Discovering Mutational Patterns in Mammals Using Comparative Genomics

Paz Polak

June 2010

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
eingereicht im Fachbereich Mathematik und Informatik
der Freien Universität Berlin

1. Referent: Prof. Dr. Martin Vingron
2. Referent: Prof. Dr. Nikolaus Rajewsky

Tag der Promotion: 17. September 2010

Preface

Acknowledgments

I would like to thank all people who have helped and inspired me during my doctoral study. My deepest gratitude goes to my advisor *Peter Arndt* whose inspiration, guidance and support enabled me to develop a greater understanding of the subject. The atmosphere of freedom to think and in particular, his accessibility and willingness to help with any problem, large or small, will never be forgotten.

Special thanks is reserved for *Nina Papavasiliou* for the discussions on somatic hyper mutation processes and to *Robert Querfurth* for assistance in the work on evolution of substitution patterns.

I heartily give gratitude to *Rosa Karlic*, *Sean O'Keeffe*, *Brian Cusack*, *Julia Lasserre*, *Yves Clement*, *Kirsten Kelleher* and *Sarah Behrens* for critically reading this thesis and for their useful comments. Large parts of my scientific education I gained during the weekly Gene Regulatory meetings and the Vingron department seminars and therefore I wish to thank all the past and present colleagues in the Vingron department and in particular the EvoGen group. I also thank the International Max Planck Research School for Computational Biology for the financial support and the coordinator of the program *Hannes Luz* who made my life easier during my PhD. Many thanks are given to *Martin Vingron* who established this school and (together with *Peter Arndt*) allowed me, at this early stage of my career, to enjoy rare and exceptional scientific conditions, which were essential for me to develop my current view on biology.

Thanks also goes to my friends and family in Berlin and Israel who supported me during the last four years. I would especially like to thank my mother without whose continuous support I could not have carried out my research.

Publications This thesis conceals within the content of three publications. The regional patterns and strand asymmetries of substitution rates along human genes appeared in *Genome Research* [121]. The evolution of strand asymmetries across vertebrates was accepted to *BMC Evolutionary Biology*. And the findings that strand asymmetries are found in intergenic regions and originate in CpG islands were published in *Genome Biology and Evolution* [122].

Polak Paz Berlin, June 2010

Contents

Preface i

1 Introduction 1
- 1.1 DNA mutations . . . 1
- 1.2 Substitutions . . . 2
- 1.3 Genome wide mutation rates . . . 5
- 1.4 Bias in substitution rates in double stranded DNA . . . 8
- 1.5 Bias in substitution rates in single stranded DNA . . . 11
- 1.6 Thesis overview . . . 14

2 Methods 15
- 2.1 Inferring substitution rates . . . 15
- 2.2 Analyzed sequences . . . 26

3 Methylation deamination rate in the vicinity of the mammalian 5'end 33
- 3.1 Analysis of nucleotide substitutions . . . 33
- 3.2 CpG methylation deamination rates . . . 34
- 3.3 Possible mechanisms to explain lower CpG loss near the TSS . . . 36

4 Transcription-associated strand asymmetries in mammals 39
- 4.1 Localized asymmetry . . . 39
- 4.2 Global asymmetry . . . 40
- 4.3 Strand asymmetries in non intronic regions of genes . . . 40
- 4.4 The impact of CpG islands on strand asymmetries in transcribed regions 43
- 4.5 Strand asymmetries are found in other mammals . . . 47
- 4.6 Regional patterns of nucleotide composition . . . 49
- 4.7 Possible mutational processes that generate localized asymmetries in genes . . . 50
- 4.8 Mutational processes that may generate global asymmetries . . . 55

5 Asymmetries in intergenic regions originate in CpG Islands 57
- 5.1 Strand asymmetries in intergenic regions in the vicinity of genes . . . 57
- 5.2 Long range strand asymmetries around CpG islands . . . 58
- 5.3 Mechanisms that can generate strand asymmetries in intergenic regions 62

6 Weak to strong bias in promoters and CpG islands **67**
6.1 Correlation of $r_{W\to S}/r_{S\to W}$ ratio with crossover rates in human CpG islands 67
6.2 Substitution signature of recombination in vertebrate promoters 69
6.3 BGC as a putative mechanism to increase GC content 71

7 Summary **75**

A Appendix A **91**

B Appendix B **97**

C Appendix C **107**

Notation and abbreviations **111**

Zusammenfassung **113**

Curriculum vitae **115**

Erklärung zur Urheberschaft **117**

1 Introduction

Deoxyribonucleic acid (DNA) is a highly stable molecule. However, changes in the sequence, called DNA mutations, continuously occur but typically at very low rates. Although mutations are seen as the fuel for evolution, little is known about their rates along chromosomes. Currently, the best way to study the patterns of mutation rates along chromosomes is via comparative genomics. In this introduction, I will review the two main lessons from comparative genomics studies on the most frequent mutations, the single nucleotide mutations. Firstly, mutation rates vary along chromosomes and are correlated with the activity of processes such as replication, recombination and transcription. Secondly, there is a mutational spectrum, which means that not all mutation rates are equal to each other.

1.1 DNA mutations

Despite their central role in evolution, there is not yet a good knowledge about the rate of neutrally occurring mutations along mammalian genomes [70]. Fundamental questions about neutral mutation rates such as the relative contribution of replication and transcription and their associated processes to mutation rates are still unanswered [70]. The two main fields that are currently concerned with the study of mutations are the study of genetic diseases and evolution [70]. Both fields are mainly concerned with the functional impact of mutations. In humans, DNA mutations can cause a variety of genetic diseases if they occur in the germline or at early developmental stages and can cause cancer if they happen in somatic tissues [45]. Mutations are also a major force of evolution, since mutagenesis generates innovations by introducing genetic variations, some of which might have phenotypic impact [58; 95].

DNA structure DNA consists of two single strands of phosphate and sugar coiled around each other in a helical manner and held together by weak hydrogen bonding between pairs of nitrogenous bases to form the double-helix structure. The building blocks of the DNA strands are four nucleotide bases: adenine (A) and guanine (G), which are purines, and thymine (T) and cytosine (C), which are pyrimidines [155]. In the DNA double helix structure adenine is joined with thymine and guanine with cytosine to form the base pairing couples. In a DNA strand every base is attached to a five-carbon sugar ring and a phosphate group. The single stranded (ss)DNA is a

chain of nucleotides that are joined to each other by the phosphate group that form phosphodiester bonds between the third and fifth carbon atoms of the sugar ring of two nucleotides. This leads to a distinction between the ends of a DNA strand, since on one end the fifth carbon of the sugar ring is attached to a free phosphate group while on the other end the third carbon is characterized by a free hydroxyl (OH) group. This suggest that strand itself has a directionality which is designated by 5' to 3' direction. Sequences are always replicated and transcribed by corresponding polymerases from 5' to the 3'end. It is a convention to write the DNA sequence of one strand from 5' (left) to 3' (right). Since the two strands of sequences are complementary, the DNA sequence is often represented only by the bases in a single strand in the direction 5' to 3'.

Types of DNA mutations Changes in the DNA sequence are called mutations. In an *in silico* view, mutations are seen as no more than a collection of editing operations on this sequence. The most frequent and most studied type of mutations are single nucleotide mutations. There are twelve possibilities to exchange one base for another. Since the DNA is a double stranded helix, a single nucleotide exchange is a base pair change mutation, for instance when A is substituted by C then there is $A : T \rightarrow C : G$ mutation, where the colons denote Watson-Crick base pairing. Deletions of DNA sequences can remove 1 base-pair (bp) to mega base-pairs (Mbps). Insertions are extra DNA sequences that can be a new sequence which was not present in the genome or a piece of DNA that is copied from one locus and pasted in another (a duplication). Translocations are a cut and paste type of mutation i.e. DNA pieces that move from one chromosome to another (inter chromosome), or to new locations within the same chromosome (intra chromosome). Inversions are a type of mutations where a part of a chromosome is reversed with respect to its flanking segments. Inversion on a ssDNA therefore results in the reverse complement of the previous sequence.

In this thesis, I will focus on single-base mutations, the most frequent mutation, although we have to bear in mind that the total number of bases that are affected by insertions, deletions and rearrangements per unit time exceeds the number of single nucleotide mutations, because insertions and deletions can impact several Mbps at a time [27].

1.2 Substitutions

Functional consequences of DNA mutations The genome encodes proteins, which are the building blocks of cells which make up an organism. It also contains information of when and where to produce RNA (ribonucleic acid) and proteins. Changes in the sequence of a protein coding gene can impact its activity since as a consequence of a mutation a protein can be misfolded, truncated, lose or gain the ability to bind to other proteins or DNA [95; 30]. Mutations in DNA in transcribed regions can disrupt

RNA secondary structure and affect splicing [81; 30]. Mutations can also alter the regulatory program since changes in regulatory DNA sequence elements will affect the time and the level of gene expression [113; 30].

Alterations to the DNA sequence can have a harmful impact on the cell or the organism [26]. When these mutations disrupt essential functions it can lead to cell death or even the death of the organism itself. But not all mutations end in death of an organism. Different individuals in a population can have different genomes [94; 64]. Therefore, different members of the population may carry different variants of sequences called alleles [58]. The allele frequency of a specific variant is the proportion of copies of this variant among all alleles at the corresponding locus in the population. In a diploid population of size N, there are 2N alleles for each locus in autosomal chromosome (i.e. non-sex chromosomes).

Fixation processes Evolution can be viewed as a change in allele frequencies from generation to generation [57]. The number of offspring is usually regarded as the fitness of the organism [57]. In this terminology alleles that increase the fitness are beneficial, the ones that reduce the fitness deleterious and the ones that do not affect the fitness are neutral [58].

Mutations in multicellular organisms with reproductive cells can be divided into two classes- somatic or germline mutations [9]. Somatic mutations occur during cell division in the process of tissue formation and can not be transmitted to the next generation (with the exception of plants). Only DNA changes in the germline are transmitted to the following generation and therefore they have the main long term evolutionary consequence. In unicellular organisms, a new generation emerges at each cell division and therefore all mutations that occur during the cell cycle are transmitted to daughter cells.

Since mutations are rare events, the general assumption is that only one mutation occurs at a specific locus at any given time. Therefore, when a mutation event occurs, a new mutation is present in one individual member of the population. For a diploid population of size N, the frequency of the mutant allele when it arises is $1/2N$ [57]. Over time, the frequency of the allele with the mutation can increase within the population until it appears in 100% of individuals in the population [57]. In the case that a mutant allele arrives at fixation, i.e. its frequency in the population is 1, the mutant allele is said to substitute the wild-type (original) allele i.e. substitution [57].

The two main forces that shape the allele frequencies at the population level are selection and chance (also called random genetic drift) [57]. Selection drives the allele frequency changes of mutations that result in a difference in fitness betwen an individual with the wild-type allele and one with the mutant allele. Positive selection drives alleles to fixation when they increase fitness relative to the wild-type allele. Increase of fitness means that the representation of the allele in the following generation is increased due to the higher number of offspring of the individual that carries the mutant

and therefore over time the mutant allele frequency will increase until it becomes 100% in finite population size (i.e. number of the members in the population is bounded over time). If a mutation decreases the fitness of an organism, then selection is said to play a negative role in shaping the allele frequency to remove the mutation, since (by the above definition) the individual has less offspring than the members of the population that carry the wild type allele. Selection is blind to neutral mutants and therefore does not play a role in shaping their allele frequencies.

For finite size populations, random genetic drift is the force that determines the fixation probabilities of neutral sites in the genome [57]. Random genetic drift assures that neutral mutations can arrive at fixation (i.e. substitution). The reasons for the drift are DNA replication, reproduction and the constrained size of the population. The random nature of changes in allele frequencies is due to the fact that the number of fertile offspring of two genetically identical individuals is not constant. Because of this, chance also serves as a counter force of positive selection since individuals that carry a strong advantageous mutation might fail to reproduce or may produce offspring without the mutated allele and therefore the mutation would be lost in the next generation.

Random genetic drift is probably the major force that shaped the mammalian genome [57; 95]. Most changes in the DNA have only a small impact (if any) on the fitness of the organism. Therefore, most mutations in the mammalian genome are subject to weak (if any) selection forces.

Molecular mutation rates The rate at which a base is exchanged in the genome per unit of time is called the molecular mutation rate (and we denote it by μ). The unit of a time that is often used in the literature is generation time and for human is typically 20 years [108]. In other words, it measures the number of *de-novo* mutations per site that a newborn inherits from its parents, which were not present in the parents' genomes when they were born.

Substitution rate New alleles continuously arise within population due to mutation processes. A new allele can replace (substitute) another allele and in turn can be replaced after same time by another. Therefore we can define the substitution rate, denoted by R, as the number of fixations of new alleles per site per unit time (e.g. each generation) [57].

