Università Degli Studi Di Torino
Facoltà di Scienze Matematiche Fisiche e Naturali
Dipartimento di Fisica Teorica

Dottorato di Ricerca in Sistemi Complessi Applicati alla Biologia Post-Genomica

CICLO XIX

Eukaryotic non coding DNA: statistical properties of 5'UTR exons and comparative genomics methods for non coding sequences

TESI PRESENTATA DA:
Dott. Loredana Martignetti

TUTORS:
Prof. M. Caselle
Dott. C. Herrmann

COORDINATORE DEL CICLO: Prof. F. Bussolino
RELATORE ESTERNO: Prof. C. Destri

Anni Accademici: 2003 - 2007
SETTORE SCIENTIFICO-DISCIPLINARE DI APPARTENENZA: FIS-02

Contents

Abstract

The availability of complete sequenced genomes has led to the crucial discovery that large amounts of non protein coding DNA is a general property of the genomes of complex eukaryotes. Altough the great increase in methodologies and approaches to analyze the unknown fraction of the genome, the rules governing its structure, function and evolution are far from being well understood. In this direction, statistical and computational methods can provide a great aid to extract information about general features of the non coding genome.

The first part of the thesis manuscript is devoted to discuss a general statistical property observed in the length distribution of a particular class of eukaryotic non coding DNA sequences, namely the 5'UTR exons. We show that both in mouse and in human these exons show a very clean power law decay in their length distribution and suggest a simple evolutionary model which may explain this finding. We conjecture that this power law behaviour could indeed be a general feature of higher eukaryotes.

In the second part of the manuscript, I introduce comparative genomics methods applied to the non coding DNA of complex eukaryotes, in particular in the Drosophila genome. Based on the recent availability of 12 Drosophila species genome sequences and annotation, we present a novel large scale comparison strategy optimized to highlight peculiar rearrangements and turnover

of non coding DNA evolution.

The statistical analysis and the comparative methods applied to DNA sequences during this work rely on extensive application and development of computational tools. The aim is considering the whole available genome–wide ensemble of non coding sequences, treated as a complex system, to give some insights into general rules behind the observed biological experimental data.

Chapter 1

Biological background

This first chapter aims at giving a concise background on eukaryotic non–protein coding DNA involved in regulation of gene expression, mainly following [1–3]. We discuss the features of DNA sequences known to play a role in transcription and processing of RNA, their function and their contribution to genome evolution.

1.1 Eukaryotic genome

Since our research focuses on eukaryotic organisms, we present a brief overview on the key features of eukaryotic genomes.

1.1.1 Genome organization in eukaryotes

The genome of most eukaryotes is much more complex than those of prokaryotes and the DNA of eukaryotic cells is also organized differently from that of prokaryotic cells. Unlike bacterial cells, eukaryotic ones contain a nucleus that accommodates the chromosomes.

The genomes of prokaryotes are contained in single chromosomes, which are usually circular DNA molecules. In contrast, the genomes of eukaryotes are composed of multiple chromosomes, each containing a linear molecule of DNA. Although the numbers and sizes of chromosomes vary considerably

between different species, their basic structure is the same in all eukaryotes.

The different levels of structural organization of the genome are depicted in Fig.1.1, where a replicated chromosome with two sister chromatids is shown. The DNA of eukaryotic cells is tightly bound to small basic proteins, called *histones*, that package the DNA in an orderly way in the cell nucleus. The complexes between eukaryotic DNA and proteins are called chromatin and the basic structural unit of chromatin is called nucleosome (Fig.1.2). The genetic material of cells fully condenses only upon entering the process of cell division. Resting cells show a varying pattern of active relaxed chromatin and *silent* condensed chromatin. The terms *active* and *silent* refer to the process of gene transcription, which will be explained later on.

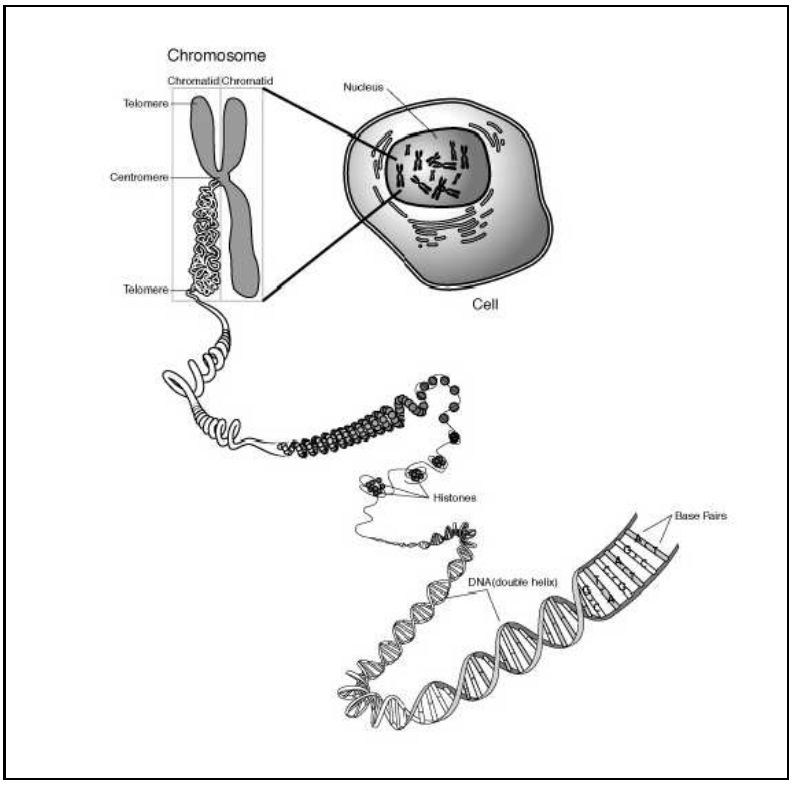


Figure 1.1: Structural organization of the genome in an eukaryotic cell (image from NIH website).

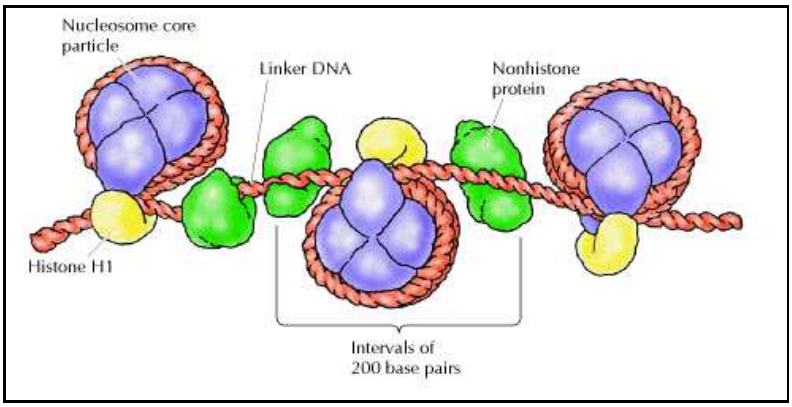


Figure 1.2: Nucleosome organization of eukaryotic genome (image from NIH website).

1.1.2 Gene structure

When a particular molecule is made by the cell, the corresponding region of the genome must therefore be accurately decoded. In molecular terms, a gene can be defined as a segment of DNA that is expressed to yield a functional product, which may be either an RNA or a polypeptide. Although most genes encode for proteins, some DNA sequences are transcribed into RNAs that do not encode proteins and they are referred to as *non coding RNA genes*.

Almost all mRNA genes have a common structure, given by the following elements.

promoter The part of a gene that contains the information to turn the gene on or off. The process of transcription is initiated at the promoter. The extent of a promoter is often difficult to determine. Proximal and distal promoter elements play a role in controlling the expression level of a gene.

exons Regions that are transcribed and exported from the nucleus as part of the messenger RNA (mRNA). The mRNA contains all information ("the message") for the formation of the final protein product of a gene.

introns Regions that are also transcribed into RNA but are excised (spliced) from the maturating RNA. Thus, these regions are absent from the mature mRNA.

UTRs Boundary mRNA regions, before the start codon and after the stop codon, which are transcribed but not translated. Namely a flanking 5' untranslated region (5' UTR) and a final 3' UTR [1]

The mRNA genes contain coding regions known as *exons*, which are expressed, with intervening sequences, known as *introns*, which are not expressed (Fig.1.3). Introns are included into the primary mRNA, also known as pre–mRNA, but they are spliced out of the mature mRNA in the cytoplasm. A cell can splice the "primary transcript" in different ways and thereby make different polypeptide chains from the same gene (a process called alternative RNA splicing) and a substantial proportion of higher eukaryotic genes (at least a third of human genes, it is estimated) produce multiple proteins in this way (isoforms), thanks to special signals in primary mRNA transcripts.

The mature mRNA usually contains not only the protein coding sequence, but also additional flanking segments, which are not translated, namely a flanking 5' untranslated region (5' UTR) and a final 3' UTR. The 5' UTR marks the start of transcription and contains an initiator codon which indicates the site of the start of translation. The 3' UTR contains a termination codon, which marks the end of translation, plus nucleotides which encode a sequence of adenosine residues known as the poly(A) tail.

During mRNA maturation, 5' and 3' UTRs can be spliced in different ways and survive in mature mRNA. For this reason, the exons in the 5' and 3' UTR regions are usually termed "non coding exons". Nucleotide patterns or motifs located in 5' UTRs and 3' UTRs are known to play crucial roles in the post–transcriptional regulation.

[1] 5' and 3' refer to the position (5' and 3' respectively) of the carbon atoms of the mRNA backbone at the two extrema of the mRNA and are conventionally used to denote the "upstream" (5') and "downstream" (3') sides of the mRNA chain.

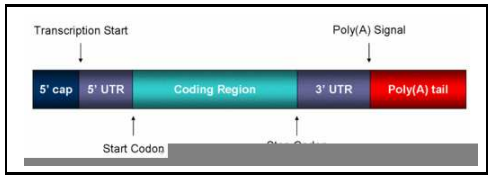


Figure 1.3: Structure of an eukaryotic genome (image from NIH website).

The translation of DNA into proteins requires a series of complex processes carefully controlled at each step by specific regulatory mechanisms activated by the cell. In particular, two crucial events in this process are the production of an intermediate molecule, the messanger RNA (mRNA) transcript, and the translation of the mRNA into proteins.
During the *transcription*, the promoter elements support the buildup of the RNA polymerase machinery. A full–length RNA copy of the genomic DNA including exons and introns is generated. The transcribed product of a gene, nuclear RNA, is subject to further modifications. A capping component is added to the 5' end and a Poly–A tail to the 3' end. The nuclear RNA is additionally shortened by the splicing process. The process of *translation* takes place after the export of the mature mRNA from the nucleus. Translation means to transfer protein–coding information on mRNAs into actual proteins by another synthesis step. Ribosomes, which are large complexes of RNA and protein, are the factories of protein biosynthesis that utilize mRNA as a template.

The flow of genetic information in cells is therefore from DNA to RNA to protein (Figure1.4). All cells, from bacteria to humans, express their genetic information in this way, a principle so fundamental that it is termed the central dogma of molecular biology.

1.2 Gene regulation in eukaryotes

The cell provides fine regulatory systems to regulate the gene expression both at transcriptional and post–transcriptional level, using several cis–acting sig-

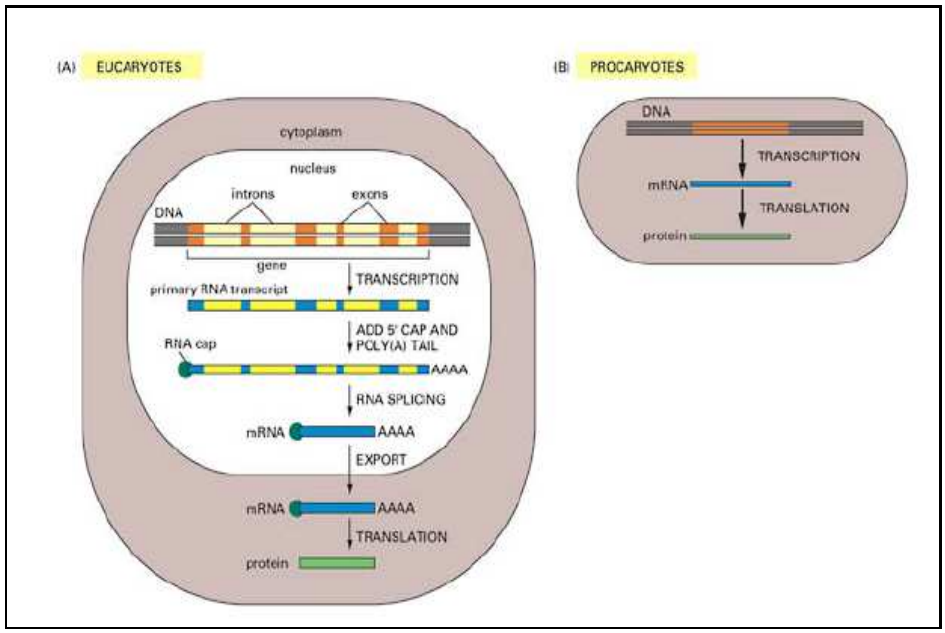


Figure 1.4: (A) In eukaryotic cells, genetic information is from DNA to RNA to protein. (B) In procaryotes, the production of mRNA molecules is simpler. The 5' end of an mRNA molecule is produced by the initiation of transcription by RNA polymerase, and the 3' end is produced by the termination of transcription. (Reproduced from http://www.accessexcellence.org)

nals located in the DNA sequence. A common molecular basis for much of the control of gene expression (whether it occurs at the level of initiation of transcription, mRNA processing, translation or mRNA transport) is the binding of protein factors and specific RNA elements to regulatory nucleic acid sequences. The absence or presence of regulatory elements has been shown to influence the initiation rate of transcription.

We can classify some general types of cis–acting elements controlling the activity of RNA polymerase II–transcribed genes:

- basal promoter sequences near the transcriptional initiation site
- long distance elements (enhancers, silencers), which regulate levels of gene activity

- Boundary elements or insulators, which can functionally separate regulatory elements.
- chromatin regulatory elements, such as those interacting with Polycomb and Trithorax proteins, which contribute to gene regulation at chromatin structure level rather than at gene level.

Since long–distance interactions are difficult to study in experiment, little is known about the underlying principles. The situation is better described for proximal elements.

1.2.1 Gene regulation at promoter level

Transcription initiates when proteins known as transcription factors bind to the promoter region and to other regulatory regions. The transcription factors contain specific structural domains, such as leucine zippers and zinc fingers, which bind to the regulatory regions that become exposed in surface folds of the double helix through the process of chromatin remodelling. There is increasing evidence that changes in chromatin structure play a key role in gene expression. Actively transcribed genes in euchromatin occur in clusters in a loop or "domain" which is unfolded.

Therefore, a prerequisite to start transcription is satisfied if a set of binding sites is occupied and/or released by the corresponding transcription factors. The RNA polymerase II complex is then recruited. Transcription factors are organized in a complex network and they act in a cooperative and combinatorial way by protein–protein interactions. A typical control region for a gene includes DNA binding sites to several factors, organized in modules. Moreover, each specific TF binding site is also often overrepresented in a given promoter sequence (cfr. [1, 4]).

TF binding sites are in general characterized by the common following features:

- they are short motifs, ranging from 5 to 20 bps

- they can be overrepresented, i.e. to appear in multiple copies in the same promoter region
- they are quite variable, meaning that are not fixed DNA strings but more often they are specified by a nucleotide pattern
- they are dispersed over long distances, over 10000 bps in the human case
- they are active in both the orientation of the DNA molecule

1.2.2 Enhancers and silencers

Enhancers/silencers were originally defined as DNA sequences that increased or decreased expression of a linked gene in an orientation- and distance- independent manner. In contemporary usage, an enhancer/silencer can refer to any (usually few Kbps long) element that binds sequence–specific transcription factors acting in a positive or negative manner. Usually more than one single type of transcription factor binds to an enhancer/silencer, creating a so called *cis–regulatory module*. They have been suggested to function in two distinct ways through remodeling of chromatin, facilitating or interfering with binding of the transcriptional machinery, and through direct interactions with the basal transcriptional machinery (see [5, 6]).

Two distinct models of enhancer action have been proposed to ascribe different computational functional roles to the enhancer (Fig.1.5). In the first one, the "enhanceosome" model, the arrangement of binding sites within the enhancer is critical to dictating the correct output of the element, so the enhancer acts as a molecular computer, leading to a single output directed to the general machinery.

In the second model, the enhancer acts as an information display, or "billboard", which is then read and interpreted by consecutive interactions with the basal machinery. In the case of a billboard enhancer, exact binding site locations are less critical, and both activating and repressing states can be represented at the same time within an enhancer.

