

**In the following I will try to give synopsis of some biological background from Genetics, Molecular Biology, and Bio-techniques point of view.**

### **Genetic Point of View**

The idea of modern genetics is to determine the transmission rules (i.e., the relationship between parents and offsprings) of measured traits. We use the terminology “phenotype” to discuss any measured trait of interest in this sense. For example, gene expression is a phenotype. More technically, some people talk of the phenotype as the result of “genotype” and “environment” interaction—meaning that it is the set of observables that result as the decoding of the genetic information under some environmental context. Ideally, we would like to know the transmission rules under all the possible complicated contexts. That is, if we are interested in “eye color” as a trait, we want to make sure we understand the transmission rules for eye color given all the other possible environments and the genetic state of the organism. In the ideal case, all the other possible measurements (or their underlying genetic factors) do not matter with respect to eye color transmission rules. Some traits behave in this ideal fashion. More interesting traits such as, say “propensity for Alzheimer’s”, are very context dependent on a lot of other genetic factors as well as the particular environment. These days, people use the catchall phrase “complex trait genetics” to describe such cases. At any rate, our goal from a genetics point of view is to understand the informational basis (i.e., the genotype) of phenotypic traits. Sometimes we say we are interested in the “genotype to the phenotype map”, however, this is meant to be a metaphoric statement and not a mathematically precise statement.

The major problem in biology in the 19th century was the generative process (so-called “development”) of how a complex adult arises from a simple egg. The idea that the somatic cells (cells that comprise most of the body) and the germline cells (the cells that give rise to next generation of somatic cells) are decoupled (= Weismann’s theory of the “germplasm”) led to the idea of looking for “determinants of traits” within the germline cells. Mendel’s observations on trait transmission rules were especially influential for postulating the idea of a “gene” as some indivisible “factor” that determines the heredity of traits.

Initially, the idea of a gene was abstract, analogized to mathematical algebra, but eventually the work of Thomas Hunt Morgan led to the “chromosomal theory of inheritance”. In this theory, chromosomes (which were visualized by light microscopes) and unknown elements within the chromosome are considered the informational basis of inheritance. Physical association of a phenotypic trait to a particular chromosome was first obtained with the fruit fly eye color and by studying the co-occurrence frequency of traits. In this particular instance, it was found that an aberrant eye color, white (normally red), was always associated with a particular sex, which in turn was always associated with a particular chromosome. Later, Morgan’s student A. H. Sturtevant produced the first “genetic map” of six traits by postulating a linear order on a “stick” (i.e., the chromosome) and association frequencies that are dependent on the distance along the stick. Today, we roughly equate the term “gene” with a substring of genomic DNA that

encodes a protein. However, it should be noted that from a genetics point of view, a gene is any necessary and sufficient inherited information factor that determines a measured trait value. This may include any segment of the genomic DNA, not necessarily a segment that encodes a protein.

The key message from the observations up to this point is that biology is determined from encoded instructions and the encoded instructions are both “indivisible” and modular suggesting a factorial combination leading to generation of biological complexity. Here by “indivisible”, we mean that that it is the total of the necessary information for a particular measured trait—that is, if it is a particular segment of the genomic DNA, then it is the smallest subsequence necessary for the phenotypic measurement of the trait. By modular, we mean that is a sufficient information source that I can “mix and match” with other information parts to produce a combination of traits.

One final note, in the genetic view-point of the organism, we first concentrate on the observable measurements of interest, say eye color, cell morphology, etc. Then, we try to figure out the necessary and sufficient transmitted informational basis for the trait. (Now we know that this information is encoded by the genomic DNA.) One might say that this is a “top down” approach where the observables determine what is relevant. This contrasts with the molecular view point, where we are interested in the physical objects such as DNA and proteins and we wish to know how they interact to produce the whole organism. One might say this is a “bottom up” approach where the components are given primacy and we are interested in how this leads to the whole organism and various set of measurements at the organism level.

### **Molecular Point of View**

As noted in class, establishing the molecular and chemical basis of life was one of the greatest achievements of the 20th century. As in many other sciences, much of the progress came from continuously ruling out that biological processes were something special and different compared to other non-biological processes. For example, it first involved ruling out that life was determined by extra ordinary physical forces such as “organic waves” and then ruling out the existence of special “organic molecules”. (Note this is not to be confused with the modern classification of organic chemistry versus inorganic chemistry. At this time, the question was whether there were really life-special molecules that can’t be created outside of an organism.)