The null hypothesis of neutral theory What is the link between substitution rates (R) and molecular mutation rates (μ)? The answer might be very surprising and useful for neutral sites in a population of constant size (the number of the members in the population in each generation remains constant over time). The substitution rate is determined by the mutation rate μ and the average fixation probability of an allele p_f. Under the assumptions that the population is diploid of size N and constant over time, the rate of mutations entering the population per unit time at a particular site

is $2N\mu$ [95]. Therefore the substitution rate is:

$$R = 2N\mu p_f. \tag{1.1}$$

In neutral site each single allele has an equal probability of fixation, for a diploid population of size N, the probability of a neutral allele, with a frequency f in the population, to drift to fixation is $f/2N$ [58]. Since mutation events are rare, when mutations arise, their initial frequency is $1/2N$. Therefore, if we substitute $p_f = 1/2N$ in Eq. (1.1) we get that

$$K = 2N\mu\left(\frac{1}{2N}\right) = \mu. \tag{1.2}$$

In words, the rate of substitutions at a specific (neutral) site is equal to the molecular mutation rate per individual in the population [78]. This fundamental result allows us to estimate the mutation rates by comparing genomes that evolved independently over millions of years and to infer from the differences in the sequence at neutral regions the actual mutation rates.

1.3 Genome wide mutation rates

Sources of mutations Most spontaneous single exchange base mutations occur during replication due to mispairing of bases. In addition, mutations are also the result of DNA damage that is not repaired by the repair machinery of the cell. There are multiple possible DNA damage types and multiple number of repair pathways that are specialized in repairing particular lesions [48]. The biochemistry of repair and mutagenesis has been extensively studied and most of the current knowledge is summarized in a book by Freidberg et al. [48]. But despite the impressive amount of work it is currently impossible to quantify the relative contribution of each of the processes to mutation rates.

Deamination This process converts cytosine to uracil and adenine to hypoxanthine. In mammals, the rate of cytosine deamination is 100 losses per double strand DNA per cell per day [47]. This rate is much higher for cytosine in single stranded DNA. For *E. coli*, the rate in ssDNA is estimated to be 140 fold higher than the rate in double stranded DNA [47; 88]. For yeast, the rate is about 4000 more frequent in ssDNA than in double-stranded DNA (dsDNA), since DNA is found in a ssDNA conformation for a large time during transcription [71]. Most of the time the DNA is in a dsDNA conformation but during transcription, replication, recombination and non-B DNA transient formation it is found in the form of ssDNA. It has been estimated that 500 C's are deaminated in a genome that is found at a mixture of 98% dsDNA and 2% ssDNA conformation per day per genome per mammalian cell [89; 47; 88; 139; 48].

This spontaneous deamination is corrected by the removal of uracil (not a natural part of DNA) by uracil glycosidase, generating an abasic site. The resulting abasic site is then recognized by enzymes, apurinic/apyrimidinic endonucleases, that cut the DNA and allow DNA to repair the lesion by replacement with another cytosine. However, the uracil might escape the repair and in the following round of replication its chemical similarity to thymine can lead to incorporation of A opposite to a U base leading to U:A base pair which is repaired to T:A.

Adenines are also deaminated but at lower rates than cytosines. The product of adenine is hypoxanthine, which selectively base pairs with cytosine instead of thymine. This results in a post-replicative transition mutation, where the original $A : T$ base pair is transformed into a $G : C$ base pair.

Direct and indirect measurement of mutation rates The rates of mutations can be derived by direct or indirect measures [162]. The most obvious and direct way to detect all the changes that recently happened in genomes is by sequencing DNA of mother and daughter cells or in germ-cells in the levels of single-cell. Although there has been enormous progress in the sequencing field, it is not ready for such kinds of tasks. The current knowledge about mutation rates is mostly based on: (1) detection of large numbers of genetic variants that are screened based on phenotypical effects of mutations [11; 162]. (2) Mutation accumulating (MA) lines is an approach that has received a major boost with the next-generation sequencing. In the MA approach mutations accumulate in the genome in several lines of species that originate from one ancestor, under laboratory conditions that reduce selection effect on the mutations to almost none [32; 97; 31]. (3) Comparing sequences that have diverged for many generations in regions which are supposed to be neutrally evolving and therefore the rates are expected to reflect the germ line mutation rates [70; 162]. In this case mutation rates are parameters of evolutionary models, which I will elaborate on in Chapter 2. (4) Lastly, mutation rates can be inferred from single nucleotide polymorphism (SNP) sites, which reflect the sampled genetic variation in a population [80; 162; 64]. The last three approaches are based on the idea that the vast majority of mutations that accumulate in the genome over time are neutral and therefore reflect the rate of neutrally occurring mutation process. One big difference is the time scale; MA lines experiments are conducted for a few generations, while the sequence divergence process of different species takes typically millions of years and lastly the diversity in a population level is established during several thousands of years. As a result, the number of mutations that can be used for rate estimation has wide range. Currently the number of mutations that are observed in MA lines is several thousands (for organism with relatively short generation time such as worms and yeast), while the number of SNPs in the human population is a few million (about 0.1% of sites in the human genome) while the density of diverged sites in genomes of two species can be much higher, for example, aligned seqeunces of human and chimpanze differ in more than 2% of the sites [29]. Until now, comparative genomics has been the most suitable to study high resolution mutational patterns. It seems that we are approaching the day that all hu-

man genomes could be sequenced and then de-novo SNPs may be used for estimating the current mutation rates in fine resolution.

Base exchange substitution rates The base mutation rate per generation in animals is in the range of 4×10^{-9} to 18×10^{-9} per base pair per generation (where the upper limit is for human). For *Saccharomyces cerevisiae* the estimated rate is 0.33×10^{-9} and in the prokaryote *Escherichia coli* it is approximately 0.26×10^{-9} [96]. This is not to say that the mutation rate per cell division is higher in humans than in other organisms. The reason for the higher mutation rate per generation is probably due to the fact that there are more DNA replication rounds during the production of human gametes compared to other species. Before a human gamete is produced there are about 216 cell divisions in the lineage that leads to the gamete compared to 36, 8.5, and 40 cell divisions during gametogensis in fly, worm and *Arabidopsis thaliana* [96]. Given the estimated rate for germline replication the mutation rate is only 0.06×10^{-9} per site per replication round [96]. This rate is lower than the somatic mutation rate of human cells which is in the range of 0.27×10^{-9} to 1.47×10^{-9} per site per cell division and it is even lower than the approximate germline mutation rates in flies, worms and *Arabidopsis thaliana* [96].

Factors that influence substitution rates Mutational processes are assumed to occur in a random fashion in the sense that mutations do not depend on their impact on the organism or the cell [58]. But the mutational processes are not random in the sense of being uniform along chromosomes and among species. The average genome wide mutation rates vary among species and between different cells. Moreover, the estimated rates of all types of mutations exhibit large heterogeneity across chromosomes. This is to be expected since mutations happen due to biochemical processes which are influenced by a large number of cellular factors [48]. Not only are mutations a consequence of the interaction of RNA polymerases and transcription associated factors, many repair proteins and mutagens, but also of their interaction with other metabolic and signaling processes [48]. The arsenal of repair proteins varies between species, their activity can be tissue specific, different repair processes are involved at different cell cycle stages [135].

The genomic environment of DNA also influences the mutagenic and repair processes. Centromere and telomere genomic regions have specific DNA polymerases and repair proteins. Epigenetic components such as DNA methylation, histone modifications, nucleosome density and DNA binding proteins impact the rate of occurrence of different types of DNA damage and also the efficiency of repair [48]. On top of that, the DNA sequence composition has different affinities to mutagens attacks and to repair proteins. Taken together this leads to a regional variation in mutational rates and spectra along the genome. When these mutation biases accumulate over hundreds of million of years they result in biases in nucleotide composition [95].

Regional variation in substitution rates along chromosomes In 1989, Wolfe, Sharp and Li [160] reported that substitution rates vary among different regions in

the genome. Since they analyzed sequences that are neutrally or nearly neutrally evolving, they concluded that at an evolutionary time scale there are regional variations in mutation rates. Genes which have similar substitution rates cluster together within chromosomes [86], but the effect is not a consequence of similarity in expression levels [85]. Later, by sliding window analysis and using repetitive sequences, Arndt, Hwa and Petrov [7], provided a spatial substitution rate map along the chromosomes (similar analysis is presented in Figure 1.1 but with multiple alignments of human, chimpanzee and rhesus intergenic sequences).

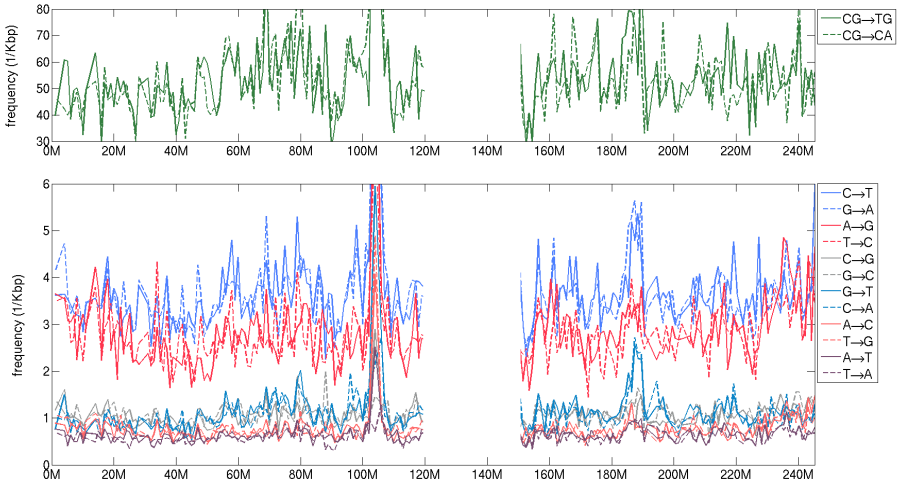

Figure 1.1: Profiles of substitution frequencies along human chromosome 1. Frequencies of 14 substitutions along chromosome 1 were estimated in non-overlapping 1 Mbp long windows slided along human chromosome 1. Only human intergenic sequences were included in this analysis. To estimate the rates I used human, chimpanzee, rhesus EPO primate alignments (Ensembl version 55). Substitution frequencies were estimated by maximum likelihood approach that is described in [36] and in Chapter 2. Similar analysis has been presented before in [7; 21]

1.4 Bias in substitution rates in double stranded DNA

Mutational spectrum If all single bases were subject to the same rates of replication errors, DNA damage and repair then all twelve single-nucleotide substitutions would occur at equal rates, but this is not the case [162]. In addition, the relative rates of different substitutions exhibit regional behavior [8].

Transition-Transversion bias It has become clear that the rate of change between arbitrary bases α and β is not equal for all different possibilities. The rate of change between nucleotides of similar chemical structures is several fold higher than

between non-similar nucleotides. Changes between similar nucleotides are called transitions and occur among the two purines (T and G) and among two pyrimidines (A and C). Changes between purine and pyrimidine are called transversions. There are eight possible transversions and four possible transitions; under the assumption that all mutations occur at the same rate one would expect to observe two fold more transversions than transitions. But this is not the case, and transition mutations are about 2-4 times more frequent than transversions in the vast majority of genomes including human [95] though not in *C. elegans* or *D. melonogaster* [96]. The biochemical explanation is not well understood. Most of mutations are suggested to be replication errors. Since base pairing is between a pyrimidine and purine the model suggests that DNA polymerases, the enzymes that copy DNA, are more likely to mispair incorrect purine to a template pyrimidine rather another pyrimidine and in the same manner, DNA polymerase is expected to insert more often a wrong pyrimidine opposite a purine. However, surprisingly, the error spectrum of DNA polymerases that participate in the bulk of DNA replication show that the rate of several purine-purine mismatches is higher than some of the purine-pyrimidine mismatches [101].

CpG bias CpG di-nucleotides are well known "hot-spots" in vertebrates for $C \to T$ and $G \to A$ transition mutations via the methylation-deamination process [48]. The cytosine base when positioned immediately 5' of a G on the same strand (denoted "CpG") is subject to a particular DNA modification adding a methyl group to the O-5 group of the cytosine resulting in 5-methyl-cytosine) [48]. In a similar fashion to the deamination of Cs to Us, 5-methyl-cytosines deaminate to T [48]. Uracil does not occur naturally in the DNA and there is a highly efficient repair process that corrects it back to the original base C [48]. On the other hand, T is a normally occurring DNA base and although there are mismatch repair pathways that deal with these types of mismatches (i.e. TpG paired to C^mpG), the repair process is less efficient than for $U : G$ mismatches [48]. Taken together, the higher deamination rate and the lower efficiency of the repair process make CpGs a hotspot for mutations [154; 48].

