Chapter 1

Introduction

Profound knowledge about the nature of mutational processes is essential for a comprehensive understanding of the evolutionary mechanisms that change genomes over time. The aim of this thesis is to elucidate the role of DNA insertions and deletions in this context. Compared with nucleotide substitutions, these types of mutations are far less understood. We perform a detailed genome-wide analysis of short DNA insertions and deletions that recently occurred in the human lineage. Our main finding is that insertions are predominately tandem duplications of adjacent sequence segments. We investigate the implications of this observation on possible molecular mechanisms of indel generation, large-scale statistical features of genomic base composition, and significance estimation of sequence alignment similarity scores. Starting with a short primer on molecular evolution, this first chapter provides a concise background on strategies to identify and analyze mutational processes with particular focus on DNA insertions and deletions.

1.1 Molecular biology and evolution

The DNA molecule Living organisms carry the genetic information needed for development, maintenance, and reproduction in their chromosomes. The essential components of chromosomes are long DNA molecules. DNA is a polymer made up of a linear chain of monomeric subunits called nucleotides. There are four different nucleotides in DNA; each is composed of a sugar-phosphate molecule and one specific base, namely adenine (A), cytosine (C), guanine (G), and thymine (T). The chemical structure of DNA is that of a right-handed double helix consisting of two intertwined polynucleotide strands that run in opposite directions (Fig. 1.1). Each type of base on one strand forms a bond with just one type of base on the other strand, constituting either an AT or a GC base pair. Both strands in a double helix therefore have complementary sequences [171].

Most prokaryotic organisms have only one chromosome. In eukaryotes, the genetic information is often subdivided into many different chromosomes. The combined DNA sequence of all different chromosomes of an organism is called its genome. Multicellular organisms carry copies of their genomes in almost every single cell. Most animals and plants are diploid species. Their cells contain two sets of chromosomes, each

inherited from one parent. The exception are sex chromosomes, which genetically determine the organism's gender. In mammals there are X and Y sex chromosomes. A pairing XY designates male organisms, XX females. Non-sex chromosomes are termed autosomes. The human genome, for example, comprises 22 autosomes and the two sex chromosomes. It is approximately 3.2 billion base pairs long.

The general principle by which the genetic information stored in DNA translates into proteins – the basic building blocks all cells are made up of – is subsumed in the central dogma of molecular biology. Originally proposed by Francis Crick, who together with James Watson was one of the discoverers of DNA's double-helix structure, it states that DNA is first transcribed into a complementary copy of a different nucleic acid called RNA, which is then translated into protein [44]. The sequential order of amino acids in the translated protein is encoded by the sequential order of nucleotides in the coding DNA segment such that three consecutive nucleotides, a so-called codon, always map to one particular amino acid.

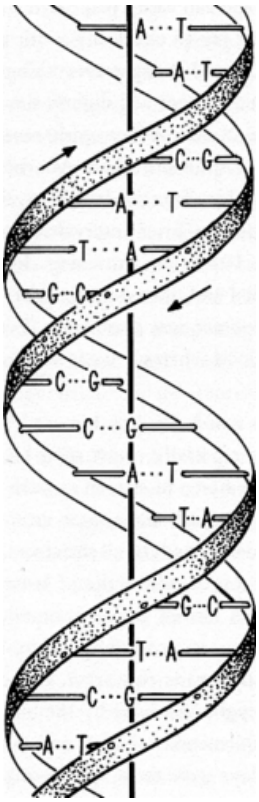

Figure 1.1: Double helix structure of the DNA molecule [170].

Only parts of a genome code for proteins. These regions are called exons. One protein can thereby be the product of several discontiguous exons, which might be spatially separated along the DNA by intermediate non-coding segments. Such intronic segments are excised from the transcribed RNA before it is translated into protein.

Organic life necessitates a plethora of different proteins. The precise set and amount of proteins required by a particular cell strongly depends on its specific role in the organism and the environmental conditions; protein synthesis thus needs to be tightly regulated. The regulation of protein expression can be extremely complex. Any step of the expression may be modulated, from DNA transcription and RNA processing to post-translational modification of a protein [61]. The non-coding regions of a genome play a crucial role in this context. For example, particular DNA regions can facilitate the binding of specific proteins to the DNA that regulate when transcription occurs and how much RNA is transcribed.

The DNA regions involved in regulating the expression of a protein are often located in the genomic vicinity of its coding sequence. This gave rise to the still widespread definition of a gene as the contiguous genomic region that comprises regulatory, intronic, and coding DNA of the gene's product [130]. For a more recent definition, which also incorporates that many genes are not clearly delimited, see e. g. [57]. The product of a gene does not always have to be a protein. For instance, it can also be an enzymatic RNA molecule. Some genes encode for more than product.

Genomes are highly heterogeneous environments where coding regions are interspersed by often large non-coding regions, especially in eukaryotic organisms. Non-coding parts of a genome can comprise besides regulatory regions also repetitive DNA, transposable elements, or simply non-coding genomic "junk". In human, for example, protein-coding segments account for only 1.2% of the total genomic sequence. The precise biological function of a large fraction of the remaining bulk of genomic material – if there always is one – has so far not been ultimately elucidated.

Genetic variation The double-helix structure of DNA provides an elegant mechanism for genome replication during cell division. As the two strands of a DNA molecule have complementary sequences, both strands can serve as templates for reproduction of the opposite strand once they are disassociated from each other during replication. Each template strand is preserved and the respective new strand is polymerized according to the rules of complementarity from new nucleotides. By this process two copies are generated from one original DNA molecule [170].

DNA replication has to be extremely accurate in order to avoid the introduction of mutations into one of the two genome copies. Mutations in exons, for example, can change the amino acid sequence of the encoded protein which then might not be able to correctly perform its designated function in the organism any longer. Sometimes mutations do, however, occur. They can result from errors during replication, effects of chemical or physical mutagens that directly alter the chemical structure of the DNA molecule, or from a variety of other cellular processes. Repair enzymes correct most of the errors, but some escape these mechanisms. Mutation events generate new mutant alleles which can differ from the original allele (wildtype).

A mutation that occurred in the germline can be transmitted to future generations. It can coexist with the wildtype in the population over a certain evolutionary period until it either substitutes the wildtype and becomes fixed, or lost. During this process, the genetic locus both alleles reside on is said to be polymorphic.

