

## Chapter 1

# Introduction to Modern Molecular Biology

### 1.1 Cells store large amounts of information in DNA

Molecular biology and bioinformatics have become the foundation of modern biology, as scientists have started to unravel the dynamic processes of molecular scale interactions in the marvelously complex cells, the basic units of all life. In essence, biological organisms package enormous amounts of information in a chemical form into microscopic packages, the cells. Cells are bounded by a double-layered lipid membrane which encloses the cytoplasm containing the biochemical machinery for cell maintenance, growth and duplication.

Cellular life depends crucially on proteins, which serve as structural components, as enzymes for chemical reactions in metabolism and as signalling messengers relaying information within and between cells. Thus cells can be said to be protein-based machines, which obtain energy from light or chemical compounds and use it to grow and multiply. The protein machinery is controlled by the macromolecules DNA and RNA which are the repositories of the large amount of genetic information, as explained below.

The main control center of the complex cellular biochemical machine is the genetic material of DNA (Deoxyribo Nucleic Acid), which contains the information about all the proteins to be produced and about the control systems in the metabolism of the cell. The information coded in the totality of DNA in one cell (called the genome) determines the identity and characteristics of

cells, and is responsible for transmitting this information to the next generation of cells. Thus the genome functions to preserve the identity and special characteristics of the self-replicating cell. An additional important property of the DNA is its mutability. Changes in DNA modify the genome information slightly and cause the next generation of cells to be slightly different. These variant cells are then exposed to natural selection to screen out the most successful cells based on the best changes in the genome. This natural selection of cells with the best genomes in each generation allows cells and organisms to adapt their cell machinery to new situations in successive generations. This in essence makes evolution possible.

The other main information-processing components of cells are RNAs (Ribo Nucleic Acid) which act as snippets of information delivered from DNA to cell protein machinery. They also have important regulatory and enzymatic activities by themselves, and represent probably the most ancient cell regulatory molecules.

In addition to proteins and DNA and RNA, cells contain a large variety of other kinds of polymers and lipids, which serve as structural components, *e.g.* supporting scaffold structures and enclosing membranes of various cell organelles. Cells also include many different kinds of organic chemicals, both simple and complex, which are processed and utilized in the cell metabolism.

The total chemical complexity in cells is huge: a single cell of a higher eukaryotic organism can contain 30,000 genes in the DNA, corresponding to at least 100,000 different RNAs, about 200,000 protein variants (including splice variants with differing amino acid sequences and variants with post-translational modifications, like phosphorylation, glycosylation *etc.*, as explained below) and at least some 20,000 other organic compounds used in the metabolism. This is discussed further in the next subchapter 1.2.

The three main branches in the taxonomy of cellular organisms consist of single-celled bacteria and archaeobacteria, and eukaryotes, most of which are multicellular. Bacteria and archaeobacteria are characterized by a single large DNA molecule

as the main chromosome within the cell cytoplasm, sometimes with smaller DNA plasmids containing additional genetic information. Higher organisms (eukaryotes) include fungi, plants and animals, and also a large number of relatively little studied unicellular microbes. Eukaryotes are characterized by a membrane-bound separate compartment, the nucleus, which serves as a storage site for the DNA which is packaged in several chromosomes. Eukaryotic chromosomes are much more complicated than in prokaryotes, because they have a compact wound structure with many associated proteins called histones. This is to package the information more efficiently and to keep some of the information locked safely away when it is not needed.

In addition to cellular organisms, there are many kinds of viruses, which are basically parasites of cells. They consist only of DNA or RNA as genetic information and some packaging proteins protecting the genetic material. Viruses are very small, lack cytoplasm and enzymes and therefore do not have their own independent metabolism. They can be considered to be “alive” (metabolically active) only when they are in contact with their cellular hosts.

Thus DNA in prokaryotic and eukaryotic cells is essentially a system for information storage and propagation. DNA is a long macromolecule having a sequence of different nucleotides coding information using four different nucleotides: adenine (A), thymine (T), guanine (G) and cytosine (C). The DNA molecules can be very long, up to hundreds of millions nucleotides, depending on the complexity of the organism. The typical ranges of information content in genomes of different organisms are given in Table 1-1.

Bacterial chromosomes normally have about 1–2 million nucleotides, sufficient for encoding all the instructions for microbial life. Scientists aiming to build synthetic microbes believe that only about 113,000 nucleotides, coding for less than a thousand genes, are necessary to control a simple self-replicating organism (Forster & Church 2006). In single-celled simple organisms like bacteria chromosomes are located in the main

Table 1-1. Typical information storage capacity in different organisms in counts of nucleotides. Data from various sources (e.g. Plant DNA C-values Database and Animal Genome Size Database, Zonneveld *et al.* 2005, Gregory 2006, Zubáčová *et al.* 2008; minimal synthetic microbe size from Forster & Church 2006).