Schleiden and Schwann in 1838 proposed that (1) all life forms are made from one or more cells, (2) cells only arise from pre-existing cells, and (3) the cell is the smallest form of life. These statements, which comprise the “cell theory”, are still largely true. A gross description of a cell is as a bag of water. The cell membrane consists of a so-called lipid bilayer, meaning that it is a molecular film assembly that is a sandwich of two identical layers. (Imagine a tape with a sticky side and a smooth side and we have stuck the sticky sides together.) The interior of the membrane is an oily environment (lipid) and the exterior is water-soluble. The inside of a cell is usually called the cytoplasm. The cell itself contains other bags including various organelles (anatomical structures specialized

for some function such as the mitochondria that produces energy for the entire cell), and the nucleus, which contains the DNA. Prokaryotes are a large group of single-celled organisms (e.g., bacteria) that do not have a well-defined nucleus while Eukaryotes have a well-defined nucleus and a very different set of molecular mechanisms.

Cells contain small molecules and macromolecules. Chemical processes in the cell involve energy-coupled reactions mediated by catalytic enzymes, which are mostly proteins and occasionally catalytic RNA. An important point to consider here is that because a cell is a mostly closed system and the reactions are energy-coupled, the reactions are extremely efficient; therefore, there is a great room for complicated reactions without a great energetic cost to the organism. For example, some organisms have extremely large genomes requiring the biosynthesis of very large DNA strings. While this of course requires energy, much of the energy comes from “recycling” the energy from other reactions through their by-products.

By 1905, proteins are known to be polymers made up of amino-acids and by 1940 a vast majority of proteins are shown to be comprised of 20 amino-acids. In the meanwhile, by 1935, all enzymes mediating the coupled reactions in the cell are shown to be proteins. Around 1936, L. Pauling established the idea that the 3D configuration of the proteins is the physical basis of protein function and by 1951 Sanger determined the amino-acid sequence of the protein insulin, which eventually led to the idea of “sequence-dependent determination of protein properties.” The sequence-dependent determination of 3D form was later established by showing that one could denature a protein so that it loses its properties and then re-fold it to regain its properties without any external sources of information.

Initially, it was thought that proteins were the informational sources for inheritance because DNA was too simple. But first an experiment by Fred Griffith showed that one could take an extract of pathogenic bacteria and transform a non-pathogenic bacteria to a pathogenic one (this means a strain that is capable of invading hosts—typical infectious microbes come in pathogenic and non-pathogenic varieties). In 1944, O. Avery followed this up with purifying only the DNA portion of the bacterial extract and showing that this was sufficient to transform the bacteria. Finally, in 1952 Hershey and Chase observed a bacteriophage (a virus that infects bacteria) attacking a bacterium and showed that only the DNA of the virus enters the host—thus is sufficient to recreate new viruses.

Up to this point, it was still not clear how the DNA carries genetic information. Possibilities still included the scenario that the long string of DNA is folded in complicated ways to form templates for each protein. However, Watson and Crick’s 1953 paper showing the double helical structure of DNA provided the evidence that genetic specification is informationally encoded rather than physico-chemically determined. That is, the linear permutation sequence determines the information rather than the folded three-dimensional structure. Both DNA and RNA have special properties that make them especially appropriate for carrying information.

Proteins are comprised of a chain of amino-acids joined by a peptide bond between the amino group and the carboxyl group of individual amino-acids—thus a protein sequence has N-terminal to C-terminal polarity. Protein structure is determined by weak interactions and the final structure is dependent on the primary sequence. Major function of the proteins is conferred by the folded geometry.

DNA strings encode protein information by disjoint triplet of letters—the so-called universal genetic code. (The code is universal in the sense that the translation table from triplets of nucleotides (64 combinations) to the 20 amino acids is generally constant for most organisms, although there are known slight variations. There are also cases of “overlapping” genes where the information for one protein is encoded in one set of triplets and the same sequence with an offset can encode a different protein.) The DNA is comprised of a chain of four different nucleotides joined by a phosphodiester bond (a chemical bond involving a phosphorus atom and some oxygen atoms.) between the 5<sup>th</sup> carbon in the deoxyribose sugar to the 3<sup>rd</sup> carbon in the sugar (note, that the deoxyribose is a sugar molecule with five carbons which are somewhat arbitrarily numbered from 1 to 5). Thus DNA strings have a 5' to 3' polarity. The double helix consists of two DNA strings running 5' to 3' in one string and 3' to 5' in the other string (so-called anti-parallel strands).