The result of these processes can be observed via comparative genomics, where in human the transition frequencies of cytosines and guanines in the context of CpG are more than 10 fold higher than genome-wide substitution rates of C or G residues in a non-CpG context [5]. It is estimated that one quarter of single-nucleotide differences between human and chimp are in CpG sites [29].

Strong to Weak bias The base pairing between G and C nucleotides is energetically more stable than between A and T bases due to three hydrogen bonds between G and C bases compares to only two bonds between A and T. Therefore, GC nucleotides are called Strong (S) bases while AT are the weak (W) bases. More than 40 years ago, Cox and Yanofsky [28] observed that in selected genomic loci of *E. Coli* strain, that harbors a mutator gene *mutT*, there is an excess in the rate of $W \to S$ mutations over $S \to W$. If the directionality in mutation towards GC bases is exerted for a long enough evolutionary time, it will induce genomes to become enriched in

GC nucleotides. Therefore, the base composition of genomes of different organisms can imply on the underlying mutational bias that has shaped these genomes. The prokaryotic genomes exhibit a wide range of GC content between 0.25 to 0.75 [162]. This imply that there is directionality in the mutation rates in prokaryotes but it is not universal

In contrast, most of eukaryotic genomes that have been sequenced so far are GC poor i.e. their GC nucleotides compose less than 50% of the total number of nucleotides in the genome [95]. This nucleotide composition reflects the corresponding universal excess of strong to weak mutations ($S \to W$) over weak to strong mutations ($W \to S$) in vertebrates, flies, nematodes and fungi as derived by comparative genomics, using polymorphism and mutation accumulating lines data [95; 96].

The biochemical source of this bias is again unknown. The common explanation is that this bias is due to hydrolytic and oxidative damage that targets mostly cytosine (deamination, see above) and guanine (oxidation [48; 95]) and leads to $C \to T$ and $G \to T$ mutations, respectively [48; 96]. The rate of the major processes that lead to weak to strong mutation is adenine deamination (which is discussed above), but it occurs at lower rates than cytosine deamination. Alternatively, DNA polymerases without proofreading tend to incorporate wrong bases opposite the template Gs or Cs about 2 fold more often than opposite As or Ts on the template strand [101].

GC biased gene conversion The ratio $r_{W\to S}/r_{S\to W}$ has been found to vary in the genomes and to be correlated with crossing-over rate [36]. As I mentioned above, in all animals genomes it is has been found that $r_{W\to S}/r_{S\to W}$ is less than one. However, comparative genomics revealed that the ratio between the substitution frequencies $W \to S$ and $S \to W$ varies and in some regions the ratio is greater than one. These fluctuations in the ratio of $W \to S$ frequencies over $S \to W$ are suggested not to be a result of mutations but rather a fixation bias which is called GC biased gene conversion (BGC) [51].

BGC occurs during recombination in regions where heteroduplexes hybrids between highly similar allele sequences is formed during meiosis [51]. DNA repair mechanisms acting on mismatches in the heteroduplex regions are suggested to correct mismatches between GC and AT bases more frequently towards GC bps than to AT bps. Under this model, at the population level, BGC mimics positive selection for higher GC content since the allele frequencies of strong bases are expected to increase at the population level from one generation to another while the frequency of alleles with weak bases is expected to decrease. Studies of SNP frequencies in human populations support selection-like models over the mutational bias model [158; 140].

CpG islands In human, the genome wide average of GC nucleotides is about 40% and the CpG odd ratio, which is the ratio between the observed number of CpG over the expected by the density of Cs and Gs, is 0.2. These are typical values for all mammalian geneomes that were analzyed so far and at some extent also in

other vertebrate genomes. But, in the vertebrate genomes, especially, in mammalian genomes there are regions which are extremely rich in CpG-dinucleotides and GC nucleotides. These regions are of typical length of several hundreds of bps and are composed with over than 50% strong bases and the CpG odd ratio over 0.6. These regions are called CpG islands (CGIs) [13; 14; 52; 145].

1.5 Bias in substitution rates in single stranded DNA

Chargaff's second parity rule The double stranded structure of DNA and Watson-Crick base pairing couples mutation processes of both strands. Every base α on one strand pairs with just one complementary base α' on the other strand: G pairs with C and A with T. Therefore, a mutation is between base pairs,i.e. a mutation of a base α to β on the forward strand, denoted by $\alpha \to_F \beta$, is a mutation $\alpha' \to_R \beta'$ on the reverse strand. The copy of a mutation from strand to its complementary strand done either by repair enzymes that ensure the right base pairing of the DNA or by replication if the mutation of one strand is not recognized by such enzymes. As a consequence, the substitution rate $\alpha \to_F \beta$ on the forward strand, denoted by $r_{\alpha \to_F \beta}$, is equal to the substitution rate of the complementary bases, α' and β' on the reverse strand: $r_{\alpha \to_F \beta} = r_{\alpha' \to_R \beta'}$. Under the assumption that mutations occur randomly on both strands and that the repair process is equally effective on both strands we further have that

$$r_{\alpha \to_F \beta} = r_{\alpha \to_R \beta}$$

and therefore

$$r_{\alpha \to_F \beta} = r_{\alpha' \to_F \beta'} \tag{1.3}$$

i.e. that complementary substitutions occur with the same rate on one strand [144; 90]. When the equality in Eq. (1.3) holds then in a single strand of DNA the complementary bases are found in equal frequencies the number of A's is equal to T's and number of G's is equal to the number of C's (this is called Chargaff's second parity rule) [144; 90]. However, this assumption and the symmetry of the substitution rates are not granted, since other cellular processes like transcription and replication can distinguish between the two strands. Over the past 20 years there has been accumulating evidence that violations of symmetries are the rule and not the exception [47; 90; 60; 148; 152]. Detecting asymmetries might be difficult since measuring the spectrum of mutations via experimental procedures is still laborious. Still, significant advances in the field were achieved though complete or partial sequencing of genomes and the use of comparative genomic approaches [90; 60; 70; 148; 152].

Strand asymmetries in bacterial genomes The first process that was shown to break the symmetry was transcription in bacteria. It was shown in the early 1990's in *Escherichia coli* that cytosine deamination rates on the non-template strand were increased as a consequence of ssDNA conformation of the non-template strand during

transcription [47]. This leads to a higher mutation rate of the type $C \to T$ over $G \to A$ on the non-transcribed strand. In 1985, it was found that lesions in expressed genes were repaired in a strand specific manner by the transcription coupled repair (TCR) process [16]. TCR has been postulated to lead to strand bias since the transcribed strand is preferentially repaired. This preference in repair will result in a strand-specific mutation spectrum.

A few years later, in the mid-1990's Lobry [90] suggested that in bacterial genomes, replication was associated with violations of Chargaff's second parity rule. As I mentioned above the rule states that if the mutational processes do not distinguish between the two DNA strands then there should be an equal frequency of complementary bases in a long piece of ssDNA, that is, $[A] = [T]$ and $[C] = [G]$ [90]. To access strand symmetries, it proposed to define the quantities of TA skew ($(S_{TA} = ([T]-[A])/([T]+[A]))$) and GC skew ($S_{GC} = ([G] - [C])/([G] + [C])$) in a single strand of DNA. Deviations of TA and GC skews from zero imply that mutational processes distinguish between the two DNA strands [91]. A positive GC skew and a nonzero TA skew are often used in bacteria to detect the leading strand [44]. A switch in the sign of a skew is indicative of an origin of bidirectional replication (OBR) or the terminus of replication [105; 115; 153].

The increasing amount of sequenced genomes enables us to derive a mutation rate approximation using multiple alignments. The results of these rate estimations were striking, because, even though the studies of skews revealed that genomes share similar skews, comparative genomic analyses revealed that the underlying single nucleotide substitution rates vary significantly between taxa. In each taxon at least one symmetry of nucleotide substitution rates is broken. Often, there is an excess of mutations $C \to T$, $A \to G$, and $C \to G$ on the leading strand compared with the complementary mutations $G \to A$, $T \to C$, and $G \to C$, respectively. However, there is not one symmetry of nucleotides that is broken in all taxa [131].

Several mechanisms have been suggested to explain the strand asymmetries in reverse complement mutation rates. Lobry [91] suggested that the asymmetries are a result of the difference in the replication of two DNA strands, the leading and lagging strands. It has been suggested that the discontinuous replication of the lagging strand by the Okazaki fragments increases the frequency with which the template of the lagging strand is in a single-strand conformation relative to the template of the leading strand, which is replicated continuously [44]. Because nucleotides on ssDNA are more prone to be damaged (in particular by adenine and cytosine deamination), this can induce differences in the mutational spectrum between the two strands [47]. Alternatively, the difference between the mutational spectra of the two DNA strands can be a result of replication of the two DNA strands by two different polymerases [84] that have different error rates [83; 101].

Strand asymmetries coupled to replication in human Studies of strand asymmetries of mutation rates in mammalian genomes are not as far advanced as in bacte-

ria. The main reason was related to difficulty of estimation of mutation rates by the methods used for bacteria. First, directly measuring of mutation rates is very difficult due to the size of the mammalian genomes which are three orders of magnitude larger than the genomes of bacteria. Second, the rate of mutations is ten-fold lower in mammals compared to bacteria. Third and last, the rate of DNA replication is also several orders of magnitude lower in human than in bacteria.

Direct measurement of nucleotide strand asymmetries of mutation rates in human has been done previously. Asymmetries which are associated with replication, have recently been found in humans and profiles of TA and GC skews have been used to identify putative origin of replications (ORIs) along human chromosomes; Touchon et al. [152] analyzed six well-studied ORIs in the human genome. Focusing on the regions on the $5' \rightarrow 3'$ strand centered at the ORI, they detected a change from a negative GC and TA skew in the 5' flanking region of the ORI to positive skews at the downstream regions, which corresponds to results from *Escherichia coli*. By detecting similar changes in compositional skews along human chromosomes, 1,000 putative ORIs (among them 280 are OBRs) were identified *in silico*, 30-fold more than the number of ORIs that were known at that time [69]. The observed TA and GC skews in the flanking regions of human ORIs imply that there are underlying strand-specific substitution patterns that are coupled to replication in the human genome.

Strand asymmetries associated with transcription in higher eukaryotes Biases in the nucleotide composition of ssDNA have been used as evidence for bias in mutational processes or selection. Green et al. [60] have observed an excess of GT nucleotides over AC nucleotides on the nontemplate strand in human genes, which was later found to be correlated with transcription levels in testis [99; 157]. TA and GC skews are found to be positive in mammalian nontemplate strands, but are close to zero in the 5' flanking sequences of genes [149]. Skews also show a regional pattern found along the transcribed DNA [92; 149]. The TA and GC skews are found to be maximal in the regions immediately downstream of the transcription start site (TSS) of genes. Nucleotide densities are furthermore observed to be dependent on the distance from the 5'end and 3'end of introns [150; 1; 49]. This regional behavior of the nucleotide composition in genes and their flanking regions implies corresponding regional behavior of substitution processes. Substitution asymmetries were also observed between the template and non-template strand. For instance, using 1.5 Mbp of orthologous regions in chimpanzee and human, Green et al. [60] calculated the rates of intronic nucleotide transitions, and found that in the nontemplate strand the purine transitions ($A \rightarrow G$ and $G \rightarrow A$) occurred at a higher rate than their complementary pyrimidine transitions. Later, a similar bias of $A \rightarrow G$ over $T \rightarrow C$ was reported in an analysis of SNPs in introns and four fold degenerates (FFD) sites [125]. This strand bias in substitution rates has been hypothesized to be a result of TCR and of different misinsertion rates for the four nucleotides [60]. Subsequently, Hwang and Green [70] carried out an analysis of context dependence of mutation rates in 1.7-Mbp genomic regions across the phylogeny of 19 mammalian species. Beside the known asymmetries

of complementary transition rates in transcribed regions, they also found similar (but weaker) asymmetries in the transversion rates. Some of these processes also showed signatures of neighbor dependencies increasing the impact of strand asymmetry to context dependent mutations.

1.6 Thesis overview

This dissertation work evolves around the interplay of cellular processes with mutational processes in mammals. As it is discussed in this chapter, the nucleotide composition profiles along genes and across CpG islands point to regional mutational forces, that have been acting on these regions. However the underlying substitution patterns that have shaped these regional nucleotide composition profiles are unknown.