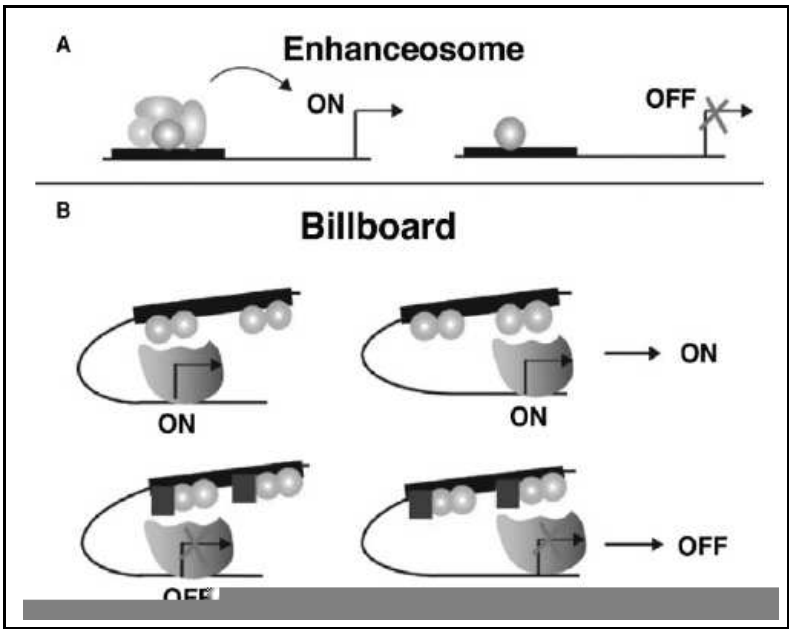


Figure 1.5: Two models of enhancer action. (A) In the "enhanceosome" model, the binding sites within the enhancer allow for a highly cooperative assembly of transcription factors(ovals), leading to gene activation. Disruption or displacment of a single binding site, or the absence of one regulatory protein, causes the element to be inactive. (B) In the "billboard" model, the enhancer contains multiple functional units that are able to independently regulate gene expression. (Image from [5]).

Recent data show that enhancers are in the same genomic region of the genes the activate and that these interactions persist during transcription, suggesting that direct interactions between enhancers and target genes may be important for activation. The level of expression from a promoter is determined by the frequency of enhancer–promoter communication, the time of enhancer–promoter interaction, and by competition between alternative enhancer–promoter interactions. Notwithstanding data showing that enhancers are in physical proximity to the genes they regulate, it is unclear how they find their targets.

1.2.3 Chromatin insulators

Active and silenced chromatin domains are often in close juxtaposition to one another and enhancer and silencer elements operate over large distances to regulate the genes in these domains. The lack of promiscuity in the function of these elements suggests that active mechanisms exist to restrict their activity [7].

Insulators are DNA elements that restrict the effects of long–range regulatory elements. Studies on different insulators from different organisms have identifed common themes in their mode of action. Numerous insulators map to promoters of genes or have binding sites for transcription factors and like active chromatin hubs and silenced loci, insulators also cluster in the nucleus. Regulatory elements such as chromatin insulators may function as nucleation sites for the initiation of active or repressed chromatin states, creating the frontier between domains 1.6.

1.2.4 Regulation of chromatin structure

The effects that chromatin structure can have on gene expression range from the modulation of transcription initiation at an individual promoter, through to the silencing of large segments of DNA in higher order chromatin structure.

One example of chromatin silencing concerns the Polycomb gene family. The proteins coded by these genes bind to DNA sequences called Polycomb response elements and induce formation of heterochromatin, the condensed form of chromatin that prevents transcription of the genes that it contains.

1.2.5 Gene regulation at post–transcriptional level

Once the mRNA is transcribed, the cell provides many subsequent mechanisms to drive is fate. 5' UTR and 3' UTR contain regulatory signals involved in mRNA transport, localization and stability.

5' UTR features, as their length, secondary structure and the presence

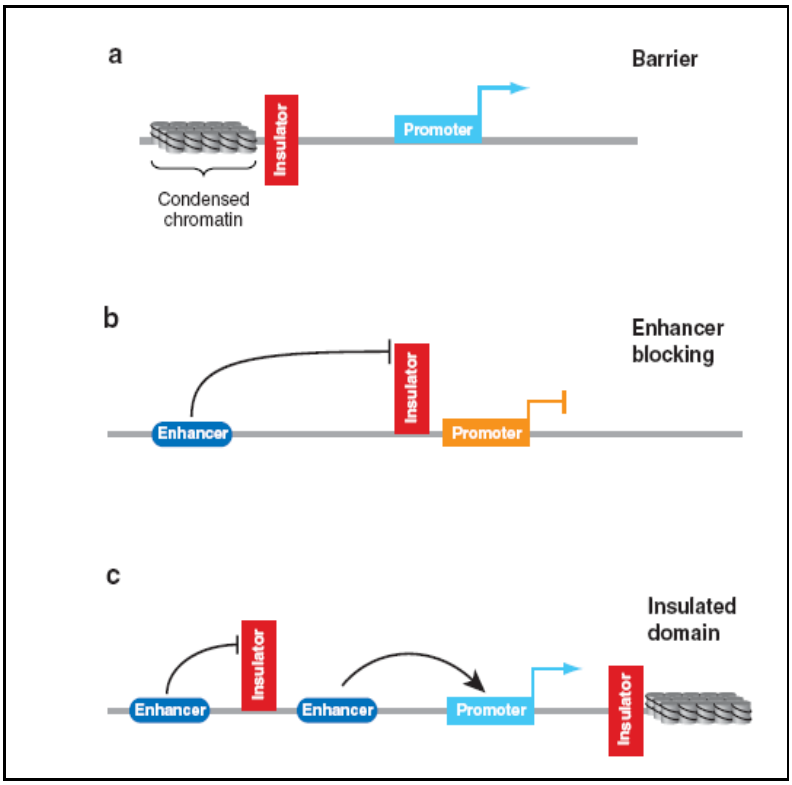


Figure 1.6: Insulators block enhancer and silencer elements in a position-dependent manner. (a) Barrier elements block the linear spread of silenced chromatin protecting the reporter gene from silencing. (b) Enhancer-blocking elements interfere with enhanced transcription when placed between an enhancer element and the promoter. (c) Flanking a transgene with insulator elements generates a functionally independent domain protected from position effects. Regulatory interactions can occur within the domain whereas the insulators block external signals. (Image from [7]).

of AUG triplets upstream of the true translation start in mRNA, known as upstream AUGs, have been shown to affect the efficiency of translation and to be preserved in the evolution of these sequences [9, 36, 40].

A recent remarkable discovery revealed the existence of an extremely important post–transcriptional regulatory mechanism, performed by an abundant class of small non coding RNA, known as microRNA (miRNA), that

recognize and bind to multiple copies of partially complementary sites in 3'UTR of target transcripts, without involving 5'UTR [12–14].

Appropriately named microRNA (miRNA), these mini-molecules are encoded by DNA like all RNA. At this stage, miRNA molecules display a very peculiar property: folding back on themselves to create a double–stranded structure known as a stem-loop (see Figure 2). This stem–loop binds special enzymes in the nucleus, that cut the miRNA molecules into smaller pieces and drive them into the cytoplasm. Once here the enzyme Dicer further cuts the miRNA, bringing it to its working size. Finally, the miRNA is able to bind a number of proteins collectively known as the RNA–Induced Silencing Complex (RISC) and assumes a linear single-stranded shape 1.7.

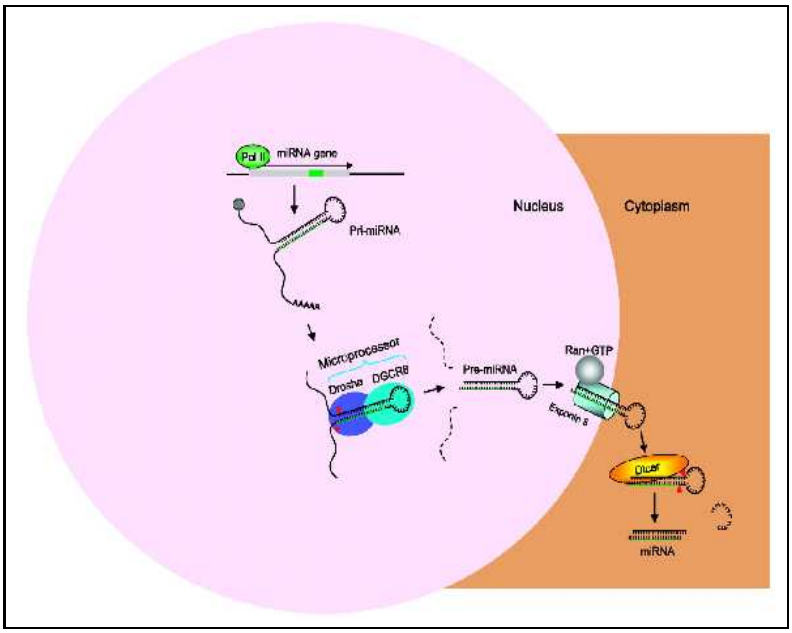


Figure 1.7: Generation pathway of a typical human miRNA.

Even though the precise mechanism of action of the miRNA/RISC complex is not very well understood, the current idea is that miRNAs are able to negatively affect the expression of a "target" gene via mRNA cleavage or translational repression, by antisense complementary base–pair matching to specif target sequences in the 3' UTR of the target gene. In plants, usually

miRNA have perfect or near perfect complementarity to their mRNA target, whereas in animals the complementarity is restricted to the 5' regions of the miRNA, in particular requiring a "seed" of 6 nucleotides, around nucleotides 2 to 7.

1.2.6 Non coding DNA and eukaryotic complexity

The genome sequencing projects have revealed an unexpected result in our understanding of the complexity in the higher organisms: a small portion of eukaryotic genome covered protein–coding genes (only up to 1.5 % in human genome), while most of eukaryotic DNA sequences that do not code for proteins.

Another striking surprise rises from the observation that complex organisms have lower numbers of protein coding genes than expected. The fruitfly Drosophila melanogaster and the nematode Caenorhabditis elegans appear to have only about twice as many protein coding genes (12-14 000) as microorganisms such as Saccharomyces cerevisiae (6200). Humans appear to have only twice as many again (30 000) (International Human Genome Sequencing Consortium, 2001). Moreover, the proteome of the higher organisms is relatively stable. Humans and mice share the 75% of their protein coding genes (cfr. [15, 16]).

Thus, phenotypic variation between both individuals and species may be based largely on differences in non–protein coding sequences and be mainly a matter of variation in gene expression, due to the control of the system. Although protein variation will also contribute, the primary source of complex phenotype and variation is due to the machanisms developed by the cell to control gene activity.

The non–coding DNA seems to have a primary role in differentiation and evolution. We described some of the them, in particular:

- the role of introns and alternative splicing process
- the information located in cis–acting gene promoters and enhancers subject to combinatorial inputs from transcription factors

- the signals devoted to the control of chromatin structure
- the activity of non–coding RNAs

The investigation of eukaryotic non–coding sequences is still at the beginning and much more about the information specified inside them has to be decipher.

1.3 DNA sequence evolution

A change in the DNA sequence that is passed on to daughter cells is called a *mutation*. Mutations - or changes to the nucleotide sequence of DNA - can arise in a number of ways. Mistakes can be made during DNA replication that result in the transmission of an incorrect base. The chemical nature of any given base can be altered either by environmental or chemical means. Once altered, these changes may then be propagated by DNA replication. Finally, large scale changes can sometimes occur in the form of chromosomal rearrangments [17, 18, 20, 21].

1.3.1 Types of mutations

We can consider the following broad categories of mutations:

- point substitution (or point mutation) involves a change in the identity of a single base. There are two types of mutation that change the identity of a base–pair:

  Transitions do not alter the chemical nature of the base. Thus, a purine base is replaced by another purine; a pyrimidine is replaced by a pyrimidine.

  The only transitions are:
  $$A \rightarrow G \text{ or } G \rightarrow A$$
  $$C \rightarrow T \text{ or } T \rightarrow C$$

Transversions change the the chemical nature of the base. Thus, a purine base is replaced by a pyrimidine; a pyrimidine is replaced by a purine.

The transversions are:

$$\begin{aligned} &A \rightarrow C \text{ or } A \rightarrow T \\ &C \rightarrow A \text{ or } C \rightarrow G \\ &G \rightarrow C \text{ or } G \rightarrow T \\ &T \rightarrow A \text{ or } T \rightarrow G \end{aligned}$$

- Deletions that result in missing DNA. These can be small, such as the removal of just one base, or longer deletions that affect a large region on the chromosome.
- Insertions that result in the addition of extra DNA.
- Inversions in which an entire section of DNA is reversed. A small inversion may involve only a few bases, while longer inversions involve large regions of a chromosome.

	Jukes-Cantor model				Kimura model			
	A	C	G	T	A	C	G	T
A	$-$	α	α	α	$-$	β	α	β
C	α	$-$	α	α	β	$-$	β	α
G	α	α	$-$	α	α	β	$-$	β
T	α	α	α	$-$	β	α	β	$-$

Figure 1.8: Models of nucleotide substitution.

Larger changes in genetic sequence (i.e. chromosomal rearrangements) place DNA chunks of several thousands to millions of nucleotides into a new genomic context. These large–scale rearrangements are often observed in cancer cells where entire chromosomes are broken up and randomly fused. Figure1.9 summarizes the most important of these changes operating at the

chromosome level.

Base pair changes are not necessarily deleterious, a change in genotype need not necessarily cause any change in phenotype. It depends on whether the change occurs within a coding region or within some other important regulatory sequence.

Silent mutations occur when a base pair change in a coding region does not affect the amino acid that is encoded. Another way in which silent mutations might occur is when the change occurs in a non–coding or non–regulatory region.

When mutations occur in a coding region, they can change one amino acid or change the entire sequence of amino acids from the point of the insertion or deletion, because the codons are now different. As the vast majority of such copying errors are *deleterious*, sophisticated error–correcting mechanisms try to avoid them. Uncorrected deleterious mutations are quickly removed from the gene pool by negative natural selection, if the fitness is decreased sufficiently. Some random mutations may have no effect on fitness, and very rarely they are *beneficial*. It seems to be a widespread belief that these very rare beneficial random mutations are the main source for the genetic variation on which positive natural selection acts, thus driving the process of evolution.

1.3.2 Evolution of genes

Gene configurations of species often resemble one another. This is rather intuitive as many biological processes like the cell cycle are shared among species. This idea is readily carried over to single gene pairs. Gene products from different species like proteins or microRNAs are far more similar than what would be expected by chance. The theory of Evolution explains this remarkable observation by postulating common ancestor sequences or species. Thus, the evolutionary history of genes or species can be represented by tree structures [22, 23].

However, gene and species trees may differ for individual gene groups. Another problem may occur when the gene studied belongs to a multi–gene family. Suppose that two related species, 1 and 2, have two duplicated genes (A1,B1) and (A2,B2), respectively. Gene duplication occurred before the speciation event. In this case, genes A1 and A2 as well as B1 and B2 are called orthologous genes. All other pairings are called paralogous. In other words, two genes are said to be paralogous if they are derived from a duplication event, but orthologous if they are derived from a speciation event. Taken together, orthologous and paralogous genes form a group of homologous genes because of their shared ancestry.

Contiguous DNA regions that encompass two or more related genes in the same order in different species (i.e. man and mouse) are an example of conserved synteny. Conserved *synteny* generalizes the concept of homology to large chromosomal regions.

1.3.3 Evolution of regulatory regions

Regulatory regions are needed to control the timing, level, and spatial location of transcription for thousands of proteins, and can have evolutionary dynamics much different from the protein–coding regions they control.

Some kinds of phenotypic difference are easier to achieve through cis–regulatory mutations than through coding mutations. Transcription is a dynamic process that can be "fine–tuned" to meet context–dependent functional demands, whereas structure is generally more static. Many aspects of organimsal phenotype require dynamic changes in gene function, including reproduction, development, behaviour, immune responses and resource utilization. Phenotypes associated to these dynamic processes might be expected to evolve to some extent more readily through regulatory rather than coding mutations [24, 25].

Gain and loss of functional transcription factor binding sites has been

proposed as a major source of evolutionary change in cis–regulatory DNA and gene expression [26,27,29]. Since binding sites in a CRM are short over–represented motifs, the regulatory module can preserve its function though after multiple point mutations. In other cases, binding sites can arise or disappeare quite easily, when a point mutation in a crucial base result in the complete loss of binding affinity.

Moreover, the modular organization of cis–regulatory regions allows a mutation in one module to affect only one part of the overall transcription profile. For instance, the effects of a cis–regulatory mutation could be limited to larval anatomy without affecting the adult, or to a single organ or tissue even when the gene is much more widely expressed. By contrast, most non–silent coding mutations change the resulting protein no matter where it is expressed.

A paradigmatic example of the complexity in cis–regulatory modules evolution is provided by the elegant work by Ludwig et al. [26], who studied a well characterized enhancer of the Drosophila even–skipped gene. They demonstrated that selection during evolution in enanhcers sequences is more likely to work at the level of en entire cis-regulatory module than on the individual binding sites. Moreover, there can be conpensatory mutations that mantain the function of the enhancer despite the loss of individual binding sites.