Genomic information is most often carried by double stranded DNA, as discussed above, however there are variants where the information is carried by single-stranded DNA, double stranded RNA, or single stranded RNA. However, all of these variant genomes usually have some kind of double stranded DNA stage in the replication process. In addition, a genome can be comprised of linear strings, circular strings, and in various pieces (e.g., multiple chromosomes). The DNA string is incredibly long and thus it is organized into chromosomes, which are compacted structures for carrying the genomic DNA. Chromosomes of Eukaryotes contain highly packed single string of DNA wound over a protein core. Prokaryotic chromosomes are not as well organized with structural proteins and, typically, the DNA string is compacted by multiple windings. The structure of the chromosome can affect reading of the information on the DNA.

The main information transfer using DNA occurs through the phenomenon of complementary hydrogen bonding of bases. Given two single strands of DNA, there is a statistical mechanical tendency of the strands to form hydrogen bonded duplexes. The bonding happens in such a way that the A residue (Adenine) in one strand matches up with the T residue (Thymine) in the other strand; and the C residue (Cytosine) matches to the G residue (Guanine). In RNA, A matches up to a U residue (Uracil), which is a chemical analogue of T. The bonds between the C-G pair are stronger than the A-T pair. During cell replication, the double strands of DNA separate and make copies of each other to produce two copies of double strands. The primary sequence information is involved in two important biochemical processes: the template-dependent synthesis and complementary hybridization. Complementary hybridization is the phenomenon just discussed where a single strand of DNA (or RNA) with a particular sequence will tend to duplex with a single strand carrying the complementary sequence. For example, “ACCGGAT” will match with the string “TGGCCTA” to form the duplex:

ACCGGAT  
TGGCCTA

Such pairing need not be perfect, but matches are energetically preferred than mismatches (but not in a simple linear relationship).

Template-dependent synthesis is the process of creating a new string of DNA (or RNA) by copying the information from an existing strand (the template). In this process, there is a pool of available DNA (or RNA) monomers, A, C, G, and T. A polymerase enzyme, a protein for catalyzing the sequence creation, “reads” the template strand and matches the appropriate monomers and creates the phospho-diester bond connecting the bases. An important part of this process is the fact that the polymerase enzyme can not de novo start template-dependent synthesis but requires a small initial starting duplex with an available free 3’ end to attach new bases. The small sequence forming the starting duplex is often called the “primer” sequence. Therefore, by supplying a specific primer sequence and using complementary hybridization, we can initiate DNA sequence production at some specified position in the genome.

The information encoded in the DNA is turned into proteins by several steps. First, a small substring of the genome is copied into a RNA molecule called the pre-mRNA (pre-messenger RNA). This is called “expression” of the gene, “induction” of the gene, or “transcription” of the gene. (Note, it actually involves a larger string than the particular substring encoding the protein sequence.) Only one of the double strands of the DNA is copied in this manner—the particular strand that is copied for a given gene is called the “sense strand”. This idea of a “sense strand” designation is local to the particular gene. Suppose we label the whole DNA duplex string 0 and 1 for a given chromosome. This duplex string may have 100’s of genes. For any particular gene, the sense strand might be strand 0 or strand 1. That is, of the two strands 0 and 1, which strand is the sense strand changes for different genes. Because DNA and RNA have 5’ to 3’ polarity, which strand is the sense strand determines the direction in which the pre-mRNA is synthesized. Thus, on a given chromosome, genes may be arrayed in different directions (sometime even in an overlapping manner but on opposite strands).

The pre-mRNA is edited in various ways including deletion of some substrings (called splicing) and attaching some more sequences, which turns out to control the molecular behavior of the resulting polymer. This is called mRNA processing and the result is a mature mRNA. The mRNA is exported to ribosomes, which are the location at which amino-acid synthesis takes place. Each triplet genetic code in the mRNA is matched to a carrier RNA molecule called tRNA (transfer RNA). Each tRNA is an adaptor molecule associating a particular amino-acid to a particular triplet of bases. This association happens because the tRNA molecule contains the complementary triplet base (so-called anti-codon); thus there are at least twenty different tRNAs, one for each amino-acid (there can be more).