Although the first draft of the human genome was released ten years ago [23], substitution rates could not be studied from this data alone, and could only be inferred after the release of mouse and rat genomes [24; 54]. The major advance in the study of human substitution patterns happened with the release of species closely related to human such as chimpanzee and macaque genomes in 2005 and 2007, respectively [29; 25]. We use comparative genomic approach to estimate the substitution rates around mammalian genes and their flanks and along human CpG islands and their intergenic flanks (the methods are described in Chapter 2).

We discuss several regional patterns of substitution in the mammalian genomes. Chapter 3 deals with a drop in methylation-deamination rates at the 5'ends of human genes and at CpG islands. In Chapter 4, a comparison of the substitution rate profiles in mammalian introns with the ones in mammalian intergenic regions reveals two types of transcription-associated asymmetries; first, strand asymmetries that start at the TSS and extend along the whole gene and beyond the 3'end, second, a localized strand asymmetry that is limited to the first 2 kbp of mammalian transcripts. The following Chapter 5 concerns intergenic asymmetries that originate at CpG islands and extend for hundreds of thousands of bps. I will discuss the possibility that these asymmetries indicate that CpG islands are origins of replication or of unknown transcripts. Chapter 6 deals with strong to weak bias in mammalian promoter regions and CpG islands. This bias seems to vary among promoters of different mammalian species. In addition, this ratio varies among different types of human CpG islands, probably due to differences in recombination rates among the different CpG island classes.

2 Methods

In this thesis, we wish to study the dependence of neutrally substitutional spectrum and frequencies in different genomic context. In other words we want to further understand the factors that have been shaping the rates of $\alpha \to \beta$ substitution i.e $r_{\alpha\to\beta}$ (where $\alpha, \beta \in \{A, C, G, T\}$) in mammalian evolution, particularly, in human since the specie diverged from chimpanzee. The main focus in the this thesis is on the dependence of single nucleotide substitution rates on the distances to transcription starting sites of genes or to CpG islands and on the dependency of the substitution spectrum on the DNA strand. Using comparative genomic approach, the substitution rates in neutrally evolving regions are inferred from multiple alignments that correspond to these regions [36].

2.1 Inferring substitution rates

To get an estimation of the neutral mutation rates we rely on the fundamental theorem of molecular evolution: Substitution rates (at the population level) are equal to the mutation rates (for an individual)in neutrally evolving sites at a constant population size (Eq. 1.2) [78]. The mathematical models for sequence evolution describe evolution as a Markov process in which states are DNA sequences i.e. at any given time the rate of mutation depends only on the genome sequence at that time point and not on the past. We can use sequence evolution models to infer the rates of mutations from multiple alignments. This approach which has been referred to as a comparative genomic approach has been proven to be very powerful in revealing novel mutational patterns along the chromosomes [70; 35]. Combining comparative genomic and biochemical knowledge should probably lead to a better understanding of the current molecular processes that have shaped our genome. In this section I will explain the sequence evolution model that is used in this thesis to infer substitution rates from multiple alignments, which is presented in a recent paper by Duret and Arndt [36].

Sequence evolution Let us consider a DNA sequence $\vec{\alpha}(0)$ at time $t = 0$ of a fixed length S

$$\vec{\alpha}(0) = (\alpha_1(0), \ldots, \alpha_S(0)) \tag{2.1}$$

where $\alpha_i(0) \in \{A, C, G, T\}$ and $i \in \{1, \ldots, S\}$.

Furthermore, let us assume that all sites are neutral. This means that for any other individual carrying a sequence

$$\vec{\beta}(0) = (\beta_1(0), \ldots, \beta_S(0))$$

where $\beta_i(0) \in \{A, C, G, T\}$ and $i \in \{1, \ldots, S\}$, there is no difference in the fitness compared to an individual that carries sequence $\vec{\alpha}(0)$. This suggests that in the space of all possible DNA sequences of length S which is of size 4^S, all states are equiprobable.

This sequence changes over time due to a stochastic process called mutation and therefore at time $t > 0$ the sequence

$$\vec{\alpha}(t) = (\alpha_1(t), \ldots, \alpha_S(t)) \tag{2.2}$$

might be different from $\vec{\alpha}(0)$. The probability to observe $\vec{\alpha}(t)$ at time t is dependent on the initial sequence i.e. we have

$$\Pr(\vec{\alpha}(t)|\vec{\alpha}(0)) \tag{2.3}$$

is a function of $\vec{\alpha}(0)$.

At this point it seems natural to represent a sequence evolution by a chain of random variables. Let $\{A(t),\ t \geq 0\}$ be a continuous time chain of random variables where t is a continuous time index and $A(t) = (A_1(t), \ldots, A_S(t))$ represents the random variable over the space of sequences of length S at time t e.g. $\vec{\alpha}(t)$. With these notations equation (2.3) read as $\Pr(\vec{\alpha}(t)\,|\vec{\alpha}(0)) = \Pr(A(t) = \vec{\alpha}(t)\,|A(0) = \vec{\alpha}(0))$.

Many models for sequence evolution assume that sites along the sequence evolve independently from one another in a context free manner, in other words we can write

$$\Pr(A_i(t_2)\,|A(t_1)) = \Pr(A_i(t_2)\,|A_i(t_1))$$

where $t_2 > t_1$ are times in evolution.

Under the conditions of neutrality and independence we treat genome evolution as S independent chains of random variables i,e. $\{A_i(t)\}_{i\in\{1\ldots S\}}$ and therefore the model can be seen as a single-site evolution process each having four states corresponding to the four nucleotides.

The next natural assumption is that all chains $\{A_i(t),\ t \geq 0\}$ hold the Markovian property i.e for each $t_1 \geq 0$ and $t_2 \geq 0$ such that $t_1 < t_2$ the following holds:

$$\Pr(A_i(t_2) = \alpha_i(t_2)\,|\{A_i(t),\ t \leq t_1\}) = \Pr(A_i(t_2)\,|A_i(t_1))\,.$$

Therefore, we just defined single site evolution as a continuous Markov chain and the sequence evolution as multiple Markov chains

Time homogeneity From practical reasons that we will explain below, the Markov chain is time homogeneous, i.e at each each site, for any given times $t_1, t_2 \geq 0$, we have

$$\Pr\left(A_i(t_2+\Delta t)=\beta\,|A_i(t_2)=\alpha\right)=\Pr\left(A_i(t_1+\Delta t)=\beta\,|A_i(t_1)=\alpha\right).$$

In other words, the probability to change one nucleotide into another within a time frame $\Delta t > 0$ is independent of time.

We can therefore define the transition matrix $P_i(\Delta t)$

$$P_{i,\beta\alpha}(\Delta t)=\Pr(A_i(\Delta t)=\beta|A_i(0)=\alpha)$$

without loss of generality we can assume that $t_1 = 0$ and therefore we replace Δt with t.

Ergodicity We call a Markov process ergodic if for every β and α there exist finite $t > 0$ such that:

$$P_{i,\alpha\beta}(t) > 0.$$

This means that it is possible to move from every nucleotide to every nucleotide in a finite time.

Substitution rate matrices Let $\vec{p_i}(t) = (p_{i,A}(t), p_{i,C}(t), p_{i,G}(t), p_{i,T}(t))$ be the probabilities to find a particular nucleotide at position i at time t. By definition we have $\sum_{\alpha\in\{A,C,G,T\}} p_{i,\alpha} = 1$ and $\vec{p_i}(0)$ is the initial distribution given by

$$p_{i,\alpha}(0)=\begin{cases}1 & \text{if } \alpha=\alpha_i(0)\\ 0 & \text{otherwise}\end{cases}.$$

Assuming time homogeneity of the Markov process implies that transition probability between two distinct nucleotides $\alpha, \beta \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ is constant over time. Then for small enough Δt

$$P_{i,\alpha\beta}\left(\Delta t\right)=\Pr\left(A_i(t+\Delta t)=\beta\,|A_i(t)=\alpha\right)=I+Q^{(1)}_{i,\beta\alpha}\Delta t+O(\Delta t) \tag{2.4}$$

where

$$Q^{(1)}_{i,\beta\alpha}=\begin{cases}r_{i,\alpha\to\beta} & \text{if } \alpha\neq\beta\\ -\sum_{\beta\neq\alpha} r_{i,\alpha\to\beta} & \text{if } \alpha=\beta\end{cases} \tag{2.5}$$

$r_{i,\alpha\to\beta}$ are the positional substitution rates from a nucleotide α to β and $Q^{(1)}_i$ is called the substitution rate matrix.

For the probabilities vector the dynamics for a small Δt are described by:

$$p_{i,\alpha}(t+\Delta t) = p_{i,\alpha}(t) - \sum_{\beta\neq\alpha} p_{i,\alpha}(t) r_{i,\alpha\to\beta}\Delta t + \sum_{\beta\neq\alpha} p_{i,\beta}(t) r_{i,\beta\to\alpha}\Delta t + O(\Delta t) \tag{2.6}$$

where the second term in the sum in (2.6) is the amount of nucleotides α that are substituted by a different nucleotide and therefore are removed from the frequency of $p_{i,\alpha}(t)$, while the last term is the number of nucleotides that have changed to α. Notice that under this dynamic the sum of the probabilities stays constant i.e. 1

$$\begin{aligned}
\sum_{\alpha} p_{i,\alpha}(t+\Delta t) &= \sum_{\alpha} p_{i,\alpha}(t) + \sum_{\alpha}\sum_{\beta\neq\alpha} \left(p_{i,\beta}(t) r_{i,\beta\to\alpha}\Delta t - p_{i,\alpha}(t) r_{i,\alpha\to\beta}\Delta t\right) \\
&= \sum_{\alpha} p_{\alpha}(t) + \sum_{\alpha}\sum_{\beta\neq\alpha} \left(p_{i,\alpha}(t) r_{i,\alpha\to\beta} - p_{i,\alpha}(t) r_{i,\alpha\to\beta}\right)\Delta t \\
&= \sum_{\alpha} p_{i,\alpha}(t) \\
&= 1.
\end{aligned} \tag{2.7}$$

The representation of Equation (2.6) in a matrix formalization is

$$\vec{p_i}\,(t+\Delta t) = \vec{p_i}(t) + Q_i^{(1)}\vec{p_i}(t)\Delta t \tag{2.8}$$

for small Δt.

Master equation Equation (2.8) can be written as a differential equation which is often called *Master Equation*:

$$\frac{\partial}{\partial t}\vec{p_i}(t) = Q_i^{(1)}\vec{p_i}(t) \tag{2.9}$$

with the solution

$$\vec{p_i}(t) = e^{Q_i^{(1)}t}\vec{p_i}(0) \tag{2.10}$$

We can also calculate the transition probability matrix $P_i(t)$ by solving the matrix form of the Master Equation (2.9):

$$\frac{\partial}{\partial t}P_i(t) = Q_i^{(1)}P_i(t) \tag{2.11}$$

whose solution is:

$$P_i(t) = e^{Q_i^{(1)}t} P_i(0)$$

where $P_i(0)$ is the 4×4 identity matrix and therefore:

$$P_i(t) = e^{Q_i^{(1)}t}.$$

Stationary distribution Due to the assumption that our sequence evolution model is an ergodic Markov process, it converges to a unique stationary state. For a equilibrated system the nucleotide composition is in a steady state and does not change over time. The stationary nucleotide composition is:

$$\lim_{t\to\infty} \vec{p}_i(t) = \vec{\pi}_i = (\pi_{i,A}, \pi_{i,C}, \pi_{i,G}, \pi_{i,T})^T$$

where $\vec{\pi}_i$ is the stationary distribution over the four DNA nucleotides. This distribution has the property that:

$$Q_i^{(1)} \vec{\pi}_i = 0$$

One can test the stationarity of DNA sequences in evolution by comparing the current nucleotide composition with the stationary one. For example, by similar comparison it has been shown that human and flies genomes are not in equilibrium [6; 141]. The non-stationarity of nucleotide composition can be explained by the fact that mutational rates and biases might have changed during evolution even in a regional way. For example, a source of mutational bias such as recombination hotspots have been shifted in evolution [124], transcriptional and methylation levels have been changed between homologous genes in different species [76; 39]. Since methylation, transcription and replication, recombination have a specific mutational signature, then mutation spectrum and levels might also change (see Chapter 1). When not enough time has passed from the time the mutational spectrum has changed, the nucleotide composition in a region does not reflect the current mutation process. The latter has been demonstrated recently by Duret and Arndt with respect to rapid changes in recombination hotspot density along the human genome [36]. Therefore, the stationary state better reflects the current mutational biases.