Similar studies highlighted the extensive turn over of binding sites during evolution, as in this work about the comparison of Ubx promoter regions between different drosophila sp ecies [27,28]. Again, they observed that the promoter is able to preserve its function, despite the high rate of rearrangements and turn over at the binding sites level.

Analogous studies have not been carried out for miRNAs due to the paucity of experimentally verified miRNA binding sites. However, several recent microarray-based studies have indicated that the rate of binding site conservation is also around 50%. Furthermore, computational miRNA target predictions indicate that many lineage-specific miRNA binding sites exist in

Drosophila and vertebrates (Rajewsky N. unpublished data)

Despite these recent observations, our current knowledge of how transcription factors, miRNAs, signalling pathways and other regulators are put together is very limited. A great challenge in modern biology is the reconstruction of the global network evolution.

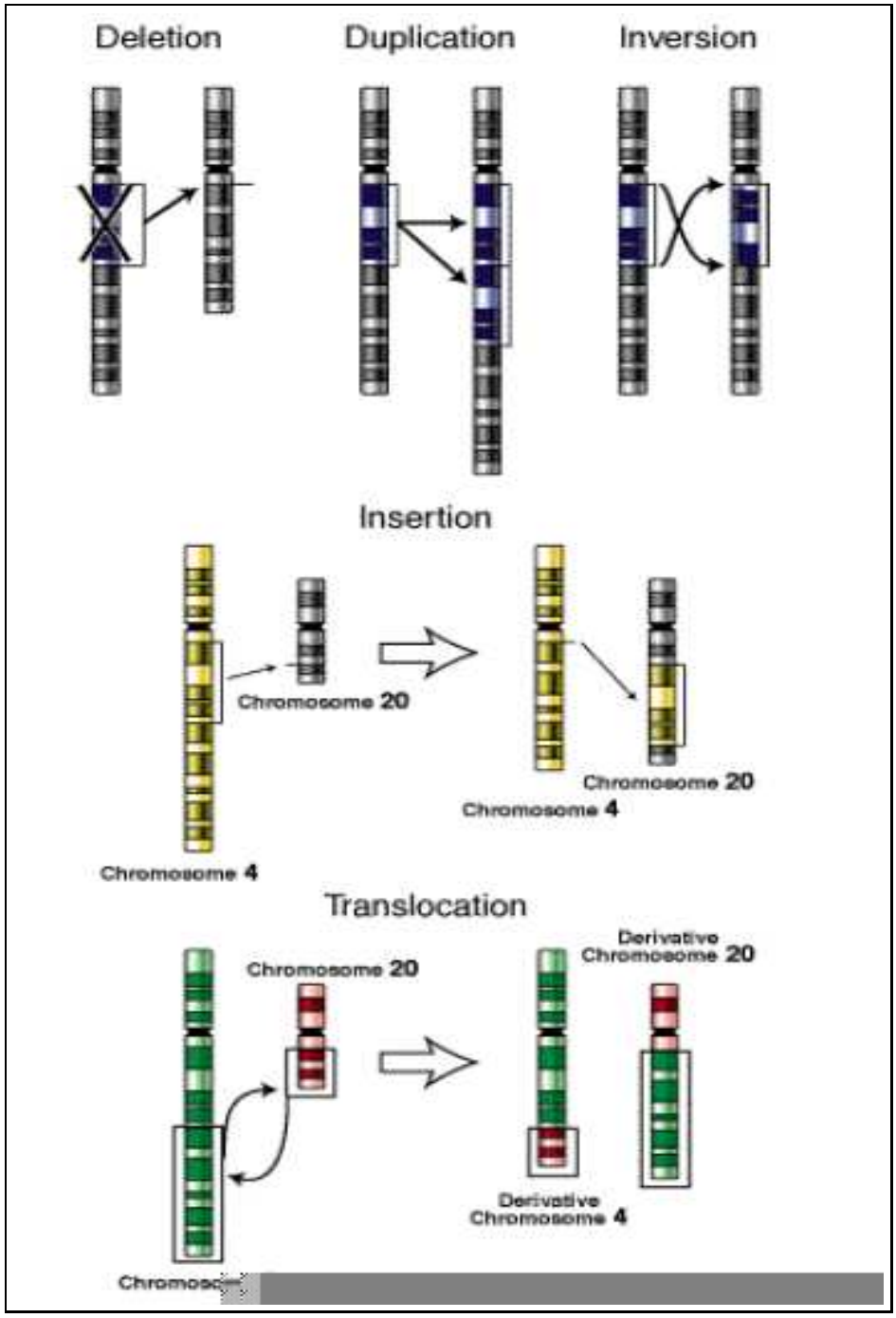


Figure 1.9: Chromosomal rearrangements. Large–scale mutations may alter the structure of a genome (i.e. number of chromosomes) and place large chunks of DNA into a new genomic context. (Image from NIH website)

Chapter 2

Statistics and bioinformatics for DNA sequences evolution

This second chapter aims at giving a brief overview of some statistical and computational methods extensively applied in the analysis of DNA sequences evolution. We will firstly discuss simple stochastic models for DNA sequences evolution and the search of general rules behind the observed genomic features. Then, we will introduce comparative genomics methods to identify DNA related sequences. Most of the fundamentals presented here are adapted from the excellent introductory text book Durbin et al. [17] and from [30].

2.1 Statistical analysis of DNA sequences evolution

Genomic sequences are a valuable source of information about the evolutionary history of species. With the rapidly growing availability of whole–genome sequence data, the evolution of genomic DNA can now be studied systematically over a wide range of scales and organisms.

Statistical analysis of the sequences can provide important biological information concerning the evolution of DNA molecules. The nucleotide sequence of a DNA molecule can be viewed as a text containing the hereditary

information of a living organism. The information content and the structure of the text is related to its evolution.

The statistical analysis is quite intricate since genomic DNA is a rather "patchy" statistical environment [31]: it consists of genes, noncoding regions, repetitive elements, etc..., and all of these substructures have a systematic influence on the local sequence composition.

2.1.1 Scaling laws in DNA sequences

In these last years lot of efforts have been devoted in trying to find universal laws in nucleotide distributions in DNA sequences. A typical example was the identification more than ten yars ago of long range correlations in the base composition of DNA (see for instance [32] and references therein). With the availability of complete sequenced genomes, the correlation property of sequences has been studied separately for coding and non coding segments of complete bacterial genomes, showing a rich variety of behaviour for different kinds of sequences [33,34]. This line of research has been recently extended to the search of similar universal distribution for more complex features of eukaryotic DNA sequences like for instance 5' untranslated regions (UTR) lengths [36], UTR introns [37] or strand asymmetries in nucleotide content [38, 39]. The main reason of interest for this type of analyses is the search of general rules behind the observed universal behaviours. The hope is to get in this way new insight in the evolutionary mechanisms shaping higher eukaryotes genomes and to understand functional role of the various portions of the genome.

The content and the organization of non–coding DNA is poorly understood and it seems to evolve by its own laws not restricted by a specific biological function. These laws are based on probabilities of various mutations that resemble the laws governing other complex systems, like systems of interacting particles such as liquids and magnets, but also purely geometric systems, such as random networks.

A wide variety of complex natural phenomena is characterized by power law behavior of their parameters. This type of behavior is also called scaling. Power laws were found to describe various systems, like the ones described above, in the vicinity of critical points. Empirical power laws are found to characterize also many physiological, ecological, and socio–economic systems.

An important step to highlight general rules in DNA sequences evolution is the construction of simplified (and possibly exactly solvable) stochastic models to describe the observed behaviours. This is the case for instance of the model discussed in [31] for base pair correlations or the model proposed in [36] for the 5'UTR length.

2.1.2 Evolutionary models of DNA sequences

From a biological point of view, the two main assumptions of any evolutionary model are:

- evolution can be decribed as a Markov process, i.e. the modifications of a DNA sequence only depend on its current state and not on its previous history.
- evolution is "shaped" by functional constraints: DNA sequences with a negligible functional role evolve at a higher rate with respect of functionally important regions. This implies that regions with different functional roles must be described by different choices of the various mutational rates. The free evolution of sequences without functional constraints is usually called "neutral evolution".

Let us see a few examples:

- protein coding exons are usually strongly constrained since the proteins they code have an important role in the life of the cell, however due to redundance of the genetic code, the third basis of each codon in the coding exons is free to mutate. On the contrary insertions and deletions are suppressed because they can dramatically affect the shape and function of the protein.

- Sequences devoted to transcriptional regulations (which very often lie outside exons) are usually so important for the life of the cell that they are kept almost unchanged over millions of years of evolution

- Regulatory sequences on the messanger RNA (mRNA) whose function often depends on the tridimensional shape of the RNA molecule and not on its exact sequence are in an intermediate situation between the above cases and the neutral evolution: they can tolerate mutations which do not modify their tridimensional shape (typically these are pairs of pointlike changes of bases and are usually called "compensatory mutations"). Most of the mRNA regulatory signals of this type are located in 3'UTR exons.

- 5'UTR regions contain sometimes regulatory sequences of the transcriptional type (which, as mentioned above, are stongly conserved under evolution) but their relative position seem not to have a crucial functional role. They can thus tolerate insertion and deletions as far as they do not affect the regulatory regions.

2.1.3 Markovian processes

A markovian process is defined as a process obeying the following rules.

1. A system at any time step t, can be in n possible states e_1, e_2, ..., e_n.

2. The probability to find a system in a certain state at any time step depends only on its state at the previous time step. Thus to fully characterize a markovian process, we must define a nxn set of transition probabilities p_{ij} which are the probabilities to find a system in a state e_i at time $t+1$ provided that a time t it was in a state e_j. Obviously, $\sum n_i = p_{ij} = 1$.

3. It is assumed that p_{ij} do not depend on time.

The evolution of DNA may be thought of as a markovian process, with mutation probabilities depending on the nature of the current state of bases.

Minimal markovian models describe evolution of the DNA sequence as a series of stochastic mutations. Three elementary processes are taken into account: changes in the nucleotide type, insertions or deletions of one or more nucleotides. The various existing models differ with each other for the different assumptions they make on the parameter which control these changes (for a review see for instance [17, 18, 42]).

2.1.4 Models of nucleotide substitution

Jukes-Cantor Model The basic assumption is equality of substitution frequency for any nucleotide at any site. Thus, changing a nucleotide to each of the three remaining nucleotides has probability α per time unit. The rate of nucleotide substitution per site per time unit is then $r = 3\alpha$. Let q be the proportion of identical nucleotides between two sequences. In a continuous time model q is given by the following equation:

$$q = 1 - \frac{3}{4}(1 - e^{-\frac{8rt}{3}}) \tag{2.1}$$

The expected number of substitutions per site (d) is approximately $2rt$. Rearranging the equation above yields:

$$d = -\frac{3}{4} ln[1 - \frac{4}{3}p] \tag{2.2}$$

where $p = 1 - q$ is the proprortion of different nucleotides.

Kimura model In DNA, the rate of transitions is usually higher than that of transversions. In the model of Kimura, both types of substitution rates are explicitly modeled with parameters α (the rate of transitions) and β (the rate of transversions). The total substitution rate per site and time unit is then $\alpha+2\beta$. Hence, the expected number of nucleotide substitutions is given by $d = 2rt = 2\alpha t + 4\beta t$ where t is the time after divergence of two sequences. Using his model, Kimura showed that the frequencies of Transitions (T_s) and Transversions (T_v) are given by

$$T_s = \frac{1}{4}(1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t}) \tag{2.3}$$

$$T_v = \frac{1}{2}(1 - e^{-8\beta t}) \tag{2.4}$$

The expected number of substitutions per site (d) is then

$$d = -\frac{1}{2}ln[1 - 2T_s - T_v]ln[1 - 2T_v] \tag{2.5}$$

Both models have an equilibrium frequency of each nucleotide of 0.25. Both models are reversible meaning that sequences evolve equally over time. In this regard, it is irrelevant whether some sequence A evolves into B or vice versa. The degree of similarity between two sequences is expressed in the expected number of substitutions per site (d). Therefore, we employ the notion of point accepted mutation per site or PAM.

One point accepted mutation ($1PAM$) is defined as an expected number of substitutions per site of 0.01. A $PAM1$ substitution matrix is thus derived from any evolutionary model by setting the row sum of off-diagonal terms to 0.01 and adjusting the diagonal terms to keep the row sum equal to 1. A substitution matrix M for any PAM distance n is then obtained by iterative multiplication of a $1PAM$ matrix: $M_n = (M_1)^n$. We are now able to model substitution processes by selecting an evolutionary model and a PAM distance, which reflects the expected degree of sequence similarity.

2.2 Sequence alignments

Once a DNA evolutionary model is defined, it is possible to compare sequences looking for evidence that they have diverged from a common ancestor. This is usually done by first aligning the sequences (or part of them) and then deciding whether that alignment is more likely to be occurred because the sequences are related, or just by chance.

An alignment can bee seen as a way of transforming one sequence into the other. The idea of aligning two sequences (of possibly different sizes) is to write one on top of the other. The example in Fig.2.1 illustrates an alignment between the sequences A="ACAAGACAGCGT" and B="AGAACAAGGCGT".

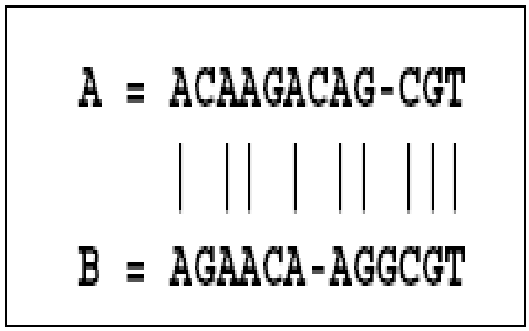


Figure 2.1: Example of nucleotide pairwise alignment.

The objective is to match identical subsequences as far as possible. In the example, nine matches are highlighted with vertical bars. However, if the sequences are not identical, mismatches are likely to occur as different letters are aligned together. Two mismatches can be identified in the example: a "C" of A aligned with a "G" of B, and a "G" of A aligned with a "C" of B. The insertion of spaces produced gaps in the sequences. They were important to allow a good alignment between the last three characters of both sequences.

The key issues of an alignment procedure are:

- what sort of alignment should be considered
- the scoring system used to rank alignments
- the algorithm used to find optimal (or good) alignments
- the statistical method used to evaluate the significance of an alignment score

2.2.1 Global vs local sequence alignment

The similarity between entire sequences was the first alignment problem that caught the attention of researchers. Needleman and Wunsch (1970) were the first to report a global alignment approach, which consists in finding the best match covering the two sequences in their entirety.

Global alignment methods are widely used to identify highly similar regions in the sequences which appear in the same order and orientation. A global alignment procedure can be suitable to find a map of genomic colinear conserved segments between the input sequences, as syntenic regions. Moreover, global alignments methods can be used for closely related sequence, as members of the same protein family, such as globins, that are highly conserved and have almost no variation in sequence length.

In many cases, the score of an alignment between substrings of two sequences may be larger than the overall score in a global alignment. In this case, detecting local sequence similarities is more informative. This is the reason why biologists are frequently interested in short regions of local similarity. Local alignment algorithms are generally very useful in finding similarity between regions that may be related but are inverted or rearranged with respect to each other. Smith andWaterman (1981) proposed an algorithmic solution for the local alignment problem. Both algorithms return optimal solutions to the respective problems. Both employ dynamic programming for this purpose, that will be discuss in section....

Global alignments are less prone to demonstrating false homology as each letter of one sequence is constrained to being aligned to only one letter of the other. Local alignments, on the other hand, can cope with rearrangements between non-syntenic, orthologous sequences by identifying similar regions in sequences; this, however, comes at the expense of a higher false positive rate due to the inability of local aligners to take into account overall conservation maps.

To compare entire genomes from different species, biologists increasingly need alignment methods that are efficient enough to handle long sequences, and accurate enough to correctly align the conserved biological features between distant species. For this reason, the novel notion of glocal alignment, a sophisticated combination of global and local methods, has been introduced [54]. This class of alignment algorithms create a map that transforms one sequence into the other while allowing for rearrangement events. This

procedure, at the base of Shuffled-LAGAN algorithm [55], is able to take into account large scale genomic rearrangments, but fails at lower scale.

2.2.2 Scoring scheme

An alignment procedure is an optimization problem and thus requires an objective function that could be maximized in terms of an alignment score. The total score we assign to an alignment will be the sum of scores for each aligned pair of residues and each gap. If the two sequences under comparison are related, we expect a match between identical residue pairs to be more likely than we expect by their single frequencies. Thus matches should contribute positively to the alignment score whereas non-conservative substitutions (i.e. transversions) and gaps should penalized. Generally, an additive scoring scheme is used under the assumption that mutations occur independently. Most alignment algorithms depend on this assumption.

Assume A and B are two sequence of length n and m over an alphabet $\Sigma := A, C, G, T$. Given a gapless alignment of those sequences, we want to assign a score to the alignment that gives a measure of the relative likelihood that the sequences are related as opposed to being unrelated. A prominent additive scoring system is based on the log-odds ratio of a residue pair (a_i, b_j) occurring as an aligned pair instead of being unaligned.

$$Score = \sum_i pairscore(a_i, b_j) \tag{2.6}$$

where

$$pairscore(a_i, b_j) = ln \frac{p_{ab}}{q_a q_b} \tag{2.7}$$

p_{ab} is the probability of seeing a residue pair (a, b) in a match model (i.e. a 1 PAM Kimura model) and q_a, q_b are the letter frequencies in a random model assuming that letters occur independently.

