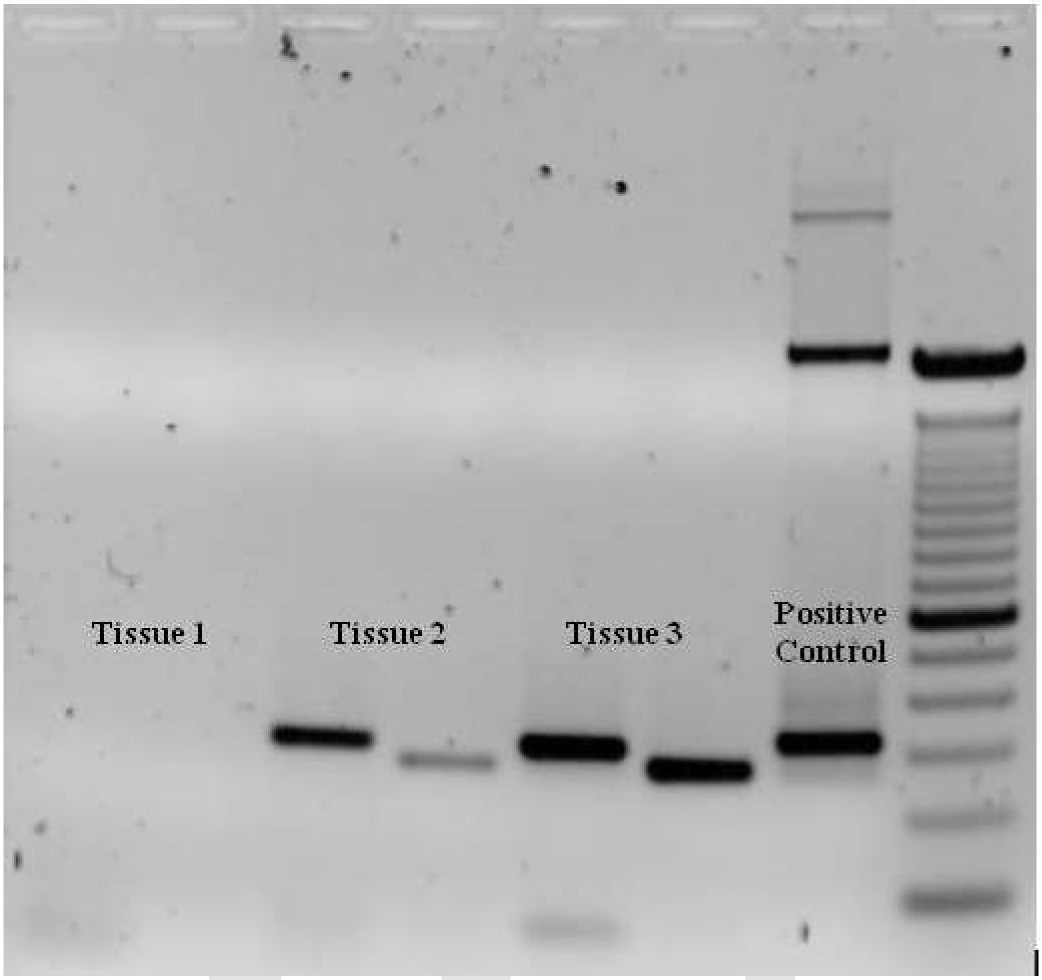


Figure 3: Ethidium bromide-stained PCR products after gel electrophoresis. Two sets of primers were used to amplify a target sequence from three different tissue samples. No amplification is present in sample #1; DNA bands in sample #2 and #3 indicate successful amplification of the target sequence. The gel also shows a positive control, and a DNA ladder containing DNA fragments of defined length for sizing the bands in the experimental PCRs.

To check whether the PCR generated the anticipated DNA fragment (also sometimes referred to as the amplimer or amplicon), agarose gel electrophoresis is employed for size separation of the PCR products. The size(s) of PCR products is determined by comparison with a DNA ladder (a molecular weight marker), which contains DNA fragments of known size, run on the gel alongside the PCR products (see Fig. 3).

PCR stages

The PCR process can be divided into three stages:

Exponential amplification: At every cycle, the amount of product is doubled (assuming 100% reaction efficiency). The reaction is very sensitive: only minute quantities of DNA need to be present.

Levelling off stage: The reaction slows as the DNA polymerase loses activity and as consumption of reagents such as dNTPs and primers causes them to become limiting.

Plateau: No more product accumulates due to exhaustion of reagents and enzyme.

PCR optimization

In practice, PCR can fail for various reasons, in part due to its sensitivity to contamination causing amplification of spurious DNA products. Because of this, a number of techniques and procedures have been developed for optimizing PCR conditions. Contamination with extraneous DNA is addressed with lab protocols and procedures that separate pre-PCR mixtures from potential DNA contaminants. This usually involves spatial separation of PCR-setup areas from areas for analysis or purification of PCR products, use of disposable plasticware, and thoroughly cleaning the work surface between reaction setups. Primer-design techniques are important in improving PCR product yield and in avoiding the formation of spurious products, and the usage of alternate buffer components or polymerase enzymes can help with amplification of long or otherwise problematic regions of DNA. Addition of reagents, such as formamide, in buffer systems may increase the specificity and yield of PCR.

Application of PCR

Selective DNA isolation

PCR allows isolation of DNA fragments from genomic DNA by selective amplification of a specific region of DNA. This use of PCR augments many methods, such as generating hybridization probes for Southern or northern hybridization and DNA cloning, which require larger amounts of DNA, representing a specific DNA region. PCR supplies these techniques with high amounts of pure DNA, enabling analysis of DNA samples even from very small amounts of starting material.