The transcription of a gene involves the action of a set of proteins that land on a specific region called the promoter of a gene and replicates a finite string of information. Prokaryotes have a reasonably conserved sequence pattern that determines the promoter region, while eukaryotes have more variable patterns. (Eukaryotic transcription also involves many more proteins than prokaryotes.) This means upstream (i.e., towards the 5' direction) of each gene, there is a special DNA sequence pattern where the transcription protein complex lands. In more complicated systems, there are many different such patterns where many different proteins land whose interactions govern the fate of transcription. Such sequence patterns are some times called *cis* regulatory factors (meaning local) as opposed to the proteins that land there which are called *trans* regulatory factors (meaning far).

As mentioned, the transcription of a segment of the genomic DNA into mRNA is not contiguous. The DNA is initially copied into a string called pre-mRNA and then edited and modified. The segments of the genomic DNA that eventually become part of the fully edited mRNA and codes for the protein sequences are called exons, while the intersegments that are edited out are called introns. (This kind of editing mostly occurs in Eukaryotes, while with most of the genes in Prokaryotes, there is a straightforward relationship between the genomic sequence and the mRNA sequence.) As mentioned, the process of editing the pre-mRNA is called splicing. The regions that are edited by the splicing process often involve sequence-dependent patterns. The sequence patterns that specify editing regions are called splice-signals. Such splicing signals are used to computationally predict the boundaries of exons and introns. However, the patterns can be variable in different genes and in different organisms. A gene is typically made up of multiple introns and exons.

mRNA is further modified by the attachment of a methylated guanine at the 5' end and poly-A sequence at the 3' end. This modification affects the chemical stability of mRNA, transport of the mRNA out of the nucleus, and translational efficiency of the mRNA. It is important to note that the presence of this poly-A tail in a mature mRNA is often used in experimental techniques to separate those mRNA from everything else. The editing pattern of pre-mRNAs may not be unique and several different mRNAs can be produced from a single genomic region. This is called alternative splicing. Other modifications can happen to the mRNA as well as to the fully translated protein.

There are multiple points of control for the copying of information from the genome to the final protein. This includes transcriptional control (DNA to pre-mRNA), RNA processing control (pre-mRNA to mRNA), RNA transport control, translational control (mRNA to protein), and protein activity control (post-translational modification). The organism is a dynamical system whose state changes both autonomously and in response to the environment. Dynamical changes are modulated by the control of protein activity at various stages as discussed above. In particular, an important control stage is at the transcription level.

An example of transcriptional control is the *lac* operon. (Operon refers to a group of sequences including genes, promoters and other factors whose transcription is regulated

together. In biology the suffix “-on” is often used to refer to a unit of something. Here operon is “operational unit”.) The *lac* operon involves the action of a specific repressor molecule that turns the transcription of *lac* metabolic genes off. The action of the repressor molecule itself is blocked when there is lactose present in the cell. A variant of the lactose called “allolactose” binds to the repressor proteins and prevents them from blocking the transcription of the *lac* metabolic genes. Thus, the system is turned “on” when lactose is available. The entire system is further modulated by a cyclic-AMP (cAMP) dependent system that governs the strength of transcription following the levels of available glucose. Cyclic-AMP is a particular molecule that accumulates in the cell when the available level of glucose is low. When there are high-levels of cAMP (low glucose) it interacts with a protein called catabolite activator protein (CAP) that enhances the rate of *lac* metabolic gene transcription. Thus, the *lac* operon control can be analogized to a electrical circuit with a on/off switch (the *lac* repressor) and a variable resistor (the cAMP system).

Gene regulation is critical in the developmental process of organisms whereby an egg turns into the full organism. One of the well studied system in development is the *Drosophila melanogaster* (a type of a fruit fly) system. Early *Drosophila* development involves gene expression control processes to establish information for front to back polarity (anterior-posterior polarity), large regions of the body (so-called “gap” genes), and segmentation pattern of an insect (so-called “pair-rule” genes). That is, the embryo starts out as a kind of homogeneous football shaped thing and various genes have to turn on and turn off to establish the information for the eventual body plan. The identity and function of these developmental genes were established by a detailed study of various mutants that had defects in their developmental process. A particularly interesting mechanism is how the segmentation pattern information is established. *Drosophila* is an insect with a segmented body plan—in fact, it has 14 segments along the body. The embryo initially contains no information about which part should be which segment. The information is provided by the “pair-rule” genes that are expressed in a regular striped pattern from front to back of the embryo thereby “painting” the location of the different segments. Each cell along the front to back of the embryo contains the full genome, thus something has to tell the pair-rule genes to turn on in certain cells and turn off in other cells to produce the stripe pattern. It turns out that this depends on a complicated *cis*-regulatory sequence pattern that provides locations for binding of the “gap” genes to the DNA. Different combinations of the various gap genes tell the pair-rule genes to turn on or turn off. Thus, the pair-rule genes are a kind of digital logic whose state depends on various different combinations of logic inputs.