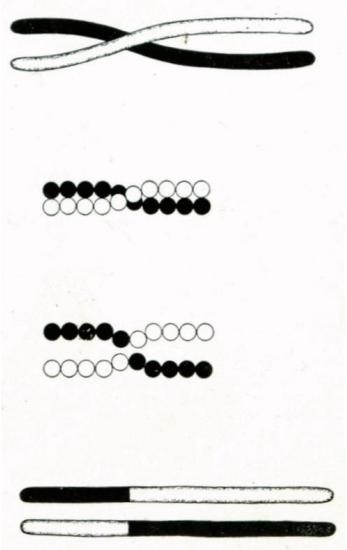

Figure 1.2: Thomas Hunt Morgan's illustration of crossing over (1916) [117].

If reproduction occurs in a way that offspring always inherits the complete genotype of its parents, alleles at distinct genomic loci would be tightly linked to each other. A beneficial combination of two specific mutant alleles would require that the second mutation arises in an individual already carrying the first mutation, which is rather inefficient, especially if the second mutation is already present in different individuals. This limitation is overcome by the mechanism of crossing over. Among sexually reproducing organisms, crossing over is an essential step during meiosis, the process by which a diploid cell divides to form four haploid gametes. Crossing over thereby occurs in prophase I of meiosis, where paternal and maternal chromosomes are

in tight formation. Both matching chromosomes have to break at least once and then reconnect to the other chromosome (Fig. 1.2). The consequence of crossing over is an exchange of DNA between both chromosomes, so-called genetic recombination. Due to the reshuffling of genetic material each gamete will have a unique combination of parental alleles, and particular mutations on distinct loci that originated in different individuals can rapidly be brought together on the same chromosome.

In large populations a substantial fraction of genomic loci will be polymorphic, implicating an enormous number of possible combinations of alleles (genotypes). This provides the source for the immense variety in observable phenotypic differences among individuals of a natural population. Recombination is one of the key processes to produce and maintain such variation [104].

Population genetics A newly arisen non-lethal mutation can coexist with the wildtype in the population over a certain evolutionary period. The decisive factor determining the probability of fixation or elimination of a new allele in the population is its relative fitness compared to the wild type. The fitness of an allele is thereby defined as the average number of offspring among individuals carrying the allele.

If fitness differences between mutant and wildtype are small, the dynamics of the mutant within the population is essentially determined by genetic drift, reflecting stochastic fluctuations that result from a finite population size (Fig. 1.3). If, on the other hand, fitness differences are sufficiently large, stochastic fluctuations are overruled by deterministic selective forces. This can lead to accelerated fixation of a beneficial mutant (positive selection), or its rapid removal as a consequence of strong selective constraints (purifying selection). These considerations have been put on a quantitative basis in the famous Kimura-Ohta theory of population genetics for finite populations evolving by stochastic fluctuations and selection [80].

Most new alleles that arise by mutational events in individual organisms have no chance of superseding the wildtype. They will simply vanish from the pool of genetic variation after some generations, especially if they cause a fitness disadvantage to the organism. Studies in the fly Drosophila melanogaster suggest that about 70% of single nucleotide mutations are deleterious [139]. However, some mutations will get fixed in the population from time to time, either by positive selection, or randomly due to genetic drift. The mutant allele substitutes the original allele becoming the new wildtype that subsequently is transmitted to future generations. Over longer periods successive substitution events accumulate in the population. This will also lead to perceivable evolutionary changes in phenotypic traits. Mutational processes and selective forces that influence the dynamics of alleles within the population together form the basis of evolution [45].

Since the advent of evolutionary studies on the molecular level in the 1960's there has been an extensive and still ongoing debate about the relative contributions of neutral evolution, purifying selection, and positive selection in molecular evolution. The dominant view over the last 40 years has been embodied by Kimura's neutral theory of molecular evolution [82, 79]. Its central proposition states that the vast majority

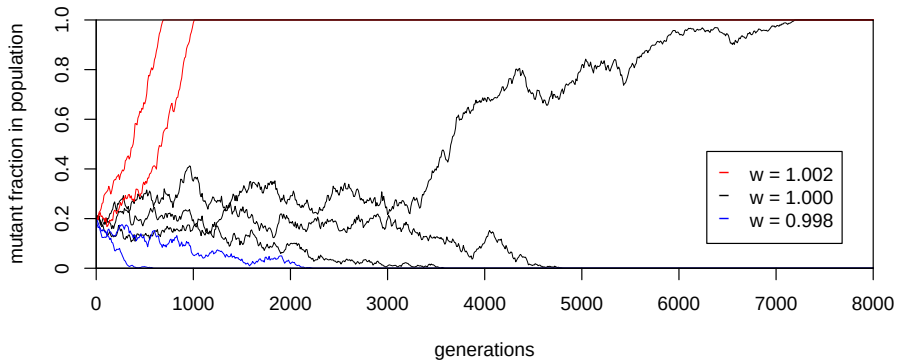

Figure 1.3: Simulated dynamics of mutant alleles in a haploid population of 10^4 individuals. New mutants always start within a 20% fraction of the whole population in our simulations. The relative fitness w of a mutant is defined as the ratio of average offspring number between individuals carrying the mutant and those carrying the wildtype. Overall population size is kept constant. Beneficial mutants become fixed rapidly (red trajectories), deleterious mutants (blue) are removed from the population. Neutral mutants conduct an unbiased random walk and can thereby eventually substitute the wildtype. Notice the profound effect small fitness differences of 0.2% have on the fixation dynamics of a mutant.

of substitutions in the genomes of natural species were in fact selectively neutral. Positive selection is acknowledged as an important force in generating phenotypic adaptation but in comparison is considered to be exceedingly rare.

Over the past few years, the validity of the neutral theory has been questioned in favor of an indeed substantial role of positive selection in molecular evolution [52, 147, 8, 138, 162, 51]. The main evidence for strong positive selection in recent studies arises from the estimates of the rate of adaptation derived from comparisons of rates of between species divergence at functional sites to those derived from the estimates of polymorphism at the same sites using reasoning of the neutral theory [105].

Besides far-reaching conceptual implications on our view of the fundamental mechanisms that drive the evolutionary process, the question of the relative contributions of selection and stochasticity in molecular evolution also has direct consequences on our potential to recover information about the characteristics of mutational processes from the analysis of substitution events. The reason for this arises from a remarkably simple result of population genetics theory. It states that the rate of substitution of selectively neutral alleles due to genetic drift is equal to the rate at which such mutations occur in individuals [81].