Other applications of PCR include DNA sequencing to determine unknown PCR-amplified sequences in which one of the amplification primers may be used in Sanger sequencing, isolation of a DNA sequence to expedite recombinant DNA technologies involving the insertion of a DNA sequence into a plasmid or the genetic material of another organism. Bacterial colonies (E. coli) can be rapidly screened by PCR for correct DNA vector constructs. PCR may also be used for genetic fingerprinting; a forensic technique used to identify a person or organism by comparing experimental DNAs through different PCR-based methods.

Some PCR 'fingerprints' methods have high discriminative power and can be used to identify genetic relationships between individuals, such as parent-child or between siblings, and are used in paternity testing (Fig. 4). This technique may also be used to determine evolutionary relationships among organisms.

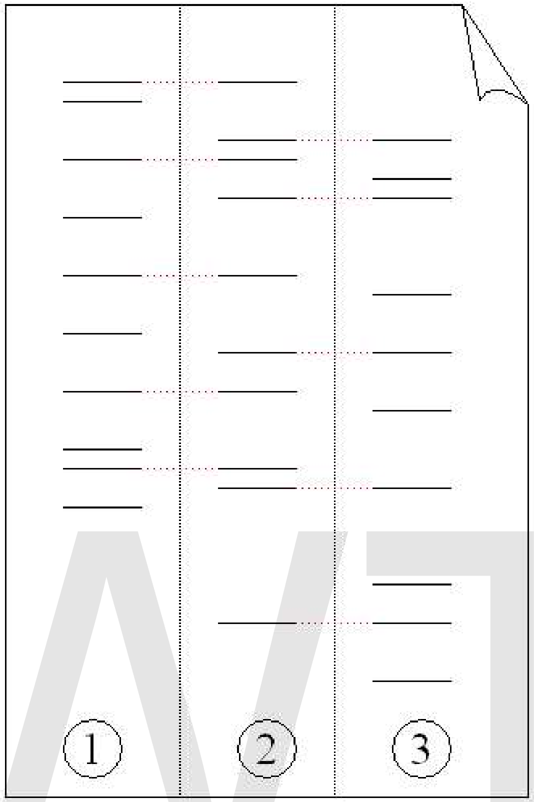


Figure 4: Electrophoresis of PCR-amplified DNA fragments. (1) Father. (2) Child. (3) Mother. The child has inherited some, but not all of the fingerprint of each of its parents, giving it a new, unique fingerprint.

Amplification and quantification of DNA

Because PCR amplifies the regions of DNA that it targets, PCR can be used to analyze extremely small amounts of sample. This is often critical for forensic analysis, when only a trace amount of DNA is available as evidence. PCR may also be used in the analysis of ancient DNA that is tens of thousands of years old. These PCR-based techniques have been successfully used on animals, such as a forty-thousand-year-old mammoth, and also on human DNA, in applications ranging from the analysis of Egyptian mummies to the identification of a Russian tsar.

Quantitative PCR methods allow the estimation of the amount of a given sequence present in a sample—a technique often applied to quantitatively determine levels of gene expression. Real-time PCR is an established tool for DNA quantification that measures the accumulation of DNA product after each round of PCR amplification.

PCR in diagnosis of diseases

PCR permits early diagnosis of malignant diseases such as leukemia and lymphomas, which is currently the highest developed in cancer research and is already being used routinely. PCR assays can be performed directly on genomic DNA samples to detect translocation-specific malignant cells at a sensitivity which is at least 10,000 fold higher than other methods.

PCR also permits identification of non-cultivatable or slow-growing microorganisms such as mycobacteria, anaerobic bacteria, or viruses from tissue culture assays and animal models. The basis for PCR diagnostic applications in microbiology is the detection of infectious agents and the discrimination of non-pathogenic from pathogenic strains by virtue of specific genes.

Viral DNA can likewise be detected by PCR. The primers used need to be specific to the targeted sequences in the DNA of a virus, and the PCR can be used for diagnostic analyses or DNA sequencing of the viral genome. The high sensitivity of PCR permits virus detection soon after infection and even before the onset of disease. Such early detection may give physicians a significant lead in treatment. The amount of virus ("viral load") in a patient can also be quantified by PCR-based DNA quantitation techniques (see below).

Variations on the basic PCR technique

- *Allele-specific PCR*: a diagnostic or cloning technique which is based on single-nucleotide polymorphisms (SNPs) (single-base differences in DNA). It requires prior knowledge of a DNA sequence, including differences between alleles, and uses primers whose 3' ends encompass the SNP. PCR amplification under stringent conditions is much less efficient in the presence of a mismatch between template and primer, so successful amplification with an SNP-specific primer signals presence of the specific SNP in a sequence.

- *Assembly PCR* or *Polymerase Cycling Assembly (PCA)*: artificial synthesis of long DNA sequences by performing PCR on a pool of long oligonucleotides with short overlapping segments. The oligonucleotides alternate between sense and antisense directions, and the overlapping segments determine the order of the PCR fragments, thereby selectively producing the final long DNA product.

- *Asymmetric PCR*: preferentially amplifies one DNA strand in a double-stranded DNA template. It is used in sequencing and hybridization probing where amplification of only one of the two complementary strands is required. PCR is carried out as usual, but with a great excess of the primer for the strand targeted for amplification. Because of the slow (arithmetic) amplification later in the reaction after the limiting primer has been used up, extra cycles of PCR are required. A recent modification on this process, known as *L*inear-*A*fter-*T*he-*E*xponential-PCR (LATE-PCR), uses a limiting primer with a higher melting temperature (Tm) than the excess primer to maintain reaction efficiency as the limiting primer concentration decreases mid-reaction.

- *Helicase-dependent amplification*: similar to traditional PCR, but uses a constant temperature rather than cycling through denaturation and annealing/extension cycles. DNA helicase, an enzyme that unwinds DNA, is used in place of thermal denaturation.

