Contents

List of Figures

List of Tables

Chapter 1

Background in Molecular Biology and Bioinformatics

1.1 From DNA to proteins

Many tasks in bioinformatics deal with the analysis of sequences because the large macromolecules that play an important role in the cell are *polymers*: sequences of linearly concatenated basic units. The most important biopolymers are nucleic acids and proteins, and the following sections describe how they are assembled, and how they relate to each other. I focus on mechanisms in eukaryotic cells and on the concepts that will be used in later chapters.

1.1.1 Structure of DNA and chromosomes

Hereditary information in the cell is stored in the form of DNA, *deoxyribonucleic acid*. DNA is usually present as a double-stranded molecule wherein the individual strands are wound around each other, forming a helix. The basic units of DNA are the *nucleotides* which consist of a sugar-phosphate backbone and one of the four bases adenine, cytosine, guanine and thymine. They are denoted by the letters A, C, G, and T; sometimes different letters are used to denote possible subsets from the set of all four, such as an N for "any nucleotide" (see appendix 3). Adenine and guanine are purines; cytosine and thymine belong to the group of pyrimidine bases. Every turn of the double helix gives room for ten nucleotides. The bases are situated in the middle of the helix and form hydrogen bonds with the bases from the other strand, with the rule that only A and T as well as C and G can complement each other (see figure 1.1). That means that all necessary information is stored in the bases on *one* strand of the helix. This enables an easy mechanism to pass on hereditary information: in cell division, the DNA double helix separates, and afterwards

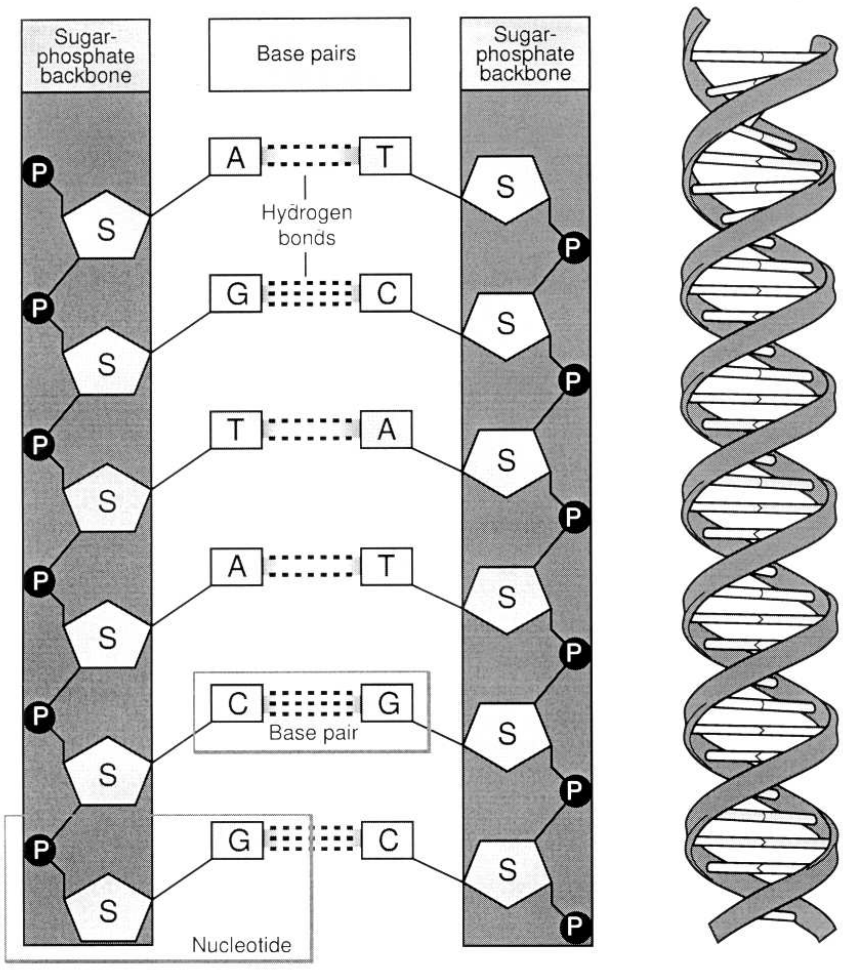

Figure 1.1: **Structure of DNA.** The left side shows the double stranded composition, the right side the famous double helix of the molecule (from The National Human Genome Research Institute (2002)).

each of the two daughter cells contains one strand of the original cell and one newly synthesized strand. Double-stranded molecules also have the advantage of increased stability.

DNA molecules are synthesized and read in a particular direction, from the 5' to the 3' end[1]. When a DNA sequence is written down, the strand which is read from left to right is called the *sense* strand whereas its counterpart in the double helix is the *anti-sense* strand because it is read in the opposite direction. A point located on the 5' side of a reference point is said to be *upstream*,

[1]The numbers 5 and 3 denote the locations on a nucleotide molecule where the previous and next one in a chain are attached.

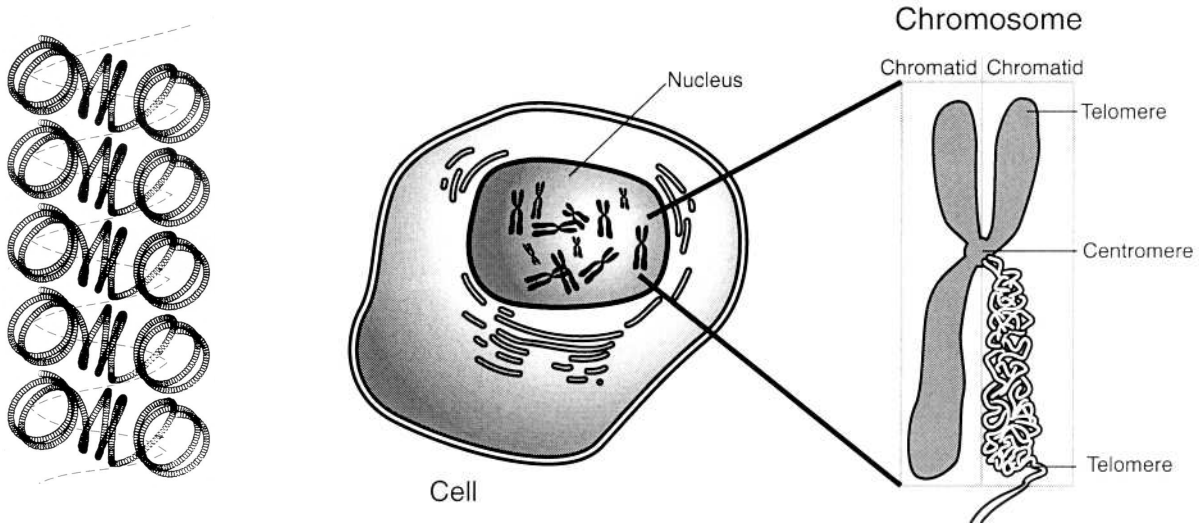

Figure 1.2: **Structure of solenoids and chromosomes.** (Left) Schematic solenoid structure, from Latchman (1998). For clarity, histones are not depicted, but one can see how the DNA loops around them (each "row" contains three nucleosomes). The dotted line denotes the higher order wrapping into solenoids. (Right) Wrapping of solenoids during cell division results in the well-known chromosomal structure of DNA (from The National Human Genome Research Institute (2002)).

a location on the 3' side *downstream*, and distances are denoted in bases or base pairs (bp).

DNA contains information on a multitude of levels. Already the simple relative frequency of A–T and G–C nucleotides plays an important role: A–T pairs have a weaker hydrogen bond than G–C pairs, and a separation of the double helix into single strands therefore requires less energy in AT-rich regions. Besides, the molecule does not only serve as a carrier of information but also contains sequences with no other function apart from the regulation of the expression of information contained in other parts.

Another variant of nucleic acids that can be found in cells is RNA, *ribonucleic acid*; this usually single-stranded molecule is very similar to DNA, the differences being a slightly modified sugar in the backbone and the base uracil (U) in place of thymine. Among other purposes, RNA can serve as a temporary transmitter of information or as a structural component of cell particles.