Combination of substitution rate matrices So far we treated sequence evolution as multiple Markov chains by substitution rate matrices $Q_i^{(1)}$. In reality, even under neutrality assumptions, different sites along the sequence evolve at different rates because multiple factors shape mutation rates such time of replication, crossing over rates, GC content, distance to telomere, expression levels or so [21]. Under different conditions, denoted by $\{F_k\}$, there are different rate matrices $Q_{F_k}^{(1)}$. The rate matrix $Q_i^{(1)}$ for a site i can be written as a linear combination of these different

rate matrices in case the different features do not contribute to mutation rates in a corporative manner:

$$Q_i^{(1)} = \sum_k a_i^k Q_{F_k}^{(1)}$$

where a_i^k is a coefficient that describes the contribution the features F_k to substitution rates in a site i. Since these coefficients can practically be any non-negative number all sites can evolve under different substitution rate matrices i.e. $\{Q_i^{(1)}\}$.

If we know the rates at each site it is not a computational problem to let the sequence evolve according to these rates. However, many times the rates are unknown but the sequences at distinct time point are given that with without a loss of generality are denoted by $\vec{\alpha}(0)$ and $\vec{\alpha}(1)$. The computational task is to infer rates by maximizing the likelihood to observe the event $\vec{\alpha}(0) \to \vec{\alpha}(1)$

$$\Pr\left(\vec{\alpha}(1) \middle| \vec{\alpha}(0), \{Q_i^{(1)}\}\right).$$

There are $S \times 12$ parameters to infer from only S sites, without no further information, that reduce the number of parameters, we can not rely on such estimations.

When $\{a_i^k\}$ are given the number of parameters that need to be estimated decreases

$$\Pr\left(\vec{\alpha}(1) \middle| \vec{\alpha}(0), \{\sum_k^K a_i^k Q_{F_k}^{(1)}\}\right)$$

and we only need to estimate the set of parameters $\{Q_{F_k}^{(1)}\}$ where K is the number of features and number of parameters is $K \times 12$ and S sites . The reliability (confidence) in the estimate is dependent on the ratio between S and $K \times 12$.

Sequence evolution as independent and identically distributed (i.i.d.) random variables In practice also $\{a_i^k\}$ are not given. And therefore, in this thesis we make another simplification and we consider all sites to evolve according to the same substitution rate matrix $Q^{(1)}$, i.e. we assume the $(A_i(t))$ to be i.i.d. From now on, we omit the subscript i that indicates a position and we treat the rate matrices as $\{Q^{(1)}\}$

where

$$Q_{\beta\alpha}^{(1)} = \begin{cases} r_{\alpha\to\beta} & \text{if } \alpha \neq \beta \\ -\sum_{\beta\neq\alpha} r_{\alpha\to\beta} & \text{if } \alpha = \beta \end{cases} \tag{2.12}$$

and $r_{\alpha\to\beta} \geq 0$.

It is easy to see that

$$\begin{aligned}\Pr\left(\vec{\alpha}(1)\,\middle|\vec{\alpha}(0),\{Q^{(1)}\}\right) &= \prod_{i=1}^{S}\Pr\left(\alpha_i(1)\,\middle|\alpha_i(0),Q^{(1)}\right)\\ &= \prod_{\alpha}\prod_{\beta}\Pr\left(\beta\,\middle|\alpha,Q^{(1)}\right)^{N(\alpha\to\beta)}\end{aligned}\tag{2.13}$$

where α and β are one of the four nucleotides and $N(\alpha \to \beta)$ is the number of sites along the sequences $\alpha_i(0) = \alpha$ changes to $\alpha_i(1) = \beta$.

We will return to this likelihood function (2.13) further below. With this formulation (or assumptions), inferring substitution rates is easier since we only need to know is only $\vec{\alpha}(0)$ and $\vec{\alpha}(t)$ in order to estimate the substitution rates.

Context dependency Until recently, most of the DNA evolution models do not assume dependency in the sequence context. For example, the probability that C changes into T is independent of whether C is flanked by G or by A. However, this assumption is inadequate for C and a G in a context of a CpG that are substituted 10 fold faster to T or A, respectively, than in a non-CpG context [5]. The CpG bias or CpG methylation deamination process has been described in Chapter 1 and occurs due to cytosine methylation in the mammalian genome. If one drops this assumption and considers neighbor dependencies then for di-nucleotides the probability of a substitution of a nucleotide α_0 in a nucleotide β_0 is dependent on on its flanking nucleotides α_{-1} and α_{+1} i.e. the probability is determined by the triplet $(\alpha_{-1}, \alpha_0, \alpha_{+1})$. The dependency on the -1 and $+1$ neighbor suggests the description to model substitution from a triplet to a triplet

$$Q^{(3)} = Q^{(1)} \otimes I \otimes I + I \otimes Q^{(1)} \otimes I + I \otimes I \otimes Q^{(1)} + I \otimes Q^{(2)}_{CpG} + Q^{(2)}_{CpG} \otimes I \tag{2.14}$$

where, I is 4×4 identity matrix and $\otimes$ is the Kronecker tensor product. The first three terms in Eq. (2.14) represent the neighbor independent single nucleotide substitution process on the the three sites. The last two terms in (2.14) represent additional neighbor dependent contributions to the dynamic. For the CpG process the 16×16 matrix $Q^{(2)}_{CpG}$ is given by

$$\left(Q^{(2)}_{CpG}\right)_{(\alpha,\beta,\alpha',\beta')} = \begin{cases} r_{CpG\to\alpha' pG} & \text{if } (\alpha\beta) = (CG),\ \alpha' \neq C \text{ and } \beta' = G \\ r_{CpG\to Cp\beta'} & \text{if } (\alpha\beta) = (CG),\ \alpha' = C \text{ and } \beta' \neq G \\ 0 & \text{otherwise} \end{cases} \tag{2.15}$$

for $(\alpha, \beta, \alpha', \beta') \neq$ (CGCG) and

$$(Q^{(2)}_{CpG})_{(CGCG)} = -\sum_{\alpha' \neq C} r_{CpG \to \alpha' pG} - \sum_{\beta' \neq G} r_{CpG \to Cp\beta'}$$

This encodes the modeling of the transitions and transversions from CpG.

Under this triplet model, the probability of finding a triplet $\beta_{-1}\beta_0\beta_{-1}$ is given by the master equation:

$$\frac{\partial}{\partial t} p_{\beta_{-1}\beta_0\beta_{-1}}(t) = \sum_{\alpha_{-1}\alpha_0\alpha_{+1}} Q^{(3)}_{\beta_{-1}\beta_0\beta_{-1},\alpha_{-1}\alpha_0\alpha_{+1}} p_{\alpha_{-1}\alpha_0\alpha_{+1}}(t)$$

In a compact way it can be written as

$$\frac{\partial}{\partial t}\vec{p} = Q^{(3)}\vec{p} \tag{2.16}$$

where $\vec{p}$ is a vector representation of all possible triplets. Its entries are $p_{\beta_{-1}\beta_0\beta_{+1}}$, the probability for a triplet β in a site.

Equation (2.16) is an ordinary differential equation and its solution is given by

$$\vec{p}(t) = P^{(3)}(t)\vec{p}(0)$$

Where $\vec{p}(0)$ is the initial distribution of triplets at time 0. As before,

$$P^{(3)}(t) = e^{Q^{(3)}t} \tag{2.17}$$

The 64×64 matrix $P^{(3)}$ encodes the frequencies or the probability that a triplet $\alpha_{-1}\alpha_0\alpha_{+1}$ is substituted with another triplet $\beta_{-1}\beta_0\beta_{+1}$ after time t.

The context evolution models can be extended for more context dependent processes besides CpG methylation deamination processes. Indeed, such extensions have been used by Hwang and Green [70]. However, Arndt and Hwa [6] have shown that using other processes beside methylation deamination for the genome-wide estimation do not significantly change the estimated single nucleotide rates. Inclusion of more rates, however, do come with the price of an increased amount of data that is needed for rate inference and of course computational time [36].

Substitution model for a tree The current view on evolution suggests that all DNA sequences evolved from one ancestral sequence and that they are related by some phylogenetic relations. This can be described by a phylogentic tree, a branching diagram showing the order of speciation events.

As a model, we start from an initial sequence, that is in the root of the tree that bifurcate (i.e. that from a node there are only two branches going out)into two different sequences which might also bifurcate into additional sequences. In a general case, the tree is composed of M leafs representing M genomic sequences, $\vec{\alpha}^{N+1}, \ldots, \vec{\alpha}^{N+M}$, of present day species, where N is the number of genomic sequences in the internal nodes $\vec{\alpha}^1, \ldots, \vec{\alpha}^N$. These sequences evolved from a sequence in the root $\vec{\alpha}^0$. As before, two sequences, an ancestral, $\vec{\alpha}_i$ and a daughter sequence $\vec{\alpha}_j$, which are the nodes of branch (i, j), are assumed to have evolved according to a substitution rate matrix $Q_{(i,j)}$.

Practical assumptions on substitution models There are two additional assumptions that have been made and are still used by the vast majority of the community, especially among communities that process limited amounts of sequences per species or that analyze hundreds of species at a time.

Time homogeneity over a tree - (the molecular clock is constant) The rate matrix does not change over time along each branch and it is the same along the branches. This assumption is relaxed by us and others, we allowed different rate matrices in all branches

Time-reversible models and stationarity assumption. For a time reversible model, there is no assumption that substitutions preferentially change in certain directions over time. The conditions for time reversibility is the detailed balance equation

$$\pi_\alpha r_{\alpha\to\beta} = \pi_\beta r_{\beta\to\alpha}$$

for some vector of probabilities $\vec{\pi}$ [141]. This assumption is very useful for the computation of likelihood functions (see in the following subsection).

Historical overview of substitution models The first model was proposed by Jukes and Cantor in 1969 [73] which assumes that all nucleotides change into another at the same rate. Eleven years later, Kimura [79] suggested that transition and transversions should be distinguished and introduced the so-called Kimura two parameter models. One year later, based on the stationarity assumptions, Felsenstein [41] modeled sequence evolution by substitution rate matrices whose entries correspond to the stationarity distribution of nucleotides which are allowed to be non uniform. In 1985, Hasegawa, Kishino and Yano (HKY) [65] suggested a model that combine the Kimura and Felsenstein models. The Generalized Time Reversible (GTR) model which is the most general neutral, independent, finite-site, time-reversible model has 9 independent parameters.

Two teams have omitted time reversible and time homogeneity assumptions [5; 70]. This formalization is very useful, since the twelve substitution rates are decoupled from each other. This is the only way to measure substitution rates and searching with biases in sequences in an unbiased manner. In this thesis, I use a model developed by Arndt [5; 36], that does not assume time reversibility and does not assume a single

substitution rate matrix over the whole tree but rather different rate matrices for the all different branches.

The probability of a model Let us assume that we are given a phylogenetic tree, T with a set of branches $E(T)$ and that all the sequences in its nodes are given. Let us also assume that we are given the substitution rate matrices $Q_{(i,j)}$ for the sequence evolution model along each one of the tree's branches (i, j). Then we can compute the probability L of the sequences given the model and the tree

$$L = \Pr(\vec{\alpha}^0, \ldots, \vec{\alpha}^{M+N} | \left\{ Q_{(i,j)} \ \forall (\mathrm{i,j}) \in E(T) \right\})$$

where (i, j) is a branch in the tree T.

The above equation can be rewritten as:

$$L = \Pr(\vec{\alpha}^0, \ldots, \vec{\alpha}^{M+N} | \{ Q_{(i,j)} \ \forall (\mathrm{i,j}) \in E(T) \}) = \Pr(\vec{\alpha}_0) \prod_{(i,j) \in E(T)} \Pr(\vec{\alpha}^j | \vec{\alpha}^i, Q_{(i,j)})$$

where $\Pr(\vec{\alpha}^0)$ denotes to probability to have a sequence $\vec{\alpha^0}$ in the root and the product is taken over all the possible branches. For a given set of sequences, the probability $\Pr(\vec{\alpha}^j | \vec{\alpha}^i, Q_{(i,j)})$ for evolution on a branch is the probability to see sequence $\vec{\alpha}^j$ in node j if the sequence in the parental node i is $\vec{\alpha}_i$ and the the sequence evolves under the substitution rate matrix $Q_{(i,j)}$.

However, in many cases we retrieve only the M sequences in leafs. Therefore, the likelihood L for a set of substitution matrices $Q_{(i,j)}$ must be a summation over all the possible configurations of the sequences in the N internal nodes

$$L = \sum_{\vec{\alpha^0}, \vec{\alpha^1}, \ldots, \vec{\alpha^N}} \Pr\left(\vec{\alpha}^0\right) \prod_{(i,j) \in E(T)} \Pr\left(\vec{\alpha}_j \left| \vec{\alpha}_i, Q_{(i,j)} \right.\right). \tag{2.18}$$

Maximization of the likelihood function The likelihood function (2.18) that has been introduced above is used to estimate substitution frequencies. Given multiple alignments, we can estimate the substitution frequencies as the frequencies that maximize the likelihood function in (2.18).