- *Hot-start PCR*: a technique that reduces non-specific amplification during the initial set up stages of the PCR. It may be performed manually by heating the reaction components to the melting temperature (e.g., 95°C) before adding the polymerase. Specialized enzyme systems have been developed that inhibit the polymerase's activity at ambient temperature, either by the binding of an antibody or by the presence of covalently bound inhibitors that only dissociate after a high-temperature activation step. Hot-start/cold-finish PCR is achieved with new hybrid polymerases that are inactive at ambient temperature and are instantly activated at elongation temperature.

- *Intersequence-specific PCR* (ISSR): a PCR method for DNA fingerprinting that amplifies regions between simple sequence repeats to produce a unique fingerprint of amplified fragment lengths.

- *Inverse PCR*: is commonly used to identify the flanking sequences around genomic inserts. It involves a series of DNA digestions and self ligation, resulting in known sequences at either end of the unknown sequence.

- *Ligation-mediated PCR*: uses small DNA linkers ligated to the DNA of interest and multiple primers annealing to the DNA linkers; it has been used for DNA sequencing, genome walking, and DNA footprinting.

- *Methylation-specific PCR* (MSP): developed by Stephen Baylin and Jim Herman at the Johns Hopkins School of Medicine, and is used to detect methylation of CpG islands in genomic DNA. DNA is first treated with sodium bisulfite, which converts unmethylated cytosine bases to uracil, which is recognized by PCR primers as thymine. Two PCRs are then carried out on the modified DNA, using primer sets identical except at any CpG islands within the primer sequences. At these points, one primer set recognizes DNA with cytosines to amplify methylated DNA, and one set recognizes DNA with uracil or thymine to amplify unmethylated DNA. MSP using qPCR can also be performed to obtain quantitative rather than qualitative information about methylation.

- *Miniprimer PCR*: uses a thermostable polymerase (S-Tbr) that can extend from short primers ("smalligos") as short as 9 or 10 nucleotides. This method permits PCR targeting to smaller primer binding regions, and is used to amplify conserved DNA sequences, such as the 16S (or eukaryotic 18S) rRNA gene.

- *Multiplex Ligation-dependent Probe Amplification* (*MLPA*): permits multiple targets to be amplified with only a single primer pair, thus avoiding the resolution limitations of multiplex PCR (see below).

- *Multiplex-PCR*: consists of multiple primer sets within a single PCR mixture to produce amplicons of varying sizes that are specific to different DNA sequences. By targeting multiple genes at once, additional information may be gained from a single test run that otherwise would require several times the reagents and more time to perform. Annealing temperatures for each of the primer sets must be optimized to work correctly within a single reaction, and amplicon sizes, i.e., their base pair length, should be different enough to form distinct bands when visualized by gel electrophoresis.

- *Nested PCR*: increases the specificity of DNA amplification, by reducing background due to non-specific amplification of DNA. Two sets of primers are used in two successive PCRs. In the first reaction, one pair of primers is used to generate DNA products, which besides the intended target, may still consist of non-specifically amplified DNA fragments. The product(s) are then used in a second PCR with a set of primers whose binding sites are completely or partially different from and located 3' of each of the primers used in the first reaction. Nested PCR is often more successful in specifically amplifying long DNA fragments than conventional PCR, but it requires more detailed knowledge of the target sequences.

- *Overlap-extension PCR*: a genetic engineering technique allowing the construction of a DNA sequence with an alteration inserted beyond the limit of the longest practical primer length.

- *Quantitative PCR (Q-PCR)*: used to measure the quantity of a PCR product (commonly in real-time). It quantitatively measures starting amounts of DNA, cDNA or RNA. Q-PCR is commonly used to determine whether a DNA sequence is present in a sample and the number of its copies in the sample. *Quantitative real-time PCR* has a very high degree of precision. QRT-PCR methods use fluorescent dyes, such as Sybr Green, EvaGreen or fluorophore-containing DNA probes, such as TaqMan, to measure the amount of amplified product in real time. It is also sometimes abbreviated to RT-PCR (*R*eal *T*ime PCR) or RQ-PCR. QRT-PCR or RTQ-PCR are more appropriate contractions, since RT-PCR commonly refers to reverse transcription PCR (see below), often used in conjunction with Q-PCR.

- *R*everse *T*ranscription PCR (RT-PCR): for amplifying DNA from RNA. Reverse transcriptase reverse transcribes RNA into cDNA, which is then amplified by PCR. RT-PCR is widely used in expression profiling, to determine the expression of a gene or to identify the sequence of an RNA transcript, including transcription start and termination sites. If the genomic DNA sequence of a gene is known, RT-PCR can be used to map the location of exons and introns in the gene. The 5' end of a gene (corresponding to the transcription start site) is typically identified by RACE-PCR (*Rapid Amplification of cDNA Ends*).

- *Solid Phase PCR*: encompasses multiple meanings, including Polony Amplification (where PCR colonies are derived in a gel matrix, for example), Bridge PCR (primers are covalently linked to a solid-support surface), conventional Solid Phase PCR (where Asymmetric PCR is applied in the presence of solid support bearing primer with sequence matching one of the aqueous primers) and Enhanced Solid Phase PCR (where conventional Solid Phase PCR can be improved by employing high Tm and nested solid support primer with optional application of a thermal 'step' to favour solid support priming).

- *Thermal asymmetric interlaced PCR (TAIL-PCR)*: for isolation of an unknown sequence flanking a known sequence. Within the known sequence, TAIL-PCR uses a nested pair of primers with differing annealing temperatures; a degenerate primer is used to amplify in the other direction from the unknown sequence.

- *Touchdown PCR* (*Step-down PCR*): a variant of PCR that aims to reduce nonspecific background by gradually lowering the annealing temperature as PCR cycling progresses. The annealing temperature at the initial cycles is usually a few degrees (3-5°C) above the T_m of the primers used, while at the later cycles, it is a few degrees (3-5°C) below the primer T_m. The higher temperatures give greater specificity for primer binding, and the lower temperatures permit more efficient amplification from the specific products formed during the initial cycles.

- *PAN-AC*: uses isothermal conditions for amplification, and may be used in living cells.