<b>Taxon</b>	<b>Small</b>	<b>Typical</b>	<b>Big</b>
<b>Viruses</b>	5,000	50,000	200,000
<b>Bacteria</b>	500,000	1.5x10 <sup>6</sup>	6x10 <sup>6</sup>
<b>Archaeobacteria</b>	500,000	1x10 <sup>6</sup>	6x10 <sup>6</sup>
<b>Minimal synthetic microbe</b>	NA	113,000	NA
<b>Eukaryotes</b>			
<b>Protista</b>	2.2x10 <sup>6</sup>	?	177x10 <sup>6</sup>
<b>Algae</b>	9x10 <sup>6</sup>	850x10 <sup>6</sup>	19x10 <sup>9</sup>
<b>Plants</b>	120x10 <sup>6</sup>	5x10 <sup>9</sup>	250x10 <sup>9</sup>
<b>Fungi</b>	12x10 <sup>6</sup>	?	65x10 <sup>6</sup>
<b>Invertebrates</b>	29x10 <sup>6</sup>	?	37x10 <sup>9</sup>
<b>Vertebrates</b>	342x10 <sup>6</sup>	1x10 <sup>9</sup>	130x10 <sup>9</sup>
<b>Human</b>	NA	4x10 <sup>9</sup>	NA

compartment of the cell, the cytoplasm. In higher organisms (eukaryotes) the DNA is kept in a special organelle, a nucleus, which is bordered by a double-layer membrane. The membrane has many small holes to allow information bearing molecules (DNA, RNA and proteins) to pass between nucleus and cytoplasm.

The information-storing DNA macromolecule is normally a double helix, with two long antiparallel molecules containing the same information in mirror copies in two directions. The two chains are joined together at the nucleotides so that the A pairs with T and C pairs with G. The unidirectional information can be read by the cellular machinery from either strand, in opposite directions. Note that the information is not identical in the antiparallel strands, but complementary, so that the opposite strand message can be obtained by biochemical copying as well as by *in silico* computation. Living cells rely on copying and interconverting this information by the complicated biochemical machinery of DNA and RNA synthesis for cell function and cell division.

In eukaryotes, DNA in the nucleus is normally wound tightly into packages called chromosomes to save space, and unwound only where and when the information is needed. In multicellular eukaryotes (*e.g.* in plants, fungi and animals) there are dozens or even hundreds of chromosomes in the nucleus, and each chromosome can be several hundred million nucleotides long. For example, the human genome (23 pairs of chromosomes) has a total length of about 4 billion nucleotides, coding for some 22,000 genes and corresponding proteins. Each gene can furthermore produce a variety of messenger RNAs (mRNAs) in the RNA processing stage, and thus many different varieties of proteins, called splice variants.

At least half or more of genes have splice variants and alternative polyA sites (termination signals for RNA copying operation), leading to at least 100,000 possible protein variants in a human cell. The complexity of messenger RNAs is increased further by non-translated variants, that is, various mRNA differences in areas which do not code for the protein, for example, alternative lengths of 3' UTRs (untranslated regions after the protein coding segment). Latest research has shown that shorter mRNA UTRs are correlated with increased expression, which may be depend on control by microRNAs (introduced later below). MicroRNAs are known to target mostly 3' UTRs, so shorter UTRs can result in less suppression by microRNAs.

This enormous complexity in the chemical compounds and controlling mechanisms is shown in a simplified form in Figure 1-1 depicting the general structure of a eukaryotic cell.

Duplication of the information in the DNA is performed within a dividing cell by opening the double helix and building another antiparallel strand to each strand of the original helix. This results in two identical copies of the original double helix (barring occasional mistakes in the copying process, causing mutations). The double-stranded helix is more stable than a single molecule, and errors on one strand can be repaired based on the correct information in the other strand.

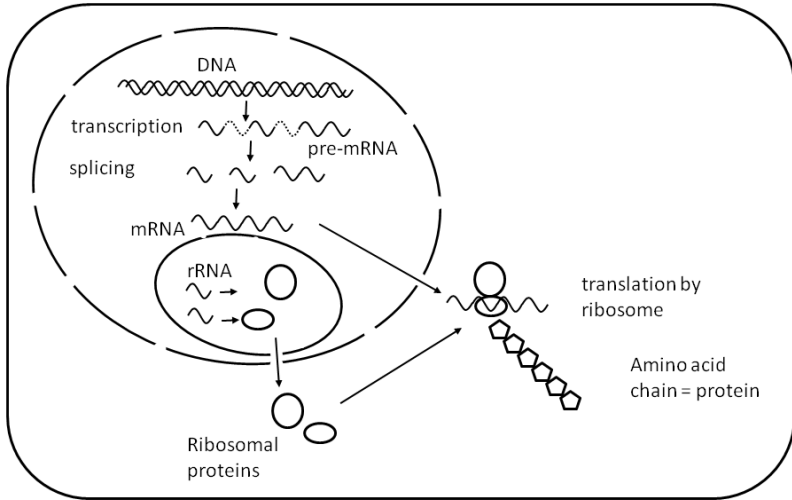


Figure 1-1. Eukaryotic cell structure and DNA transcription and translation according to the main dogma of molecular biology (DNA  $\Rightarrow$  RNA  $\Rightarrow$  protein). mRNA = messenger RNA carrying the information of coded protein from inside the nucleus to the cytoplasm to the translation site and binds to the ribosomes for translation. rRNA = ribosomal RNA which assembles with proteins to ribosomes, which serve as a scaffold for the translation of RNA to protein.