In prokaryotic organisms that do not have a nucleus, the DNA is present as a naked double stranded helix. In eukaryotes, the DNA is divided into several molecules, the *chromosomes*, and wrapped up in *chromatin*. One reason for this is the limited space in the cell: A linear double helix of DNA of the entire chromosome set of a human cell, for example, would be two meters in length. Chromatin consists of the DNA itself and protein complexes, mainly *histones*, around

Organism	No. chromos.	Total size (bp)	No. chromos. sets	Est. no. genes
Simian Virus 40	1	$5,243$	1	6
E. coli	1	$4.6 \cdot 10^6$	1	4,400
S. cerevisiae	16	$12 \cdot 10^6$	1 or 2	6,000
D. melanogaster	4	$140 \cdot 10^6$	2	13,000
A. thaliana	5	$120 \cdot 10^6$	2	25,000
H. sapiens	23	$3 \cdot 10^9$	2	35,000

Table 1.1: **Genome sizes of selected organisms.** The genomes of the simian virus 40 and the prokaryote *E. coli* contain one double-stranded DNA molecule; the eukaryote genomes are organized into a number of chromosomes. The yeast *S. cerevisiae* has a different number of chromosome sets depending on the state of proliferation. In contrast to the model plant organism *A. thaliana*, many other plants are polyploid, i. e. they contain more than two chromosome copies. Note the decreasing gene density in higher eukaryotes.

which the DNA is coiled up forming *nucleosomes*. This structure is subsequently folded into a more compact *solenoid*, where specific histones seal the DNA around one nucleosome and associate with each other (see figure 1.2). This tight packing is able to regulate the accessibility of regions in the genome on a high level and is therefore important for gene regulation (see section 2.2.2).

During cell division, the solenoids are further compacted by extensive looping, thus forming the well-known structure of the chromosomes (figure 1.2). As opposed to prokaryotes which are *haploid*, i. e. they own only one copy of their genes, eukaryotes are usually polyploid and own multiple copies of their chromosomes. Vertebrates and *Drosophila* are both diploid; one set of chromosomes is inherited from the male and one from the female ancestor. The ensemble of molecules bearing hereditary information is called the *genome*. Table 1.1 shows the genome sizes of some organisms whose DNA has already been sequenced.

1.1.2 Proteins, transcription and translation

Another important class of biopolymers, proteins, are macromolecules made up by polymerization of basic units, the *amino acids*. In general and throughout all species, a cell uses 20 different amino acids, although there are additional rare ones. The relationship between DNA and proteins is as follows: A *gene* denotes a discrete segment of DNA which encodes the sequence of one (or possibly more than one) protein. Three nucleotides within the coding part of a gene, a *codon* or *triplet*, encode one particular amino acid. As there are 64 different triplets and only 20 different

amino acids, this is a many-to-one relationship: the genetic code is degenerate. Three codons (UAG, UAA, UGA) serve as a stop signal without an amino acid counterpart, and the codon AUG encoding methionine always starts a protein.

Proteins are active components of a cell and serve different purposes: For example, they can form part of the cell membrane, serve as catalytic compounds (enzymes), or influence the expression of genes. The function of a protein is determined by its three-dimensional structure which occurs when the linear sequence is folded into its most energetically favorable state. The *secondary structure* of a protein describes the arrangement of some of its amino acids into basic three-dimensional units such as α-helices or β-sheets. The *tertiary* structure then refers to the three-dimensional conformation of the whole protein.

Proteins which are derived from one common ancestor and thus serve the same purpose are called *homologous*. As the sequence determines the structure, and similar structure implies similar function, homologous proteins show considerable sequence conservation, i. e. a large number of residues[2] have the same or chemically related amino acids. A set of proteins that serve the same purpose is called a *protein family*. Homology is also observed on the level of protein *domains* — protein parts which carry out the same function, for example, interaction with DNA or integration into a membrane. A recurrent sequence pattern such as a domain is called a *motif* and often described by means of a *consensus sequence* which shows the most frequent residue(s) at every position.

The information within a gene is used to synthesize a protein in the following way:

1. During *transcription*, the so-called messenger RNA is generated — an RNA copy of the gene sequence. The enzyme which generates the copy of protein encoding genes is called RNA polymerase II; polymerase I and III transcribe RNA genes that do not encode proteins. The polymerase recognizes the start of the gene by means of a specific promoter sequence, starts its work at the transcription start site, and stops it at terminator sites about which hardly anything is known so far.

2. Eukaryotic *translation* takes place in the cytoplasm outside of the nucleus and synthesizes a sequence of amino acids using the sequence of nucleotides in an mRNA molecule as a template. This process takes place at active cell components called *ribosomes*. In this process, the non-translated RNA gene products play essential roles: ribosomal RNA sequences are structural components of the ribosomes, and transfer RNAs serve to guide the amino acids to the ribosomes and the right nucleotide triplet.

[2]The term *residue* denotes a basic unit in a biopolymer.

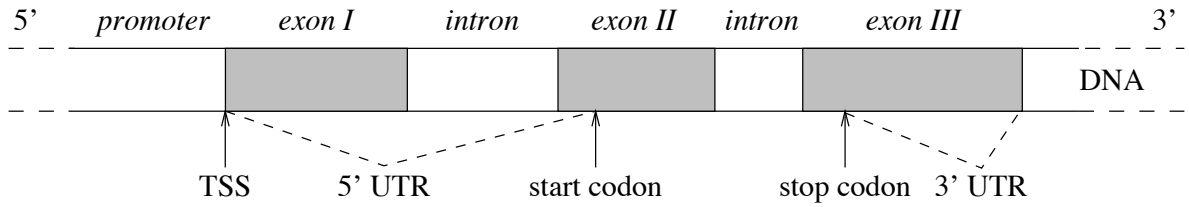

Figure 1.3: **A multi-exon gene structure.** In this artificial example, the start codon is contained in the second exon, so the 5' untranslated region (UTR) spans the complete first (non-coding) exon, the first intron, and a part of the second exon. The location of the start and stop codon, the promoter region and the transcription start site (TSS) is depicted.

The first AUG in an mRNA does not necessarily serve as a start codon, and only the mRNA part between start and stop codon encodes a protein sequence; thus, the mRNA contains untranslated regions (UTRs) on both ends. In eukaryotes, the story is even more complicated: Stretches of coding nucleotides (the *exons*) are interrupted by stretches of non-coding nucleotides (the *introns*). At the beginning and end of intron sequences, so-called *splice sites* are found, characteristic sequence patterns of about 15 base pairs. Figure 1.3 shows an example gene containing two introns.

After transcription, the introns are spliced out of the pre-mRNA, and the ribosome only sees an mRNA made up from exon sequences. There is not always a unique way to splice a gene, and therefore one gene is able to encode more than one protein. Alternative splicing becomes more the rule than the exception when we look at a highly complex organism (see the recent review by Graveley (2001)). In humans, at least one third of the genes are alternatively spliced, and at the moment this is believed to be the main reason why the number of genes does not grow linearly with the complexity of an organism. On the contrary, the relatively complicated organism of the fruit fly contains considerably fewer genes than the simple worm *C. elegans*.

1.2 Computer-based annotation of genomes

As mentioned above, the raw DNA sequence data that are determined in the course of a sequencing project offer few new insights. During the process of annotation, these raw data are interpreted into useful biological information (see the reviews by Rouzé et al. (1999) for plant genomes and Lewis et al. (2000) for a more general introduction). In a large-scale sequencing project for a model organism such as *Drosophila*, annotation integrates computational analyses with a lot of biological knowledge about specific genes. It therefore is a semi-automated pro-

cess in which the results of many algorithms are integrated and presented to a human curator, who then decides on the final annotation. Currently, one can identify two steps in the annotation task: Structural annotation, which deals with the identification of biologically relevant sites in the sequence, and functional annotation, which attributes specific biological information to the genome as a whole and to the sites found in the first step. Annotation provides a broad overview and description of the features contained in a genome. A deep analysis is still left to the biologist working in the lab.

Annotation has become a daunting task for large genomes because it is begun while the sequencing process is far from being finished. It must deal with a constantly changing target, update and re-annotate new or changed sequences, and track the changes over time. An example for such a project is the ensEMBL pipeline to annotate the public version of the human genome (Birney et al., 2001).

1.2.1 Structural annotation

Structural annotation usually comprises locating protein and RNA-encoding genes along with their control elements, translating the putative genes into proteins, and identifying global genomic features important for chromosome organization such as matrix attachment regions (Singh et al., 1997) or CpG islands (see section 2.2.3).

The most crucial step in structural annotation is gene finding. This is a complex task and involves the identification of patterns in the sequence as well as grouping these putative patterns to meaningful interpretations. The most recent reviews were written by Stormo (2000); Haussler (1998); Burge and Karlin (1998), but they focus on the first of the following three different approaches to gene finding: *ab initio*, alignment, and homology based methods.