- *Universal Fast Walking*: for genome walking and genetic fingerprinting using a more specific 'two-sided' PCR than conventional 'one-sided' approaches (using only one gene-specific primer and one general primer - which can lead to artefactual 'noise') by virtue of a mechanism involving lariat structure formation. Streamlined derivatives of UFW are LaNe RAGE (lariat-dependent nested PCR for rapid amplification of genomic DNA ends), 5'RACE LaNe and 3'RACE LaNe.

History

A 1971 paper in the Journal of Molecular Biology by Kleppe and co-workers first described a method using an enzymatic assay to replicate a short DNA template with primers *in vitro*. However, this early manifestation of the basic PCR principle did not receive much attention, and the invention of the polymerase chain reaction in 1983 is generally credited to Kary Mullis.

At the core of the PCR method is the use of a suitable DNA polymerase able to withstand the high temperatures of >90 °C (194 °F) required for separation of the two DNA strands in the DNA double helix after each replication cycle. The DNA polymerases initially employed for in vitro experiments presaging PCR were unable to withstand these high temperatures. So the early procedures for DNA replication were very inefficient, time consuming, and required large amounts of DNA polymerase and continual handling throughout the process.

The discovery in 1976 of Taq polymerase — a DNA polymerase purified from the thermophilic bacterium, *Thermus aquaticus*, which naturally lives in hot (50 to 80 °C (122 to 176 °F)) environments such as hot springs — paved the way for dramatic improvements of the PCR method. The DNA polymerase isolated from *T. aquaticus* is stable at high temperatures remaining active even after DNA denaturation, thus obviating the need to add new DNA polymerase after each cycle. This allowed an automated thermocycler-based process for DNA amplification.

When Mullis developed the PCR in 1983, he was working in Emeryville, California for Cetus Corporation, one of the first biotechnology companies. There, he was responsible for synthesizing short chains of DNA. Mullis has written that he conceived of PCR while cruising along the Pacific Coast Highway one night in his car. He was playing in his mind with a new way of analyzing changes (mutations) in DNA when he realized that he had instead invented a method of amplifying any DNA region through repeated cycles of duplication driven by DNA polymerase. In *Scientific American*, Mullis summarized the procedure: "Beginning with a single molecule of the genetic material DNA, the PCR can generate 100 billion similar molecules in an afternoon. The reaction is easy to execute. It requires no more than a test tube, a few simple reagents, and a source of heat." He was awarded the Nobel Prize in Chemistry in 1993 for his invention, seven years after he and his colleagues at Cetus first put his proposal to practice. However, some controversies have remained about the intellectual and practical contributions of other scientists to Mullis' work, and whether he had been the sole inventor of the PCR principle (see below).

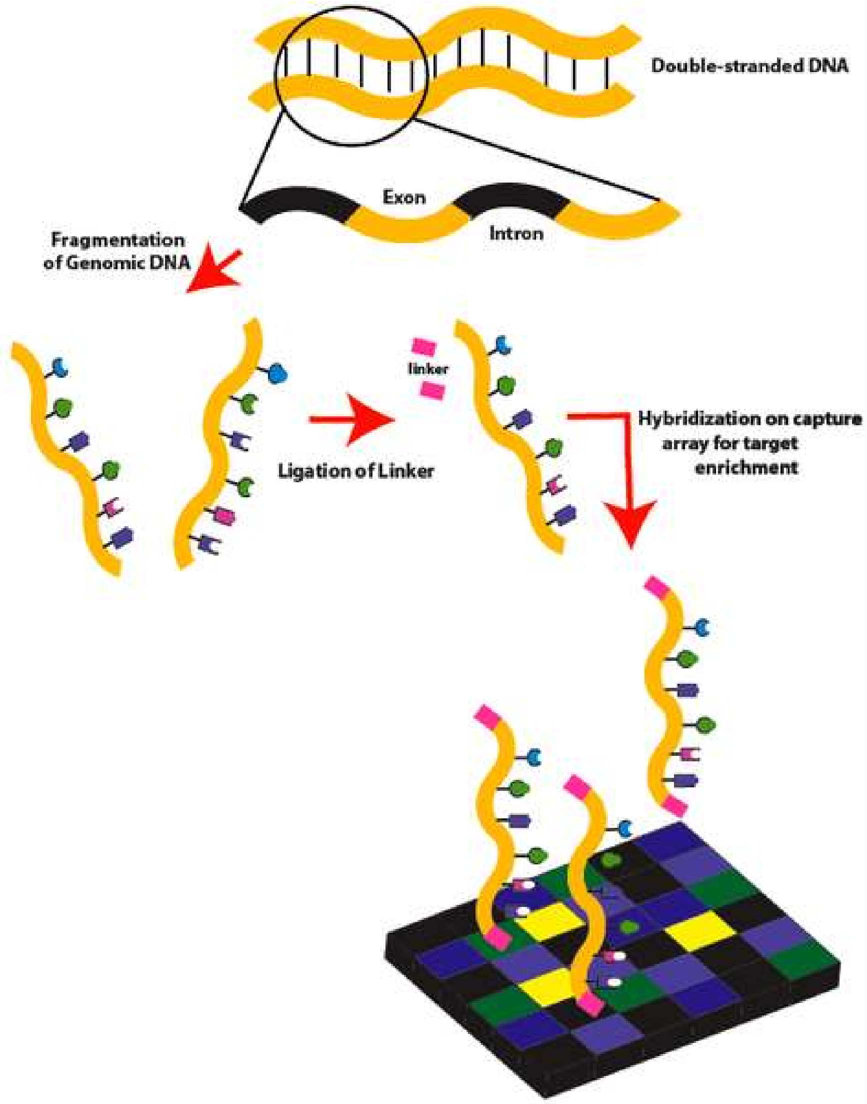
Patent wars

The PCR technique was patented by Kary Mullis and assigned to Cetus Corporation, where Mullis worked when he invented the technique in 1983. The *Taq* polymerase enzyme was also covered by patents. There have been several high-profile lawsuits related to the technique, including an unsuccessful lawsuit brought by DuPont. The pharmaceutical company Hoffmann-La Roche purchased the rights to the patents in 1992 and currently holds those that are still protected.