Hence, we can in principle derive molecular mutation rates by measuring substitution rates in the population if evolution is predominantly neutral. Moreover, if we assume that mutation rates along the genome are approximately constant, differences in substitution rates between particular genomic regions should be indicative of distinct levels of selective pressure in the compared regions. Both approaches can also

be used to investigate rates and regional selective constraints for particular classes of mutational processes. However, genomic regions that evolve under approximate selective neutrality are crucial for calibrating approaches of this type.

Macroevolution Environmental changes such as migration events or developing barriers between habitats can sometimes cause a population to be split into geographically separated subpopulations, where genetic exchange is prevented between individuals of different subpopulations. Over the course of evolution, substitution events will then progressively alter the genotype of each subpopulation. Genetic differences between groups will eventually become so pronounced that organisms of different subpopulations cannot interbreed any longer.

As the concept of a species is often defined by the capability of a group of organisms to interbreed and produce fertile offspring, evolutionary events of the above described nature mark the emergence of two distinct species from one ancestral species. They are therefore called speciation events. Although spatial segregation is likely to constitute the predominant mode of speciation in evolution, it is not a requirement for speciation to occur. In sympatric speciation, distinct species can also be formed in the absence of geographical barriers, for example as a result of assortative mating [18].

Over longer time-scales, evolution can hence be regarded as a branching process. Species change by successive genetic alterations, new species occasionally emerge through speciation, and entire species can also become extinct [45, 21]. In an analogous manner as the genealogical history of a mutant allele on the population level is represented by its coalescent tree (microevolution), evolution on the level of species (macroevolution) can also be described in terms of a tree. This phylogenetic tree reflects the large-scale evolutionary relationships among a set of species (Fig. 1.4).

Phylogenetics If one wants to reconstruct phylogenetic trees, one inevitably encounters the problem that species data is only available for present-day species. Fossil records are sporadic and often less reliable. We cannot decidedly infer the precise course of evolution of a species solely from investigating its current characteristics. Early trees were therefore based on the general notion of a hierarchy of relationships between species and higher taxa deduced from morphological differences.

This changed dramatically with the understanding that molecular sequences pose a perfect resource for investigating the phylogenetic relationships between different species. Although already suggested by Francis Crick in 1958 [43], the possibility of using molecular sequences for phylogenetic tree reconstruction was first realized by Emile Zuckerkandl and Linus Pauling [178, 179]. Their approach is based on the hypothesis that the rate of evolutionary change in genomic sequences is approximately constant over time and over different lineages. This concept of a "molecular clock" was supported by the observation that the number of amino acid differences in proteins between lineages scales roughly with speciation times estimated from fossil evidence. Measurement of divergence between pairs of genomic regions in two species that originated from the same genomic region in the most recent common ancestor (orthologous regions) can thus be used to infer speciation times.

In practical application, major challenges in estimating the evolutionary distance between species are usually of three kinds: First, the molecular clock needs to be calibrated. This can turn out to be quite intricate because different classes of mutational processes proceed at different rates, and rates will also vary between distinct genomic compartments [174]. Second, it is often not possible to decide beyond doubt whether similar sequence regions in two species are actually orthologous [55]. The third major problem is related to the fact that sequence divergence is not an absolute measure. It strongly relies on our model of the mutational processes [98], as will be discussed in Section 1.3,

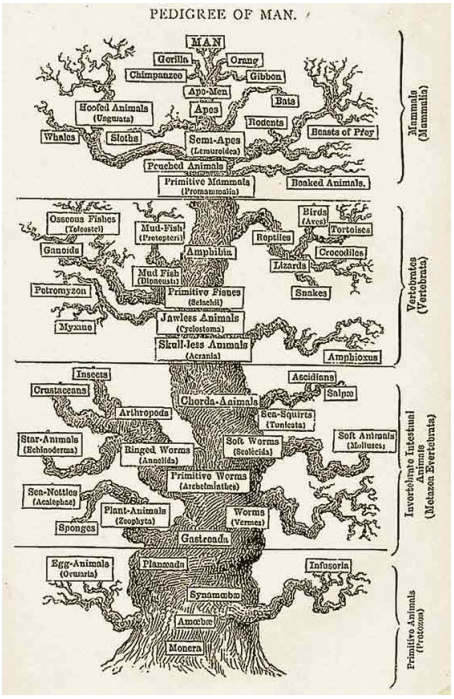

Figure 1.4: Ernst Haeckel's phylogenetic "tree of life" (1874) reflecting 19th century knowledge based on morphological differences. Present day methods utilizing DNA comparisons suggest a slightly different tree topology. Note the position of man at the very top of the tree, characteristic for the early ages of evolutionary thinking.

Despite of such problems, the molecular clock hypothesis has proven extremely successful in modern evolutionary biology. With the availability of fully sequenced genomes for a rapidly increasing number of species, divergence can nowadays be estimated on a whole-genome level. If evolutionary distances are calculated between every species pair of a given set, reconstruction of the phylogenetic tree that connects all species in the set can be reduced to the mathematical problem of finding the likeliest tree given the observed pairwise distances. Although this problem becomes highly nontrivial with increasing number of species, sophisticated computational methods have been developed during recent years enabling us to reconstruct the evolutionary history of present-day species in great detail [54].

1.2 Mutational processes

Molecular processes that generate mutant alleles provide the raw material for evolutionary change. When they initially occur in an individual, mutations are not teleologic in a sense that advantageous events are favored over deleterious ones. Our impression that on longer time-scales the evolution of species pursues a path of consecutive adaptations to altering environmental conditions first emerges from the different probabilities of substitution for beneficial, neutral, and deleterious alleles.

Mutations generate new random samples from the vast space of possible genotypes, which are then "tested" with respect to their viability for the particular species. From a broader perspective, the specific fashion in which mutant alleles are generated in evolution can also be regarded as a biological trait because the emergence of mutations is strongly linked to molecular processes inherent to the organism, e. g. DNA replication, recombination, and the organism's ability to correct for thereby induced errors. In principle, one could imagine that a species might simply invest a larger proportion of its resources in transmitting a more accurate genome copy to its descendents. Yet this would occur at the price of a reduced rate at which the species could adapt to changing ecosystems, eventually causing its extinction in periods of rapid ecological alteration. Hence, to some degree rates of mutation are also subject to natural selection on longer evolutionary timescales.