In context independent models, the sites along sequences evolve independently of each other. Therefore, the likelihood is factorized and can be written as

$$L = \prod_{k=1}^{S} \sum_{\alpha_k^0, \alpha_k^1, \ldots, \alpha_k^N} \Pr\left(\alpha_k^0\right) \prod_{(i,j) \in E(T)} \Pr\left(\alpha_k^j \left| \alpha_k^i, Q_{(i,j)} \right.\right)$$

where $\Pr(\alpha_k^0)$ is the nucleotide distribution at the root node at position k.

The maximization of this likelihood function is done by Powell's algorithm as in [36].

The neighbor dependent case For the neighbor dependent case, the sites do not evolve independently of each other and therefore the likelihood function does not factorize. This leads to computational problems in estimating the rates. For this Peter Arndt applies the Monte-Carlo Maximum-Likelihood (MCML) to maximize the likelihood function. The algorithm is an iterative Expectation-Maximization approach with E- and M-steps. In the M-step we estimate the substitution frequencies along the tree for given ancestral sequences in internal nodes. In the E-step we update the ancestral sequences in internal nodes.

In the M-step, we maximize the likelihood function given in (2.18). Since we use the sequences in the all nodes, the maximization is done separately for all branches. The likelihood for context dependent models between two sequences, ancestral $\vec{\alpha}$ and daughter $\vec{\beta}$ is given by:

$$\begin{aligned} L = \Pr\left(\vec{\beta}\,|\vec{\alpha}\right) &\approx \prod_{i=2}^{S-1} \Pr\left(\cdot\beta_i\cdot\left|\alpha_{i-1}\alpha_i\alpha_{i+1}, Q_{(i,j)}\right.\right) \\ &= \sum_{\alpha_{-1}\alpha_0\alpha_{+1}\beta_0} \Pr\left(\cdot\beta_i\cdot\left|\alpha_{i-1}\alpha_i\alpha_{i+1}, Q_{(i,j)}\right.\right)^{N(\alpha_{-1}\alpha_0\alpha_{+1}\to\cdot\beta_0\cdot)} \end{aligned}$$

where $\alpha_{-1}\alpha_0\alpha_{+1} \to \cdot\beta_0\cdot$ is the event that a nucleotide α_0 flanked by 5' neighbor nucleotide α_{-1} and 3' neighbor nucleotide α_{+1} has changed to nucleotide β_0 and $N(\alpha_{-1}\alpha_0\alpha_{+1} \to \cdot\beta_0\cdot)$ is the observed number of triplets $\alpha_{-1}\alpha_0\alpha_{+1}$ in ancestral sequence that change to $\cdot\beta_0\cdot$. The probability of such events is given by

$$\Pr\left(\cdot\beta_i\cdot\left|\alpha_{i-1}\alpha_i\alpha_{i+1}, Q_{(i,j)}\right.\right) = \sum_{\beta_{i-1},\beta_{i+1}} \Pr\left(\beta_{i-1}\beta_i\beta_{i+1}\left|\alpha_{i-1}\alpha_i\alpha_{i+1}, Q_{(i,j)}, t\right.\right) \qquad (2.19)$$

where $\Pr\left(\beta_{i-1}\beta_i\beta_{i+1}\left|\alpha_{i-1}\alpha_i\alpha_{i+1}, Q_{(i,j)}, t\right.\right)$ is an element of the transition matrix in Eq. (2.17), without loss of generality, we take $t = 1$ in Eq. (2.19).

Substitution analysis in this thesis We used the maximum likelihood approach which has been explained above. It is able to reliably estimate substitution frequencies from given aligned sequences along all branches that are not directly connected to the root node (the node that represents the last common ancestor of all species in a given tree). In contrast to many models, we assume that the molecular clock might differ between different branches. Therefore, we estimate the 14 or 18 substitution frequencies along each branch of the phylogeny. Another difference between our model and other commonly used sequence evolution models is that we do not assume that the DNA nucleotide composition is at equilibrium. These relaxations make our model very general in comparison with other models.

We measure substitution frequencies per base pair, estimating the (fractional) number of nucleotide exchanges from one nucleotide to another along each branch of the phylogeny. We may compute the corresponding substitution rates (measured per bp and time) by dividing frequencies by the time that passed along a branch. However, in this theses we are interested in the spectra of substitutions i.e. the relative difference between two substitution processes and therefore work with the frequencies only.

2.2 Analyzed sequences

Neutrally evolving regions Estimation of mutation rates in germline via comparative genomics can be done only using regions which are neutrally evolving. The vast majority of sites in these regions are not or only slightly subject to evolutionary constraints. In evolutionary studies the common practice is to choose FFDs, intergenic regions and introns as regions which are presumably neutrally evolving or at least less constrained than coding and UTR sequences. FFDs which were used for several decades as the gold standard for neutral sites lost their status because of the finding that they are functional as splicing sites.

In this thesis we analyzed mainly intronic and intergenic regions. They both contain regulatory elements but unfortunately not very much is known about these elements. We used them as our standard for neutrally evolving sites.

Sequence data and annotation During the time this work was carried several out Ensembl versions have been released. The analysis that is presented in this thesis was done using different genome versions, sequence annotations and multiple alignments.

Estimating substitution rates along average human genes In Chapter 3, the estimation of the substitution rates was mainly done using triple human-chimp-rhesus alignments which were retrieved from the Ensembl database, version 41 from October 2006 (49). They are based on the releases *homo sapiens* (41, 36c), *pan troglodytes* (41, 21), and *macaca mulatta* (41, 10a), and were generated by MLAGAN [24]. The annotation for genes, exons, and translatable exons are according to Ensembl version 41, which uses the NCBI36 annotation of the human genome.

CpG islands definition We retrieved 21353 CGIs from the Ensembl database, which defines CpG islands using the following cut offs; minimum length is 400 bps long; minimum GC content is 50%; and minimum of observed over expected CpG ratio of 0.6.

Evolution of substitution rates and separation to CGI and nonCGI-genes To study the specie impact on substitution rates, we analyzed the substitution frequencies of ten species: Human (*Homo sapiens*), Chimpanzee (*Pan troglodytes*), Orangutan

(*Pongo pygmaeus*), Mouse (*Mus musculus*), Rat (*Rattus norvegicus*). Dog (*Canis familiaris*), Cow (*Bos Taurus*), Horse (*Equus caballus*). Stickleback (*Gasterosteus aculeatus*) and Medaka (*Oryzias latipes*). For the purpose of our analysis; mammals were grouped into the three clades of primates, rodents and laurasitheria (including dog, cow and horse). For mammals, gene annotation and multiple species alignments were downloaded from ensemble v55 [42] . For primates we used 4-way catarrhini-specific EPO alignments [117], for rodents and laurasiatheria the twelve amniota vertebrates EPO alignments [117]. The Enredo and Pecan (which together with Ortheus) comprise the EPO pipe line assures the consistency of the alignments with paralogs [117]. For fish gene annotation and Multiz 3-way alignments were downloaded from UCSC [128]. In each clade analyzed, gene annotation of just one species was used to determine the regions of interest. That is for primates: human (Ensembl v55), rodents: mouse (Ensembl v55), laurasitheria: dog (Ensembl v55) and in fish: stickleback (Ensembl v55). CGI coordinates for human, mouse, dog and stickleback were taken from the Ensembl (v55).

In the analysis of evolution of substitution rates only protein coding genes were analyzed and were divided into two classes: (1) CGI-genes, where the 5' end is found within a CGI; (2) And nonCGI-genes, where 5' end is not located in a CpG island. The number and proportion of CGI-genes and nonCGI-genes vary among the references species. In human we used 9021 CGI-genes and 4786 nonCGI-genes and in mouse 6448 and 6708, respectively. Among the dog genes 2882 were CGI-related and 9262 nonCGI-related. The frequencies were then estimated from multiple alignments and the following phylogenies: primates (((human, chimpanzee), orangutan), rhesus), rodents ((mouse, rat),human), laurasiatheria (((dog, horse),cow),human) and ((stickleback, medaka), tetraodon).

Substitution analysis around and within CGI Sequence annotation and multiple alignments. Triple human-chimp-rhesus alignments were retrieved from the Ensembl database, version 50 from July 2008 [42] using the Ensembl API. They are based on the releases *Homo sapiens* (50, 36c), *Pan troglodytes* (50, 2.1), and *Macaca mulatta* (50, 10), and they were generated using the EPO (Enredo Pecan Ortheus) pipeline [117] for 4 catarrhini species. The annotation for genes, exons, and translatable exons are from to Ensembl version 50.

The position of the 5'end of an Ensembl gene The 5' end of a gene which is coded on the forward (backward) strand, is defined as the lowest (highest) position among all 5' chromosomal locations of its transcripts. To ensure a high quality of the TSS annotation, a gene is only included into the analysis if one of the transcripts defining the 5' end is also in the RefSeq transcript or peptide database [82]. We further included only the first (most 5') TSS if a gene has multiple TSSs. This way we minimized effects of transcription on intergenic regions upstream of the 5'ends of genes. In addition, genes, which were located on sex chromosomes, were filtered out.

After applying all these filters, we are left with 15552 genes that are used for estimation of the substitution rates.

Classification of CpG islands We also wanted to check if there is a difference between CGIs which are or are not associated to transcription. Therefore, we classified CGIs in the human genome into two main classes according to their distance from the TSS. The first class of CGIs is associated with transcription; CGIs in this class harbor one or more TSSs, but do not overlap with genes that are transcribed from different strands. This class is denoted by tCGI and includes 8249 CGIs. The second class of CGIs encompasses 3378 distal CGIs (denoted by dCGIs) which are intergenic and found at least 10kbp away from any annotated TSS of a human gene (Figure 2.1). We also defined two additional subclasses of CGIs; proximal CGIs (pCGIs) which are intergenic CGI and are found at a distance of up to 10 kb from some annotated TSS in Ensembl; and genic CGIs (gCGIs), which are CGIs which are completely contained within a gene.

Region of analysis: Gene centric view The primary analysis was done on DNA sequences in the vicinity of the 5'ends of genes. For each gene, we determined a region of analysis, which was defined as the region 5000 bps up and downstream to the TSS. However, in order to avoid the twofold analysis of one region, the upstream region was truncated to the middle position between two genes, if the next upstream gene was closer than 10000 bp (Figure 2.1). For genes shorter than 5 kbp the downstream region was also truncated and included the sequence up to the 3'end of the gene (Figure 2.1). A similar procedure has been applied to determine sequences for the regional analysis surrounding the 3'end of genes.

Pooling procedure in gene analysis The next step involved retrieving the mutiple alignments for each region of analysis. In order to reduce effects of selection, we excluded all exons using the annotation of the reference genome in each of the studies. Masking out exonic sequences we kept the positions of intronic sequence segments relative to the TSS unchanged (Figure 2.1). The resulting sequences have been further partioned into non-overlapping 200 bp long windows, where the reference point was the 5'end of the gene (Figure 2.1). For each window, we extracted the appropriate triple alignment for all genes and concatenated them to estimate substitution frequencies as outlined below. The length of these triple-alignments in different windows varies due to different restrictions on the gene sets or sequence characteristics.

We estimated the substitution frequencies for all genes with respect to the non-template strand (i.e. the not transcribed or coding). We estimated the profiles of twelve single nucleotide substitution frequencies $\alpha \rightarrow \beta$. In addition, due to the impact of methylation on mutation frequencies in vertebrate cells, six context dependent substitution processes of CpGs into TpG, CpA, ApG, CpT, CpC and GpG, have been taken into account and their frequencies quantified.