A related patent battle over the Taq polymerase enzyme is still ongoing in several jurisdictions around the world between Roche and Promega. The legal arguments have extended beyond the lives of the original PCR and Taq polymerase patents, which expired on March 28, 2005.

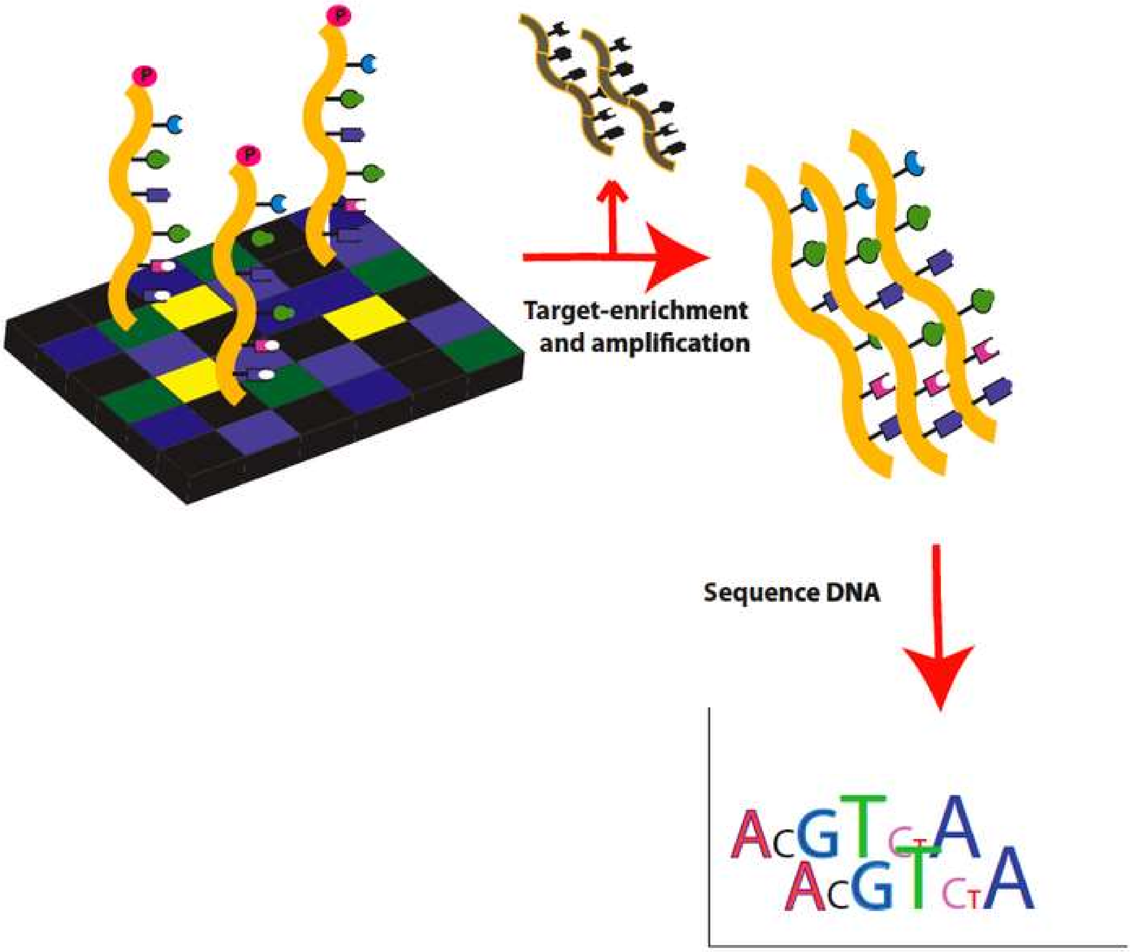
Chapter- 21

Exome Sequencing


Double-stranded DNA
Exon
Intron
Fragmentation of Genomic DNA
linker
Ligation of Linker
Hybridization on capture array for target enrichment

Exome Sequencing Workflow: Part 1.

Exome sequencing (also known as **targeted exome capture**) is an efficient strategy to selectively sequence the coding regions of the human genome to identify novel genes associated with rare and common disorders. Routine whole genome sequencing of large numbers of individuals is still not feasible partly due to the high cost associated with the technique. At present, it is necessary to use an alternative approach, in which certain regions of the genome, such as the "exome", are targeted, enriched and sequenced, which requires ~5% as much sequencing as a whole genome. The "exome" represents all the exons in the human genome (i.e., the transcribed region of the genome). Exons are short, functionally important sequences of DNA which represent the regions in genes that are translated into protein and untranslated region flanking them (UTR). UTRs are usually not included in exome studies. In total there are about 180,000 exons found in the human genome. These protein coding regions constitute about 1% of the human genome which translates to about 30 megabases (Mb) in length. It is estimated that the protein coding regions of the human genome constitute about 85% of the disease-causing mutations.



Exome Sequencing Workflow: Part 2.

The robust approach to sequencing the complete coding region (exome) has the potential to be clinically relevant in genetic diagnosis due to current understanding of functional consequences in sequence variation. The goal of this approach is to identify the functional variation that is responsible for both mendelian and common diseases such as Miller syndrome and Alzheimer's disease without the high costs associated with whole-genome sequencing while maintaining high coverage in sequence depth.

As an efficient strategy

Exome sequencing is an efficient strategy to identify these rare causal variants of mendelian disorders over whole genome sequencing due to few factors:

1. Positional cloning strategies have reduced power to successfully identify causal rare variants
2. The majority of genetic variants that underlie mendelian disorders disrupt protein-coding sequences
3. A large number of rare nonsynonymous substitutions are predicted to be deleterious
4. Splice sites also represent sequences in which there is high functional variation

The exome represents an enriched portion of the genome that can be used to search for variants with large effect sizes.