An additional benefit of the double-stranded molecule is that the information on opposite strands can be used to code for different functions. For example, the meaning of a short string like ATCGGCTTT and its reverse reading AAAGCCGAT on the opposite strand in the opposite direction can encode different meanings for the cell. This efficient packing of information in both directions is utilized by some viruses with very small genomes, when the amount of available DNA is a limiting factor in the life-cycle of the virus.

There can be different types of mistakes, *i.e.* mutations, in the original DNA code, like mistakes in a printed book. The three types of mutations are: transitions (conversions of one or more nucleotides to another alternative of the four nucleotides A, T, C or G), deletions (losses of one or more nucleotides) or insertions (additions of one or more nucleotides). These mutations are essential in generating diversity in the inherited information, thus

generating new variants in successive generations and making natural selection and evolution possible. This is akin to the evolution of words and sentences in human languages, when the commonly used vocabulary and grammar changes in time to encode new semantic functions.

## 1.2 Cells process complex information

Like any real-world information processing automaton (like a modern personal computer), a cell needs a mechanism to read the programming information from storage to active use. In cells this is performed by RNA molecules, which act as messengers from DNA in the chromosome (which is in the nucleus in eukaryotes) to the cellular working machinery in the cytoplasm. RNA molecules are typically quite short, not exceeding a few thousand nucleotides and are less stable than the DNA molecules. Their short lifespan compared to DNA is helpful in producing short-term messages, which are then destroyed, when their data is no longer needed.

The RNA molecules are copies of short messages from DNA, often coding for proteins to be produced. These messenger RNA molecules are then transported from the nucleus through the nuclear membrane holes into the cytoplasm. There the RNA molecules are used to translate the information into specific amino acid sequences to form functional proteins, which run the various machineries for growth, metabolism, cell division and so on.

DNA is first copied to RNA and then triplets of RNA are used to code for the 21 amino acids needed as building blocks for natural proteins. Four nucleotides in triplets can code for 64 messages, so there is ample redundancy in the genetic code and some of the triplets are used for start and stop signals in the production of proteins. Redundancy of the encoding from DNA via RNA to protein means that some mutations do not actually change the encoded amino acid incorporated into the specified protein amino acid chain. These are called silent mutations. Even

though such mutations can be silent for the produced protein, they may still have a biological effect. For example, a nucleotide change (transition) may cause DNA-binding proteins to bind differently to the changed DNA, affecting the production of the RNA molecule from the DNA. The change in resulting RNA may also affect the protein production from the RNA due to *e.g.* RNA stability or action of RNA-binding proteins.

This whole system of transcribing of information from DNA to RNA to proteins is called the main dogma of molecular biology, constituting the main direction of the flow of information in the cell from chromosomes to cytoplasm. The functioning of the genetic code is shown in Figure 1-2.

The information flow according to the central dogma is from DNA to RNA to proteins, unraveled in the fifties and sixties. Later research then has found that information flows in various other ways also. Until recently, it has been thought that most of the information in DNA is “junk”, as only a very small proportion of sequence codes for proteins. The proportion of such non-coding DNA in various organisms is shown in Figure 1-3.

There is, however, further complexity and feedbacks in the system. For example, RNA molecules can bind to DNA, affecting the DNA function. Special proteins can bind to DNA, also affecting DNA function; an example of this kind of proteins is the class of proteins called transcription factors. These transcription factor proteins move from their production site in the cytoplasm into the nucleus and bind to specific DNA sequences upstream of protein coding regions in the DNA, and control the production of RNAs from the DNA, thus influencing protein production from specific genes. RNA can also bind to proteins, to affect protein function or to form riboproteins, which have many important functions in the cell, like protein translation and chromosome duplication.

Transcription factor proteins were earlier thought to be the main controlling mechanism for gene expression, but recently non-coding RNAs have been shown to be another major layer of

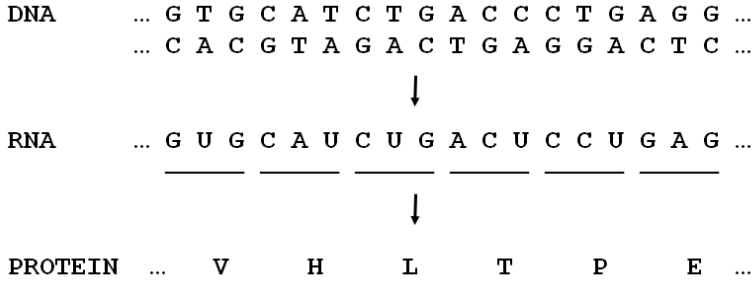


Figure 1-2. Transcription of DNA to RNA and translation to protein according to the genetic code by transcription from DNA to RNA and translation from RNA to amino acid chain of a protein using the triplet code.