***Ab initio* gene finding.** This group of gene finders uses no information but the genomic sequence itself to find a gene. According to Burge (1997), there are four generations of *ab initio* gene finders. The first generation used statistics on coding and non-coding regions to approximately locate exons; the second generation combined these statistics with models for splice sites to exactly locate exons; the third generation combined multiple exons in a model for a single gene; and the fourth generation was finally able to predict multiple and partial genes on both sides of a long genomic sequence. Recent gene finders thus consist of *signal sensors* that identify patterns with positionally conserved nucleotides such as the splice sites found at the intron/exon boundaries, and of *content sensors* that identify regions with a statistically significant composition such as coding exons. Exon sequences have distinct oligonucleotide statistics because of the three-periodicity caused by the triplets, and also because of a bias in codon usage: not all codons

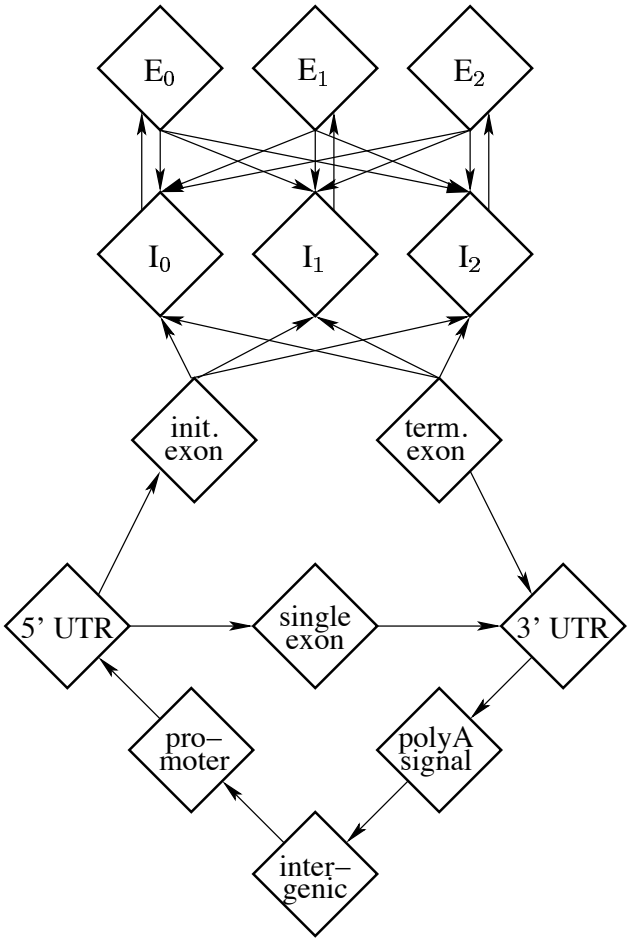

Figure 1.4: **A probabilistic *ab initio* gene finder.** The picture shows the model structure for the forward strand that is used in the GenScan system by Burge and Karlin (1997). A comparison with figure 1.3 reveals that each state represents a particular pattern or region of a gene. GenScan contains a specific state for single-exon genes and for initial and terminal exons, as their length distributions differ considerably from internal exons. The three states each for internal exons and introns are necessary to ensure that the total length of the coding sequence is a multitude of three.

are used equally frequent. A model then describes admissible combinations of these patterns and serves to calculate an optimal parse of a DNA sequence. As such, almost all current gene finders use a framework of hidden Markov models (HMMs) or generalized HMMs (see section **??**). The first simple HMM based gene finder was developed by Krogh et al. (1994) to analyze the bacterial genome of *E. coli*; later, Kulp et al. (1996) and Burge and Karlin (1997) pioneered the application of so-called generalized HMMs for eukaryotic genomes. In contrast to simple HMMs, where a state emits a *single* symbol each time it is visited, a state in a generalized HMM

models a *sequence* of symbols. This formalism is perfectly suited for gene structures, as it allows arbitrary sub-models in the states and therefore integrates both content and signal sensors within a probabilistic model (see figure 1.4).

Alignment gene finding. An *alignment* generally refers to the local or global matching of two biopolymer sequences. Usually, the residues are superimposed in such a way as to minimize the distance between the two sequences, measured by (possibly negative) scores for matches, mismatches, and insertions/deletions. Alignments can be efficiently computed by dynamic programming algorithms; Durbin et al. (1998) provide an excellent introduction (cf. also section **??**). It is possible to find genes and their structure by means of an alignment of complementary DNAs (cDNAs) with a genomic sequence. A cDNA is obtained by reversely transcribing an mRNA found in the cytoplasm, and its complete sequence is assembled from short, sequenced segments called expressed sequence tags (ESTs). The set of all cDNAs from a certain tissue is called a library. Because of the low quality of cDNA sequences, wrong nucleotides as well as insertions and deletions are likely to occur. Therefore, gene finding based on cDNAs employs dynamic programming for the alignment to the genomic sequence, along with suitable gap costs for the intronic sequences that are not present in the cDNA, and a model for splice sites (see the sim4 program by Florea et al. (1998) as a widely used example). Theoretically, cDNA alignments provide the best way to find genes — *ab initio* gene finders are only able to find the coding part of genes along with the intervening introns, and usually miss the pattern-less non-coding exons and UTRs. Nevertheless, there are a number of pitfalls: Apart from contamination, a cDNA library contains only sequences of genes that were actively transcribed under certain conditions; otherwise, no mRNA would be found. Besides, cDNA sequencing does hardly ever span the complete mRNA; the single-stranded RNA is easily disrupted or digested before the reverse transcription has reached the opposite end. So-called *full-length* cDNAs try to circumvent this problem by selecting the longest cDNA of a whole set that all refer to the same gene, or by selecting cDNAs that contain the cap structure (section 2.1).

Homology gene finding. Two variants of homology based gene prediction exist. The first one attempts to identify a gene in a DNA sequence based on known proteins. This can be done by a modified dynamic programming approach which takes the one-to-many relationship between a protein sequence and all DNA sequences that can be translated into that particular protein into account. Different algorithms use either data bases of known proteins (TBLASTX, a variant of the popular BLAST alignment algorithm by Gish and States (1993)), or a library of protein family HMMs (the GeneWise approach by Birney and Durbin (2000)).

The second approach to homology based gene prediction makes use of genomic sequences

of two species, known to contain homologous genes, and aligns them taking possibly different splicing into account (Bafna and Huson, 2000; Batzoglou et al., 2000). This follows the observation that coding sequences are usually conserved across species because they are translated into a protein with similar function, whereas the intervening non-coding sequences may accumulate mutations without affecting the product.

Homology based gene finding has the disadvantage that the homology might not span the complete gene but could be limited to a part of the protein. A similar observation is true for partial cDNAs, too. On the other hand, false positives are hardly ever made with these approaches, and if something about the function of one gene product is already known, the other one can be assumed to serve the same purpose. Therefore, recent gene finders usually integrate *ab initio* with alignment and/or homology approaches (Reese et al., 2000; Krogh, 2000; Yeh et al., 2001).

Promoter recognition. The recognition of regulatory regions, namely of promoters, also belongs to the first step of annotation. Similar to the different approaches for gene finding, we can also distinguish between *ab initio* and homology based methods. Because eukaryotic mRNAs usually contain only one transcribed gene at once, it is tempting to use a suitable promoter model as one state of a probabilistic model for *ab initio* gene finding, as in figure 1.4. In this way, the admissible search region is restricted to upstream regions of detected genes, and on the other hand, a reliable promoter recognition could help to recognize the border between two neighboring genes.

In practice, this idea is hampered because of the lacking ability of gene finders to predict the non-coding exons at the 5' and the 3' end of a gene which do not contain specific patterns. It has also turned out that promoter recognition is a problem that equals if not exceeds the complexity of gene recognition. This does not come as a real surprise if one considers that promoters are located within double stranded DNA in chromatin, whereas the patterns used in gene finding are still present in linear single stranded mRNAs. Therefore, a simple promoter recognition module as it is used in the GenScan system in figure 1.4 (see section **??** for details) is much less reliable than the other modules. Gene finders with integrated cDNA alignment are able to considerably restrict the admissible region of promoter predictions and are therefore more successful (Reese et al., 2000).

1.2.2 Functional annotation

Once the genes and other functional parts of the DNA sequence have been identified, the next step consists of the functional annotation of those features. Teichmann et al. (1999) provide a recent overview of this field.

The first step is the assignment of an isolated function to each individual gene, either by pairwise sequence similarity or similarity to a model of a protein family — this is implicitly carried out in gene finding by homology. The currently best way to do this appears to be a bootstrap approach: using an up-to-date protein database, similar sequences to the query sequence are pulled out, and a model is constructed from this new set. This step can be iterated and will thus find more and more distant homologues (Park et al., 1998).

If no homologue to a protein arising from a complete gene structure can be found, it may still be possible to provide some information on the domain level. For this task, large databases of domain models have been collected (e. g. InterPro (Fleischmann et al., 1999)) that integrate different resources. Also, some properties of a protein, such as its integration into a membrane, can be predicted reliably (Krogh et al., 2001), and secondary structures such as helices and sheets can be assigned to some protein regions. This annotation then enables us to look at cross-species conservation on a genome-wide level: how many protein families and which of the protein domains are present in all species, how many of the proteins contain transmembrane components, and so forth. The Gene Ontology Consortium (2000), among others, aims to provide a controlled vocabulary for such large-scale annotations to ease comparisons of annotation results for different species.