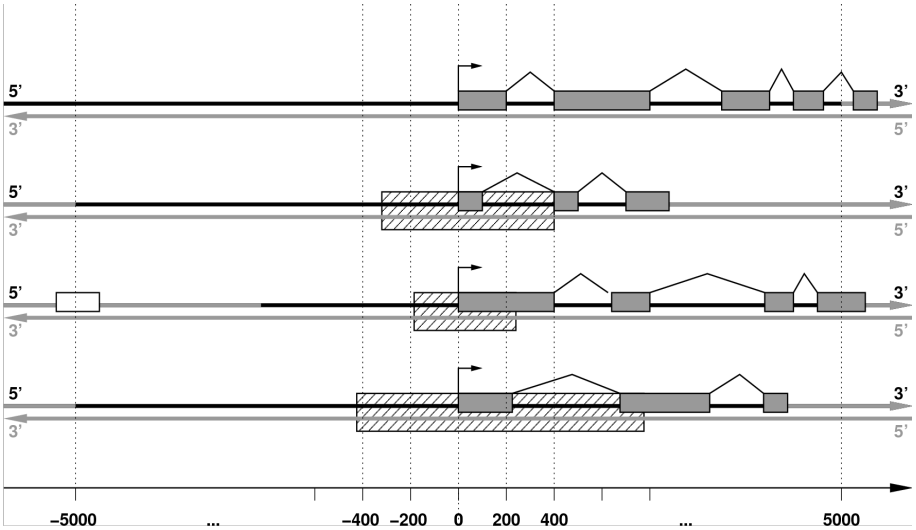

Figure 2.1: Sketch of the analyzed regions around the 5' ends of two classes of genes. Genes starting within a CGI, denoted by striped boxes, and genes without CGI. CGIs are denoted by striped boxes. The template strand for transcription (also denoted by the coding strand) is the reference strand that is used for the substitution analysis and for defining the directionality $5' \to 3'$ relative to the 5' end (which is denoted by 0) gene start (broken arrow). The substitution analysis was done in the 10000 bp long regions centered on the 5'end of gene (denoted by the two outmost vertical lines). This region of analysis was further truncated if the next upstream gene was closer than 10000 bp (white box) or the 3'end of the gene was closer than 5000 bp. Furthermore, exons were excluded (gray boxes). Bold black lines depict the finally analyzed sequences. The substitution frequencies are estimated relative to the 5'end of genes using a sliding window analysis.

Region of analysis: CGI centric view One of the aims of our analysis is to derive the profile of nucleotide substitutions rate around and within a CGI. In particular, we estimate the dependence of rates relative to the 5' and 3'end of CGI on the reference strand (see Figure 2.2). The sequence for the 5' side analysis starts up to 1 Mbp upstream of the 5' end of the CGI and ends either at the middle point of the CGI for distal CGIs or at the TSS inside of tCGIs (see dashed line in Figure 2.2). In order to avoid twofold analysis the upstream region was truncated to the middle position between two CGI, if the next CGI was closer than 2 Mbp. A similar procedure has been applied to determine sequences for the regional analysis surrounding the 3'end of CGIs (Figure 2.2).

Sliding window analysis and pooling around CGIs In order to get high resolution of the dependence of 18 nucleotide substitution rates in the distance from CGI, we wished to estimate the rates in sliding windows, of length 10 kbp each, along the flanking region of individual CGI (Figure 2.2). However, since the divergence in DNA sequence between human and chimp is about 1%, the 18 estimated rates in a single 10 kbp long DNA sequence are noisy. Therefore, in order to perform this

kind of analysis we had to pool data for all CGIs in a similar fashion as before for TSSs . We estimated the 18 substitution rates in genome-wide pooled 10 kbp long non-overlapping windows, which are located at fixed distances from individual CGIs up to a distance of 1Mbp for a CGI. The coordinate of a window in the 5' and 3' analyzed regions is the distance from the center of the window to the 5' and 3' end of CGIs, respectively (Figure 2.2). Such analysis is possible with the availability of genome-wide human-chimpanzee-macaque alignments that cover about 85% of the human genome.

Reference strand for substitution analysis around CGI For estimation of substitution rates one has to choose one of the two DNA strands as a reference (Figure 2.2). For dCGIs we analyzed the forward strand in the NCBI annotation, since we cannot *a priori* distinguish the two DNA strands. For tCGI the two DNA strands are distinguishable since on one strand a gene is transcribed (Figure 2.2). Therefore, for tCGI we estimated the substitution rates with respect to the non-template (i.e. the non-transcribed) strand of the associated gene.

Nucleotide composition We calculated the nucleotide composition of DNA sequences in non-overlapping windows. The sequences that were used were the same as for the estimation of substitution rates. This estimation was not done for all human genes but only in human sequences that were part of the analysis. This means that we picked human sequences in which triple alignments were found for gapless alignment. There for the nucleotide composition is for sites that are aligned in all species in a given alignment.

Estimation of $W \to S$ and $S \to W$ frequencies The frequencies of substitutions of a weak base (W=A or T)in a strong base (S= C or G) and vice versa are calculated as follows:

$$r_{W\to S} = \frac{[A](r_{A\to G} + r_{A\to C}) + [T](r_{T\to G} + r_{T\to C})}{[A] + [T]} \tag{2.20}$$

$$\begin{aligned} r_{S\to W} = &([C](r_{C\to A} + r_{C\to T}) + [G](r_{G\to A} + r_{G\to T}) \\ &+ [CpG](r_{CpG\to CpT} + r_{CpG\to CpA} + r_{CpG\to ApG} + r_{CpG\to TpG}))/([C] + [G]) \end{aligned} \tag{2.21}$$

where $r_{\alpha\to\beta}$ is the frequency of substitutions of base α in β. The density of base α in a bin is denoted by $[\alpha]$; $[CpG]$ is the density of the CpG di-nucleotides.

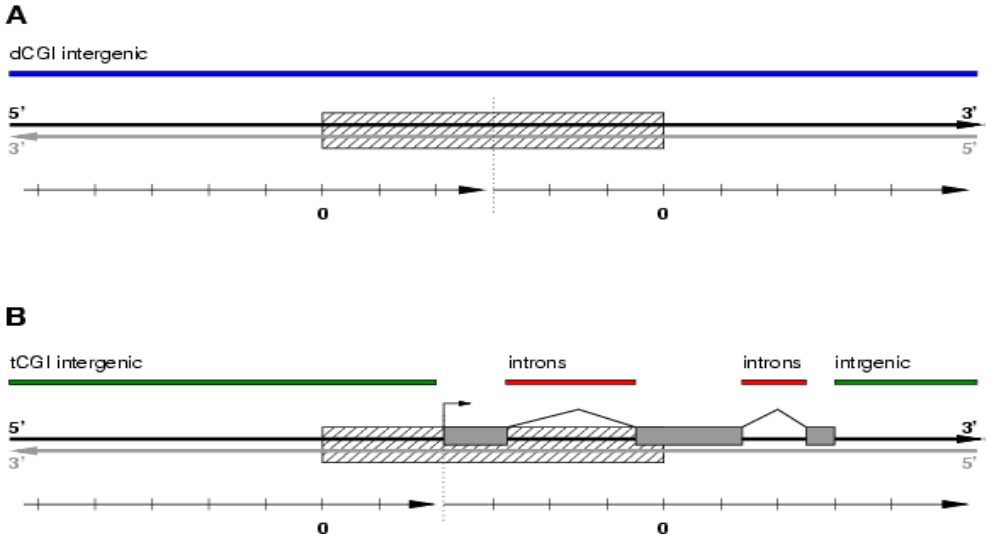

Figure 2.2: Sketch of the analyzed regions around and within two classes of CGIs: dCGIs (A) and tCGIs (B). CGIs are denoted by striped boxes. The bold strand is the reference strand that is used for the substitution analysis and for defining the directionality $5' \rightarrow 3'$ relative to the CGI. The substitution rates are estimated relative to the 5' end (3' end) of the CGI using a sliding-window analysis. The left (right) coordinate system described the distances of the windows from the 5' end (3' end), which is denoted by the left (right) origin (0 k). The left (right) coordinated system starts (ends) at the middle position to the next CGI upstream (downstream) to the 5' (3') end of CGI and ends (starts) in the dashed line. However, the analyzed regions in both sides of the CGI are restricted to 1 Mbp. (A) dCGIs are CGIs that are intergenic and found at distance of at least 10 kbps from a TSS. The reference strand is the NCBI forward strand, and the position of the dashed line is in the middle of the dCGI. (B) A tCGI is a CGI that harbors TSS of a transcript (exons are denoted by shaded areas). The reference strand is chosen to be the nontemplate (or coding) strand of this gene. The dashed line coincides with the TSS. The colored bars indicate regions that were analyzed in figure 2: dCGIintergenic regions (blue); tCGIintergenic regions (green); and tCGIintrons (red) of genes whose TSS are inside of the tCGI.

3 Methylation deamination rate in the vicinity of the mammalian 5'end

The nucleotide composition along transcripts, in the vicinity of 5'ends of genes and CpG islands are different than in other parts of the genome. Therefore, we investigate the dependence of nucleotide substitution frequencies on three factors: transcription, distance from the 5'ends of genes and distance from CpG islands. The result of this analysis reveal a sharp decrease of CpG deamination rates at the 5'ends of genes and in CpG islands.

3.1 Analysis of nucleotide substitutions

Study of regional patterns of substitutions around 5' and 3'ends of averaged human genes One is tempted to estimate substitution rates at different distances from the TSS on a single gene level. However, the low divergence between human and chimpanzee genomic sequences prevents us from getting reliable estimates of mutation rates in small windows, in particular because we wish to estimate 14 different mutation rates. To overcome this problem we estimate mutation rates in genome-wide pooled 200 bp long non-overlapping windows, which are located at fixed distances from individual TSS (see Figure 2.1 and Chapter 2 for further details). Such an analysis is possible because of availability of genome-wide human-chimpanzee-rhesus alignments that cover about 85% of the human genome. In order to minimize effects of selection, we analyze only the intronic parts of genes and their 5' and 3' flanking intergenic sequences. Finally, we estimate the strand dependency by estimating substitution rates in the non-template strand only. Due to the dependency of substitution frequencies on the distance from CGIs' ends, we also estimate the substitution frequencies separately for genes whose 5'end resides within a CGI (CGI-genes, see Chapter 2) and those genes, whose 5'end is located outside of CGI (nonCGI-genes).

Study of regional patterns of substitutions around 5' and 3'ends of CpG islands To explore the impact of CGIs, we adopted a CpG island centered view (see Figure 2.2), where for each CGI we defined a region centered at the CGI that spreads from this CGI to the middle points of the adjacent CGIs. The average distance between two CGI is about 100 kbp but the distance can be more than 1 Mbp. We estimated the substitution rates in intergenic regions in the 5' and 3' sides relative to

CGIs. To control for the impact of transcription on substitution we divided the CGI into two groups, these that are associated with a transcription start site, tCGIs, and those that are at least 10 kbp away from a TSS, dCGIs, (Figure 2.2 , panels A and B). In addition in the analysis of intergenic regions, the sequences that are located in the 5 kbp flanking regions of genes were removed.

Study of the evolution of substitution patterns in vertebrates Although our main focus is on the human genome it is useful to study substitution patterns in other species. By comparing human patterns to those of other species we can find whether the human patterns are specific or universal to some extent. Therefore we estimated substitution frequency profiles along genes of 9 other vertebrates: chimpanzee, orangutan, mouse, rat, dog, cow, horse, stickleback and medaka.

Study of regional patterns of substitutions

Substitution frequencies have been estimated from pooled triple alignments of genomic sequences from human, chimp, and rhesus. We used a maximum likelihood approach which correctly handles effects due to back-mutations and is able to reliably estimate substitution frequencies from given aligned sequences as described in the Methods section. A similar regional analysis was also done in the surrounding regions of the 3'end of genes.

3.2 CpG methylation deamination rates

Sharp drop in CpG methylation deamination rates Our analysis revealed three types of regional substitution patterns (Figure 3.1). In this chapter we will focus on CpG methylation deamination rate pattern. The reduction of loss of CpG's is the strongest regional behavior. In our framework the cytosine in a CpG di-nucleotide may undergo two mutual independent processes to become a T. The first is the common $C \to T$ transition which does not depend on the neighboring bases; the second is the neighbor dependent CpG methylation deamination process $CpG \to TpG$ ($CpG \to CpA$ on the reverse strand). The frequencies of the latter process are reduced by a factor of 15 near the TSSs. This decrease in the CpG loss rate occurs symmetrically down- and upstream of the TSS, and both processes, $CpG \to TpG$ and its reverse complement substitution $CpG \to CpA$, are affected in the same way.

CpG loss frequencies are lower in CpG islands As we discussed in Chapter 1 the high frequencies of substitutions of CpG are due to methylation-deamination processes. However, in CGIs the rate of this process is low and therefore these regions tend to lose their CpG content at lower rates. Most of human gene 5'ends are located within CGIs. Therefore, the observed drop in methylation-deamination rates near the TSS of average human genes can be explained by their association with CGIs.