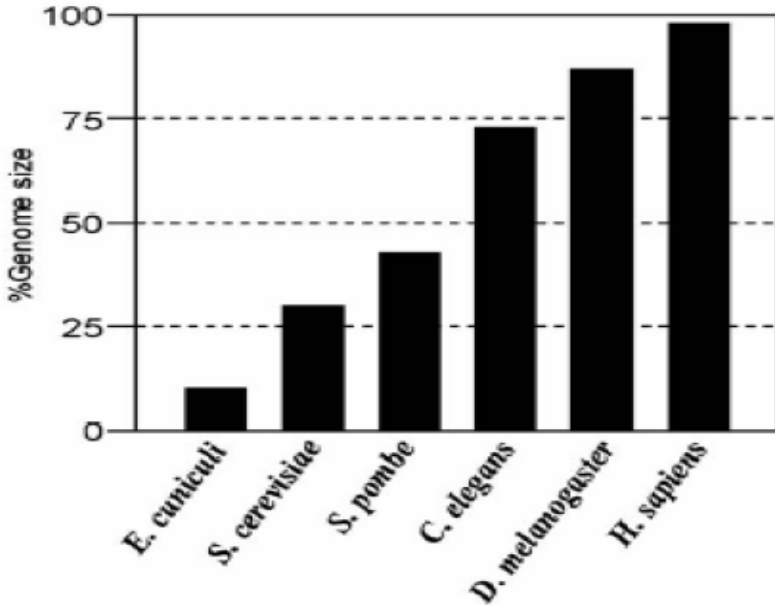


Figure 1-3. Proportions of non-coding DNA in various organisms (Taft *et al.* 2007). Reproduced by permission from John Wiley and Sons.

control, adding to the complexity of the whole system. These non-coding RNAs include a large number of only recently found novel entities (Dinger *et al.* 2008). Examples are shown in Table 1-2.

Table 1-2. Examples of non-coding RNAs, their sizes and functions.

Acronym	Name	Size nt	Function
miRNA	microRNA	21–23	single-stranded, regulate mRNAs
siRNA	small interfering RNA	20–25	double-stranded, regulate mRNAs
snoRNA	small nucleolar RNAs	70–240	guide methylation or pseudouridylation of <i>e.g.</i> ribosomal RNAs
ncRNAs	long ncRNAs	>200	regulate gene expression
piRNAs	Piwi-interacting RNA	27–30	germline chromatin modifications and transposon silencing
	Riboswitches	varies	bind small molecules, enzymatic function or modification of mRNA splicing or translation

Especially notable are microRNAs, which are 20-24 nucleotides long tiny RNA molecules that work by binding to messenger RNAs or to genomic DNA to control gene activity. The roles of microRNAs keep expanding, as this is one of the hottest topics in current molecular biology. MicroRNAs have already been demonstrated to be an essential layer of gene expression regulation in addition to small metabolites and transcription factor proteins binding to promoters of genes. It is even considered that small RNAs may be trafficking between cells and act as intercellular messengers in both plants and higher eukaryotic animals (Dinger *et al.* 2008). Non-coding RNAs are also thought to be involved in epigenetic mechanisms in chromatin modification, *e.g.* paramutation (Chandler 2007) and therefore control chromosome function in eukaryotes.

Thus the information processing in cells is very complex and integrates information from DNA, RNA, proteins and the many biochemical compounds of the metabolism. The current view of the network of genetic information flows is in Figure 1-4 below. The role of small non-coding RNAs is becoming increasingly prominent and is even considered to underlie long-term memory (Mercer *et al.* 2008). Interestingly, the importance of non-coding RNA is supported by theoretical calculations of the proportion of regulatory mechanisms needed in a complex metabolic system, based on the lower cost of developing short RNA based control mechanisms compared to the more complicated protein based DNA and RNA control systems (Ahnert *et al.* 2005).