Transcriptional regulation involves factors that help the transcriptional machinery assemble, locate the promoter, access the DNA, and progress along the DNA. Therefore, the process heavily depends on DNA protein interaction. We discussed above, sequence patterns that help regulate transcription. These patterns are called promoters and enhancers and as a group were called *cis*-regulatory factors. An important part of the regulatory process is the proteins that bind to the DNA. As a group, these are called trans-regulatory factors. Of course, the transcriptional complex that synthesizes the RNA is an important member of the trans-regulatory factors. However, in more complicated

regulation, other accessory proteins that bind to DNA are involved. These proteins may block the progression of the polymerase or help the progression as well as help the assembly of the protein involved in the polymerization. These are called transcription factors (or transcription activators). These proteins fall into several families delineated by their 3D structure. Prominent families include homeobox proteins, zinc finger proteins, bHLH (basic-Helix-Loop-Helix) proteins, and leucine zipper protein. These proteins often need to be multiplexed to be biologically active.

In multi-cellular organisms, different cells need to communicate with each other. Typically, this communication involves special small molecules that are produced by one cell and received by another cell. The production and reception often involves specific proteins. Communication of events from outside the cell to inside of a cell can involve several distinct mechanisms. *Transporters* move outside molecules or ions to inside, usually with high specificity for the particular molecule. Ions such as calcium and potassium are often moved in this manner. *Linkers* are heterogeneous proteins that are in the membrane and interact (become complexed) with one another when certain molecules are present in the environment. The transition from single to complexed state signals information to the interior of the cell. *Receptors* are special proteins situated in the membrane, spanning inside and outside, which bind to small molecules (called ligands). The transition from the free to bound state induces a structural change in the protein that is detected by other molecules inside the cell. There are also enzymes that penetrate the membrane and catalyze reactions inside the membrane and thus transmitting information. Once some kind of signal is transferred from outside to inside of a cell, it often triggers a cascade of molecular events called *signal transduction*. Here a sequence of molecular state changes is induced starting from the receptor. For example, the receptor might change its conformation and then catalyze the state change (usually by modifying some residues) of protein A, which in turn modifies protein B, and so on. Typically, this signal transduction cascade eventually triggers the expression/repression of several genes—that is, the signal eventually reaches the genome and causes some of the genes to turn on or off. Something called the *ras* signal transduction pathway is a well-studied system for signal transduction. The putative reason to have a cascade (i.e., a chain of molecular changes) is to amplify the outside signal and to make sure the signal triggers all the required components of the cell. Cell communication is some times classified into four modes based on whether it involves long range transfer of small molecules (called endocrine signaling), medium range small molecules (paracrine signaling), long range transfer within a very large cell with a specific receiving cell (neuronal signaling), and contact-dependent signal that requires two cells to be in direct physical contact.

### **Bio-techniques View**

Much of techniques used in biology have to do with the fact that we are dealing with small molecules that are present in small numbers. Thus, we need techniques to first segregate the molecules into homogeneous sets or have tools that specifically recognize only certain molecules, and second to make copies so that we have sufficient numbers for study, and third to have ways of recording their presence (usually through some kind of visualization).

*Techniques for purification or segregation*

- (1) **Dilution.** Given a heterogeneous set of molecule in some solution, sufficient dilution can result in a homogeneous population (albeit at low concentrations) of molecules. This usually requires starting out with a lot of material and can be inefficient. The technique is often used in isolating pure cultures of microorganisms—because once a low concentration of homogeneous cells are obtained, they can be made to self-replicate.
- (2) **Sieving.** Sieving involves trying to move molecules through some kind of a semi-solid matrix. The matrix might be long polymer chains (e.g., agarose, starch, acrylamide), small beads, fibers (paper), etc. A heterogeneous mixture of molecules will move through the matrix (movement can be induced by electrical field, capillary action, or gravity) at rate proportional to their relative hindrance (typically size of the molecule). Sieving is the basis of gel electrophoresis, filtration columns, sequencing gels, and it is the most commonly used separation technique.
- (3) **Movement in a field.** This is related to sieving as well, but relative rate of movement in a field without a hindrance mechanism can be directly used for separation. Examples include centrifugation with a density gradient and mass spectroscopy.
- (4) **Phase separation.** If the heterogeneous sample contains molecule that have different solvent properties such as a mixture of water soluble and oil soluble particles, a mixture of the different solvents can be used to separate the sample. In this technique, the different solvents are initially vigorously mixed with the heterogeneous sample and then allowed phase separate. The sample then separates into the different solvent layers. This technique is used for DNA preparation and many first-stage purification protocols.