Introducing gaps in alignments raises the question of evaluating them. Biologists have long recognised that insertions and deletions generally do not occur a single base at a time. Therefore, when biomolecular sequences are

compared, it is commonly accepted that a gap of k spaces is more likely to appear than k incidences of a single gap spread across the sequences. In order to make this distinction, a general gap penalty function is needed so that the cost of a gap is a function of its length. The most common function is called affine gap penalty function. For affine gap penalties, the score for a gap of length x is $(g + ex)$, where $g < 0$ is the penalty for introducing a gap (gap open) and $e < 0$ is the penalty for each gap symbol (gap extension).

2.2.3 Dynamic programming alignment

Dynamic programming is a strategy of building a solution gradually using simple recurrences to compute the similarity between two sequences A and B of lengths m and n [8]. The key observation for the alignment problem is that the similarity between sequences $A[1..n]$ and $B[1..m]$ can be computed by taking the maximum of the three following values:

- the similarity of $A[1..n-1]$ and $B[1..m-1]$ plus the score of substituting $A[n]$ for $B[m]$;
- the similarity of $A[1..n-1]$ and $B[1..m]$ plus the score of deleting aligning $A[n]$;
- the similarity of $A[1..n]$ and $B[1..m-1]$ plus the score of inserting $B[m]$.

From this observation, the following can be derived:

$$sim(A[1..i], B[1..j]) =$$

$$max[sim(A[1..i-1], B[1..j-1]) + sub(A[i], B[j]);$$
$$sim(A[1..i-1], B[1..j]) + gap;$$
$$sim(A[1..i], B[1..j-1]) + gap]$$

where $sim(A, B)$ is a function that gives the similarity of two sequences A and B, $sub(a, b)$ is the scoring function that gives the score of a substitution of character a for character b and gap is the gap scoring. This is complete

with the following base case:

$$sim(A[0], B[0]) = 0;$$

where $A[0]$ and $B[0]$ are defined as empty strings.

To solve the problem with this , the algorithm build an $(n+1)(m+1)$ matrix M where each $M[i, j]$ represents the similarity between sequences $A[1..i]$ and $B[1..j]$ (Fig.2.2).

		0	1	2	3	4	5	6	7	8	9	10	11	12
		-	A	G	A	A	C	A	A	G	G	C	G	T
0	-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
1	A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
2	C	-2	0	0	-1	-2	-1	-2	-3	-4	-5	-6	-7	-8
3	A	-3	-1	-1	1	0	-1	0	-1	-2	-3	-4	-5	-6
4	A	-4	-2	-2	0	2	1	0	1	0	-1	-2	-3	-4
5	G	-5	-3	-1	-1	1	1	0	0	2	1	0	-1	-2
6	A	-6	-4	-2	0	0	0	2	1	1	1	0	-1	0
7	C	-7	-5	-3	-1	-1	1	1	1	0	0	2	1	0
8	A	-8	-6	-4	-2	0	0	2	2	1	0	1	1	2
9	G	-9	-7	-5	-3	-1	-1	1	1	3	2	1	2	1
10	C	-10	-8	-6	-4	-2	0	0	0	2	2	3	2	1
11	G	-11	-9	-7	-5	-3	-1	-1	-1	1	3	2	4	3
12	T	-12	-10	-8	-6	-4	-2	0	0	0	2	2	3	5

Figure 2.2:

The first row and the first column represent alignments of one sequence with spaces. $M[0, 0]$ represents the alignment of two empty strings, and is set to zero. All other entries are computed with the following formula:

$$M[i, j] = max[M[i-1, j-1]+sub(A[i], B[j]); M[i-1, j]+del(A[i]); M[i, j-1]+ins(B[j])]$$

The matrix can be computed either row by row (left to right) or column by column (top to bottom). In the end, $M[n, m]$ will contain the similarity score of the two sequences. Since there are $(m + 1)(n + 1)$ positions to compute and each take a constant amount of work, this algorithm has time complexity of $O(n^2)$. Clearly, it has also quadratic space complexity since it needs to keep the entire matrix in memory.

Once the matrix has been computed, the actual alignment can be retrieved by tracing a path in the matrix from the last position to the first (Figure 2.2). The trace is a simple procedure that compares the value at each $M[i, j]$ to the values of its left, top and diagonal entries according to the formula given above. For instance, if $M[i, j] = M[i, j - 1] + gap$, the trace reports an insertion of character $B[j]$ and proceeds to entry $M[i, j-1]$. In the matrix of Figure 2.2, two optimal alignments can be retrieved (Figure 2.1 and Figure 2.3).

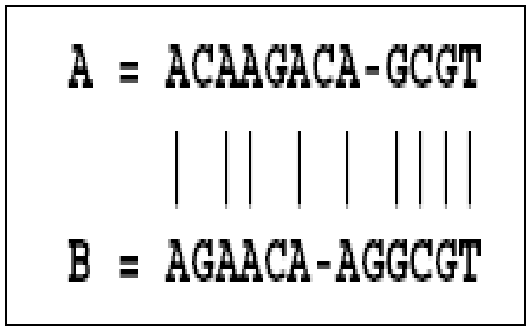


Figure 2.3: A second example of nucleotide pairwise alignment obtained comparing sequences A and B.

2.2.4 Smith–Waterman algorithm

A local alignment was defined as the problem of finding the best alignment between substrings of both sequences. In 1981, T. F. Smith and M. S. Waterman [29] showed that a local alignment can be computed using essentially the same idea employed by Needleman and Wunsch. The main difference is that $M[i, j]$ contains the similarity between suffixes of $A[1..i]$ and $B[1..j]$. As

a result, the relation is slightly altered because an empty string is a suffix of any sequence and, therefore, a score of zero is always possible. The formula for computing $M[i, j]$ becomes:

$$M[i,j] = max[0; M[i-1, j-1]+sub(A[i], B[j]); M[i-1, j]+del(A[i]); M[i, j-1]+ins(B[j])]$$

Another important distinction is that the score of the best local alignment is the highest value found anywhere in the matrix. This position is the starting point for retrieving an optimal alignment using the same procedure described for the global alignment case. The path ends, however, as soon an entry with score zero is reached. We could observe that the Smith-Waterman algorithm has the sametime and space complexity as the Needleman-Wunsch.

2.2.5 Suboptimal alignments

The best local alignment of two DNA sequences does not always capture the biological meaning. The best local alignment would be either misleading (e.g. returns only one particular element) or non-specific (e.g. covers regions of poor conservation) in such a setting. A better way of handling such comparisons is to retrieve more than one alignment from the alignment space. Waterman and Eggert (1987) proposed an algorithm for finding non–trivial local similarities, which are called suboptimal local alignments.

2.2.6 Heuristic approach

The dynamic programming algorithms are guaranteed to find the optimal solution for a given scoring scheme. However, these algorithms are not a feasible solution to the comparison of long sequences in the order of 100 Kb to several Mb. Heuristic approaches mitigate this problem by trying to reduce the search space, while still maintaining a high sensitivity level.

The most famous tool is certainly the BLAST package (Altschul et al., 1990). The BLAST algorithm exploits the idea that meaningful alignments contain short identical subsequences, or very high scoring matches. BLAST compiles a list of all words of a fixed length (e.g. 11 nucleotides for DNA),

that would match the query sequence with scores higher than some threshold. BLAST assumes collinearity of the local similarities.

2.2.7 Alignment statistical significance

To assess whether a given alignment constitutes evidence for homology, it helps to know how strong an alignment can be expected from chance alone. Distinguishing between "true" alignments (those resulting from homology) and spurious alignments (resulting from chance similarities) is a notable issue in extensive searches, where a large number of sequences are compared to a single query sequence.

The significance of an alignment ($p - value$) is the probability that an equal or better alignment score could be attained by aligning two unrelated (random) sequences. Generally, the $p - value$ of a random variable T is the probability $P(T \leq t_{observed})$.

It is possible to generate scores from random sequences and compute p–values for our observed alignments based on those. This simulation approach has the disadvantages that the sample size directly affects the accuracy of the computed p–values. Traditionally, one would generate a random DNA sequence by sampling letters from an alphabet A with probabilities $\Sigma_{i=A,C,G,T} p_i = 1$. However, this is a rather simplistic view of randomness in a biological setting. Global sequence properties like length and nucleotide composition are preserved, but regions of low complexity and local compositional biases are not represented. An alternative practical solution could be reversing one nucleotide sequence in a pairwise comparison.

While it is then possible to express the score of interest in terms of standard deviations from the mean, it is a mistake to assume that the relevant distribution is normal and convert this Z-value into a P-value; the tail behavior of gapped alignment scores is unknown. The most one can say reliably is that if 100 random alignments have score inferior to the alignment of interest, the P-value in question is likely less than 0.01.

Chapter 3

Universal power law behaviours in genomic sequences and evolutionary models

This chapter presents our study about the length distribution of a particular class of DNA sequences, namely the 5'UTR exons. These exons belong to the messanger RNA of protein coding genes, but they are not coding and are thus less constrained from an evolutionary point of view. We observe that both in mouse and in human these exons show a very clean power law decay in their length distribution and suggest a simple evolutionary model which may explain this finding.

3.1 Motivation

In these last years lot of efforts have been devoted in trying to find universal laws in nucleotide distributions in DNA sequences. A typical example was the identification more than ten yars ago of long range correlations in the base composition of DNA (see for instance [32] and references therein). With the availability of complete sequenced genomes, the correlation property of length sequences has been studied separately for coding and non coding segments of complete bacterial genomes, showing a rich variety of behaviour for different kinds of sequences [33, 34]. This line of research has been recently

extended to the search of similar universal distribution for more complex features of eukaryotic DNA sequences like for instance 5' untranslated regions (UTR) lengths [36], UTR introns [37] or strand asymmetries in nucleotide content [38, 39]. The main reason of interest for this type of analyses is the search of general rules behind the observed universal behaviours. The hope is to get in this way new insight in the evolutionary mechanisms shaping higher eukaryotes genomes and to understand functional role of the various portions of the genome. An intermediate important step of this process is the construction of simplified (and possibly exactly solvable) stochastic models to describe the observed behaviours. This is the case for instance of the model discussed in [31] for base pair correlations or the model proposed in [36] for the 5'UTR length. In this letter we describe a similar universal law for the exon length in the 5'UTR of the human and mouse genomes. Looking at the 5'UTR exons collected in the existing genome databases for the two organisms we shall first show that they follow with a high degree of confidence a power law distribution with a decay exponent of about 2.5 and then suggest a simple solvable model to describe this behaviour.

We shall also compare the impressive stability of the power law decay of 5'UTR exons with the distributions in the case of the 3'UTR and coding exons which turn out to be completely different. This is most probably due to the different evolutionary pressures to which are subject the three types of sequences.

We think that the behaviour that we observed should indeed be a general feature of higher eukaryotes, however its identification requires a very careful annotation of 5'UTR regions which exist for the moment only for human and mouse (see tab. below).

3.2 5'UTR exons biological constrains

In eukaryotic organisms, DNA information stored in genes is translated into proteins through a series of complex processes, carefully controlled at each step by specific regulatory mechanisms activated by the cell. In particular, two crucial events in this process are the production of an intermediate molecule, the messanger RNA (mRNA) transcript, and the translation of the

mRNA into proteins. The cell provides fine regulatory systems to regulate the gene expression both at transcriptional and post–transcriptional level, using several cis–acting signals located in the DNA sequence. A common molecular basis for much of the control of gene expression (whether it occurs at the level of initiation of transcription, mRNA processing, translation or mRNA transport) is the binding of protein factors and specific RNA elements to regulatory nucleic acid sequences.

Once mRNA is transcribed, it usually contains not only the protein coding sequence, but also additional segments, which are transcribed but not translated, namely a flanking 5' untranslated region (5' UTR) and a final 3' UTR Nucleotide patterns or motifs located in 5' UTRs and 3' UTRs are known to play crucial roles in the post-transcriptional regulation. Most of the primary transcripts of euKaryotic genes also contain sequences (named "introns") which are eliminated during a maturation process named "splicing". The sequences which survive this splicing process are named "exons" they are glued together by the splicing machinery and form the mature mRNA transcript. Both the UTRs and the coding portions of the mRNA are usually composed by the union of several exons. It is thus possible to classify the exons as coding, 3'UTR and 5'UTR depending on the portion of the mRNA to which they belong.

A cell can splice the "primary transcript" in different ways and thereby make different polypeptide chains from the same gene by alternative RNA splicing process and a substantial proportion of higher eukaryotic genes (at least a third of human genes, it is estimated) produce multiple proteins in this way (isoforms), thanks to special signals in primary mRNA transcripts.

Some hints about the 5' and 3' role in gene expression can be derived from a quantitative analysis of UTR length.

Recent large scale databases suggest that the mean 3' UTR length in human transcript is nearly four times longer than the mean human 5'UTR length [8] and that the evolutionary expansion of 3'UTR in higher vertebrates, not observed in 5' UTR, is associated to their peculiar regulatory role. Very recent works revealed the existence of an extremely important post-transcriptional regulatory mechanism, performed by an abundant class of small non coding RNA, known as microRNA (miRNA), that recognize and

bind to multiple copies of partially complementary sites in 3'UTR of target transcripts, without involving 5'UTR [12–14].

Differently, 5'UTR sequences are expected to be constrained mainly by splicing process and translation efficiency. The exons in the 5' UTR regions are usually termed "non coding exons", since they are not included in the protein coding portion of the transcript. However, their characteristics, as their length, secondary structure and the presence of AUG triplets upstream of the true translation start in mRNA, known as upstream AUGs, have been shown to affect the efficiency of translation and to be preserved in the evolution of these sequences [9, 36, 40]. 5'UTR exons length can vary between few tens until hundreds of nucleotides, without typical length scale around favourite size, and the lower and upper bounds of this distribution is likely to be shaped by splicing and translation efficiency: exons that are too short (under 50 bp) leave no room for the spliceosomes (enzymes that perform the splicing) to operate [41], while exons that are too long can contain signals that affect translation efficiency. 5'UTR "non coding exons" are also free from selective pressure acting on coding exons, which strongly preserves the amminoacid information written in triplets of nucleotides in the protein coding exons.

3.3 Length distribution of 5'UTR exons

In our analysis we decided to construct strictly disjoint subsets of exons, according to their position in the transcript (5'UTR exons, protein coding exons or 3'UTR exons)[1]. Moreover, we created a non redudant genome-wide datasets of exons, considering only one isoform for each gene, the most extended one.

Curated information about DNA sequences and annotation of eukaryotic organisms are provided by the Ensembl project, based on a software system which produces and maintains automatic annotation on selected eukaryotic genomes [16].

[1]Obviously in several cases one can have exons which are partially included in one of the two UTR regions and partially in the coding portion of the mRNA. These mixed exons were excluded from our analysis.

We downloaded from the Ensembl database (release 40 [16]) all the available transcripts annotated as protein coding for different organisms, and we created a filtered dataset of non redundant exons, considering the most extended transcript for each gene. We eliminated all the exons with mixed annotations and grouped the remaining ones in three classes: 5'UTR, protein coding exons, and 3'UTR.

Plotting the length distribution of exons, separately for 5'UTR, coding exons and 3'UTR, we clearly observe different behaviours, which we think should reflect different evolutionary constrains acting on these classes of DNA sequences (Fig.1 a,b,c). In particular, the 5' UTR exons size distribution shows a remarkably smooth power decay for large enough values of the exon length. To assess this point and to evaluate the threshold above which the power law behaviour starts, we fitted the observed distributions with a power law:

$$N(l) = l^{-\alpha} \tag{3.1}$$

where $N(l)$ is the number of exons of length l.

In order to evaluate the goodness of the fits that we performed, we divided the set of all exons into 18 equivalent bins and then assumed the variance of these bins as an indication of the statistical uncertainty of our estimates (results are independent from the binning choice). This allowed us to perform a meaningful χ^2 test on the fits. This test is commonly used when an assumed distribution is evaluated against the observed data [19]. The quantity χ^2 may be thought of as a measure of the discrepancy between the observed values and the respective expected values. It is convenient to compute the reduced chi square $\widetilde{\chi}^2$ (i.e. the ratio $\chi^2/(N_p - N_f)$ where N_p is the number of points included in the fit and N_f the number of parameters of the fit). With this normalization one can immediately see if the fitting function correctly describes the data (which requires $\widetilde{\chi}^2 \leq 1$). When instead $\widetilde{\chi}^2 > 1$ the absolute value of $\widetilde{\chi}^2$ gives a rough estimate of how inaccurate is the tested distribution to describe the data.