DNA and protein microarrays provide a completely different view of the genome (DeRisi et al., 1997; Haab et al., 2001). This technology monitors the activity of many, up to all known, genes of an organism under certain experimental conditions, either on the mRNA or protein level. Analyses such as the activity of genes related to the cell cycle (Spellman et al., 1998) provide a wealth of information. Recently, Shoemaker et al. (2001) have even released an annotation of the draft human genome based on microarray data. From a number of different experiments, correlations between the expression of several genes may become visible, and may thus serve to reconstruct genetic networks depicting the flow of information in different metabolic or regulatory pathways of the cell (Bower and Bolouri, 2001; Friedman et al., 2000).

The outcome of microarray analysis also enables further functional annotation. The promoter regions of co-regulated genes can be analyzed for common patterns (see chapter **??**); on the other hand, such common patterns may also serve to provide functional annotation based on the promoter regions of genes, especially for those for which no homologous sequences could be found (Pavlidis et al., 2001). If the proteins interacting with regulatory patterns are known, this approach is a promising way to elucidate regulatory networks.

As a summary for the annotation process, figure 1.5 gives a schematic overview of the flow of information in an annotation pipeline.

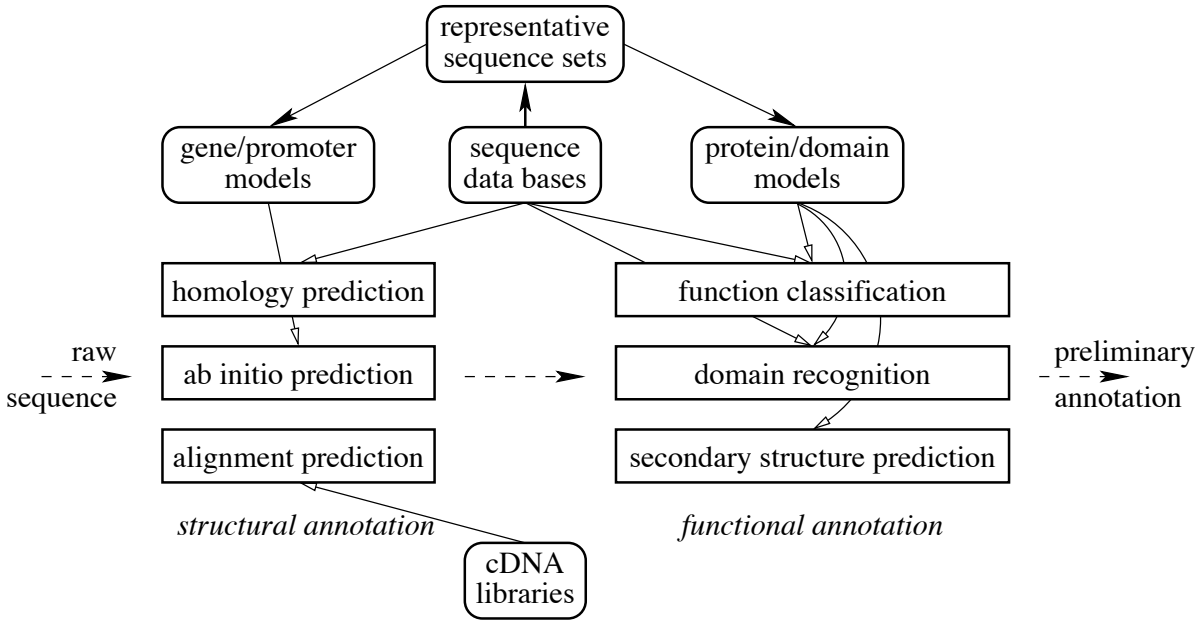

Figure 1.5: **Schematic overview of sequence annotation.** The picture shows only some important tasks of structural and functional annotation to exemplify the relations between data bases, models, and the sequence to be annotated. The preliminary annotation is in many cases validated by human curation.

Chapter 2

Promoters and Promoter Recognition

The topic of gene regulation has always received great attention because the key for the development of complex organisms does not lie as much in the mere number of genes but rather in their specific regulation and interaction. In the following, I will give a brief description of the biology of gene regulation, particularly of DNA transcription control and the organization of eukaryotic promoter regions. Again, this text cannot go into all necessary details but will focus on the concepts relevant to this thesis. A comprehensive yet easy to read introduction to this fascinating topic was written by Latchman (1998) from which much of the following description is inspired. Other, mostly more recent references are cited throughout where appropriate. In the final section, I will turn to computational approaches for promoter recognition published so far and discuss what aspects of promoters are taken into account in current algorithms.

2.1 Gene regulation in eukaryotes

It was observed rather early that a loss of DNA content occurs only in some notable exceptions and therefore cannot offer a general explanation for the individual protein levels found in different developmental stages or tissues. Rather, the process whereby DNA produces mRNA (and subsequently proteins, see section 1.1) must be responsible for the regulation of gene expression in eukaryotes. A number of stages leads from the initial transcription to the final protein product (see figure 2.1 for a schematic overview). In theory, any of these stages could be used to regulate the expression of a gene, and it has been shown that indeed all of them are targeted under one condition or another. I will explain the stages shown in figure 2.1 and indicate how they can be regulated, before I turn to a more detailed description of transcription control. Even though this description suggests that all steps have to be performed in a rigorous order, evidence shows that they are at least partly concurrent.

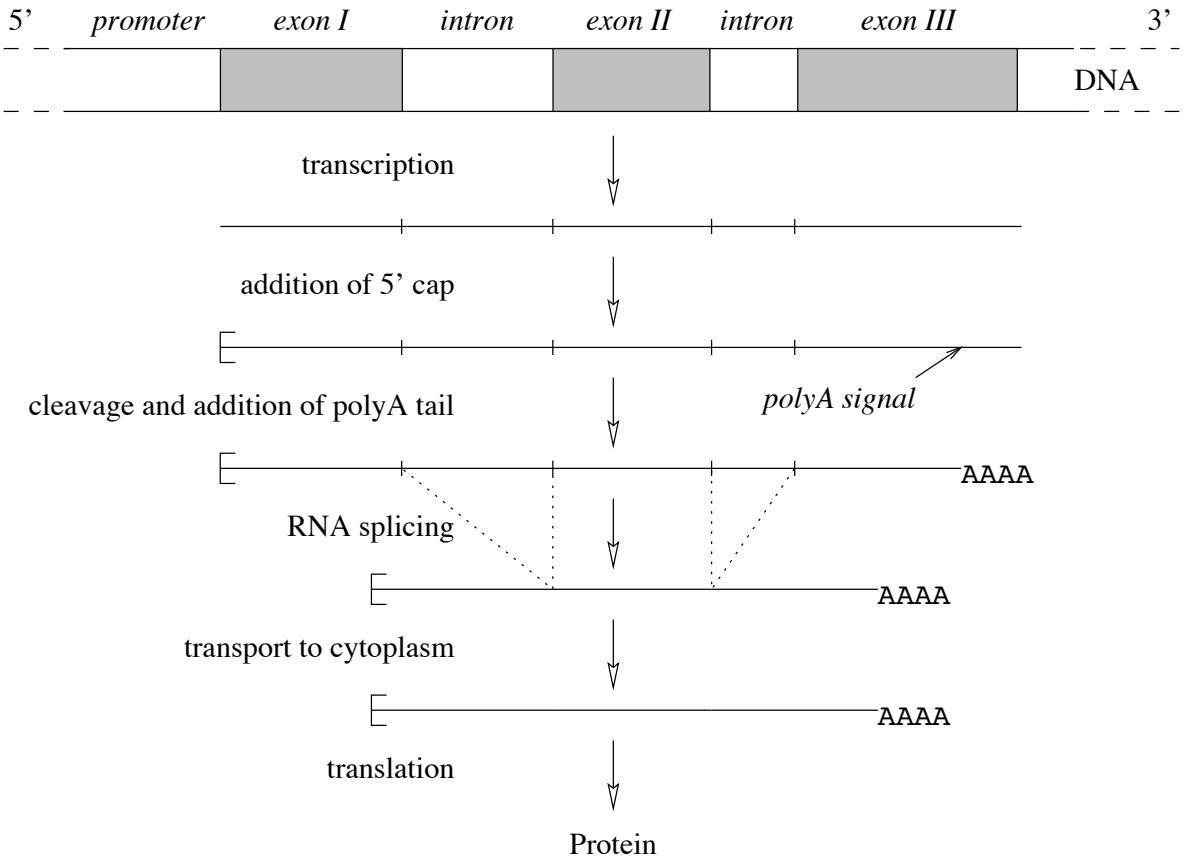

Figure 2.1: **Stages of gene regulation**, after Latchman (1998). See the text for details.