Mutational processes can be classified into different categories according to the type of sequence change caused by the event. A broad classification distinguishes three prevalent types of events [95]: (1) Single nucleotide mutations; (2) insertions of DNA segments into the genome; (3) deletions of DNA segments.

A more detailed classification can also take into account the particular nature of the exchanged, inserted, or deleted segment. Because DNA is made-up of four different nucleotides, there can be twelve different single nucleotide mutations that exchange two specific nucleotides. Insertions can be further specified with regard to the origin of the inserted DNA segment. This can for example be a duplication of an already existing sequence segment. Often one also observes genomic rearrangement events where particular DNA regions are deleted and inserted elsewhere in the genome, or inversions where an entire section of DNA is reversed.

Different types of mutation are likely to originate from different molecular causes. Whereas single nucleotide mutations may often be generated by "external" factors like radiation or mutagenic chemicals, insertions and deletions are mainly assumed to result from errors during replication and recombination, viruses that can paste copies into the genome by a mechanism called retrotransposition, and transposable elements – sequences of DNA which can move around to other positions within the genome [61]. Again, as was discussed above for the overall rate of mutation, selection will also be acting between the different types of mutational processes.

The generation of new mutant alleles yields samples from the space of possible genotypes, but different sampling strategies will be more or less efficient in optimizing the trade-off between preserving established functionality, and allowing for adaptation by generating new function or removal of molecular heritage that became dispensable in the course of evolution. We can expect that over several billion years of life on earth extremely sophisticated and optimized sampling strategies have emerged. Several general questions can be posed in this context, for example: Where is the optimal balance between the rate of single nucleotide mutations and the rate of DNA insertion and deletion, and what are the predominant characteristics of inserted or deleted sequence segments? Addressing the second question will constitute a major focus of this thesis.

The identification and precise characterization of the fundamental molecular processes that induce genomic variation will shed light on evolution's key mechanisms underlying the emergence of genetic innovation and adaptive evolution. The role of single nucleotide mutations in this context has been investigated in great detail and is nowadays described in most standard textbooks on molecular genetics, see for instance [95, 61, 89, 58]. In contrast, the nature of DNA insertions and deletions (indels) is far less understood. Ubiquitous throughout evolution indels occur on all scales ranging from single nucleotides up to whole genome duplications [68, 24, 77, 156, 177, 35, 33, 60, 15, 135, 31]. Comparative studies between human and chimp revealed that indels comprise approximately 3-5% in alignments of the two genomes, and therefore clearly outnumber the 1.23% divergence resulting from single nucleotide substitutions between these two species [32, 24, 166, 35]. Understanding more about indels is also important because they are associated with a variety of human genetic diseases [40, 165, 30, 144, 85, 152, 125, 36].

Irrespective of the various molecular causes, DNA insertions of larger segments in eukaryotic genomes typically involve duplications of parts of the genome. Examples include insertions of transposable elements, gene duplications, or large-scale segmental duplications. From a mechanistic point of view, the ubiquity of duplications reflects intrinsic features of the prevalent molecular processes generating insertions of DNA segments, such as retrotransposition, replication slippage (RS), or unequal crossing over (UCO). While the generation of duplications is obvious for the case of retrotransposition, in Fig. 2.4 of Section 2.2 we illustrate why UCO and RS generically generate duplication insertions too.

An evolutionary approach, on the other hand, focuses on a possibly beneficial role of duplication events. For example, following the duplication of a selectively constrained gene, one copy is allowed to evolve freely and can possibly acquire new functions, whereas the remaining copy will continue to perform the original task. Initially established by Ohno in his seminal work on proteome evolution by gene duplication [124], the concept of duplication-driven evolution has nowadays been extended from genes to also larger segmental duplications [15]. Consequently, this raises the question whether in a similar fashion duplications might also play an important evolutionary role on smaller length scales ranging down to single nucleotides.

Although such small DNA insertions comprise by far the largest number of all insertion events, for example throughout recent human evolution [24, 77, 156, 35], profound knowledge about their characteristics, underlying molecular processes, and evolutionary role is sparse. Yet, it is commonly believed that short indels are primarily generated by RS or UCO [88], and both processes generate tandem duplication insertions. Additional indication for a duplication mechanism on small length scales is provided by the overrepresentation of short paired duplicates in mammalian genomes [1, 155]. In Chapter 2 of this thesis, we will investigate whether tandem duplication is indeed the prevalent mode of small DNA insertions by a detailed analysis of the characteristics of short indels that recently occurred in the human lineage.

1.3 Inferring mutation characteristics

How can one infer the characteristics of mutational processes that occurred in the genome of a species during a particular evolutionary period? Obviously todays humans were not present to directly observe the mutational changes when they took place. However, Zuckerkandl and Pauling stated already in the 1960's in the first sentence of their seminal paper "Molecules as Documents of Evolutionary History" that "Of all natural systems, living matter is the one which, in the face of great transformations, preserves inscribed in its organization the largest amount of its own past history" [179]. They argued further that among all different elements of an organism its genomic DNA poses the best candidate to reveal such information.

In this section, we want to focus on two complementary approaches to derive information about the characteristics of mutational processes. The first aims at identifying mutation events by means of comparative genomics. This way, one can try to recover a set of mutational changes that have occurred between species since speciation from their common ancestor by analyzing alignments of their genomes. In Chapter 2, we will apply this "backward" approach to identify and investigate the characteristics of recent insertions and deletions in the human lineage.

The second, more indirect approach is based on the assumption that statistical features of genomes can reveal signatures of the repeated action of mutational processes during long-term evolution. By "forward" modeling one can thus test whether particular mutation models are compatible with the observed statistical features of genomic sequences. This second approach will be applied in Chapter 3 to investigate the effects of tandem duplications in genome evolution. The two approaches are however highly entangled as they both rely on stochastic models of sequence evolution.

Identifying mutational events by sequence comparison When comparing the genomes of two species, mutational events that have occurred in either of the two lineages since speciation from their most recent common ancestor will have lead to differences in both DNA sequences. In a reverse manner, by analyzing such differences we should hence gain insight into the mutational events that have taken place. Unfortunately this is not possible without ambiguity because different mutation scenarios can yield equal outcome (Fig. 1.5).

The standard procedure theoreticians usually apply to tackle problems of this kind is that they will try to come up with a probabilistic model. In our case, this needs to be a stochastic model of sequence evolution. Under the assumption of such a model, each mutation scenario that transfers ancestral sequence into present-day sequence can then be assigned a likelihood specifying the probability of the particular scenario according to the assumed model. Different scenarios can be compared with respect to their likelihoods.