We fitted the data for the 5'UTR exons setting a minimum threshold on the exon length and then gradually increasing this threshold until a reduced $\widetilde{\chi}^2$ value smaller than one was obtained. The rationale behind this choice is that (as we shall see below) the power law decay is likely to be an asymptotic

behaviour which is violated for short exon lengths. Starting from $l_{min} \sim 150$ both in human and in mouse good $\widetilde{\chi}^2$ values were obtained and we could estimate the critical index to be $\alpha \sim 2.5$. Detailed results of the fits are reported in Tab.1. The $\widetilde{\chi}^2$ values that we found support in a quantitative way the power law behaviour of the data, which was already evident looking at Fig.1a.

On the contrary, the coding exons and the 3'UTR exons length histograms display (on a log-log scale) non linear distributions with peaks of population around favourite sizes. In the range, where we are able to fit the power law decay of 5'UTR exons length, $\widetilde{\chi}^2$ values for linear fit in the other classes of exons are completely unacceptables (Tab. 2).

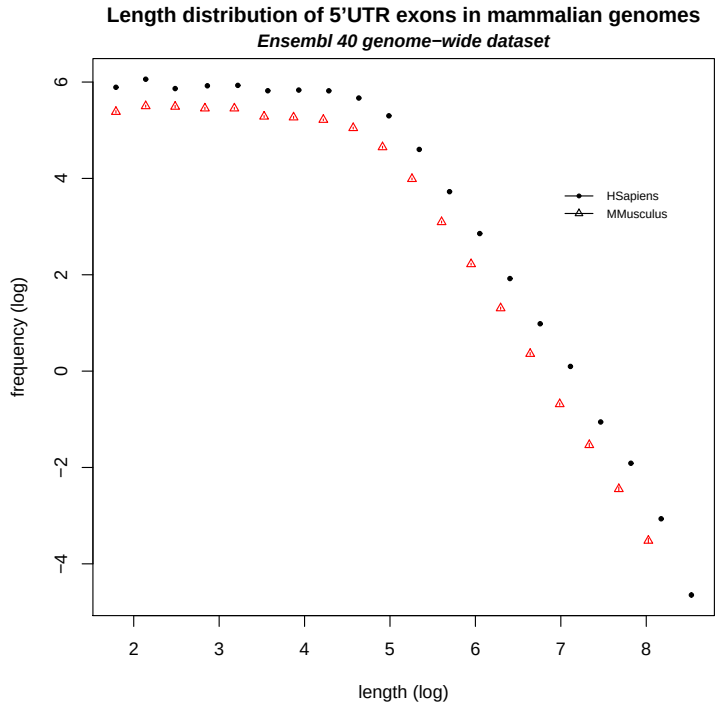
Species	$\widetilde{\chi}^2$	α index	*lmin* (bps)
H.Sapiens	0.52	2.56(2)	150
M.Musculus	0.74	2.61(2)	140

Table 3.1: Extimate of critical index α and length threshold *lmin* for the power law distribution of 5'UTR exons in human and mouse

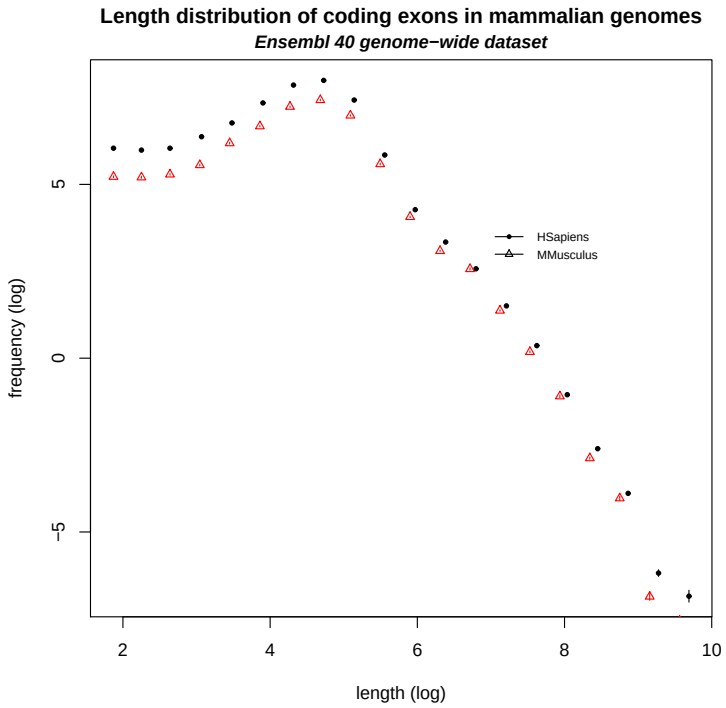
Species	protein coding exons	3'UTR exons	*lmin* (bps)
H.Sapiens	84.37	13.46	150
M.Musculus	153.31	5.91	140

Table 3.2: $\widetilde{\chi}^2$ values for the linear fit of protein coding exons and 3'UTR exons length distribution, in the same range where we are able to fit the power law decay of 5'UTR exons length

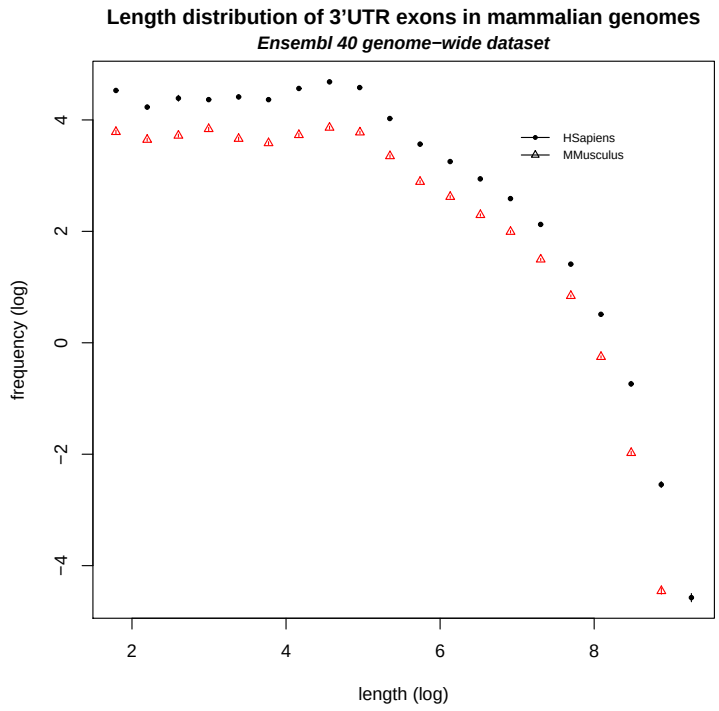
The same plots for other organisms show exactly analogous trend, but they are affected by poor annotation of 5' and 3' UTR, which are very difficult to identify entirely (see Tab.3). In Tab. 3 we reported the total number of annotated protein coding genes, annotated 5'UTR and annotated 3'UTR for 4 different mammalian genomes, according to Ensemble database release 40. These data underline the current lack in the annotatation of 5'UTR and


Length distribution of 5'UTR exons in mammalian genomes
Ensembl 40 genome−wide dataset
frequency (log)
length (log)
HSapiens
MMusculus
6
4
2
0
−2
−4
2
3
4
5
6
7
8

(a)


Length distribution of coding exons in mammalian genomes
Ensembl 40 genome−wide dataset
frequency (log)
length (log)
HSapiens
MMusculus
5
0
−5
2
4
6
8
10

(b)



(c)

Figure 3.1: Exons length distribution in 5'UTR (a), protein coding exons (b) and 3'UTR (c) in human and mouse genome reported in log-log histograms (with bin size growing logaritmically). Plot errors are derived dividing the complete dataset in subsets of comparable dimension, avoiding biological biases, and averaging the length distribution of each subset.

3'UTR for other mammals, besides H. Sapiens and M. Musculus. For this reason, the same analysis performed for H. Sapiens and M. Musculus exon length distribution is prevented for other organisms.

Species	Annotated protein coding genes	Annotated 5'UTR	Annotated 3'UTR
H.Sapiens	23735	18333	18592
M.Musculus	24438	15945	16429
C.Familiaris	18214	5925	6298
G.Gallus	18632	7463	7670

Table 3.3: Annotated protein coding genes, 5'UTR and 3'UTR in Ensembl database release 40

3.4 The model

In order to understand this peculiar behaviour of the 5'UTR exons we propose and discuss a simple model of exon evolution. Our goal is to understand if it is possible to associate the different behaviour that we observe to the greater freedom from selective pressure of the 5'UTR exons with respect to the coding and 3'UTR ones.

Evolutionary models describe evolution of the DNA sequence as a series of stochastic mutations. There are three major classes of mutations: changes in the nucleotide type, insertions or deletions of one or more nucleotides. The various existing models differ with each other for the different assumptions they make on the parameter which control these changes (for a review see for instance [17, 18, 42]). From a biological point of view the two main assumptions of any evolutionary model are:

- evolution can be decribed as a Markov process, i.e. the modifications of a DNA sequence only depend on its current state and not on its previous history.

- evolution is "shaped" by functional constraints: DNA sequences with a negligible functional role evolve at a higher rate with respect of func-

tionally important regions. This implies that regions with different functional roles must be described by different choices of the various mutational rates. The free evolution of sequences without functional constraints is usually called "neutral evolution".

Let us see a few examples:

- protein coding exons are usually strongly constrained since the proteins they code have an important role in the life of the cell, however due to redundance of the genetic code, the third basis of each codon in the coding exons is free to mutate. On the contrary insertions and deletions are suppressed because they can dramatically affect the shape and function of the protein.

- Sequences devoted to transcriptional regulations (which very often lie outside exons) are usually so important for the life of the cell that they are kept almost unchanged over millions of years of evolution

- Regulatory sequences on the messanger RNA (mRNA) whose function often depends on the tridimensional shape of the RNA molecule and not on its exact sequence are in an intermediate situation between the above cases and the neutral evolution: they can tolerate mutations which do not modify their tridimensional shape (typically these are pairs of pointlike changes of bases and are usually called "compensatory mutations"). Most of the mRNA regulatory signals of this type are located in 3'UTR exons.

- 5'UTR regions contain sometimes regulatory sequences of the transcriptional type (which, as mentioned above, are stongly conserved under evolution) but their relative position seem not to have a crucial functional role. They can thus tolerate insertion and deletions as far as they do not affect the regulatory regions.

Since in our model we are only interested in the exon length distribution we may neglect the nucleotide changes and concentrate only on insertions and deletions. From this point of view, according to the above discussion both coding and 3'UTR should behave as highly constrained sequences while

the 5'UTR ones should be more similar to the neutrally evolving ones. With this picture in mind we decided to model the neutral evolution of a DNA sequence under the effect of insertions and deletions only, to see which general behaviour one should expect for the length distribution and then compare it with the data discussed in the previous section.

To this end let us define n_j as the number of 5' UTR exons of length j in the genome and let N be the total number of such exons. Let $x_j \equiv n_j/N$ be the fraction of exons of length j.

If we assume that the exon length distribution evolves as a consequence of insertions and/or deletions of single nucleotides we find the following evolution equation for the $x_j(t)$ (where t labels the time step of this process)

$$x_j(t+1) = x_j(t) + (j-1)\alpha x_{j-1}(t) - j\alpha x_j(t) + (j+1)\beta x_{j+1}(t) - j\beta x_j(t) \quad (3.2)$$

where α and β denote the insertion and deletion probabilities respectively and we have kept into account the fact that for an exon of length j there are exactly j sites in which the new nucleotide can be inserted (i.e. that the insertion and deletion probabilities are linear functions of j, since the implied assumption is that all sites in our sequences are independent of one another).

At equilibrium the exon length distribution must satisfy the following equation (we omit the t dependence which is now irrelevant)

$$(j-1)\alpha x_{j-1} - j\alpha x_j + (j+1)\beta x_{j+1} - j\beta x_j = 0 \quad (3.3)$$

It is easy to see that the only solution compatible with this equation is a power law of this type: $x_j = cj^\eta$ with c a suitable normalization constant. Inserting this proposal in eq.(3.3) one immediatly finds $\eta = -1$.

This result is very robust, it does not depend on the values of α and β and, what is more important, it holds also if instead of assuming the insertion (or deletion) of a single nucleotide, we assume the insertion or deletion of oligos (i.e. small sequences of nucleotides) of length k, with any choice of the probability distribution for the oligos length as fas as k is much smaller than the typical exon length. Moreover one can also show that the power law decay still holds if we add to the process a fixed background probability of creation of new exons of random length as far as this probability is smaller than $x_{j_{max}}(\alpha - \beta)$ where j_{max} is the largest exonic length for which the

power law is still observed. This is rather important since it is known that retrotransposed repeats (in particular of the Alu family) may in some cases (with very low probability) become new active exons and represent one of the major sources of evolutionary changes in the transcriptome.

On the contrary this power law disappears if we assume that there is a finite probability that, as a consequence of the new insertion of deletion, the exon is eliminated. In this case the power law changes into a exponential distribution. This may explain why the power law decay is not observed in the coding and 3'UTR portion of the genes which are under a much stronger selective pressure (in the 3'UTR region are contained lot of post-trascriptional regulatory signals).

Since the critical index that we observe in the actual exon distribution in human and mouse is much larger than 1 it is interesting to see which type of evolutionary mechanism could lead to a $\eta > 1$ behaviour while keeping a power law decay. It is easy to see that this can be achieved assuming that the insertion (or deletion) probability is not linear with the length of the exon but behaves, say, as $p_{insertion} = \alpha j^{\lambda}$ with $\lambda > 1$. Then, following the same derivation discussed above, we find at equilibrium an exon length distribution $x_j = cj^{-\lambda}$.

A possible explanation for such non-linear insertion rate comes from the observation that the transcribed portions of the genome (like the 5' UTR exons in which we are interested), besides the normal mutation processes typical of the intergenic regions, are subject to specific mutation events due to the transcriptional machinery itself (see for instance [38]).

It is clear from the above discussion that in this case the critical index of the exon distribution, strictly speaking, is not any more an universal quantity, but depends on the particular biological process leading to the $p_{insertion} = \alpha j^{\lambda}$ probability discussed above. However it is conceivable that similar mechanisms should be at work in related species. This in our opinion explains why the critical indices associated to the mouse and human distributions are so similar and led us to conjecture that similar values should be found also in other mammalians as more and more 5'UTR sequences will be annotated.

Let us conclude by noticing that this whole derivation is based on the assumption that the system had reached its equilibrium distribution. This is

by no means an obvious assumption and it is well possible that the fact that we observe a critical index larger than 1 simply denotes that the system is still slowly approaching the equilibrium distribution. There are three ways to address this issue. First one should extend the analysis to other organisms (however, as we discussed above, this will require a better annotation of the UTR regions in these organisms). Second one could reconstruct, by suitable aligning procedures, the UTR exons of the common ancestor between mouse and man and see if they also follow a power law distribution and, if this is the case, which is the critical index. Third one could simulate the model discussed above and look to the behaviour of the exon distribution as the equilibrium is approached.

3.5 Derivation of the power law

Inserting the distribution $x_j = cj^\eta$ in eq.(3.3) we find

$$\alpha(j-1)^{\eta+1} - \alpha(j)^{\eta+1} + \beta(j+1)^{\eta+1} - \beta(j)^{\eta+1} = 0 \tag{3.4}$$

which can be expanded in the large j limit as

$$j^{\eta+1}\left[\alpha\left(1-\frac{\eta+1}{j}\right) - \alpha + \beta\left(1-\frac{\eta-1}{j}\right) - \beta\right] = 0 \tag{3.5}$$

which implies:

$$(\beta-\alpha)\frac{\eta+1}{j} = 0 \tag{3.6}$$

which (assuming $\beta \neq \alpha$) implies, as anticipated, $\eta = -1$.