1. The *transcription* of protein encoding genes is done by the RNA polymerase II enzyme, and control at this level involves guiding the polymerase to the right places as well as inhibiting its activity (see section 2.2). The result of transcription is the pre-mRNA or *primary transcript*. A single gene can be transcribed starting from different promoters that are active only under specific conditions, giving rise to two or more (partially) different gene products.

2. The *post-transcriptional events*, which lead from the primary transcript to the final mRNA serving as a template for translation, start with the *capping* of the pre-mRNA. A cap structure consists of a guanosine residue linked in an unusual way to the 5' end of the RNA. The cap is the place where a ribosome binds to the RNA and is also necessary to protect an RNA from degradation enzymes.

3. Contrary to the modification at the 5' end which involves the adding of a single nucleotide, the 3' end is cleaved, a large RNA is stretch removed, and up to 200 adenosines are added.

The site of this *poly-adenylation* is flanked by two conserved sequence patterns where two protein factors bind, interact, and finally cut the mRNA. Similar to the cap at the 5' end, the polyA tail serves as protection against degradation, and it appears to have an effect on the translation efficiency of the mRNA. Also, more than one polyA signal can be present, leading to different possibilities to truncate an mRNA on its 3' end.

4. The next step in RNA processing is *splicing*. Apart from the splice sites at both ends of an intron (see section 1.1), an additional less well-conserved pattern around the *branch point* can be found close to the splice site at the 3' end of an intron. Splicing occurs in a complex structure known as the spliceosome which involves a number of RNA and protein components and holds the upstream and downstream parts of the mRNA in the correct place while cutting out the intron. Alternative splicing as discussed in section 1.1 has emerged to be a crucial regulatory step, complementing transcription control and serving to deliver variants of a single protein needed under specific conditions. Alternative splicing is regulated by tissue specific factors promoting a certain splice site as well as by the ratio balance of several proteins belonging to the spliceosome.[1]

5. After the mRNA has been brought in its final shape, it is *transported* from the nucleus through the nuclear membrane into the right place in the cytoplasm. A number of proteins have been identified that are believed to mediate this transport. It appears that a nuclear export signal in such a protein is crucial to guarantee export of itself and its associated mRNA. A few examples show that regulation may also happen at this stage, e. g. promoting the transport of a certain splice variant of a viral mRNA in HIV infected cells.

6. In the cytoplasm, *translation* takes place at the organelles known as ribosomes. It is initiated by the binding of a ribosome at the cap structure on the 5' end. A subunit of this ribosome then migrates along the mRNA until it finds an appropriate start codon. In rare cases, an mRNA may contain more than one functional start codon. A key role in the subsequent translation is played by transfer RNA (tRNA) molecules which deliver the correct amino acid to the currently considered nucleotide triplet. tRNAs have a common characteristic secondary structure and are bound to the mRNA by means of *anti-codons* complementary to the triplet for which they carry the appropriate amino acid. Subsequently, one tRNA after another is recruited, and a polypeptide is synthesized until the first stop codon is encountered. General control at this stage is possible by inhibiting components of the ribosomes; specific mechanisms interact with patterns in the 5' and 3' UTR of the

[1] Different proteins that are derived from the same primary transcript can also be a cause of *RNA editing* which modifi es single bases, thus replacing one amino acid by another or introducing a stop codon.

mRNA that form characteristic secondary structures, the latter sometimes also preventing poly-adenylation which is necessary for a correct translation.

Translation is influenced by the *stability* of the mRNA which determines the number of times that it is translated. A working model of this mechanism involves digestion enzymes attached to the ribosome that either recognize the beginning of the synthesized polypeptide or short regions in the 3' UTR that fold into a secondary structure. RNA stability is an effective means to control the rate of protein synthesis, especially in cases where a rapid and transient change of a specific protein level is necessary, and is often accompanied by a change in transcription rate.

To summarize, gene expression controls which genes are used, which modifications are carried out to the transcript, and how efficiently the final product is synthesized. A clear point should be made that gene regulation, and therefore transcription, is not a yes/no activity: Genes which are "switched on" do not all produce the same amount of mRNA. Although analogies from the terminology of engineering might suggest it, a cell is not a simple machine, not even at the level of individual genes. It is a viable precondition for the correct development of an organism that a subtle control is possible for every single product of biosynthesis. Also, if a gene is found to be "active" under certain *in vitro* conditions, its activity might be dramatically enhanced or suppressed by interactions only observable *in vivo*[2].

2.2 Regulation at the transcriptional level

Even though regulation occurs at all stages of protein synthesis, the control on the transcriptional level is clearly the most important. This intuitively makes sense: Why should a cell generally sacrifice valuable energy to synthesize products whose activity is subsequently repressed, again under the consumption of energy?

The RNA polymerase II (pol-II) enzyme has 12 subunits; but it despite its structural complexity, it cannot carry out transcription by itself. It requires auxiliary factors to recognize its target promoters, and to modulate production to react on specific environmental conditions.

The promoters of protein encoding genes can be seen to consist of a core promoter, a proximal promoter region, and distal enhancers, all of which contain *transcription elements*, short DNA sequence patterns that are targeted by specific auxiliary proteins called transcription factors. Transcription initiation by pol-II is regulated by those factors interacting with transcription

[2] *"In vitro"* usually means "simplified conditions in experiments at a lab bench", whereas *"in vivo"* refers to the living cell.

elements, pol-II, and also with each other, and by an open chromatin structure that enables the factors to access the DNA.

2.2.1 The basal transcription machinery

The common and best characterized part of promoters is the *core* promoter which is responsible for guiding the polymerase to the correct transcription start site (TSS). Accurate initiation of transcription depends on assembling a pre-initiation complex (PIC) containing pol-II and at least six transcription factors, the *general* initiation factors, which have been identified over the past 20 years (see the general review by Roeder (1996), and the one by Nikolov and Burley (1997) focusing on a detailed view on the protein structures). This complex machinery is immensely well preserved throughout all species.

Inspection of the sequences immediately upstream of the transcription start sites showed that a large group of eukaryotic promoters share an AT-rich sequence element around position -30.[3] This so-called TATA box is the most prominent sequence element in eukaryotic promoters. Detailed *in vitro* studies have elucidated the fundamentals of PIC assembly, deriving a minimal set of factors that suffice for transcription from a strong viral promoter containing a well-conserved TATA box.

The TATA box is a target of transcription factor (TF) IID, or more specifically, of one component of TFIID, the TATA binding protein (TBP). TFIID contains at least a dozen other components known as TAFs (TBP associated factors) which also interact, directly or indirectly, with other sequence elements. Upon binding of TBP, the DNA is strongly distorted, and sequences up- and downstream of the TATA box are brought in close proximity. Transcription factor IIA stabilizes this complex, even though it is not essential in all cases as originally thought and only vital for promoters with weaker TATA boxes.

After binding of TFIID (and possibly TFIIA), this complex is recognized by transcription factor IIB. This orients the growing complex towards the transcription start site and maybe guides the polymerase to the exact start position. TFIIB also recruits the pre-formed TFIIF–pol-II complex through direct interactions with both components. The binding of transcription factors IIE and IIH to the polymerase completes the assembly of the pre-initiation complex. With the exception of TFIID and possibly TFIIB (see below), all TFs are recruited by protein-protein interactions, and no interactions with specific DNA motifs has been observed so far. TFIIH finally triggers the start of transcription by modification of a pol-II subunit, and unwinds the DNA double helix in a 10 base pair long stretch downstream of the TSS. This step-wise assembly is

[3]Positions upstream of the TSS are counted backwards starting at -1, and positions downstream, including the TSS itself, are started counting at +1.

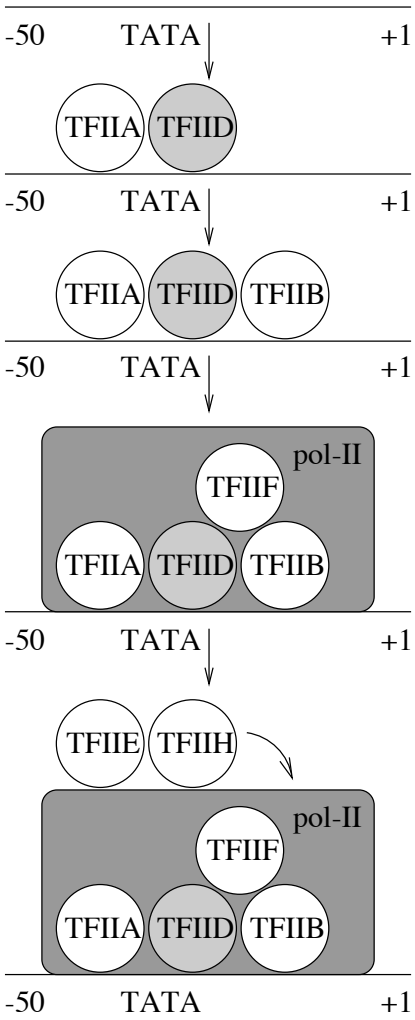

Figure 2.2: **Step-wise assembly of the pre-initiation complex**, after Latchman (1998).

summarized in figure 2.2.