Mendelian disorders

Rare diseases affect less than 200,000 individuals in the United States and are of interest because the identification of the genetic basis can provide knowledge about biological pathways and therapeutic targets. It is suspected that there are more than 7,000 rare mendelian diseases which affect millions of people in the US. The majority of mendelian diseases studied to date are known to be caused by rare mutations that affect protein function. The majority of mutations that are known to cause mendelian disorders are located in protein-coding regions while non-coding regions on the other hand are likely to have weak or neutral effects.

To date, less than half of all rare monogenic disorders have been discovered. The identification of genetic variants for rare disorders is limited by a number of factors. These include sample size of affected individuals, reduced penetrance, locus heterogeneity, and alleles that impair reproductive fitness. These factors make it difficult to map these traits by linkage analysis and they reduce the power of traditional positional cloning strategies to identify these variants. For both dominant and recessive traits finding an excess of independent mutations in the same locus will provide evidence that a disease gene has been identified. Exome sequencing is a powerful technique to identify genes in rare mendelian disorders because it requires only a small number of unrelated cases to identify a causal gene.

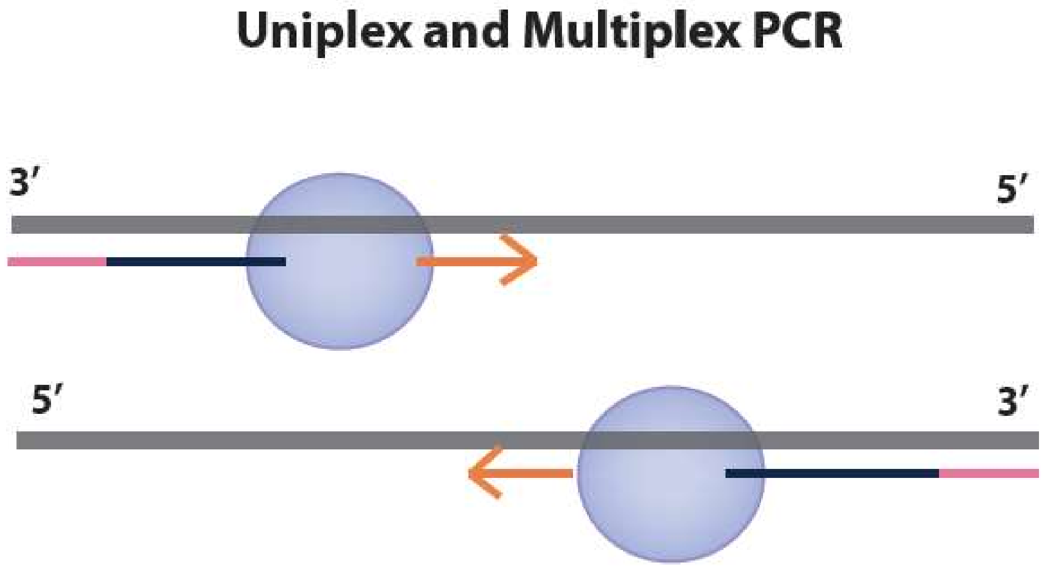
Technological platforms

The technical platforms used to carry out exome sequencing are DNA microarrays and magnetic bead based systems for the enrichment of the exome DNA and next-generation sequencing technologies.

Target-enrichment strategies

Target-enrichment methods allow to selectively capture genomic regions of interest from a DNA sample prior to sequencing. Several target-enrichment strategies have been developed.

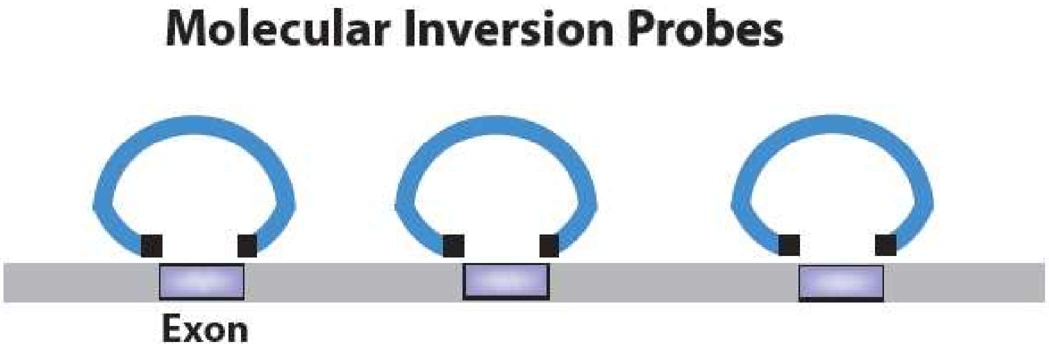
PCR



Uniplex and Multiplex PCR

PCR is one of the most widely used enrichment strategies for over 20 years. This approach is known to be useful in classical Sanger sequencing because a uniplex PCR used to generate a single DNA sequence is comparable in read length to a typical amplicon. Multiplex PCR reactions which require several primers are challenging although strategies to get around this have been developed. A limitation to this method is the size of the genomic target due to workload and quantity of DNA required. The PCR based approach is highly effective, yet it is not feasible to target genomic regions that are several megabases in size due to quantity of DNA required and cost.