However, when trying to establish a probabilistic model of sequence evolution, one usually relies on sequence comparison to estimate realistic values of the model's parameters. Apparently there seems to be a problem of circular reasoning. In practice

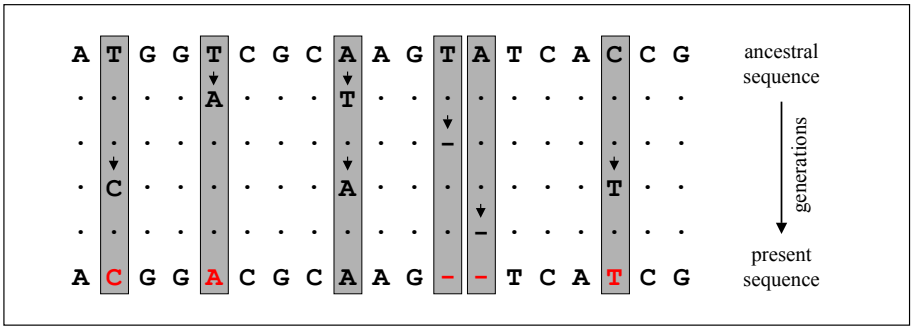

Figure 1.5: Illustration of a mutation scenario that over six generations transferred an ancestral sequences into the present sequence. The mutational processes of the particular scenario feature five single nucleotide mutations and two deletions of a nucleotide. Note that the observed differences between ancestral and present sequence (red) do not allow us to reconstruct the mutational scenario. There is a back mutation, and also the two deletion events might erroneously be regarded as one deletion of two nucleotides.

this can be dealt with by several approaches. First, we can incorporate additional *a priori* knowledge on the nature of mutational processes obtained from experiments and biochemical considerations. We expect, for example, that many mutational events are substitutions of single nucleotides, and that also insertions, deletions, or inversions of DNA segments can occur. With an external estimate of the speciation time and enough sequence data at hand, one can then try to infer the rates of these processes from sequence comparison in a maximum likelihood framework.

A second approach aims at reducing the number of sequence differences that resulted from more than one mutation event by investigating species that are very close. Why this works becomes clear from the following simple considerations. If we assume that mutational processes are Markov processes (see below), the expected number of events that have occurred in a sequence segment during time t is proportional to t. The number of sequence positions where two events have occurred is proportional to t^2. Among all positions where at least one event occurred, the fraction of positions where actually more than one event occurred is therefore approximately proportional to t (neglecting higher order terms). In more closely related species, we will hence observe a higher relative fraction of differences reflecting elementary mutational events. When trying to recover mutational processes from sequence differences between closely related species, it is consequently also justified to favor the least complex (most parsimonious) explanation of the observed differences that involves the smallest overall number of mutational events.

Markov models of sequence evolution In theoretical genetics, sequence evolution is usually modeled as a Markov process acting in sequence space (for a comprehensive introduction to the topic, see e. g. [122]). In the most general form, a state of the Markov process is a sequence $\vec{s} = (s_1, \ldots, s_N)$ with letters $s_i \in \{A, C, G, T\}$. Over the course of evolution the sequence can be considered a multidimensional stochastic random variable $\vec{S}(t)$. At subsequent time points $t_3 \leq t_2 \leq t_1$, we will observe sequences $\vec{s}_3, \vec{s}_2, \vec{s}_1$. They can differ from each other as a consequence of mutational

events. The Markov assumption states that the conditional probability to observe $\vec{s}_1$ is entirely determined by the knowledge of the most recent condition,

$$\Pr[\vec{s}_1, t_1 \,|\, \vec{s}_2, t_2\,;\, \vec{s}_3, t_3] = \Pr[\vec{s}_1, t_1 \,|\, \vec{s}_2, t_2], \tag{1.1}$$

meaning that the process is "memoryless". If this probability does not explicitly depend on t_1 and t_2 but only on the difference $T = t_1 - t_2$, the Markov process is said to be time-homogeneous and we denote the transition probability (1.1) from state $\vec{s}_2$ to $\vec{s}_1$ during a time interval T by $P_{12}(T)$. In practice, explicit forms for general transition probabilities can only be obtained for idealized mutational models.

To illustrate the power of Markov models in sequence evolution we want to shortly outline a simple but – because widely used – important example, which incorporates only single nucleotide mutations. It further assumes that all sites in a sequence evolve independently of each other according to the same time-homogeneous mutational model. The transition probability of a sequence then factorizes in the probabilities of its individual sites, where each site can have only four possible states $S(t) \in \{A, C, G, T\}$. Hence, there are 12 possible transitions between different states. Their probabilities can be described in terms of a 4×4 matrix $\mathbf{P}(T)$ with elements $P_{ij}(T) = \Pr[S(t+T) = i \,|\, S(t) = j]$. For a small time interval dT, we can write

$$\mathbf{P}(T + dT) = (\mathbf{I} + \mathbf{Q}dT)\mathbf{P}(T), \tag{1.2}$$

where $\mathbf{I}$ is the identity matrix. The matrix $\mathbf{Q}$ is also know as the instantaneous rate matrix as its off-diagonal entries Q_{ij} are the rates of mutations of nucleotides j to i. Knowledge of $\mathbf{Q}$ allows one to calculate transition probabilities for all T.

The mutational processes of our model can also change the average nucleotide composition of the sequences they are acting on. If we denote with $\vec{\rho}(t) = (\rho_A, \rho_C, \rho_G, \rho_T)$ the average composition of a sequence at time t, where ρ_i is the frequency of nucleotide i in the sequence, the time evolution of $\vec{\rho}(t)$ is given by

$$\frac{d\vec{\rho}(t)}{dt} = \mathbf{Q}\vec{\rho}(t). \tag{1.3}$$

In the long-time limit and for long enough sequences, $\vec{\rho}(t)$ will converge to an equilibrium distribution $\vec{\pi}$ because all nucleotides evolve independently of each other in our model. This equilibrium distribution is characterized by $0 = \mathbf{Q}\vec{\pi}$. If the equilibrium distribution is reached, the Markov process is said to be stationary.

The most general rate matrix $\mathbf{Q}$ has 12 independent parameters. For practical application, one can often make use of symmetries that effectively reduce the number of independent parameters. Complementary strand symmetry, for example, arises from the fact that there are only two different types of pairings in the DNA molecule, AT and GC base pairs. If an A mutates to a C on one strand, this will result in an exchange of a T by a G on the other strand, and so on. One thus expects that certain pairs of substitutions occur at similar rates.