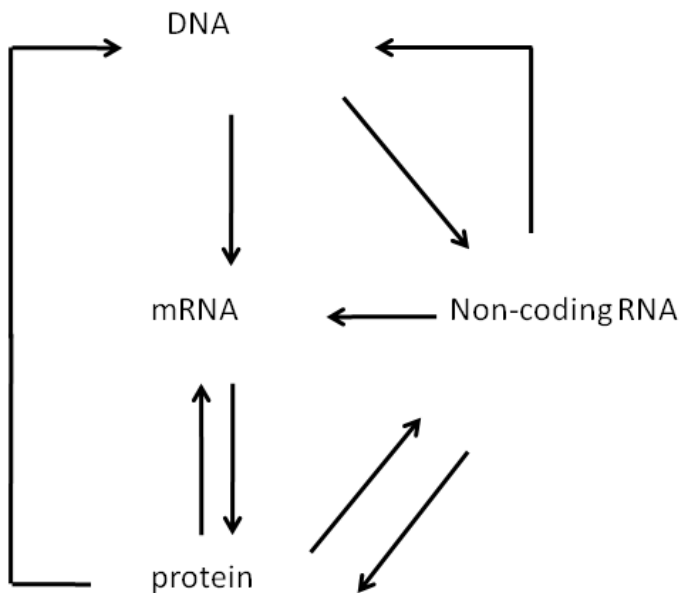


Figure 1-4. Genetic information flows in eukaryotes, including the important role of non-coding RNAs. Proteins bind to mRNAs (*e.g.* ribosomal proteins), to nucleus DNA (*e.g.* transcription factors) and to non-coding RNA (*e.g.* microRNA controlling factors).

### 1.3 Cellular life is chemically complex and somewhat stochastic

This book is about analyzing the complex web of interactions and information processing that forms the basis of cellular functions. The complexity of life is being studied by an increasing army of researchers, working on DNA, RNA, proteins, metabolism, organic compounds and so on. This has given rise to the large number of “omics” (see Table 1-3). The basic “ome” is the genome as the repository of information from one generation to next, but the so-called epigenome is also now recognized as an essential source of information during the development of a eukaryotic organism and involves DNA modifications, chromatin modifications, histone code, methylation, *etc.*, which do not affect the DNA sequence itself, *i.e.* the sequence of A's, T's, C's and G's (Callinan & Feinberg 2006).

Information storage in cells is more complex than just having a long DNA molecule and a large number of short RNAs floating inside a cell interacting randomly. There are many hardware components necessary to keep the information packaged safely from degradation, and complex mechanisms for its use, as we have seen above, including transcription (DNA to RNA) and translation (RNA to protein). The eukaryotic cell is in fact full of membrane extensions and vesicle systems and macromolecular scaffolds acting as organizing structures for well synchronised and accurately scheduled transport of various materials to their sites of targeted interactions. The complex infrastructure within the cells is evident in the electron microscope pictures of cells showing a large number of vesicles and membrane enclosures.

Genotype variation from one individual to another is a recently emerging complex data domain, when high-throughput methods of recording such variations have become available. Various mutations (transitions, deletions or insertions) in genomic DNA ultimately account for the majority of the metabolic or structural variation in different individuals belonging to the same species. For example, most of the individual differences in humans are due

Table 1-3. Some “omes” of modern biology.

<b>Genome</b>	All genetic information in an organism
<b>Epigenome</b>	Other than DNA-based heritable information
<b>Orfeome</b>	All open reading frames (ORFs) coding for proteins in a genome
<b>Methylome</b>	Pattern of methylated nucleotides in a genome
<b>Transcriptome</b>	All transcribed mRNAs in a cell
<b>Proteome</b>	All proteins coded for in a genome
<b>Interactome</b>	All interactions between macromolecules in a cell
<b>Regulome</b>	All regulatory networks of a cell
<b>Physiome</b>	All physiological functions in a whole organism
<b>Metabolome</b>	All small molecules in a cell in a specific physiological state
<b>Metabonome</b>	Metabolic responses to drugs, environmental changes and diseases.
<b>Glycome</b>	All carbohydrate molecules in a cell
<b>Secretome</b>	All gene products secreted to outside of a cell
<b>Localizome</b>	All locations of various proteins in cell types and subcellular compartments
<b>Unknownome</b>	All genes of unknown function

to individual nucleotide mutations, called Single Nucleotide Polymorphisms (SNPs), and sets of such SNPs in limited areas of genome are called haplotypes. The length of the area containing one haplotype indicates the time that has elapsed since the last recombination (crossing over between maternal and paternal chromosomes) in previous generations. The comparative information between haplotypes can be used *e.g.* to track ancestry of mutations and to map disease genes.

In addition to the purely genetic DNA information (*i.e.* the cell's genome or genotype), there exists a large amount of additional epigenetic information based on non-DNA molecules associated with the DNA sequence. The term “epigenetic” means that the underlying DNA sequence can be unmodified, but the additional modifications of chromosomes and their information context is still heritable and can affect gene expression.

Epigenetic mechanisms include various chemical modifications of the DNA, for example methylation and the histone code. The

individual DNA nucleotides C and G can be methylated, which affects their message handling (*e.g.* by DNA binding proteins and enzymes processing DNA). The histone code is based on modifications in the histone proteins (mostly acetylation and deacetylation of amino acids), which attach to DNA to package it into supercoiled segments in the eukaryotic chromosomes. Both the DNA methylation and histone modifications are known to be very important in regulating gene expression information. Summary of these various levels of information variants is in Figure 1-5, showing the combination of haplotype and epigenetic variation into so-called heptypes (= haploepitypes).