While the polymerase moves off down the gene, TFIIF remains associated with the polymerase and TFIID remains bound to the core promoter, alleviating further cycles of PIC assembly. Large multi-protein complexes containing pol-II and some of the transcription factors have been reported from purification experiments. This suggests the existence of a so-called holo-enzyme in which much of the PIC is already pre-assembled, which allows the process of transcription initiation to happen much faster than an individual step-wise assembly would require.

As tempting as it sounds, the above description is by far a general mechanism of pol-II recruitment. For example, the promoters of house-keeping genes (i. e. genes that are always

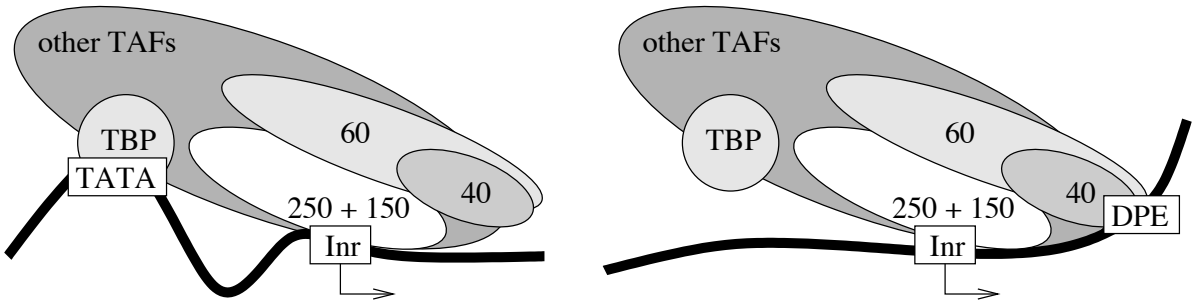

Figure 2.3: **Interaction of TFIID with the core promoter elements.** Two distinct interactions with TATA-driven (by TBP) and DPE-driven (by TAFs 60 and 40) promoters are shown in this model after Kutach and Kadonaga (2000) (Inr: initiator).

"switched on") do not contain anything resembling the TATA box. In these cases, the binding of TFIID is mediated by the *initiator* sequence element right at the TSS, but which is not only present in TATA-less promoters. Chalkley and Verrijzer (1999) recently reported that some of the TAFs are able to directly recognize this element.

In *Drosophila* as well as vertebrates, sequences downstream of the initiator were also found to have influence on basal transcription activity. Arkhipova (1995) showed that a number of short sequence patterns are significantly over-represented in downstream sequences of *Drosophila*, but found considerably weaker conservation in vertebrates. According to recent findings by Kutach and Kadonaga (2000), a specific *downstream promoter element* (DPE) appears to be as widely used as the TATA box but is less well-conserved. Its core motif is located exactly from 28 to 33 base pairs downstream of the TSS and was earlier shown to be recognized by two factors of the TFIID enzyme (Burke and Kadonaga, 1997). A striking preference for the initiator consensus in promoters that contain a DPE suggests a strong co-dependency of both elements. Although evidence for downstream vertebrate elements exists, current knowledge suggests that DPEs play a less important role in these organisms. It should finally be noted that in TATA-less promoters, different transcription starts from several neighboring bases have been observed, and a transcription start "site" as such does not exist. If the promoter elements are not well conserved, it possibly is a general rule that the transcription start differs within a small range.

Sequence patterns in the core promoter. To summarize, the main sequence patterns by which interactions with transcription factors occur in the core promoter, and which could be exploited in a computational promoter finding system, are the TATA box, the initiator, and the downstream promoter element (see figure 2.3). These are all known to be directly targeted by TFIID compo-

nents. Bucher (1990) was the first to systematically study the patterns of TATA box and initiator in vertebrates, and Arkhipova (1995) extended this to the sequence elements of *Drosophila*, including DPE. On the one hand, she found that the TATA box is present in at most 50% of the *Drosophila* promoters which is less frequent than in vertebrates. On the other hand, the initiator is better conserved in fly promoters. Also, as stated above, the DPE is much more frequent in *Drosophila* and appears to play the role as a downstream counterpart of the TATA box. Hence, the machinery of transcription is well conserved throughout the whole eukaryotic kingdom, but the ways in which it is employed in transcription regulation are not. This makes it vital to use different models for the prediction of promoters in different organisms. It shall also be noted that a working binding site such as the TATA box is not defined by some absolute strength but also by the context in which it appears: Using the best hit of a TATA box model within each promoter, instead of all above a threshold, results in a much better sensitivity and specificity of the detection of known TATA boxes (Audic and Claverie, 1998).

TFIIA and TFIIB also have direct contact with DNA, but it has been widely believed that these contacts are not sequence-specific. Recent experiments (Lagrange et al., 1998) that were published during the course of this thesis suggest that TFIIB binding in humans is at least partly influenced by a sequence motif directly upstream of the TATA box, but this is not well characterized so far. Detailed studies of sequence motifs in *Drosophila* (Arkhipova, 1995; Kutach and Kadonaga, 2000) revealed the TATA, initiator, and DPE motifs, but failed to detect any motif resembling the TFIIB response element. So even if it might be present in a small number of cases, it does not play an overall important role, at least in *Drosophila*. It is striking that this putative sequence pattern consists almost exclusively of guanines and cytosines: Human promoters have a very high overall GC content. This gives rise to the suspicion that the TFIIB element is to some extent reflecting the overall human promoter sequence composition. Thus, the proven *in vitro* binding of TFIIB to a sequence pattern might not play a specific role *in vivo*.

2.2.2 Chromatin structure in promoter regions

The large number of genes found in eukaryote genomes would render it very impractical should all of them compete for the components of the basal transcription machinery at the same time. Most of them are transcribed only inside a specific tissue or under rarely occurring circumstances. Evolution has therefore found a way to effectively shut down large regions of the genome that are not needed within a certain tissue. This also guarantees that all cells of a tissue stay committed to expressing the same genes without actually losing parts of the genome.

Experiments have shown that even transcribed genes are still wrapped up around nucleosomes (cf. figure 1.2), but that the higher order condensation into solenoids is lost in active or

potentially active genes. Such genes exhibit a heightened sensitivity to a DNA digestion enzyme that even extends for some distance up- and downstream of the transcribed regions. These less condensed regions are not dependent on the act of transcribing but remain stably established and thus reflect the ability to be transcribed.

Methylation and CpG islands. In vertebrates, the open solenoid structure is closely associated with *DNA methylation*: some cytosines are chemically modified and bear an additional methyl group. In 90% of the cases, the methylation occurs in cytosines that are part of the di-nucleotide CG.[4] It was found that some CG sites are always methylated whereas for others, this pattern keeps changing in a tissue-specific manner, and active genes appear to be un-methylated. Furthermore, Antequera and Bird (1993) postulated that the upstream regions of all constitutively (i. e., constantly) expressed genes, and also a substantial portion of other genes, are correlated with clusters of CG dinucleotides, so-called CpG islands (Gardiner-Garden and Frommer, 1987). Methylated CG dinucleotides are a hot spot for mutations in which the cytosine is wrongly replaced by a thymine, which over the course of time leads to CG depleted regions. Indeed, the CG di-nucleotide occurs much less frequently in vertebrate genomes than expected from the mono-nucleotide composition. CpG islands with a high number of CG di-nucleotides therefore hint at generally low methylated regions.

DNA methylation has no direct effect on the chromatin structure, and no direct evidence for specific protein interactions with methylated regions that are associated with chromatin structure has been reported. On the other hand, DNA methylation is known to be stable during cell division because the CG di-nucleotide on the opposite strand of a methylated one is also methylated. Methylation can therefore explain the stable commission to certain groups of active genes within specific tissues.

Histone modification. Methylation can possibly explain the majority of cases of tissue-specific commitment in vertebrates, but in invertebrates such as *Drosophila* it hardly occurs (Lyko, 2001). A number of vertebrate cases are also known where differences in methylation between expressing and non-expressing tissues cannot be detected. Other features of active chromatin structure concern chemical modifications of the histones. Histone modifications either affect histone association with each other or the DNA, or proteins interacting with histones. It is known that one component of the TFIID enzyme as well as other transcription factors have the ability to acetylate histones which leads to an opening of chromatin. The opposite case of de-acetylation

[4]The notation "CpG" for a CG di-nucleotide is used to resemble the phosphate bridge between adjacent bases. This avoids the possible mis-interpretation of CG as a complementary pair in the double helix.

and therefore a negative effect on regulation is also observed. In either way, chromatin structure could thus be changed.