There is in fact a large variety of different rate matrix models used in evolutionary studies. They cover various levels of complexity down to the simple one parameter model introduced by Jukes and Cantor in 1969, where all transitions have equal rate [72]. For a comprehensive review of the major commonly used models and their characteristics see e. g. [98].

The single nucleotide mutation model with independently evolving sites has been successfully employed in many evolutionary studies. Even in the presence of indels, it can still be used on the reduced set of sites which were not affected by indel events, provided that indels have occurred rarely between the sequences under comparison and their positions can thus be reliably identified from a pairwise sequence alignment. The site-independence assumption of the model can also be relaxed to some degree, for instance by incorporation of neighbor-dependent mutation processes [9, 11].

DNA insertions and deletions can also be considered memoryless stochastic events in sequence evolution. In principle, they can therefore be modeled as Markov processes too. The number of parameters required for specification of the model will however be huge. Insertions and deletions of segments of different lengths presumably occur at different rates, each of which would need to be specified separately (and there does not seem to be an upper bound of indel lengths). It is also not clear how to realistically model the sequence of inserted DNA segments. Moreover, there is a multitude of different indel-generating processes which can substantially differ in their rates and the nature of inserted or deleted sequence segments. Of course one can again try to come up with simplified models. Differences between processes could be neglected, the combined rate distributions of all processes could be approximated by analytic functions defining overall insertion and deletion rates for each particular segment length, and inserted sequence segments could generally be modeled as stretches of independently drawn random nucleotides.

Sequence alignment When aiming at the identification of elementary mutational events from pairwise comparison of orthologous sequence segments in present-day species, one requires a recommendation of the pairs of sequence positions in both sequences which are likely to share a common evolutionary origin in the most recent common ancestor. If our model of sequence evolution does only take into account single nucleotide substitutions, the problem can easily be solved by writing one sequence underneath the other with the particular shift that yields the highest number of vertical columns with equal nucleotides in both sequences. The resulting two-row matrix is then called an alignment of the two sequences. Columns with differing nucleotides indicate mutation events.

However, this simple approach will be inadequate if insertions and deletions events may have occurred. A model of sequence evolution that additionally takes into account such processes should also allow for gaps in the alignment, i. e. positions in one sequence with no corresponding orthologs in the other sequence (see Fig. 1.5). The major challenge of biological sequence alignment is to place matches, mismatches, and gaps in a way that is most likely to correspond with the actual mutational events that have occurred in evolutionary history of the two sequence.

Most commonly used alignment algorithms approach this problem by optimizing a scoring function that penalizes the number of mismatches and gaps in the alignment. The particular choice of the scoring function reflects our model of the evolutionary mutation processes. For example, the expected relative frequency between indel events and single nucleotide substitutions can be incorporated into the relative penalties for gaps and mismatches. The scoring function can further be refined by assigning different mismatch scores for each particular type of mismatch commensurate with the transition probabilities of the model's rate matrix. Presumptions of the expected gap-length distribution can be accounted for by length-dependent gap costs. A variety of computational algorithms have been applied to the sequence alignment problem, including slow but formally optimizing methods like dynamic programming, as well as efficient heuristic or probabilistic methods designed for large-scale database search [48]. Sequence alignment is one of the most commonly used computational tools in today's molecular biology.

Having constructed an alignment of two orthologous sequences, we can investigate its mismatches and gaps to gain more detailed insight into the characteristics of the underlying mutational events. From the frequency of mismatches involving nucleotides i and j, for example, we can obtain a better estimate of the average mutation rate $(Q_{ij} + Q_{ji})/2$ (additional knowledge about the actual direction of mutation is needed to decouple both rates). Length and number of gaps can be used to approximate indel rate-distributions, and the corresponding sequence segments in one row where gap segments were placed in the other sequence can be analyzed to reveal information on the nature of inserted/deleted sequence.

Our hope is that the mutational model used for the alignment process was already close enough to reality such that orthologous nucleotides were correctly identified in the alignment. The problem of circular reasoning in this context already mentioned earlier is partly alleviated by the discrete nature of an alignment. Two nucleotides in both sequences can either be aligned in the same row, or not, and changes in the alignment at one position will always entail changes at other positions too. We can thus suppose that a correct alignment will hopefully be stable over a broader range of model parameters. In fact, the sensitivity of an alignment to varying parameters is in a reverse manner often used to assess its quality and trustworthiness.

The amount of information we can retrieve from pairwise alignments of present-day sequences is limited by the problem that we cannot infer the direction of a mutation due to the unknown status of the ancestral sequence. A mismatch $i \leftrightarrow j$ in the pairwise alignment can result from a substitution $i \rightarrow j$ in the first lineage, or an $j \rightarrow i$ substitution in the other lineage, depending on whether the ancestral state was i or j. The same applies to the distinction between insertions and deletions, which are therefore unspecifically denoted as indels in pairwise alignments.

However, we can incorporate additional information from comparison with an outgroup species if we have knowledge of the orthologous sequence region in this species. In this case, one will aim at constructing a so-called multiple alignment of all three sequences, where one again assumes that each column of the alignment shares a

common origin in the ancestor of all three species. The state of the out-group sequence at a particular alignment position can then be used to infer statistical information about the ancestral state before speciation of the two in-group species. Generally, this can be done within a maximum-likelihood framework. If sequence divergence is low and the out-group state coincides with one of the states in the two in-group sequences, maximum-parsimony might turn out to be a sufficient approximation. This way on can identify direction and particular branch of the phylogenetic tree in which the mutation occurred. We will use this approach in Chapter 2 to explicitly distinguish insertions from deletions in the human lineage. Several powerful algorithms have been developed in recent years to efficiently align more than two sequences [167, 123]. The resulting multiple alignments provide the core-datasets for many of today's comparative genomics analysis [63, 115].

Statistical genome features as possible signatures of mutational processes Over longer evolutionary periods, the persistent action of mutational processes will inevitably impinge on the elementary statistical properties of genomic sequences. Effects could be as simple as changing average genomic base frequencies due to the particular rates of single nucleotide mutations according to Eq. (1.3). But in the presence of different classes of mutational events including insertions and deletions or neighbor-dependent processes, one might expect that effects will be more complex.