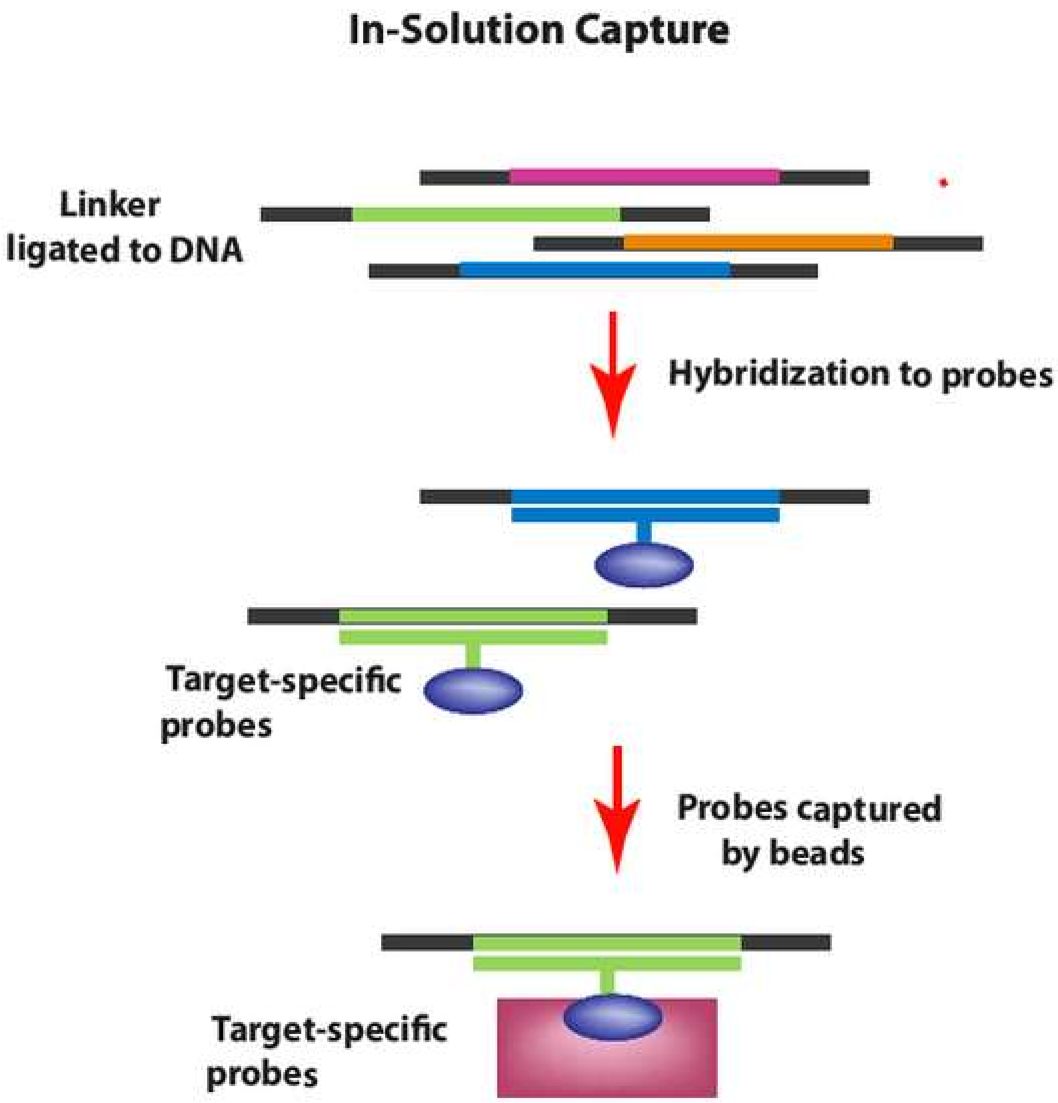
Molecular Inversion Probes (MIP)



Molecular Inversion Probes

This is an enzymatic technique that targets the amplification of genomic regions by multiplexing based on target circularization. Accurate genotypes can be achieved from massively parallel sequencing using this method. This method is suggested to be useful for small numbers of targets in a large number of samples. Major disadvantage of this method for target enrichment is the capture uniformity as well as the cost associated with covering large target sets.

Hybrid capture



In-Solution Capture

This technique involves hybridizing shotgun libraries of genomic DNA to target-specific sequences on a microarray. Roche NimbleGen was first to take this technology and adapt it for next-generation sequencing. They developed the Sequence Capture Human Exome 2.1M Array to capture ~180,000 coding exons. This method is both time-saving and cost-effective compared to PCR based methods. The Agilent Capture Array and the comparative genomic hybridization array also other methods that can be used for hybrid capture of target sequences. Limitations in this technique include the need for expensive hardware as well as a relatively large amount of DNA.

In-solution capture

To capture genomic regions of interest using in-solution capture, a pool of custom oligonucleotides (probes) is synthesized and hybridized in solution to a fragmented genomic DNA sample. The probes (labeled with beads) selectively hybridize to the genomic regions of interest after which the beads (now including the DNA fragments of interest) can be pulled down and washed to clear excess material. The beads are then removed and the genomic fragments can be sequenced allowing for selective DNA sequencing of genomic regions (e.g. exons) of interest. Several companies (e.g. FlexGen) offer custom pools of oligonucleotides or instruments to synthesize these oligopools in-house.

This method was developed to improve on the hybridization capture target-enrichment method. In solution capture as opposed to hybrid capture, there is an excess of probes to target regions of interest over the amount of template required. The optimal target size is about 3.5 Mb in length and yields excellent sequence coverage of the target regions. The preferred method is dependent on several factors including; size (bp) of region of interest, demands for reads on target, equipment in house, etc.

Sequencing

There are several sequencing platforms available including the classical Sanger sequencing. Other platforms include the Roche 454 sequencer, the Illumina Genome Analyzer II and the Applied Biosystems SOLiD, which have both been used for exome sequencing.

Significance

A study published in September 2009 discussed a proof of concept experiment to determine if it was possible to identify causal genetic variants using exome sequencing. They sequenced four individuals with Freeman-Sheldon syndrome (FSS) (OMIM 193700), a rare autosomal dominant disorder known to be caused by a mutation in the gene MYH3. Eight HapMap individuals were also sequenced to remove common variants in order to identify the causal gene for FSS. After exclusion of common variants, the authors were able to identify MYH3, which confirms that exome sequencing can be used to identify causal variants of rare disorders. This is the first reported study that used exome sequencing as an approach to identify an unknown causal gene for a rare mendelian disorder.