*Techniques for specific recognition*

- (1) **Restriction enzyme.** Restriction enzymes are particular enzymes that catalyze the cleavage of DNA based on particular sequence motifs. That is, the protein recognizes specific DNA motifs and cleaves the DNA wherever the motif occurs. There are thousands of varieties of restriction enzymes that are typically isolated from various different bacteria.
- (2) **Complementary hybridization.** See above
- (3) **Antibody:** Antibodies are specific proteins whose 3D structures recognize specific molecular shapes and binds to them. Different antibodies can be generated to recognize different molecular structures.
- (4) **Protein-Protein/DNA protein interaction.** Many proteins specifically interact with other proteins. This property can be used to manipulate specific molecules. For example, a protein may specifically recognize and bind to the presence of multiple Histidine amino-acid. A synthetic protein can be produced with a sequence of Histidine added which then become “handles” for binding.

### *Amplification*

Note, most amplification involves DNA. That is, we either directly want amplification of specific DNA, or use the amplification of specific DNA to generate protein copies if we want protein amplification.

#### **(1) In vivo amplification. (cloning)**

Typically amplification involves incorporating a piece of target DNA into a live organism (usually a single-cell organism) then letting the host organism multiply. The key here is getting the DNA into the organism and making sure that the foreign DNA does not interfere with the biology of the organism. For technical and biological reasons, foreign DNA is often included into accessory genetic elements rather than the main genome itself. There are many different accessory genetic elements, but as a group we call these elements where we insert our target DNA, *vectors*. For example, a plasmid is a kind of accessory genetic element in a bacteria. Particular plasmids can be co-opted to insert our target DNA. Such a plasmid is called the vector for our target DNA.

The commonly used vectors are plasmids, phages (viruses that have bacteria as hosts), YACs (Yeast Artificial Chromosome), and BACs (Bacterial Artificial Chromosomes). The choice of a vector depends on a lot of things but one important consideration is their capacity for accepting foreign DNA. A plasmid can be used to insert about 0.1kb-10kb of foreign sequence, a phage 8kb- 20kb, BAC about 75k-300kb, and YAC about 100kb-1Mb.

The process of putting foreign DNA into a vector is called cloning (for obvious reasons; however in recent years this term has been confounded with the notion of discovering the sequence basis of a particular trait or function as in the phrase, “the gene for olfactory reception was cloned”. This is because identification of a particular sequence piece usually requires cloning that piece—but in the modern age of genomics such identification can be carried out without ever physically cloning the sequence.) A clone then represents a particular foreign DNA inserted into a vector. We often talk of a BAC library or YAC library and so on. These are sets of foreign DNA that have been inserted into vectors. A library may be individually identified, that is, each clone in the library is known (usually by sequencing the whole library), or simply represent a pool of inserted DNA but we don't know which vector pieces hold which DNA.

#### **(2) in vitro amplification.**

In vitro amplification of DNA involves separating the duplex DNA into single strands (usually by applying heat), attaching specific primers (typically in pairs to we have one primer for each strand), and using a DNA polymerase enzyme to carry out template-dependent synthesis. This process can be automatically repeated to exponentially amplify the targeted piece of DNA in a reaction called Polymerase

Chain Reaction (PCR) amplification. The key to this process is a DNA polymerase that is thermo-stable such that the whole reaction mix can be heated (to form single stranded DNA) and then cooled (to let the primers attach) over and over again without externally adding anything. Typical PCR can only amplify relatively short stretches of DNA 100bp~10kb. There are new novel in vitro amplification techniques that attempts to amplify entire genomes.

#### *Markers and Detection*

Many different techniques are used to measure small molecules. Typical techniques involve attaching some kind of molecule that provides for detection by electromagnetic radiation including x-rays and visible light. Marker molecules may be radioactive isotopes, fluorescent molecules, or chemiluminescent molecules. Other techniques include attaching heavy metals or specific peptides (small amino-acid chains).