A classical example of neighbor-dependent mutation is the CpG methylation-deamination process, which dominates point mutations in vertebrate genomes. CpG dinucleotides are often methylated. A spontaneous deamination of the cytosine in a methylated CpG will result in a mutation CpG→TpG [42, 134]. Presumably due to this process, point substitutions occur 10 times more frequently at CpG sites compared to non-CpG sites [20, 65]. This leads to a steady removal of CpG dinucleotides from genomes. Indeed, in the human genome the observed frequency ρ_{CG} of CpG dinucleotides is substantially smaller than one would expect from the product of the individual frequencies $\rho_{\mathrm{C}}\rho_{\mathrm{G}}$ under the assumption that neighboring nucleotides are independent of each other [140]. It has been shown that this effect can be explained by incorporating neighbor-dependent mutation rates in the stochastic model of sequence evolution, and that such rates can to some degree be recovered in a maximum-likelihood framework from measuring genomic dinucleotide frequencies [11].

The value of approaches of this kind is that they establish connections between statistical properties of genomes and mutational models. For the independent single-nucleotide mutation model, this was a linkage of the elements of the rate-matrix $\mathbf{Q}$ to the stationary nucleotide frequencies via $0 = \mathbf{Q}\vec{\pi}$. For the model including neighbor-dependent mutation rates, a corresponding connection has been established in [9] between such rates and stationary genomic dinucleotide frequencies π_{ij}. However, in both cases it cannot be concluded from data that it was actually generated by the model. What can be said is that differences in nucleotide frequencies are in principle compatible with certain single-nucleotide mutation models, and that nontrivial dinucleotide frequency distributions can result from neighbor-dependent mutation rates.

The question is whether approaches of this kind can also be extended to other classes of mutational processes, for example DNA insertions and deletions, and what could be promising statistical features to investigate for this purpose.

Moving from single nucleotide frequencies to dinucleotide frequencies represented an increase of complexity in the measure of interest because dinucleotide frequencies also feature a spatial component regarding the arrangement of nucleotides next to each other along the sequence. In the limit of large sequence lengths, dinucleotide frequencies can be interpreted as joint probabilities of observing a neighboring pair $(s_x = i, s_{x+1} = j)$ of nucleotides i and j at a randomly picked position x in a genomic sequence $\vec{s} = (s_1, \ldots, s_N)$. In terms of the general probabilities of finding nucleotides i and j separated by an arbitrary distance r along the genome,

$$P_{ij}(r) = \Pr[s_x = i \wedge s_{x+r} = j], \tag{1.4}$$

dinucleotide frequencies refer to the special cases $P_{ij}(r = 1)$. Analyzing $P_{ij}(r)$ for arbitrary distances r thus poses a natural extension along the line of single- and dinucleotide frequencies when aiming at gradually increasing feature complexity.

If we subtract from $P_{ij}(r)$ the joint probability $\rho_i \rho_j$ one expects to observe in a sequence of independent nucleotides, we obtain a measure of the correlation between nucleotides i and j at distance r. In genomic sequences one observes that usually $P_{ii}(r) > \rho_i^2$ for $r > 1$, indicating that equal nucleotides are positively correlated along the genome. To obtain an impression of correlation strength and functional dependence on the distance r, it is convenient to define a general correlation function

$$C(r) = \sum_i [P_{ii}(r) - \rho_i^2] \tag{1.5}$$

by summing over all individual correlations between equal nucleotides. With the rapidly growing availability of whole-genome sequence data the correlations $C(r)$ along genomic DNA can nowadays be studied systematically over a wide range of scales and organisms. A striking observation in this field was the finding of *long-range correlations* in the base composition of genomes more than a decade ago [94, 131, 163]. They are characterized by a power-law decay of the correlation function,

$$C(r) \propto r^{-\alpha}, \tag{1.6}$$

for large r. By now it is well established that long-range correlations in base composition appear in the genomes of most eukaryotic species [13, 23, 92, 150, 93, 26]. Two examples are shown in Fig. 1.6. The form of genomic long-range correlations is often more complex than simple power-laws. Within one genomic region there can be distinct scaling regimes with different effective exponents. Correlations might be restricted to specific distance intervals $r_{\min} < r < r_{\max}$, sometimes no clear scaling is observed at all. Amplitudes and decay exponents also differ considerably between species and between different genomic regions of the same species [93].

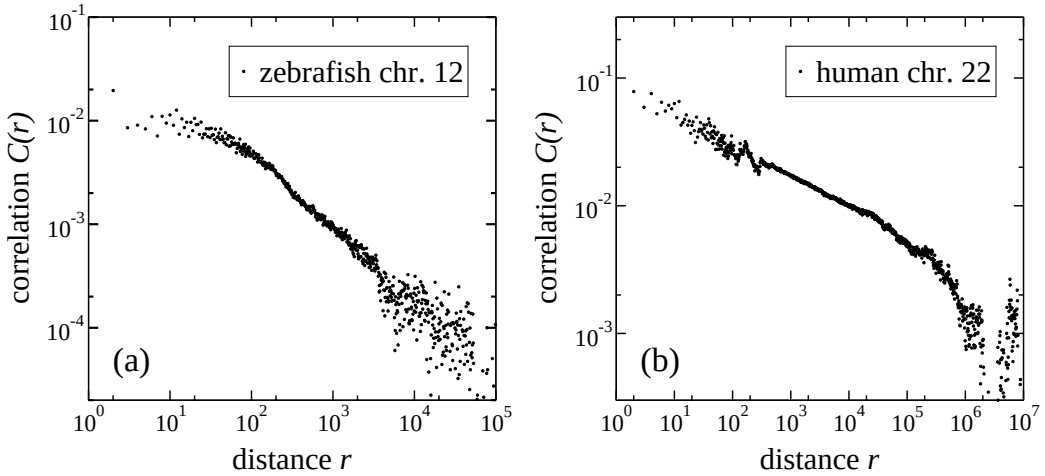

Figure 1.6: Long-range correlations in the base composition of two eukaryotic chromosomes. In the double-logarithmic plots, power-law correlations $C(r) \propto r^{-\alpha}$ show up as straight lines with slope $-\alpha$. They extend over distances of several orders of magnitude.