Subsequently, another group reported successful clinical diagnosis of a suspected Bartter syndrome patient of Turkish origin. Bartter syndrome is a renal salt-wasting disease. Exome sequencing revealed an unexpected well-conserved recessive mutation in a gene called SLC26A3 which is associated with congenital chloride diarrhea (CLD). This molecular diagnosis of CLD was confirmed by the referring clinician. This example provided proof of concept of the use of whole-exome sequencing as a clinical tool in evaluation of patients with undiagnosed genetic illnesses. This report is regarded as the first application of next generation sequencing technology for molecular diagnosis of a patient.

A second report was conducted on exome sequencing of individuals with a mendelian disorder known as Miller syndrome (MIM#263750), a rare disorder of autosomal recessive inheritance. Two siblings and two unrelated individuals with Miller syndrome were studied. They looked at variants that have the potential to be pathogenic such as non-synonymous mutations, splice acceptor and donor sites and short coding insertions or deletions. Since Miller syndrome is a rare disorder, it is expected that the causal variant has not been previously identified. Previous exome sequencing studies of common single nucleotide polymorphisms (SNPs) in public SNP databases were used to further exclude candidate genes. After exclusion of these genes, the authors found mutations in DHODH that were shared among individuals with Miller syndrome. Each individual with Miller syndrome was a compound heterozygote for the DHODH mutations which was inherited as each parent of an affected individual was found to be a carrier.

This is the first time exome sequencing has been shown to identify a novel gene responsible for a rare mendelian disease. This exciting finding demonstrates is that exome sequencing has the potential to locate causative genes in complex diseases, which previously has not been possible due to limitations in traditional methods. Targeted capture and massively parallel sequencing represents a cost-effective, reproducible and robust strategy with high sensitivity and specificity to detect variants causing protein-coding changes in individual human genomes.

Comparison with genotyping

There are multiple technologies available to undertake methods to identify causal genetic variants associated with disease. Each technology has its own technical, financial and throughput limitations. Microarrays for example, require hybridization probes of known sequence and are therefore limited by probe design and thus prevent the identification of genetic changes that can be detected. Massively parallel sequencing technologies used for exome sequencing on the other hand makes it now possible to identify the cause of many unknown diseases by screening thousands of loci at once. This technology addresses the present limitations of hybridization genotyping arrays and classical sequencing.

Although, exome sequencing is an expensive method relative to other technologies (e.g., hybridization-based technologies) currently available, it is an efficient strategy to identify the genetic bases that underlie rare mendelian disorders. This approach has become increasingly practical with the falling cost and increased throughput of whole genome sequencing. Even by only sequencing the exomes of individuals, a large quantity of data and sequence information is generated which requires a significant amount of data analysis. Challenges associated with the analysis of this data include changes in programs used to align and assemble sequence reads. Various sequence technologies also have different error rates and generate various read-lengths which can pose challenges in comparing results from different sequencing platforms.

Limitations

Exome sequencing is able to only identify those variants found in the coding region of genes which affect protein function. It is not able to identify the structural and non-coding variants associated the disease which can be found using other methods such as whole genome sequencing. There remains 99% of the human genome that is not covered using exome sequencing. Whole genome sequencing will eventually become a standard approach and allow us to gain a deeper understanding of genetic variation found in populations. Presently, this technique is not practical due to the high costs and time associated with sequencing large numbers of genomes. Exome sequencing allows sequencing of portions of the genome over at least 20 times as many samples compared to whole genome sequencing. For translation of identified rare variants into the clinic, sample size and the ability to interpret the results to provide a clinical diagnosis indicates that with the current knowledge in genetics, exome sequencing may be the most valuable.

The statistical analysis of the large quantity of data generated from sequencing approaches is a challenge. False positive and false negative findings are associated with genomic resequencing approaches and it is a critical issue. A few strategies have been developed to improve the quality of exome data such as:

- Comparing the genetic variants identified between sequencing and array-based genotyping
- Comparing the coding SNPs to a whole genome sequenced individual with the disorder
- Comparing the coding SNPs with Sanger sequencing of HapMap individuals

Rare recessive disorders would not have single nucleotide polymorphisms (SNPs) in public databases such dbSNP. More common recessive phenotypes may have disease-causing variants reported in dbSNP. For example, the most common cystic fibrosis variant has an allele frequency of about 3% in most populations. Screening out such variants might erroneously exclude such genes from consideration. Genes for recessive disorders are usually easier to identify than dominant disorders because the genes are less likely to have more than one rare nonsynonymous variant. The system screen common genetic variants relies on dbSNP which may not have accurate information about the variation of alleles. Using lists of common variation from a study exome or genome-wide sequenced individual would be more reliable. A challenge in this approach is that as the number of exomes sequenced increases, dbSNP will also increase in the number of uncommon variants. It will be necessary to develop thresholds to define the common variants that are unlikely to be associated with a disease phenotype.

Genetic heterogeneity and population ethnicity are also major limitations as it may increase the number false positive and false negative findings which will make the identification of candidate genes more difficult. Of course it is possible to reduce the stringency of the thresholds in the presence of heterogeneity and ethnicity, however it will reduce the power to detect variants as well.

Ethical implications

New technologies in genomics has changed the way researchers approach both basic and translational research. With approaches such as exome sequencing it is possible to significantly enhance the data generated from individual genomes which has put forth a series of questions on how to deal with the vast amount of information. Should the individuals in these studies be allowed to have access to their sequencing information? Is it possible to interpret theses results for these individuals and are the identified genetic variants clinically relevant? This data can lead to unexpected findings and complicate clinical utility and patient benefit.This area of genomics still remains a challenge and researchers are looking into how to address these questions.