Epigenetic complexity is just one facet of the total cellular complexity and variety of regulatory systems in cells. Various other complex data domains in the modern biology era are listed in Table 1-4.

Table 1-4. Examples of complex data domains, which are responsible for the diversity of information processing and molecular interactions in living cells.

<b>DNA replication (in nucleus)</b>	Mutations during DNA replication and RNA transcription
<b>Non-coding RNA</b>	Variety in non-coding RNAs produced
<b>RNA translation machinery in cytoplasm</b>	Splicing of introns/exons to produce a variety of mRNAs
<b>Epigenetics</b>	Histone code, methylation, acetylation
<b>Genetic variation between individuals</b>	Single Nucleotide Polymorphisms as haplotypes
<b>Protein splice variants</b>	Multitude of different proteins arising from variations in mRNAs
<b>Metabolic pathways</b>	A single enzyme can produce a wide variety of end-products from a single precursor, <i>e.g.</i> terpenoid synthases
<b>Regulatory pathways, signalling systems</b>	Interactions in the alternative receptor-ligand reactions, microRNAs, transcription factors
<b>Cellular fate and tissue differentiation pathways</b>	Variety of developmental pathways, <i>e.g.</i> in development of blood cell types from stem cells.

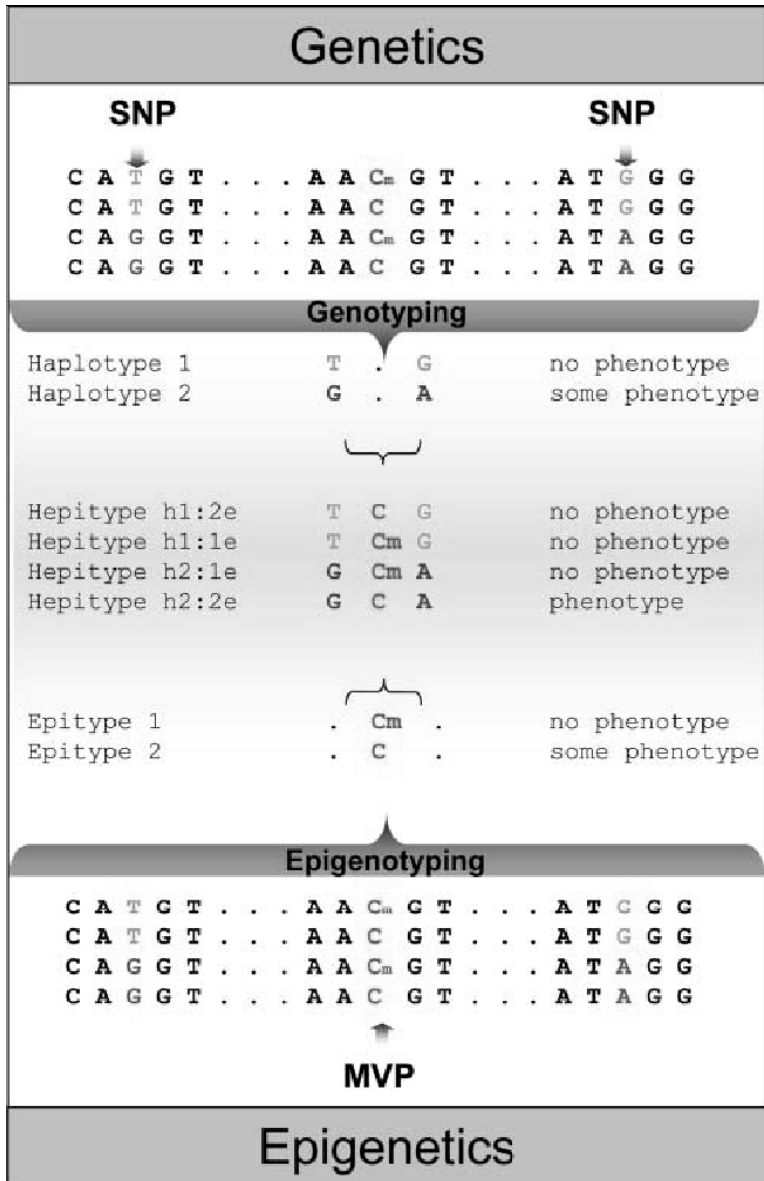


Figure 1-5. Relationships of genetic and epigenetic variations in eukaryotic cells. (Murrell 2005). Reproduced with permission from Oxford University Press.

Our information about cellular life has also stochastic aspects. As in any large and complex physical system, there are errors and random variations in basic life processes. Correct understanding of biological systems is further obscured by inaccurate data obtained from our laboratory experiments.