A few observations are in order at this point:

a] It is clear from the derivation that the result is independent from the specific values of α and β as far as they do not coincide. This independence from the details of the model holds also if we assume at each time step a finite, constant (i.e. not proportional to j) probability α' (β') of random insertion (deletion) of a nucleotide. In this case the evolution equation becomes:

$$x_j(t+1) = x_j(t)+(j-1)\alpha x_{j-1}(t)-j\alpha x_j(t)+(j+1)\beta x_{j+1}(t)-j\beta x_j(t)+\alpha'(x_{j-1}(t)-x_j(t))+\beta'(x_{j+1}(t)-x_j(t)) \tag{3.7}$$

which still admits the same asymptotic distribution $x_j = cj^{-1}$

b] If we include a fixed exonization probability p_e to create new exons from, say, duplicated or retrotransposed sequences the evolution equation changes trivially by simply adding such a constant contribution. The solution becomes in this case $x_j = cj^{-1} + d$ where the constant d is related to p_e as follows $d = p_e/(\alpha - \beta)$ and is negligible as far as it is smaller than $x_{j_{max}}$

c] Remarkably enough the above results are still valid even if the inserted (or deleted) sequence is composed by more than one nucleotide. Let us study as an example the situation in which we allow the insertion of oligos of length k with $0 < k < L$ and L smaller than the typical exon length. Let us assume for simplicity to neglect deletions and let us choose the same insertion probability α for all values of k. The evolution equation becomes:

$$x_j(t+1) = x_j(t) + \alpha \left[\sum_{k=1}^{L} x_{j-k}(t)(j-k) - Ljx_j(t) \right] \tag{3.8}$$

which implies

$$jx_j = \frac{1}{L} \sum_{k=1}^{L} (j-k)x_{j-k} \tag{3.9}$$

In the large j limit this equation admits again a power law solution $x_j = cj^{\eta}$. Inserting this solution in eq.(3.8) we find

$$j^{\eta+1}\alpha \left[\frac{1}{L} \sum_{k=1}^{L} \left(1 - \frac{k(\eta+1)}{j} \right) - 1 \right] = 0 \tag{3.10}$$

which is satisfied, as above, if we set $\eta = -1$.

d] On the contrary, if we assume a finite probability $(1 - \gamma)$ of elimination of an exon as a consequnce of the insertion (or deletion) event (as one would expect if the sequence is under strong selective pressure) we find the following evolution equation:

$$x_j(t+1) = x_j(t) + [(j-1)\alpha x_{j-1}(t)\gamma - j\alpha x_j(t)] \tag{3.11}$$

where α is, as above, the insertion probability and we are assuming for simplicity single base insertions. This equation does not admit any

more a power law solution at equilibrium but requires an exponential distribution: $x_j = e^{-\lambda j} j^{\eta}$ with $\eta = -1$ and $\lambda = log(\gamma)$.

e] It is instructive to reobtain the result discussed in [a] above by looking at the equilibrium equation as a recursive equation in j:

$$x_{j+1} = \frac{j}{j+1}\left(1+\frac{\alpha}{\beta}\right)x_j - \frac{\alpha}{\beta}x_{j-1} \qquad (j > j_{min}) \tag{3.12}$$

and

$$x_{j+1} = \frac{j}{j+1}(1+\frac{\alpha}{\beta})x_j \qquad (j = j_{min}) \tag{3.13}$$

and construct recursively the solution for any j starting from $x_{j_{min}} = c/j_{min}$. The recursion can be solved exactly and gives:

$$x_j = x_{j_{min}} \frac{j_{min}}{j} \frac{1-\left(\frac{\alpha}{\beta}\right)^{j-j_{min}+1}}{1-\frac{\alpha}{\beta}} \tag{3.14}$$

which (assuming $\alpha < \beta$) [2] leads asymptotically to the solution $x_j = c/j$ with $c = x_{j_{min}} \frac{j_{min}}{1-\alpha/\beta}$. This result allows to understand exactly the "finite size" corrections with respect to this asymptotic solution which turn out to be proportional to $\left(\frac{\alpha}{\beta}\right)^{j-j_{min}+1}$ and vanish if only deletions (i.e. $\alpha = 0$) or only insertions (i.e. $\beta = 0$) are present. In these cases the asymptotic solution is actually the *exact* equilibrium solution of the stochastic model.

[2] If $\beta < \alpha$ one should study the inverse recursion relation starting from $x_{j_{max}}$.

Chapter 4

DrosOCB: a high resolution map of conserved non coding sequences in Drosophila

In this chapter we introduce comparative genomics methods applied to the non coding DNA of complex eukaryotes, in particular in the Drosophila genome. In particular, this part of the manuscript presents our novel large scale alignment strategy, which aims at drawing a precise map of conserved non coding regions between genomes, even when these regions have undergone small scale rearrangments events and a certain degree of sequence variability.

4.1 Motivation

The functional annotation of eukaryotic DNA sequences represents a great challenge in post–genomic biological research. The identification of functional non–coding elements, such as untranslated regions (UTRs), genes for non–protein-coding RNAs, and cis-regulatory elements, is extremely difficult, as the rules governing their structure and function are far from being well undertood.

A great aid to functional annotation of genome sequences is provided by

comparative genomics methods which, since a few years, have been extended also to non coding DNA regions. The basic assumption of comparative genomic approach is that common features of two organisms are encoded within the DNA that is conserved between the species, due to purifying selection during evolution. According to the same assumption, the DNA sequences controlling the expression of genes that are regulated similarly in two related species should also be selected during evolution.

However, comparison of non coding sequences requires new algorithms and strategies to take into account the different evolutionary mechanisms affecting regulatory sequences. Recent studies examining the evolution of cis–regulatory modules in Drosophila, reveals that regulatory sequences may frequently evolve through compensatory gain and loss events in transcription factors binding sites, that produces little functional change [26], [27]. Great plasticity in the arrangement of binding sites within cis–regulatory modules is another remarkable evolutionary feature revealed to occur in vertebrates [29].

Once complete genomes from different species are available, a global alignment procedure is suitable to find a map of colinear conserved segments between the input sequences, descarding alignments that overlap or cross over. Global alignment methods are widely used to identify highly similar regions in the sequences which appear in the same order and orientation. On the contrary, local alignment algorithms are generally very useful in finding similarity between regions that may be related but are inverted or rearranged with respect to each other.

Recently, the novel notion of glocal alignment, a sophisticated combination of global and local methods, has been introduced [54]. This class of alignment algorithms create a map that transforms one sequence into the other while allowing for rearrangement events. This procedure, at the base of Shuffled-LAGAN algorithm [55], is able to take into account large scale genomic rearrangments, but fails at lower scale.

Here, we present an novel large scale alignment strategy which aims at drawing a precise map of conserved non-coding regions between genomes, even when these regions have undergone small scale rearrangement events.

Our procedure is optimized to take into account the great plasticity of non coding DNA, such as shuffling and sequence variability of binding sites within functional modules, low scale translocations, inversions and duplications. We used a "gene-centric" approch, in that it starts with a list of orthologous genes between two species, and applies a local alignment algorithm to the corresponding flanking intergenic regions and intronic regions of these orthologous pairs. Hence, it is a local alignment strategy but applied systematically on a genome-wide scale and, for this reason, we decided to call it "lobal".

The recent availability of 12 Drosophila species sequences and annotations [57] offers a complete and reliable genomic dataset for developing and testing methods for comparative genomics of non coding DNA. We applied our lobal alignment approach to align Drosophila melanogaster to several other drosophila species (D. yakuba, D. pseudoobscura, D. virilis, ...), for which a reliable genome build and annotation is available.

4.2 A comparative genomics procedure for non–coding DNA

4.2.1 Gene–centric comparative approach

For each Drosophila species examined (listed in Tab.1 and referenced to as D.xxx), we compile a list of genes orthologous to a D.melanogaster (D.mel) gene, according to the "12 drosophila genomes project" data (Tab.1 and Material and Methods). For each pair of D.mel/D.xxx orthologous genes, we extract in both species the upstream, downstream and intronic regions. Upstream and downstream regions are extracted up to the next neighboring gene (see Material and Methods for more details), taking the longest transcript as a reference in case of multiple transcripts. All sequences have been previously masked for repeats using the RepeatMasker program [71]. At this stage, the comparison procedure crucially depends on the availability of genomic annotations (i.e. gene coordinates and orthology relationships). The orthologous regions are then aligned using a local alignment procedure

described later. For the alignment, the orthologous regions are oriented such that the corresponding genes are in the same orientation. Using this gene-centric approach, most intergenic regions are considered twice. For example, the region chr4:64404-68333 in D.melanogaster is first considered as the upstream region of the *PlexB* gene, and then as the downstream region of the *ci* gene. This redundancy is taken care of in the post-processing step, described later.

4.2.2 Alignment procedure

For each pair of orthologous D.mel/D.xxx genes, we respectively align their upstream regions, downstream regions and introns. This is done by orienting the transcripts in the same direction, such as to distinguish same from opposite strand. Local pairwise alignments between orthologous sequences was performed using CHAOS [54], which is an heuristic alignment algorithm with some peculiar features optimized for large non coding DNA sequences. CHAOS works by chaining small words (called seeds) that match between the two input sequences. Unlike BLAST, it is a double seed technique which allows some degeneracy in seeds. It chains toghether seeds that are closer than a maximum distance d and it returns the highest scoring chains, according to a standard NeedlemanWunsch metric. These highest scoring chains constitue the conserved noncoding blocks (CNBs). Because it is a local alignment, it is able to identify nonsyntenic CNBs order with a very high resolution. Moreover, because the alignment is performed on both strands, we also identify CNBs resulting from inversion events. Also, it is able to rapidly align large sequences with a better specificity than purely local aligners, thanks to the double seed technique. We choose a quite sensitive set of parameters in CHAOS (see Material and Methods). An assessment of statistical significance of alignment scores is introduced to discriminate true from random alignments. The scoring cutoff is calculated by aligning randomly selected non-orthologous sequences, and setting a false discovery rate (FDR) of $2{\cdot}10^{-3}$
.

4.2.3 Processing of Drosophila sequences

Sequences and annotations have been downloaded from the AAA site [57] as fasta and GFF3 files respectively. The Drosophila melanogaster sequences and annotations correspond to version 4.3. For the other drosophila, the sequences correspond to the CAF1 assemblies and are now available from GenBank. The annotations result from a reconciliation procedure of various annotations, whereas the homology maps are built using a fuzzy reciprocal blast. For details, see [57].

We rely on the gene annotations to extract the D.melanogaster introns. When several transcripts exist for a single gene, we consider the longest transcript and its introns. For other drosophila species, no annotations exist for intron/exon structure. Hence, we extract the locus corresponding to the full gene, and align it using CHAOS to each intron of the orthologous gene. For intergenic regions, we applied a conservative definition. We define the upstream region as the longest consecutive sequence of nonexonic, nonintronic nucleotides on the 5' end of the longest transcript, and similarly for the downstream region. While this is an intuitive definition in general, it has particular implications in the case of nested genes. For a gene A nested inside the intron of a gene B, the intergenic regions associated to gene A will start at the 5'/3' extremities of gene B, in order to respect the previous definition.

We use CHAOS with the following set of parameters: chaos -wl 7 -co 12 -b -v -rsc 1500. The last parameter is a very loose lower threshold on the alignment score, but we apply more stringent thresholds in the postprocessing step.

4.2.4 Post–processing and availability

As mentioned previously, sequences are often considered and aligned twice, resulting in redundant CNBs. We eliminate this redundancy by scanning the output of the alignments, and merging overlapping CNBs. More precisely, we merge two CNBs if they meet all of the following requirements: (i) they

overlap in D.melanogaster, (ii) they overlap in the other species, (iii) both blocks are in the same orientation in D.melanogaster, and in the same orientation in the other Drosophila species. Conditions (i) and/or (ii) are for example not fulfilled in the case of duplications; in this case, the CNBs are not merged and appear as distinct blocks. Each block is assigned a unique identifier and is labelled with its score, percentage identity, as well as with the name of the gene(s) in the surrounding of which it is located. For the reasons mentioned previously, a block often refers to its two flanking genes.

The full collection of CNBs for all eleven pairwise comparisons is available as a queryable database, named DrosOCB (for Drosophila Conserved Blocks). It can be accessed through a userinterface which allows to query a particular gene or a genomic region. Our database is linked with the UCSC genome browser [72], such that CNBs can be displayed in their genomic context with the browser.

4.3 DrosOCB database content

In Table 1, the content of the database is summarized for each species compared with D.melanogaster. Their phylogenetic relationship is shown in Fig. 4.1. The cumulated size of the D.melanogaster sequences (intronic and intergenic) which are aligned varies in the range between 78.6 Mbp and 87.7 Mbp, depending on the total number of orthologous genes between D.melanogaster and the other species. Considering that the D.melanogaster genome size is around 120 Mbp, this means that we aligned between 65% and 73% of the D.melanogaster genome. Analyzing the catalog of CNBs, we can make some observations about the conservation features of Drosophila genus at large scale. The estimated percentage of non coding sequences evolutionary constrained in Drosophila genome is reported in Tab. 4.1 and displayed in Fig. 4.2. As expected, the percentage of conservation follows the evolutionary distance. It varies between 16% (13%) for D.melanogaster/D.virilis intergenic (intronic) sequences, the most evolutionary distant species in the phylogenetic tree, and 68% (54%) for D.melanogaster/D.sechellia. These estimations are lower than the ones obtained for D.melanogaster compared

with Drosophila D.virilis, D.pseudoobscura and D.yakuba from previous work [911]. However, we applied a rather conservative threshold on the scores of the CNBs, such as to reduce the number of spurious alignments. These conservation percentages are always higher in intergenic regions as compared to intronic regions. However, these figures should be taken with some care, as some regions, labelled as "intergenic" in some drosophila species (and thus not aligned as introns) might well turn out to be intronic, as distant exons will become better annotated. In fact, whereas the mean size of genes in D.melanogaster is 6.1 kb, it ranges from 2.9 kb (D.sechellia) to 4.1 kb (D.virilis) for the other species, indicating that some gene annotations might still miss distant exons.

Interestingly, these proportions are roughly constant inside the melanogaster subgroup (around 50%), indicating that the difference in the evolutionary distance between D.melanogaster and D.simulans/D.sechellia on one hand (about 5 My), and D.yakuba/D.erecta (about 10 My) on the other hand is too small to affect the conservation of noncoding DNA. There is a important decrease outside this branch (roughly 25% for D.ananassae, the closest species outside the melanogaster subgroup). The percentages for species outside the Sophophora subgenus (D.virilis, D.mojavensis and D.grimshawi) are again very comparable (about 15%). The mean size of CNBs obtained in our output is comprized between 50 bp and 94 bp, and increases with decreasing evolutionary distance, as expected (cf. column 6 in Tab.1). It is shorter for intronic regions than for intergenic regions. The lower threshold of 50 bp indicates that, although the alignment procedure is sensitive in allowing some degeneracy in the compared sequences, it preserves a certain degree of selectivity, discarding very short isolated CNBs with a score below the cutoff threshold.

4.4 Peculiarity of Drosophila non–coding DNA evolution

Due to the fact that we use a sensitive local alignment procedure, we are able to spot small scale genomic rearrangements that are not visible in standard alignments (see Fig. 4.3). As an illustration, we will focus on a particular feature, namely inverted CNBs. By this, we mean CNBs that lie on opposite strands in the orthologous regions, a situation which might result from local genomic inversions in one of the two species. Since the drosophila genomes are known to have an extreme plasticity at large/medium scales [68], it is interesting to verify whether this is also true in or below the kb range. Fig. 4.4 plots the percentage of CNBs that are inverted, for all eleven pairwise alignments, for intronic and intergenic regions. Depending on the evolutionary distance, this percentage ranges from 15% to almost 30%. Interestingly, these percentages are very comparable for intergenic and intronic regions, indicating that the evolutionary dynamics is similar for these regions [25]. In the custom track provided for UCSC genome browser, we use a particular color coding to distinguish between CNBs on the same strand (grey boxes) and the inverted CNBs (red boxes). Figure 4.5 shows an interesting example of such an event, in one of the introns of the white gene on chromosome X. The central region is highly conserved in all pairwise alignments, but the CNBs are inverted in the pairwise alignments with D.mojavensis, D.virilis and D.grimshawi. This is not due to an inversion of the full gene locus, since the transcripts are all taken in the same orientation when performing the alignment. The inverted CNBs have a very high score and a conservation of 90% over 66 bp, excluding that they might be spurious alignments. Hence, one can speculate that there exists a local inversion inside this intron which appeared in the common branch of these three Drosophila species. Note that we applied a high threshold on the score of the DrosOCB blocks (2200, $FDR = 0.8 \cdot 10^{-3}$) such as to reduce the number of displayed blocks. Interestingly, the ORegAnno track [69] in the lower part of the UCSC window indicates the presence of a regulatory element, more precisely an enhancer, underlying that even functional elements are subject to extensive rearrangements, as previously noted [26, 27, 29].

	Genome size* (Mbps)	**Number orthologous genes**	**Size of aligned sequences (Mbps)**		**% of conserved sequences**		**mean size of CNBs (bps)**	
			intronic	intergenic	intronic	intergenic	intronic	intergenic
D.sim	139,8	11540	36,9	55,4	50%	61%	103	118
D.sec	168,9	12074	39,2	56,7	54%	68%	102	115
D.yak	168,0	13005	45,7	62,5	52%	64%	84	94
D.ere	154,9	12459	44,1	60,7	51%	65%	88	95
D.ana	234,3	11530	47,7	61,7	25%	33%	57	59
D.pse	154,9	11795	66,9	44,1	19%	24%	52	54
D.per	191,0	10657	39,4	58,3	17%	21%	53	54
D.wil	238,9	10907	52,5	79,9	15%	18%	49	49
D.moj	196,6	11213	51,1	76,6	13%	15%	49	49
D.vir	209,0	11346	51,6	77,2	13%	16%	49	49
D.gri	203,3	11541	48,9	71,7	13%	16%	49	49

Table 4.1: DrosOCB database content summary From the left, columns indicate the Drosophila species, the size of the genome (defined as the total size of the genome fasta files downloaded), the number of orthologous genes between D.melanogaster and the second species, the size of intronic and intergenic sequences aligned, the percentage of conserved sequences (i.e. the total nonredundant size of intronic and intergenic CNBs divided by the size of intronic/intergenic sequences aligned), the mean size of intronic and intergenic CNBs.