Chromatin structure in regulatory regions. Following the discovery that a change in the chromatin structure of genes is necessary for their (potential) activation, further studies showed that the DNA in the regulatory regions of active genes is even more sensitive to DNA digestion. These hypersensitive sites are a result of either loss or modification of nucleosomes and hint at less tightly packed DNA compared to active genes. Hypersensitive sites are furthermore not only concentrated in the core promoter region, but exist also in other regulatory regions described in section 2.2.3. Widely used transcription factors that bind to those regions associate with specific proteins that indeed have the capability of displacing or modifying the nucleosomes. A common mechanism in gene regulation is therefore the attraction of nucleosome displacing factors which enables the binding of other factors and finally the PIC itself. The observation that nucleosome displacing proteins have also been found in some of the holo-enzymes of pol-II perfectly fits in that picture.

The DNA in promoter regions is furthermore likely to exist in an alternative super-coiled conformation, the so-called *Z-DNA*. This conformation occurs in DNA with alternating purine and pyrimidine nucleotides and offers an increased accessibility to the single strands of DNA, which means that it is easier for proteins to interact with Z-DNA than with normal DNA.

2.2.3 Specific gene regulation: Sequence elements and transcription factors

So far, I have dealt with the basal transcription machinery, describing how the transcription start site is recognized and which proteins are involved in this process, and with the chromatin structure that enables the access to genes in the first place. In eukaryotes, where each mRNA that is transcribed encodes for only one gene[5], this cannot explain how genes whose protein products are needed in parallel are co-regulated. Very often, coordinately expressed genes do not even reside at close positions in the genome, but rather on different chromosomes. Such a system reflects the greater need for flexibility in eukaryotes; for example, human α-globins on chromosome 16 are expressed at the same time as γ-globins on chromosome 11 to form working globins in the fetus, but in adults the γ-globins are replaced by β-globins which also reside on chromosome 11.

Britten and Davidson (1969) published an early working model of such coordinated gene expression that, at an abstract level, still holds (see figure 2.4). They proposed that genes regulated

[5]There is no rule without exception: Note the *Drosophila Adh/AdhR* genes that are transcribed on one mRNA.

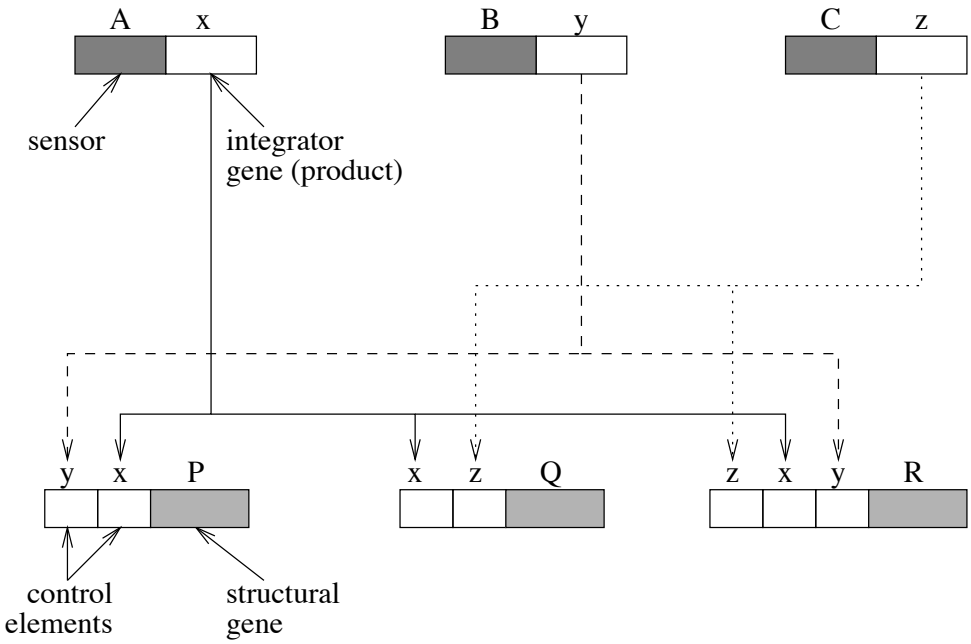

Figure 2.4: **The Britten and Davidson model** for coordinated gene regulation, after Latchman (1998). Sensor elements A, B, and C detect changes that require a different expression and therefore switch on appropriate integrator genes x, y, and z. The products of genes x, y, and z then interact with control elements, coordinately switching on appropriate genes P, Q, and R. Alternatively, x, y, and z can be proteins undergoing a conformational change under the presence of specific signals which enables them to interact with the control elements.

in parallel, in response to a particular signal, would contain a common regulatory element which would cause the activation of these genes. Moreover, genes could contain more than one element, each shared with a different group of genes. A signal would then act by stimulating a specific "integrator gene" whose product would interact with a specific sequence element in several genes at once. A gene would finally be activated if all its sequence elements had been "switched on" by integrator gene products. Using current terminology, the integrator gene is considered as encoding a transcription factor which binds to regulatory sequence elements, the transcription factor binding sites, and activates or suppresses a specific group of genes. Supplementing the original theory, a transcription factor can be activated not only by *de novo* synthesis but also by changing the inactive state of the pre-existing protein into an active one, often by means of post-transcriptional regulation (see section 2.1). The latter possibility is actually the more frequent one, as a regulation of transcription factors by transcriptional control simply pushes the problem onto a higher level.

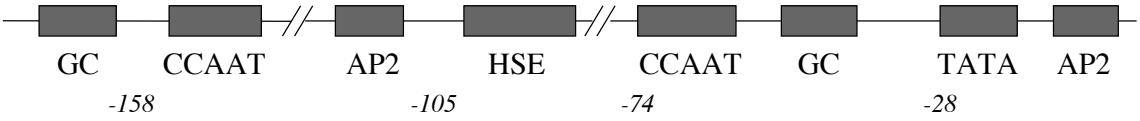

Figure 2.5: **The promoter of the human hsp70 gene.** As an example, this promoter contains non-specific (CCAAT, GC, AP2 boxes) as well as specific (HSE, heat shock element) control elements in its sequence (Latchman, 1998). The numbers in this schematic structure refer to the position relative to the transcription start site.

The proximal promoter region. Many of the regulatory elements serving as transcription factor targets are located in the proximal promoter region, i. e. directly upstream of the core promoter. These factors can either influence (both suppress or alleviate) the binding of the core promoter components, or the chromatin structure (see section 2.2.2), or both at the same time. The first group thus interacts with components of the general initiation factors, such as the TATA box binding protein or its associated factors, or alleviates the binding of other factors which then interact with the basal machinery. This only works when considering the 2-d or 3-d DNA structure — in a linear DNA sequence, the binding sites are too far away from each other to enable direct contacts of their TFs. An example for a synergistic interaction of two transcription factors with two TAFs is reported by Verrijzer and Tjian (1996): The two *Drosophila* factors bicoid and hunchback interact with specific TAFs and lead independently to an already improved transcription activation, which is nonlinearly increased when both factors are present. Therefore, the complex structure of TFIID is not necessary to bind to the DNA and recruit the polymerase, but rather serves as a modular machinery that offers a vast number of possibilities to interact with.

Some transcription factors work in a non-specific way, i. e. they merely serve to increase the production rate of the basal machinery and can thus be found in a variety of genes. Sequence elements that interact with these factors are the CCAAT and GC boxes in vertebrates, or the GAGA box in *Drosophila*. Other factors work in a very specific way and are contained in only a small number of promoters (see figure 2.5 as an example for a human promoter). Transcription elements can be present in several copies in one promoter and are very often organized in dyad symmetry, i. e. one of two identical sequence parts is contained on the sense and the other on the anti-sense strand, thus forming a palindrome. Orientation therefore does not matter, but this is also true for many non-palindromic patterns. In some cases, elements responding on related stimuli are also related on the sequence level, such as in the case of hormone receptor binding sites, some of which are made up by the same repeat of the sequence GGTCA, but with variable spacing and either as a direct or palindromic repeat (see the examples in table 2.1).

Signal	Regulatory element
Palindromic repeats	
Oestrogen	RGGTCAN^{3}TGACCY
Glucocorticoid	RGRACAN^{3}TGTYCY
Direct repeats	
Vitamin D3	AGGTCAN^{3}AGGTCA
Thyroid hormone	AGGTCAN^{4}AGGTCA

Table 2.1: **Various hormone response elements.** An N indicates any base; R indicates a purine, Y a pyrimidine (see appendix 3). Note that these are consensus sequences — i. e. the most frequent base is given at each position — but that individual binding sites may have mutations differing from the consensus.