However, nature has produced some amazingly accurate biological mechanisms for maintenance of biological information. The most accurate mechanism by far is the DNA replication to copy the genetic information between dividing cells. Its accuracy has been developed during hundreds of millions of years to a degree, which is envy for modern nanotechnologists working on controlled molecular assembly and nanomotors. A multitude of enzymes and large protein complexes are involved in unwinding the double helix of DNA and synthesizing the new strand of DNA in the replication fork (see Figure 1-6, top). This process reaches average synthesis accuracy of  $10^{-7}$  to  $10^{-8}$  (Kunkel 2004) both in living cells *in vivo* and in test tubes using isolated DNA replicating enzymes *in vitro*. This is still further improved by several repair mechanisms that make corrections after copying, leading to an overall accuracy of up to  $10^{-9}$  in eukaryotic multicellular organisms.

Local variations along the genome in the fidelity of translation occur widely, especially in the so-called hotspots, where mutations are very abundant, and cold spots, which are more carefully kept intact. These have important roles for the evolution of organisms, both to develop new variation and new genes and to preserve crucial function unaltered, when necessary. Thus evolution has found ways of utilizing stochasticity to its advantage, when new mutations and new information needs to be created for the needs of the organism to adapt to new environments by modifying its genome in successive generations.

The DNA is not copied simply continuously from one end to the other on both strands. One strand of the double helix is 5' to 3' based on the orientation of the sugar backbone, and the other strand is 3' to 5', in an antiparallel orientation. New DNA copying

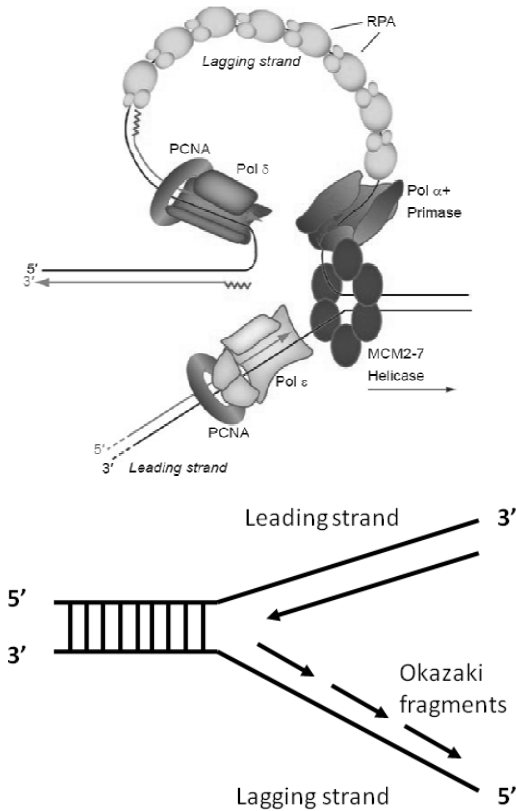


Figure 1-6. Schematic diagram of the protein complexes which control DNA replication (top, McCulloch & Kunkel 2008) and the replication of leading and lagging strands of DNA (bottom). Reproduced with permission from Nature Publishing Group.

operation by DNA polymerase only works in the 3' to 5' direction, so that synthesis is continuous in one direction (3' to 5' direction, leading strand), but occurs in bursts of about 200 base pairs (called Okazaki fragments) in the other, lagging strand, as shown in Figure 1-6 (bottom).

Due to the multiple starting points in the Okazaki fragments, there are more possibilities of things going wrong in the lagging strand, due to more enzymatic operations, and joinings of DNA pieces together. Indeed, in bacteria the lagging strand mutations

are 3- to 10-fold more abundant than mutations in the leading strand (Fijalkowska *et al.* 1998) and in yeast, 2- to 8-fold (Pavlov *et al.* 2002). Even in the leading strand, the DNA in mammalian chromosomes is not completely continuous either. DNA synthesis in a medium-sized human chromosome for example is initiated in up to 40,000 locations in the leading strand (at so-called replication origins), but in the lagging strand initiation occurs at up to 20 million locations (Hubscher *et al.* 2002).

In bioinformatics analyses, it is useful to bear in mind these basic biological sources of nucleotide data variation. Also, the current DNA sequencing methods often have much larger experimental error rates than the natural accuracy of DNA replication of  $10^{-7}$  to  $10^{-8}$ . Modern DNA sequencing machines (including the new Roche 454 GS20 sequencer introduced in 2006) can have sequencing error rate of  $10^{-3}$  to  $10^{-4}$ , depending on equipment and protocols. When sequencing is done repeatedly and results pooled in multiply aligned sequences, a much more reliable consensus sequence can obviously be built, provided reliable copies of the same molecule are at hand to start with.