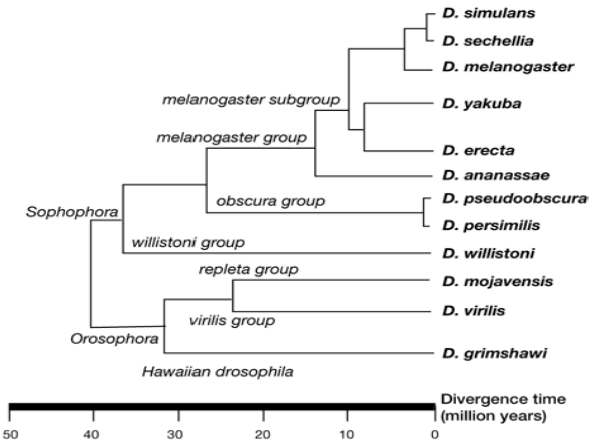


Figure 4.1: The evolutionary tree of Drosophila genus, according to the "12 drosophila genomes project". This is taken from the AAA web site [6].

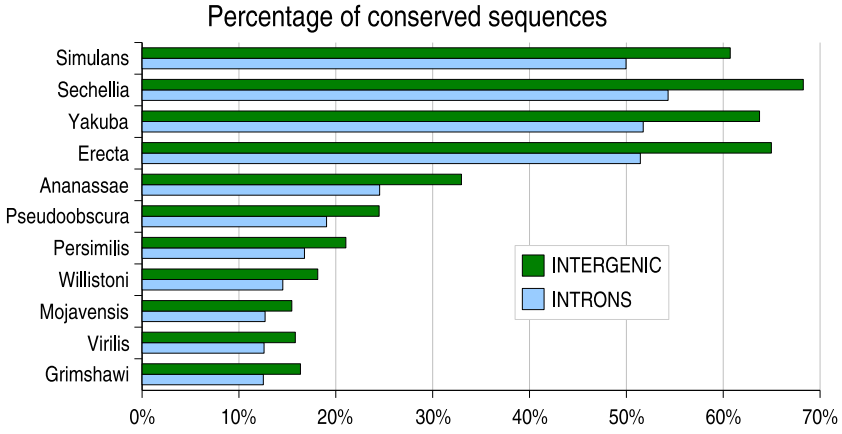


Figure 4.2: Percentage of conserved sequences in intergenic and intronic regions for each of the 11 species compared with D.melanogaster The percentages are determined taking the total length of the intergenic/intronic CNBs (redundant CNB portions are counted only once), and dividing by the total nonredundant length of the aligned intergenic/intronic sequences. This correponds to columns 5 and 6 of Table 1.

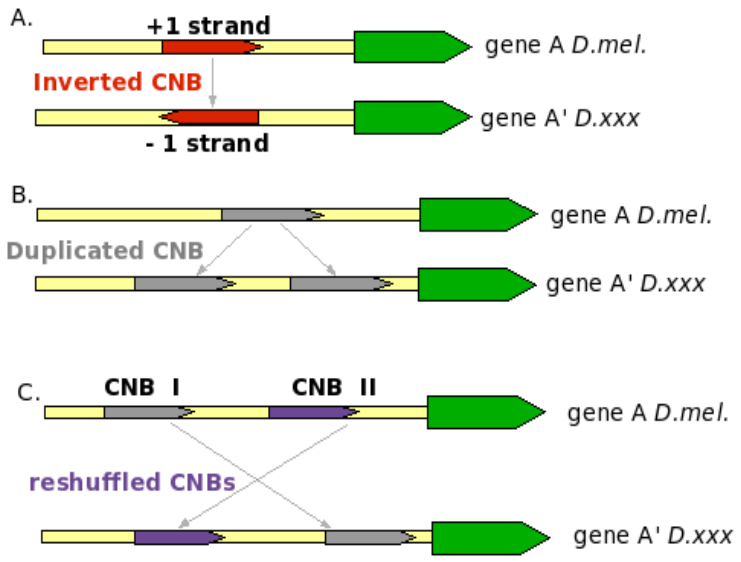


Figure 4.3: Rearrangements events of CNBs The picture represents typical genomic rearrangements that can be observed in DrosOCB. In the case A, the CNB (red box in the picture) is a conserved sequence that has changed is orientation in one of the two species, taking as common reference the orientation of the transcript. Case B represents a CNB which is duplicated in the second species compared with D.melanogaster. In this case, the same D.melanogaster sequence matches two different regions in the compared species, and appears as two overlapping blocks. Case C shows a third case of genomic rearrangement that can be detected in the DrosOCB content. We called "reshuffled CNBs" two sequences that are conserved in non collinear order in the two species (purple box in the picture). Note that we have not depicted here the most frequent configuration of noninverted, collinear CNBs.

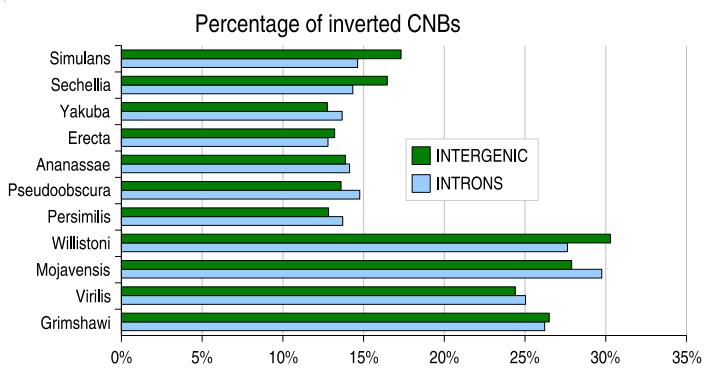


Figure 4.4: Percentage of inverted CNBs in intergenic and intronic sequences in DrosOCB. These percentages are computed by taking the ratio of the number of inverted CNBs divided by the total number of blocks, in intergenic and intronic regions respectively. Note that the length of the blocks is not taken into account here.

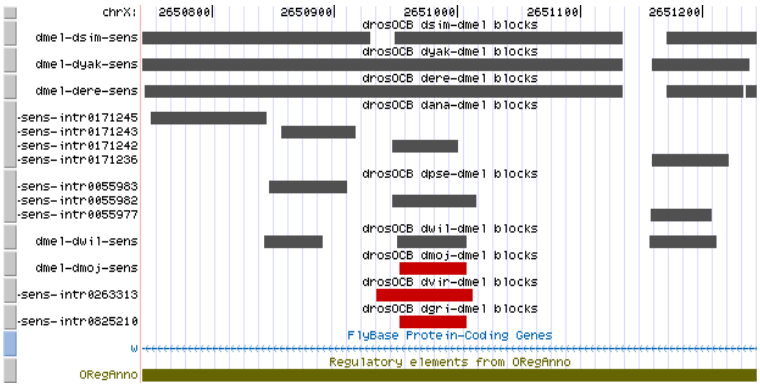


Figure 4.5: Example of inverted CNBs UCSC genome browser window with our custom tracks, showing an example of specific lineage inversion event in Melanogaster region X:5,494,877-5,495,622. Grey colour coded boxes represents DrosOCB CNBs conserved across all species in the same orientation respect to the Melanogaster locus. The red boxes (tracks dmoj-dmel, dvir-dmel, dgri-dmel) represent inverted CNBs in all species with respect to Melanogaster, highlithing an inversion event specific of the branch common to D.mojavensis, D.virilis and D.grimshawi (see also Figure 1).

Chapter 5

Conclusion

In the previous chapters, we presented two proposed applications of statistical and computational methods for investigating eukaryotic non–coding DNA features. We considered the whole available genome–wide ensemble of non coding sequences, treated as a complex system, to give some insights into general rules behind the observed biological experimental data. The hope is to get in this way new insight in the evolutionary mechanisms shaping higher eukaryotes genomes and to understand functional role of the various portions of the genome.

The first part focuses on the length distribution of a particular class of DNA sequences, namely the 5'UTR exons. We observe that both in mouse and in human these exons show a very clean power law decay in their length distribution and suggest a simple evolutionary model which may explain this finding. A simple stochastic model based on base pairs insertion and deletion events is able to reproduce the observed experimental power law decay. These results are obtained analytically and the implications for genome evolution are discussed.

The obtain solution is very robust and it does not depend on the values of insertion and deletion probabilities. Moreover, it holds also if instead of assuming the insertion (or deletion) of a single nucleotide, we assume the insertion or deletion of oligos (i.e. small sequences of nucleotides).

We show that the power law decay still holds if we add to the process a fixed background probability of creation of new exons of random length, with a superior limit in the this exonization probability. Finally we derived a possible modification of our model, leading a critical index greater than 1, in agreement with that observed in the actual exon distribution in human and mouse.

It is conceivable that similar evolutionary mechanisms should be at work in related species. This in our opinion explains why the critical indices associated to the mouse and human distributions are so similar and led us to conjecture that similar values should be found also in other mammalians as more and more 5'UTR sequences will be annotated.

Let us conclude by noticing that this whole derivation is based on the assumption that the system had reached its equilibrium distribution. This is by no means an obvious assumption and it is well possible that the fact that we observe a critical index larger than 1 simply denotes that the system is still slowly approaching the equilibrium distribution. There are three ways to address this issue. First one should extend the analysis to other organisms (however, this will require a better annotation of the UTR regions in these organisms). Second one could reconstruct, by suitable aligning procedures, the UTR exons of the common ancestor between mouse and man and see if they also follow a power law distribution and, if this is the case, which is the critical index. Third one could simulate the model discussed above and look to the behaviour of the exon distribution as the equilibrium is approached.

The second part of this thesis is devoted to discuss bioinformatics methods for comparative genomics of non–coding DNA. Once a DNA evolutionary model is defined, it is possible to compare sequences looking for evidence that they have diverged from a common ancestor. This is usually done by first aligning the sequences (or part of them) and then deciding whether that alignment is more likely to be occurred because the sequences are related, or just by chance.

Comparison of non coding sequences requires new algorithms and strategies to take into account the different evolutionary mechanisms affecting regulatory sequences. We present an novel large scale alignment strategy which aims at drawing a precise map of conserved non-coding regions between genomes, even when these regions have undergone small scale rearrangement events. Our procedure is optimized to take into account the great plasticity of non coding DNA, such as shuffling and sequence variability of binding sites within functional modules, low scale translocations, inversions and duplications.

The recent availability of 12 Drosophila species sequences and annotations [57] offers a complete and reliable genomic dataset for developing and testing methods for comparative genomics of non coding DNA in complex eukaryotes.

We have described a new, local but genomewide alignment procedure for binary comparisons of Drosophila melanogaster with eleven currently available drosophila genomes. We have shown that the resulting collection of CNBs, organized in the DrosOCB database, constitutes a highresolution collection of noncoding DNA conservation in drosophila. The small size of the blocks, and the local nature of the alignment highlights small scale genomic rearragement events, that are not apparent from other approaches.

As a preliminary study, we have focused on inverted CNBs which might correspond to small scale inversions. A more detailed analysis of this phenomenon and other rearragements is left for a future more complete investigation. An interesting aspect of our preliminary analysis is that the localization of these inverted blocks is not evenly distributed among chromosomes. Chromosome X seems to have a much higher than average proportion of these inverted blocks, indicating that it has undergone more extensive rearragements than the autosomal chromosomes, as noted previously [70].

The alignment procedure described in this work provides an optimal tool for a high resolution comparison of non coding DNA sequences. The content

of the database, and the observed high rate of low scale reshuffling suggest that this database of CNBs can constitute the starting point for several investigations, related to the evolution of regulatory DNA in Drosophila, the in silico identification of unannotated functional elements and the search for transcription factor binding sites.

Statistical and computational approaches presented in this thesis work aim at analyzing the unknown fraction of the genome, the rules governing its structure, function and evolution. The objective is considering the whole available genome–wide ensemble of non coding sequences, treated as a complex system, to give some insights into general rules behind the observed biological experimental data. In this direction, statistical and computational methods can provide a great aid to extract information about general features of the non coding genome.

Appendix A

Biological glossary

In this chapter, we report some useful biological definitions, mainly from [81].

chromatin DNA complexed with histones and other chromosomal proteins.

ChIP *chromatin immunoprecipitation* is a method for extract DNA that specifically interact with proteins of interest. Bound proteins are chemically linked to DNA and then selectively precipitated by using cognate antibodies.

coding sequence a DNA sequence that encode for a gene product such as a protein.

coding strand the strand of the DNA duplex that is really translated into a protein through the genetic code.

codon a trinucleotide (triplet, or 3-word) that specifies for a particular aminoacid through the genetic code.

differential expression the expression of one or more genes to different extents, depending upon growth conditions, treatments applied, or the state of the cell cycle.

euchromatin the portion of a chromosome that is less condensed and more transcriptionally active.

eukaryote organisms characterized by a true membrane-bounded nucleus containing chromosomes complexed with histones, a cytoskeleton, and mebrane-bound organelle such as mithocondria. Humans and yeasts are eukaryotes.

exon a contigous segment of DNA that is represented in a processed mature RNA molecule after splicing has removed intronic sequences. Exon sequences may be translated or untranslated.

gene a genomic locus (or DNA segment) specifying or contributing to an heritable trait associated with an organism. A gene usually encode for a RNA species and (after translation) to polypeptide chains, the proteins.

genome the entire genetic complement of an organism.

heterochromatin the portion of chromatin that is highly condensed and less commonly transcribed throughout the cell life.

homologs related genes or loci whose similarity is a consequence of descent from a shared common ancestor. Homologs in different species are *orthologs*, and homologs within a species are *paralogs*.

intron a segment of non-coding DNA separating exons within genes. Introns are removed by splicing from precursor RNA molecule to form mature mRNA.

locus a position on a genetic map or genome defined by a certain gene or DNA sequence appearing at that position.

mutation a heritable change in DNA relative to a defined "wild-type" reference sequence.

non coding sequence DNA sequence that does not appear in the final gene product. This include intergenic sequences, introns, untraslated regions.

open reading frame *ORF* a succession o triplet not interrupted by STOP codons.

ortholog homologs appearing in different species.

paralog homolog arising from a gene duplication within a single lineage or species instead of arising by descent in diverging lineages.

microRNA or *miRNA* a particular class of short $\sim$ 22 RNA molecules displaying post-transcriptional regulatory features trough binding to specific loci in the 3' UTR region of the regulated genes.

prokaryote an organism that does not contain a true nucleus, membrane-bound organelle, or complex cytoskeleton.

promoter a DNA sequence element required for initiation of transcription of a gene, including sites where transcription factors bind to control the time and cell type in which transcription occurs.

proteome the complete set of protein encoded by a certain organism.

repeated sequence a DNA sequence that appears more than once in a genome.

splicing processing of primary RNA transcript to remove introns and to produce mature mRNA molecules containing a continuous coding sequence composed by joined exons.

spottet microarray a collection of DNA probes, used for measuring gene expression levels.

transcription factors proteins involved in the starting of transcription for a certain gene trough binding to specific loci in the promoters regions (transcription factor binding sites, TFBSs) of the regulated gene.

transcriptome the complete collection of mature transcripts in a particular cell type under a specified set of physiological and environmental conditions.

UTR *UnTRaslated region* DNA segment transcribed into mRNA but not translated into amminoacids.

Appendix B

Bioinformatic glossary

In this chapter, we report some useful bioinformatics definitions, mainly from [81].

alignment the procedure by which two (or more) nucleic (or protein) sequences are arranged to establish a relationship between them.

assembly given numerous short sequences of DNA, the assembly is the procedure to merge them into a larger one or into a genome.

binomial distribution the binomial is the probability distribution describing the number of successes and failures in a fixed number of independent trials when only two outcomes are possible, often called "success" and "failure". The number of heads in some fixed number of tosses of a coin is an example of a binomial random variable. If we denote with Y the total number of success in n trials, so the probability distribution of Y is given by the formula

$$P_Y(y) = \binom{n}{y} p^y (1-p)^{n-y} \quad y = 0, 1, 2, ..., n.$$

where:

$$\binom{n}{y} = \frac{n!}{(n-y)!y!}$$

BLAST *Basic Local Alignment and Search Tool* a program for rapidly searching protein or DNA sequences in a database and to detect statistically significant local alignments throught heuristic procedures.

Bonferroni correction the bonferroni is a multiple testing correction method. If N multiple independent hypotheses are tested for significance, the correct $cutoff$ should be written as $\simeq \frac{cutoff}{N}$.

clustering the procedure of grouping together objects based upon some kind of similarities or distance measure between them.

consensus sequence a short DNA or protein sequence having, at each position, the most probable letter at that position.

conserved sequence a DNA or protein region found nearly identical in two or more genomes, after alignment.

deletion an alteration in DNA sequences resulting from removing one or more contiguous bases from the sequence string.

dissimilarity a measure of the degree of difference between objects with respect to a certain distance measure.

false discovery rate or *FDR*, in classification, the fraction of those features identified as positive that are in fact false positives.