Enhancers and silencers. Apart from the proximal promoter regions, it has been discovered that sequences as far away as several kilobases have a major influence on transcription. Although these sequences cannot act as promoters on their own, they are able to enhance or suppress the activity of transcription up to three orders of magnitude. Interestingly, such a sequence cannot only be far away from the promoter it affects, but also both upstream or downstream and even within an intron of the gene which promoter it enhances. As with many transcription factors, the orientation of the sequence, i. e. whether it is on the sense or anti-sense strand, is also not important for its functionality. These *enhancers* or *silencers* often exhibit a tissue-specific activity, and they are often composed of the same sequence elements found in (proximal) promoters that mediate tissue-specific expression. Like transcription factors binding to promoters, factors binding to enhancer elements influence gene expression both by changing the chromatin structure and by interaction with proteins of the transcription apparatus. Because of the very large distance of the enhancers from the affected promoters, the second mechanism is especially puzzling, and the most commonly accepted explanation in concordance with experimental results involves the looping out of intervening DNA. A particular enhancer can affect more than one promoter, and can exert its influence also on the transcription of other genes when transferred into their neighborhood.

Locus control regions. A high level of control of the expression of several genes at once is achieved by so-called *locus control regions* (LCRs). These regions were found to be crucial for the activity of all the genes in a cluster, e. g. the α- or β-globin genes. They act independently of their position and over a large distance, and without them no single promoter in a cluster

can attract the polymerase *in vivo*. As with enhancers, some elements that are present in promoter regions are also found in LCRs, and they are likely to have a long-range influence on the chromatin structure. Several LCRs were also found to contain sequences which are involved in the attachment of chromatin domains to a protein scaffold, the so-called nuclear matrix. An LCR controlled region may therefore constitute one solenoid loop, the structure of which — and therefore general accessibility to transcription factors — is regulated as a single unit. It also serves as an insulator to block the activity of outside enhancers.

Chapter 3

Ambiguous Nucleotide Letters

To deal with incomplete specification of bases in nucleic acid sequences, the Nomenclature Committee of the International Union of Biochemistry (NC-IUB) issued a nomenclature where single letter symbols are assigned to groups of nucleotides. This is useful in cases where two or more bases are permitted at a particular position, or where uncertainty exists as to extent and/or identity. These ambiguous codes are often used to describe consensus sequences, i. e. the common denominator of several instances of a binding site. They are given in table 3.1.

G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	G or C	Strong interaction (3 H bonds)
W	A or T	Weak interaction (2 H bonds)
H	A or C or T	not-G, H follows G in the alphabet
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	aNy

Table 3.1: Ambiguous nucleotide letter code

Bibliography

F. Antequera and A. Bird. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. U.S.A.*, 90:11995–11999, 1993.

I. Arkhipova. Promoter elements in *Drosophila melanogaster* revealed by sequence analysis. *Genetics*, 139:1359–1369, 1995.

S. Audic and J.-M. Claverie. Visualizing the competitive recognition of TATA-boxes in vertebrate promoters. *Trends Genet.*, 14:10–11, 1998.

V. Bafna and D. H. Huson. The conserved exon method for gene finding. In *Proc Int Conf Intell Syst Mol Biol.*, volume 8, pages 3–12, 2000.

S. Batzoglou, L. Pachter, J. P. Mesirov, B. Berger, and E. S. Lander. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, 10(7):950–958, 2000.

E. Birney, A. Bateman, M. E. Clamp, and T. J. Hubbard. Mining the draft human genome. *Nature*, 409: 827–828, 2001.

E. Birney and R. Durbin. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.*, 10(4): 547–548, 2000.

J. M. Bower and H. Bolouri, editors. *Computational Modeling of Genetic and Biochemical Networks*. MIT Press, Cambridge, MA, 2001.

R. J. Britten and E. H. Davidson. Gene regulation for higher cells: a theory. *Science*, 165:349–358, 1969.

P. Bucher. Weight matrix description of four eukaryotic RNA polymerase II promotor elements derived from 502 unrelated promotor sequences. *J Mol Biol.*, 212:563–578, 1990.

C. Burge. *Identification of Genes in Human Genomic DNA*. PhD thesis, Stanford University, 1997.

C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.*, 268: 78–94, 1997.

C. Burge and S. Karlin. Finding the genes in genomic DNA. *Current Opinion in Structural Biology*, 8: 346–354, 1998.

T. W. Burke and J. T. Kadonaga. The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev*, 11:3020–3031, 1997.

G. E. Chalkley and C. P. Verrijzer. DNA binding site selection by RNA polymerase II TAFs: a $\mathrm{TAF}_{II}250$–$\mathrm{TAF}_{II}150$ complex recognizes the initiator. *EMBO J*, 18:4835–4845, 1999.

J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–685, 1997.

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, 1998.

W. Fleischmann, S. Moller, A. Gateau, and R. Apweiler. A novel method for automatic functional annotation of proteins. *Bioinformatics*, 15:228–233, 1999.

L. Florea, G. Hartzell, Z. Zhang, G. M. Rubin, and W. Miller. A computer program for aligning a cDNA sequence with a genomic sequence. *Genome Res.*, 8:967–974, 1998.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *J Comp Biol.*, 7:601–620, 2000.

M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *J Mol Biol.*, 196:261–282, 1987.

W. Gish and D. J. States. Identification of protein encoding regions by database similarity search. *Nature Genet*, 3:266–272, 1993.

B. R. Graveley. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, 17: 100–107, 2001.

B. B. Haab, M. J. Dunham, and P. O. Brown. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biology*, 2:research0004.1–0004.13, 2001.

D. Haussler. Computational gene finding. *Trends supplement*, pages 12–15, 1998.

A. Krogh. Using database matches with HMMGene for automated gene detection in Drosophila. *Genome Res.*, 10(4):523–528, 2000.

A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.*, 305:567–580, 2001.

A. Krogh, I. S. Mian, and D. Haussler. A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res.*, 22:4768–4778, 1994.

D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman. A generalized hidden Markov model for the recognition of human genes in DNA. In *Proc Int Conf Intell Syst Mol Biol.*, volume 4, pages 134–142, 1996.

A. K. Kutach and J. T. Kadonaga. The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol Cell Biol.*, 20:4754–4764, 2000.

T. Lagrange, A. N. Kapanidis, H. Tang, D. Reinberg, and R. H. Ebright. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev.*, 12:34–44, 1998.

D. S. Latchman. *Gene Regulation — A Eukaryotic Perspective*. Stanley Thornes Ltd, 3rd edition, 1998.

S. E. Lewis, M. Ashburner, and M. G. Reese. Annotating eukaryote genomes. *Curr Opin Struct Biol*, 10: 349–354, 2000.

F. Lyko. DNA methylation learns to fly. *Trends Genet.*, 17:169–172, 2001.

D. B. Nikolov and S. K. Burley. RNA polymerase II transcription initiation: A structural view. *Proc. Natl Acad. Sci. U.S.A.*, 94:15–22, 1997.

J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. Sequence comparison using multiple sequences detect twice as many remote homologues as pairwise methods. *J Mol Biol.*, 284:1201–1210, 1998.

P. Pavlidis, T. S. Furey, M. Liberto, D. Haussler, and W. N. Grundy. Promoter region-based classification of genes. In *Pac Symp Biocomput.*, volume 6, pages 151–164, 2001.

M. G. Reese, D. Kulp, H. Tammana, and D. Haussler. Genie —gene finding in *Drosophila melanogaster*. *Genome Res.*, 10:529–538, 2000.

R. G. Roeder. The role of general initiaton factors in transcription by RNA polymerase II. *Trends Biochem. Sci.*, 21:327–334, 1996.

P. Rouzé, N. Pavy, and S. Rombauts. Genome annotation: which tools do we have for it? *Curr Opin Plant Biol*, 2:90–95, 1999.

D. D. Shoemaker et al. Experimental annotation of the human genome using microarray technology. *Nature*, 409:922–927, 2001.

G. B. Singh, J. A. Kramer, and S. A. Krawetz. Mathematical model to predict regions of chromatin attachment to the nuclear matrix. *Nucleic Acids Res.*, 25:1419–1425, 1997.

P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell-cyle regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*, 9:3273–3297, 1998.

G. D. Stormo. Gene-finding approaches for eukaryotes. *Genome Res.*, 10(4):394–397, 2000.

S. A. Teichmann, C. Chothia, and M. Gerstein. Advances in structural genomics. *Curr Opin Struct Biol*, 9:390–399, 1999.

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat. Genet.*, 25: 25–29, 2000.

The National Human Genome Research Institute. Glossary of genetic terms. http://www.nhgri.nih.gov/DIR/VIP/Glossary/, 2002.

C. P. Verrijzer and R. Tjian. TAFs mediate transcriptional activation and promoter selectivity. *Trends Biochem. Sci.*, 21:338–342, 1996.

R.-F. Yeh, L. P. Lim, and C. B. Burge. Computational inference of homologous gene structures in the human genome. *Genome Res.*, 11:803–816, 2001.