Already in 2003 the most accurate human genome assembled genomic sequences from Sanger Centre in UK had a reported error rate of only  $10^{-5}$  per nucleotide (Anonymous 2003), using the traditional Sanger method sequencing machines. Note that current methods rely on sequencing of multiple copies of the same molecule. Novel technologies relying on single-molecule sequencing are likely to become available in a few years, possibly increasing the accuracy of primary data. Until that time though, detailed comparisons of single nucleotide differences between DNA sequences need to take into account both the possible errors in biological replication and in *in vitro* amplification of the original DNA molecules that have been used as well as the experimental errors occurring in the sequencing machine and sequence assembly stage.

Thus it is good to remember both the stochastic nature of the “natural” errors present in imperfect operation of cellular machines and the experimental errors in any biodata acquisition.

One should not try to attempt to predict things more accurately than what happens in nature or is warranted by the accuracy of the raw biodata inputted into the prediction system or algorithm.

## 1.4 Challenges in analyzing complex biodata

As we have seen in this cursory introduction to the complexities of cellular life, the biological systems are akin to very complex physical systems, like weather systems or galaxy formation, with the difference that the number of types of interactions is larger and that there are many simultaneous interacting chains of deterministic causalities which makes the cellular life so efficient.

In nanoscale the biological systems can have strong randomizing processes (including Brownian motion and random destruction of biomolecules by oxidants) affecting how individual molecules move and function within cytoplasm. However, in a slightly larger scale relevant to living cells the overall functions are surprisingly deterministic. The challenges for biodata analysis lie both in selecting the relevant data from the correct physical scale, as well as measuring the correct output and dependent variables for the problem in hand.

The other main issue facing bioinformaticians is the reliability of experimental data and even metadata (*e.g.* names and functions of genes, proteins *etc.*). Even the best algorithms cannot retrieve the correct conclusions from summarized and integrated multi-domain data, if the underlying raw data has unquantified uncertainties. Therefore proper sensitivity analysis for possible natural variation in the raw data is always warranted.

## References

- Ahnert, S.E., Fink, T.M.A. & Zinovyeva, A. 2008. How much non-coding DNA do eukaryotes require? *Journal of Theoretical Biology* 252(4): 587–592.
- Anonymous. 2003. Wellcome Trust, Sanger Center Press Releases: 14th April 2003. The Finished Human Genome — Wellcome To The Genomic Age. <http://www.sanger.ac.uk/Info/Press/2003/030414.shtml> (accessed 8 July 2007)

- Callinan, P.A. & Feinberg, A.P. 2006. The emerging science of epigenomics. *Human Molecular Genetics* 15(1): R95–101.
- Chandler, V.L. 2007. Paramutation: from maize to mice. *Cell* 128: 641–645.
- Dinger, M.E., Mercer, T.R. & Mattick, J.S. 2008. RNAs as extracellular signaling molecules. *J Mol Endocrinol.* 40(4): 151–159.
- Fijalkowska, I.J., Jonczyk, P., Tkaczyk, M.M., Bialoskorska, M. & Schaaper, R.M. 1998. Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. *PNAS USA* 95: 10020–10025.
- Forster, A.C. & Church, G.M. 2006. Towards synthesis of a minimal cell. *Molecular Systems Biology* 2: 45.
- Gregory, T.R. 2006. Animal Genome Size Database. <http://www.genomesize.com>.
- Hübscher, U., Maga, G. & Silvio Spadari, S. 2002. Eukaryotic Dna Polymerases. *Annual Review of Biochemistry* 71: 133–163.
- Kunkel, T.A. 2004. DNA Replication Fidelity. *J. Biol. Chemistry* 279: 16895–16898.
- Mattick, J. S. 2007. A new paradigm for developmental biology. *Journal of Experimental Biology* 21: 1526–1547.
- McCulloch, S.D., Kunkel, T.A. 2008. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Research* 18: 148–161.
- Mercer, T.R., Dinger, M.E., Mariani, J., Kosik, K.S., Mehler, M.F & Mattick, J.S. 2008. Noncoding RNAs in Long-Term Memory Formation. *The Neuroscientist* 14(5): 434–445.
- Murrell, A., Rakyar, V.K. & Beck, S. 2005. From genome to epigenome. *Human Molecular Genetics* 14(1): R3–R10.
- Pavlov, Y., Newlon, C. & Kunkel, T. 2002. Yeast Origins Establish a Strand Bias for Replicational Mutagenesis. *Molecular Cell*, 10(1): 207–213.
- Taft, R.J., Pheasant, M. & Mattick, J.S. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays* 29(3): 288–299.
- Zonneveld, B.J.M., Leitch, I.J. & Bennett, M.D. 2005. First nuclear DNA amounts in more than 300 angiosperms. *Annals of Botany* 96: 229–244.
- Zubáčová, Z., Cimbůrek, Z. & Tachezy, J. 2008. Comparative analysis of trichomonad genome sizes and karyotypes. *Molecular and Biochemical Parasitology* 161(1): 49–54.