Little is known about the origin of genomic long-range correlations, so far. However, considerations from statistical physics suggest that their ubiquity among eukaryotic genomes might be related to universal mechanisms. Long-range correlations are a hallmark of systems with many degrees of freedom throughout physics. Their appearance is often associated with the existence of so-called universality classes referring to the observation that "macroscopic" properties of an entire class of systems are to a large extent independent of the "microscopic" dynamical details. In equilibrium condensed matter systems, long-range correlations mark critical points or phases with a particular symmetry. Out of equilibrium, long-range correlations are more generic but the classification of universality classes becomes more difficult. Well known examples are surface growth, reaction-diffusion systems, and self-organized criticality [150].

Following the line of thought that particular statistical properties of genomes can reflect the persistent action of mutational processes throughout long-term evolution, it is an alluring conjecture to propose that long-range correlations in genomic sequences might also result from the local stochastic processes of molecular evolution. A potential candidate could be the interplay between duplication, deletion, and mutation processes in genome evolution. Indeed, it has been shown in [90, 91, 111] that already a simple stochastic process consisting of duplications and mutations of single letters leads to generic power-law correlations in the sequence composition. Investigating the precise connection between statistical sequence properties and dynamical models of sequence evolution that comprise the major local mutational processes including segmental tandem duplications will be the main issue in Chapter 3 of this thesis.

1.4 Background models of DNA sequences

Recent years have witnessed an impressive advance of bioinformatics sequence analysis tools, aiming at deeper insight to the functional organization and evolutionary dynamics of genomic DNA sequences. Popular examples include algorithms for genome annotation, homology detection between genomic regions of different organisms, or the prediction of transcription factor binding sites [167, 48]. Bioinformatics methods frequently yield probabilistic statements. Usually the statistical significance of a computational prediction is characterized by a p-value, specifying the likelihood that this prediction could have arisen by chance. The calculation of p-values requires an appropriate null model of DNA, which reflects our assumptions about the "background" statistical features of the sequence under consideration. The challenging task is to decide on the set of statistical features a suitable null model should obey. Ideally, one incorporates those features into the null model which describe the background "noise" of the DNA sequence, but still allow to discern the specific signal the computational analysis tries to detect.

The simplest and most-widely used DNA background model is an *iid* model, given by a random sequence with letters drawn independently from an identical distribution [48]. The iid model can incorporate the length and the average composition of the sequences under consideration, but it lacks any specific structure concerning the arrangement of the nucleotides along the DNA. In particular, it is not capable of incorporating correlations in base composition along the sequences. However, as we have discussed in the previous section, such correlations are ubiquitous in the genomes of most eukaryotic species. The inherent statistical features of their DNA sequences hence differ substantially from those generated by an iid model.

In Chapter 3, we will show that correlations in genomic base composition inevitably arise by the long-term action of basic mutational processes with tandem duplication insertions constituting the driving force in this context. The generic origin of such correlations by fundamental evolutionary processes provides a likely explanation for their widespread presence in eukaryotic genomes, but it also raises the question if such correlations need to be incorporated into an accurate null model of eukaryotic DNA and how that would change the p-value calculations of bioinformatics tools [38].

Up to a certain degree, the additional complexity resulting from correlations compared to a simple iid model can be taken into account by an nth order Markov model specifying the transition probabilities $\Pr[s_{x+1} \,|\, s_{x-n+1}\,; \ldots ;\, s_x]$ in a genomic sequence $\vec{s} = (s_1, \ldots, s_N)$ [48]. Notice the conceptual difference in application of the Markov framework compared to Section 1.3, where Markov models were used to describe the evolution of DNA sequences in time. Elements of the Markov chain (the sequence of random variables generated by the Markov process) were DNA sequences at successive time-points. The Markov process acted "vertically" along the time axis. Here, the elements are the different nucleotides in one sequence. The Markov process is acting "horizontally" along the sequence generating the next nucleotide with respect to the previous neighboring nucleotide to its left.

Assuming sequences to be generated by such Markov models allows to incorporate a multitude of spatial statistical features into the model, e. g. preferential occurrence of DNA motifs, local peculiarities in genomic composition, or specific dinucleotide frequencies. In contrast to iid sequences, where all letters are uncorrelated, Markov processes can generate *short-range correlations* in the nucleotide composition [131]. They are characterized by exponentially decaying correlations along the sequence.

In the base composition of eukaryotic genomes, however, one usually observes algebraically decaying long-range correlations, $C(r) \propto r^{-\alpha}$, which therefore decay much slower compared to short-range correlations. Long-range correlated sequences cannot be modeled as an nth order Markov chain with finite n [131]. Yet, their effect on p-value calculations of bioinformatics sequence analysis tools – if incorporated into the DNA null model – will presumably be even more distinct. In Chapter 4, we will investigate this issue in the context of sequence alignment, which constitutes the most commonly used computational tool of molecular biology today [5, 6]. As the main result of this analysis, it will turn out that long-range correlations in the sequences indeed lead to considerable deviations in sequence alignment score statistics.

1.5 Thesis organization

The aim of this thesis is to investigate nature, origin, and consequences of short DNA insertions and deletions in genome evolution. In Chapter 2, we present a genome-wide analysis of 1-100 bp long indels in the human genome since its split from the common ancestor with chimpanzee. Insertions are explicitly distinguished from deletions by comparison with an out-group species. We show that the majority of identified short DNA insertions are actually tandem duplications of adjacent sequence segments and discuss possible molecular mechanisms of indel generation. In the second part of Chapter 2, we shift our focus from a genome-wide perspective to the analysis of indel characteristics in protein-coding regions of the human genome.

In Chapter 3, we study the stochastic dynamics of genome sequences evolving by single site mutations, duplications, deletions, and random insertions of sequence segments from a theoretical point of view. We apply sophisticated analytical tools from nonlinear dynamics, as well as detailed numerical simulations to show that these processes generate distinct statistical features of the sequences they are acting on, including long-range correlations and large-scale fluctuations in genomic base composition. A possible connection between the major local mutation processes in evolution and the commonly observed long-range composition correlations along eukaryotic genomes is discussed at the end of the chapter.

The prevalence of duplications among DNA insertions raises the question whether such processes – and the resulting correlations in genomic base composition – should be incorporated into DNA null models needed for significance estimation of bioinformatics sequence analysis tools. In Chapter 4, we address this question in the context

of sequence alignment. We introduce a novel analytic approach, the Gaussian approximation, which allows us to calculate the corrections to the scale parameter λ of the alignment score distribution for correlated sequences. We find that incorporation of long-range correlation into the DNA null model leads to considerable deviations in the score statistics of sequence alignment. The magnitude of this effect for correlations of genomic scale and its implications in a bioinformatics context are discussed at the end of the chapter.