FASTA a rapid local alignment method based upon locations of k-words in a alignment matrix. This allow a more detailed examination of regions where hits are frequent.

gene expression matrix for n genes whose expression is measured for m conditions, the $n \times m$ matrix of expression levels (typically ratios of treatment to control conditions) is the gene expression matrix.

global alignment alignment between two sequences such that all letters of both sequences are aligned opposite letters or indels.

hit a database entry matching a query sequence after a database search.

hypergeometric distribution suppose that an urn contains N objects, of which n are red and $N - n$ are white. Of these, m objects are taken out of the urn at random, in particular without reference to the color and without replacement. The number of red objects taken out is a random variable Y, with probability distribution given by the formula

$$P_Y(y) = \frac{\binom{n}{y}\binom{N-n}{m-y}}{\binom{N}{m}} \quad y = A, A+1, ..., B.$$

where $A = max(0, n + m - N), B = min(n, m)$.

indel an insertion or deletion of letters applied to either of two sequences string being aligned.

IUPAC-IUB symbols symbols for DNA combination bases:

A = Adenine	**R = A or G (purine)**	**M = A or C**
C = Cytosine	**Y = T or C (pyrimidine)**	**B = T,G or C**
G = Guanine	**S = G or C**	**V = A,G or C**
T = Thymine	**W = A or T**	**H = A,T or C**
U = Uracil	**K = G or T**	**F = A,T or G**
	N = any base	

information of a sequence: a measure of it nonrandomness. Can be measure *i.e.* using relative entropy or Shannon's entropy.

insertion the addition of one or more nucleotides into a nucleic acid sequence.

local alignment alignment of substrings taken from each of two different sequence strings.

Markov chain a probabilistic model for a sequence of dependent random variables. The probability distribution of the next outcome depends only on the identity of the k previous outcomes. The case $k = 1$ is called one-step Markov chain.

mismatch non-identity between two letters, each derived from o ne of two sequences strings being aligned or compared.

motif a short local sequence pattern found among a set of proteins or DNA sequences.

multiple alignment alignment of more than two sequences strings.

multiple hypothesis testing the simultaneous testing of two or more alternative hypotheses.

pairwise alignment alignment of two sequence strings.

PAM or *Point Accepted Mutations* a set of matrices for scoring amino acid or DNA substitutions in alignments.

phylogenetic footprinting a DNA sequence pattern recognized to appear similar in aligned regions of related genomes.

position specific scoring matrix or *PSSM* a matrix whose rows correspond to letters that occur at positions in a DNA signal and whose columns corresponds to the positions. Matrix element are the log-odds scores for each letter at each position, computed relative to an appropriate null model. A PSSM is a particular type of a PWM.

positional weight matrix or *PWM* a matrix whose rows correspond to letters that occur at positions in a DNA signal and whose columns corresponds to the positions. Elements of this matrix are related to the probability of occurrence of each letter at each position.

substitution matrix a matrix specifying scores to be applied for matching DNA sequences in an alignment. An example is the PAM matrix.

Appendix C

Publications

Publications directly related to the thesis work:

- [82] Martignetti L., Caselle M.
  Universal power law behaviors in genomic sequences and evolutionary models
  Phys Rev E Stat Nonlin Soft Matter Phys. Aug:76 2007

- [83] Martignetti L., Caselle M., Jacq B., Herrmann C.
  DrosOCB: a high resolution map of conserved non coding sequences in Drosophila
  arXiv:0710.1570 2007

Bibliography

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. Watson.
Molecular biology of the cell.
ed. Garland Science (2002).

[2] Benjamin Lewin.
Genes VIII.
ed. Pearson Prentice Hall (2004).

[3] Brown TA.
Genomes 2.
ed. Garland Science (2002).

[4] Kadonaga JT.
Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors.
Cell. Jan 23;116(2):247-57 (2004).

[5] Arnosti DN, Kulkarni MM
Transcriptional Enhancers: Intelligent Enhanceosomes or Flexible Billboards?
Journal Cell Biochem 94:890-898 (2005)

[6] Blackwood EM, Kadonaga JT.
Going the distance: a current view of enhancer action.
Science. Jul 3;281(5373):60-3 (1998).

[7] Valenzuela L, Kamakaka RT *Chromatin Insulators* Anns Rev Genet 40:107-138 (2006)

[8] G. Pesole, S. Liuni,G. Grillo, F. Licciulli,F. Mignone, C. Gissi, C. Saccone
UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002
Nucleic Acids Res. 2002 (30)

[9] M. Iacono, F. Mignone, G. Pesole
uAUG and uORFs in human and rodent 5'untranslated mRNAs
Gene 2005 (349)

[10] A. Churbanov,I.B. Rogozin,V.N. Babenko,H. Ali,E.V. Koonin
Evolutionary conservation suggests a regulatory function of AUG triplets in 5'–UTRs of eukaryotic genes
Nucleic Acids Res. 2005 (33)

[11] He L. , Hannon GJ.
MicroRNA; small RNAs with a big role in gene regulation.
Nature Review Genetics 5, 522 - 531 (2004).

[12] D.P. Bartel
MicroRNAs: genomics, biogenesis, mechanism, and function
Cell 2004 (116)

[13] L. He,G.J. Hannon
MicroRNAs:small RNAs with a big role in gene regulation
Nat Rev Genet. 2004 (5)

[14] N. Rajewsky
microRNA target predictions in animals
Nat Genet. 2006 (38)

[15] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor

J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler

G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; International Human Genome Sequencing Consortium.
Initial sequencing and analysis of the human genome.
Nature. Feb 15;409(6822):860-921 (2001).

[16] Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T et al.
Ensembl 2006.
Nucleic Acids Res. 34 D556-561 (2006).

[17] R. Durbin, S. Eddy, A. Krogh, G. Mitchison
Biological Sequence Analysis: probabilistic models of DNA and protein sequences
Cambridge University Press 1998

[18] C. Kosiol, L. Bofkin, S. Whelan
Phylogenetics by likelihood: evolutionary modeling as a tool for understanding the genome
Journal of Biomedical Informatics 2006 (39)

[19] "Statistical methods in bioinformatics"
W.J. Ewens, G.R. Grant
Springer

[20] Hardison RC.
Comparative Genomics.
PLoS Biology Nov;1(2):E58. Epub 2003 Nov 17. Review (2003).

[21] Delsuc F, Brinkmann H, Philippe H.
Phylogenomics and the reconstruction of the tree of life.
Nat Rev Genet. May;6(5):361-75. Review (2005).

[22] Y. Liu, X. S. Liu, L. Wei, R. B. Altman, and S. Batzoglou
Eukaryotic regulatory element conservation analysis and identification

using comparative genomics
Genome Res, 14(3):451–458, (2004).

[23] W. Makalowski, J. Zhang, and M. S. Boguski.
Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences.
Genome Res, 6(9):846–57, Sep 1996

[24] Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM.
Phylogenetic shadowing of primate sequences to find functional regions of the human genome.
Science. Feb 28;299(5611):1391-4 (2003).

[25] Bergman,CM and Kreitman, M
Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences
Genome Research, 11(8):1335-45 (2001)

[26] Ludwig, M Z and Palsson, A and Alekseeva, E and Bergman, C M and Nathan, J M and Kreitman, M
Functional evolution of a cis–regulatory module
Plos Biology, 3(4):e(93) (2005)

[27] Moses, A M and Pollard, D A and Nix, D A and Iyer, V N and Li, X Y and Biggin, M D and Eisen M B
Large-scale turnover of functional transcription factor binding sites in Drosophila.
Plos Computational Biology 2(10) e130 (2006)

[28] Li L, Zhu Q, He X, Sinha S, Halfon MS
Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses.
Genome Biol. 5;8(6) (2007)

[29] Sanges, R and Kalmar, E and Claudiani, P and D'Amato, M and Muller, F and Stupka, E

Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage
Genome Biol,7(7) R56 (2006)

[30] Haubold B, Wiehe T.
Comparative genomics: methods and applications.
Naturwissenschaften. Sep;91(9):405-21. Epub 2004 Jun 25. Review (2004).

[31] P.W. Messer, P.F. Arndt, M. Lassig
Solvable sequence evolution models and genomic correlations
Phys Rev Lett. 2005 (94)

[32] W. Li, *The Study of Correlation Structures of DNA Sequences: A Critical Review*
Comp. & Chem., 21, 257-272 (1997)

[33] Zu-Guo Yu, V. V. Anh, and Bin Wang
Correlation property of length sequences based on global structure of the complete genome
Phys Rev E,63 11903 (2000)

[34] Zu-Guo Yu, V. V. Anh , Ka-Sing Lau *Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome*
Physica A 2001 (301) 351–361

[35] A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy
Characterizing long-range correlations in DNA sequences from wavelet analysis
Phys Rev Lett. 1995 (74)

[36] M.L. Lynch, D.G. Scofield, X. Hong
"The evolution of transcription initiation site",
Mol Bio Evo 2005 (22)

[37] X. Hong, D.G. Scofield, M.L. Lynch
"Intron size, abundance and distribution within untranslated regions of

genes"
Mol Bio Evo 2006 (23)

[38] Touchon M, et al.
"Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes."
Nucleic Acids Res. 2004 (32) 4969-78

[39] M. Touchon,S. Nicolay,B. Audit,E.B. Brodie of Brodie,Y. d'Aubenton-Carafa, A. Arneodo, C. Thermes
"Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins"
Proc Natl Acad Sci U S A. 2005 (102)

[40] A. Churbanov,I.B. Rogozin,V.N. Babenko,H. Ali,E.V. Koonin
"Evolutionary conservation suggests a regulatory function of AUG triplets in 5'–UTRs of eukaryotic genes"
Nucleic Acids Res. 2005 (33)

[41] R. Sorek, R. Shamir,G. Ast
"How prevalent is functional alternative splicing in the human genome"
Trends Genet. 2004 (20)

[42] G. Mitchison "A probabilistic treatment of phylogeny and sequence alignment." J. Mol. Evol. 1999 (49)

[43] D. Holste, I. Grosse, S. Beirer, P. Schieg, H. Herzel
"Repeats and correlations in human DNA sequences",
Phys Rev E 2003 (67)

[44] Needleman SB, Wunsch CD.
A general method applicable to the search for similarities in the amino acid sequence of two proteins.
J Mol Biol 48:443-453 (1970).

[45] Smith TF, Waterman MS.
Identification of common molecular subsequences.
J Mol Biol 147:195-197 (1981).

[46] Pearson WR, Lipman DJ.
Improved tools for biological sequence comparison.
Proc Natl Acad Sci U S A. Apr;85(8):2444-8 (1988).

[47] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.
Basic local alignment search tool.
J Mol Biol. Oct 5;215(3):403-10 (1990).

[48] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
Nucleic Acids Res. Sep 1;25(17):3389-402. Review (1997).

[49] Delcher AL, Phillippy A, Carlton J, Salzberg SL.
Fast algorithms for large-scale genome alignment and comparison.
Nucleic Acids Res. Jun 1;30(11):2478-83 (2002).

[50] Bray N, Dubchak I, Pachter L.
AVID: A global alignment program.
Genome Res. Jan;13(1):97-102 (2003).

[51] Brudno M, Do CB, Cooper GM, Kim MF, Davydov E; NISC Comparative Sequencing Program; Green ED, Sidow A, Batzoglou S.
LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.
Genome Res. Apr;13(4):721-31. Epub 2003 Mar 12 (2003).

[52] Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W.
Human-mouse alignments with BLASTZ.
Genome Res. Jan;13(1):103-7 (2003).

[53] Waterman MS, Eggert M.
A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons.
J Mol Biol. Oct 20;197(4):723-8 (1987).

[54] M. Brudno, M. Chapman, B. Gottgens, S. Batzoglou, B. Morgenstern, Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, 4:66, 2003

[55] M. Brudno, S. Malde, A. Poliakov, C.B. Do, O. Couronne, I. Dubchak, S. Batzoglou,
Glocal alignment: finding rearrangements during alignment.
Bioinformatics, 19 Suppl 1:i54-62, 2003

[56] Waterman MS, Vingron M.
Sequence comparison significance and Poisson approssimation.
Stat. Sci. 9, 367-381 (1994).

[57] https://rana.lbl.gov/drosophila/

[58] Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J.
Abstract Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity.
Cell. Jan 13;124(1):47-59 (2006).

[59] Li E.
Chromatin modification and epigenetic reprogramming in mammalian development.
Nat Rev Genet. Sep;3(9):662-73 (2002).

[60] Rajewsky N.
microRNA target predictions in animals.
Nat Genet. Jun;38 Suppl:S8-13. Review (2006).

[61] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ.
miRBase: microRNA sequences, targets and gene nomenclature.
Nucleic Acids Res. 34, Database Issue, D140-D144 (2006).

[62] Pesole G, Liuni S, Grillo G, Ippedico M, Larizza A, Makalowski W, Saccone C.

UTRDB: a specialized database of 5' and 3' untranslated regions of eukaryotic mRNAs.
Nucleic Acids Res. Vol 27, 00 (1999)

[63] Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES.
Sequencing and comparison of yeast species to identify genes and regulatory elements.
Nature 423: 241 (2003).

[64] Dieterich C, Wang H, Rateitschak K, Luz H, Vingron M.
CORG: a database for Comparative Regulatory Genomics.
Nucleic Acid Res, 31:55-57 (2003).

[65] Dieterich C, Cusack B, Wang H, Rateitschak K, Krause A, Vingron M.
Annotating regulatory DNA based on man-mouse genomic comparison.
Bioinformatics, 18 Suppl 2, S84 (2002).

[66] Xie X., Lu J., Kulbokas EJ., Golub TR., Mootha V., Lindblad-Toh K., Lander ES. and Kellis M.
systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.
Nature 434, 338 - 345 (2005) .

[67] Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert F.
Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.
Genome Res. May;16(5):656-68. Epub 2006 Apr 10 (2006).

[68] Ranz J, Casals F, Ruiz A,
How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus Drosophila
Genome Research, 11:230–239 2001.

[69] Montgomery S, Griffith O, Sleumer M, Bergman C, Bilenky M, Pleasance E, Prychyna Y, Zhang X, Jones S

ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation.
Bioinformatics, 22(5):637–640 (2006).

[70] Gonzalez J, Ranz J, Ruiz A,
Chromosomal elements evolve at different rates in the drosophila genome.
Genetics, 161:1137–1154 2002.

[71] Smit, A.F.A.,
Origin of interspersed repeats in the human genome, Curr. Opin. Genet. Devel. 6 (6), 743-749 1996

[72] Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pedersen JS, Hsu F, Hinrichs AS, Harte RA, Diekhans M, Clawson H, Bejerano G, Barber GP, Baertsch R, Haussler D, Kent, WJ.
The UCSC genome browser database: update 2007.
Nucleic Acids Res, 35(Database issue):D668-73, 2007.

[73] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L,

Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y; FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group).
The transcriptional landscape of the mammalian genome.
Science. Sep 2;309(5740):1559-63 (2005).

[74] Caselle M, Di Cunto F and Provero P.
Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes.
BMC Bioinformatics 3:7 (2002).

[75] Corá D., Di Cunto F., Provero P., Silengo L. and Caselle M.
Computational identification of transcription factor binding sites by functional analysis of set of genes sharing overrepresented uptream motifs.
BMC Bioinformatics May 11;5(1):57 (2004).

[76] Corá D., Herrmann C., Dieterich C., Di Cunto F., Provero P. and Caselle M.
Ab initio identification of putative human transcription factor binding sites by comparative genomics.
BMC Bioinformatics May 2;6(1):110 (2005).

[77] Corá D., Di Cunto F., Caselle M. and Provero P.
Identification of candidate regulatory sequences in mammalian 3'-UTR regions by statistical analysis of oligonucleotide distributions.
In preparation (2006).

[78] Re A., Corá D., Puliti AM., Caselle M. and Sbrana I.
Correlated fragile site expression allows the identification of candidate fragile genes involved in immunity and associated with carcinogenesis.
BMC Bioinformatics Sep 18;7(1):413 (2006).

[79] Sara Zanivan, Davide Corà, Michele Caselle and Federico Bussolino
VRG: a database of vascular dysfunctions related genes.
accepted in Computers and Mathematics with Applications (2006).

[80] Arthur M. Lesk.
Introduction to Bioinformatics.
Oxford University Press (2002).

[81] Richard C. Deonier, Simon Tavaré, Michael S. Waterman.
Computational Genome Analysis - An introduction.
Springer Science (2005).

[82] Martignetti L., Caselle M.
Universal power law behaviors in genomic sequences and evolutionary models
Phys Rev E Stat Nonlin Soft Matter Phys. Aug:76 2007

[83] Martignetti L., Caselle M., Jacq B., Herrmann C.
DrosOCB: a high resolution map of conserved non coding sequences in Drosophila
arXiv:0710.1570 2007

WEB address of the Ensembl database
`http://www.ensembl.org`

WEB address of the UCSC database
`http://genome.ucsc.edu/`

WEB address of the NCBI - NIH database
`http://www.ncbi.nlm.nih.gov/`

WEB address of the RSAT tools
`http://rsat.ulb.ac.be/rsat`

WEB address of the CORG database
`http://corg.molgen.mpg.de`