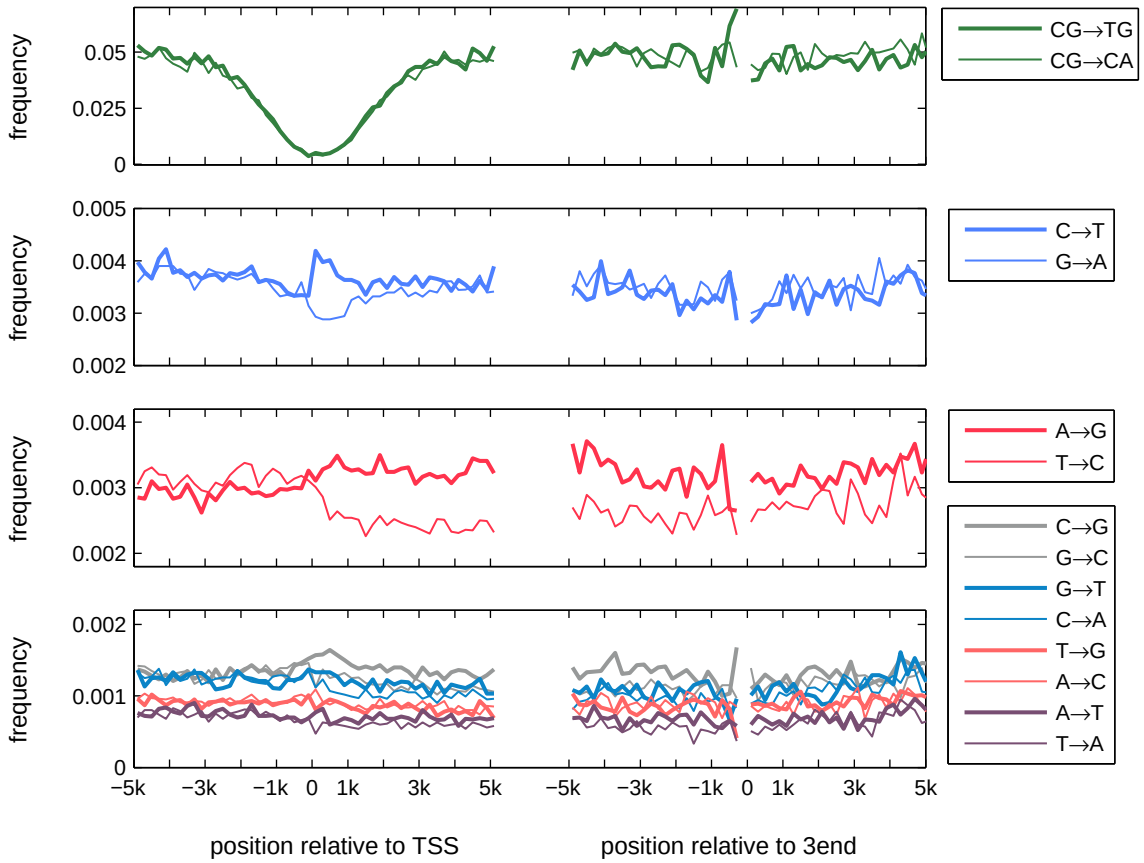

Figure 3.1: Substitution rates in introns and in intergenic regions in the vicinity of 5' and 3'ends of human genes. The plots show the estimated twelve single nucleotide substitution rates and the CpG deamination rates in non-overlapping 200 bp long windows along the non template strand. The distances of the windows' centers from the 5'end or 3'end are indicated on the x- axes. The estimation of substitution frequencies has been performed using the non-template strand.

The CpG methylation-deamination (transition) rates are almost two orders of magnitudes lower within CGIs compared with outside CGIs (Figures 3.2 and C.1). In the flanking regions of tCGI and dCGI the average surplus of the $CpG \rightarrow TpG$ and $CpG \rightarrow CpA$ substitution rate is 0.08 per CpG while in both tCGI and dCGI this rate is about 0.001 and 0.002 per CpG, respectively (Figures 3.2 and C.1). The sharp increase of CpG methylation-deamination rates when the distance from CGIs increase implies that it is the association of 5'end with CGIs that causes the decrease in the rates. Moreover, the fact that both in tCGI and dCGI there is a decrease implies that it is not transcription *per se* that lowers the mutation rates of CGIs. Interestingly, transversion rates, $CpG \rightarrow CpA/CpT$, are higher than the other transversion rates in non-CpG context but are similar within and in the flanks of CGIs (Figures 3.2 and C.1).

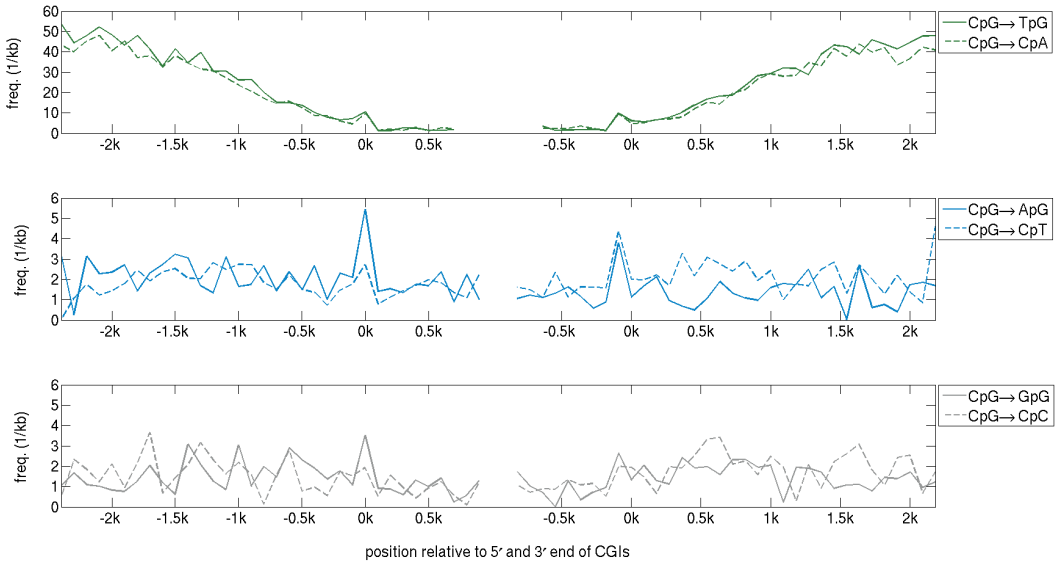

Figure 3.2: CpG methylation-deamination rates in the proximity of 5 and 3 ends (left and right 0k, respectively) of tCGI. In the top two panels, the current values of GC content and CpG odds (continuous lines) are compared with the corresponding stationary quantities (dashed lines). The data points between the left and the right 0k are calculated within the CGIs.

The impact of CGIs on mutation rates near the 5'end of genes We repeated the analysis of mutational patterns that was carried out separately for gene-promoters overlapping with CpG islands (CGI-genes) and non associated with CGI genes (nonCGI-genes). The results of this analysis show that the drop in methylation deamination rates is more pronounced in CGI-genes (Figure B.1) then in nonCGI-genes (Figure B.2). The drop in the CpG mthylation deamination is lower near the TSS of CGI-genes. The region around the TSS, where we observe a low methylation-deamination transition rate, is narrower. However, since there is a drop of CpG loss in nonCGI-genes which suggests that not all CGIs have been annotated and there are also CGI regions shorter than the minimal cut off, that is set by Ensembl, of 400bp [15].

3.3 Possible mechanisms to explain lower CpG loss near the TSS

The observed reduction of the CpG methylation deamination rate near the TSS can be explained by two mechanisms. First, the lack of methylation in CGIs in germline cells [156] decreases the probability of $C \to T$ transitions (in CpGs). Second, purifying selection might counteract the loss of CpGs in order to preserve the existence of a CGI

for regulatory processes in somatic cells [100]. A loss of CpGs due to mutations would lead eventually to the loss of the CGI property of a gene promoter and change the gene expression pattern [72]. However, the similar rates of methylation mediated transversions of CpGs in CGIs and in their flanks imply that there are additional factors that shape the profile of CpG-mutations in CGIs beside the methylation levels themselves.

4 Transcription-associated strand asymmetries in mammals

Transcription and mutations are key processes in evolution. The effects of mutations on transcription of genes have been extensively investigated in recent years. In this chapter, we use comparative genomics to study how transcription has influenced the rates of single nucleotides mutations in the germline. We observe two types of asymmetries. First, a strand asymmetry in complementary substitution rates, which extends from the 5'end to 1 kbp downstream from the 3'end. Second, a localized strand asymmetry, an excess of $C \to T$ over $G \to A$ substitution in the nontemplate strand confined to the first 2 kbp downstream of the 5'end of genes and restricted to the CpG islands in the beginning of genes. These asymmetries can lead to base composition asymmetries in intronic regions i.e violations of Chargraff's second parity rule. The global asymmetries is suggested to be a result of transcription coupled repair. The localized asymmetry is hypothesized to be due to a higher exposure of the nontemplate strand near the 5'end of genes coupled to a higher cytosine deamination rate. These asymmetries are found in 7 more mammalian genomes and can generate nucleotide composition asymmetries in introns of mammalian genes.

4.1 Localized asymmetry

An excess of $C \to T$ over $G \to A$ substitutions is restricted to the first 1-2 kbp downstream to the TSS For the neighbor independent single nucleotide substitutions we find much richer patterns by estimating and comparing the complementary substitutions on the non-template strand. First, reverse complement substitution processes do not occur at the same rates on the non-template strand, i.e. the transcription process singles out one strand breaking the symmetry between the two strands in untranscribed regions. Second, this violation of the symmetry occurs only downstream of the TSS. Strikingly, the first such asymmetry, an excess of the $C \to T$ over $G \to A$, is confined to the first 1-2 kbp long region downstream of the TSS (Figure 3.1). For this localized asymmetry we observe an elevation of the $C \to T$ transition rate by about 20% in the first kbp of the transcript compared to the rate in promoter regions upstream to the TSS, whereas the $G \to A$ rate decreases by about the same percentage in transcribed regions. The difference between these two rates

reaches up to 40% and the gap between these rates is closed at a distance of about 1.5 kbp to the TSS (Figure 4.1). This asymmetry is specific to the 5'end of genes, and it is not detected in the vicinity of the 3'end (Figure 3.1), which confirms the localized nature of this strand bias. The trend of the localized asymmetry that we found is opposite to the one that had been reported by Green et al. [60]. This discrepancy between the results may very well be explained by the fact that in contrast to the previous study we surveyed the whole genome and included a broader spectrum of transcripts into our analysis (see below).

4.2 Global asymmetry

Global asymmetry: $A \to G$ exceed $T \to C$ transitions along the transcript In contrast to the above process, other processes show a global asymmetry defined as a bias in complementary nucleotide substitutions that extends along the whole transcript. There are four pairs of nucleotide substitution processes that show global asymmetries: $r_{A\to G}/r_{T\to C}$, $r_{C\to G}/r_{G\to C}$, $r_{A\to T}/r_{T\to A}$, and $r_{G\to T}/r_{C\to A}$. In Figure 3.1, it is shown that the asymmetry in the substitution frequencies extends from the TSS down to the end of the analyzed region (5 kbp downstream of the TSS). In order to check whether the (global) asymmetries extend along the whole transcript, we also analyzed the 10 kbps long region centered on the 3'end of genes (Figure 3.1). Interestingly, these asymmetries extended not only until the 3'end, but on average also into the 1 kbp region downstream of the 3'end of genes (Figure 3.1). This extension is similar to the extension of the TA bias in the nucleotide composition as far as 1 kbps downstream to the 3'end, which reflects the fact that the termination position of the transcription process is not always at the annotated 3'end of genes and might continue several hundreds of bases further [38; 92]. The bias in the transition frequencies $A \to G$ over $T \to C$ was also reported previously by Green et al. [60], who analyzed 1.5 Mbp of human chromosome 7. Our genome-wide analysis reveals similar biases for three out of the four transversions (Figure 3.1), which confirms a prediction from theoretical considerations [150].

4.3 Strand asymmetries in non intronic regions of genes

The substitution asymmetries in introns can be shaped either by mutational processes or by selection pressures on introns. These two processes should leave two different mutational signatures. While mutational processes should not distinguish between the different parts of the transcript, selection effects are limited to functional elements. If these functional elements are intron specific, the asymmetries should be restricted to

introns. We tested this hypothesis by analyzing the substitution rates in three different parts of transcript, 5' untranslated regions (UTR), 3'UTR and Four fold degenerate sites.

UTR We started by estimating the substitution rates in 5'UTR regions; however, since the average length of 5'UTR is about 200 bp long, we had enough data for estimating the rates in 5'UTR sequences, which overlap the first two 200 bp long windows downstream to the TSS. We found out that in 5'UTR sequences that are located within the first 200 bp of the transcript the rate of $C \to T$ (0.0034 per nucleotide) exceeds $G \to A$ (0.0025), whereas, the rates of $G \to A$ and $T \to C$ are similar (0.00311 and 0.003, respectively). In the next 200 bp long window, the $C \to T$ transitions occur in higher frequencies than $G \to A$, but the degree of asymmetry between these two rates is lower than in the first window (0.00314 and 0.0029 respectively), and the asymmetry of $G \to A$ (0.00284) versus $T \to C$ (0.0025) becomes more pronounced. Comparison with intronic regions at similar distances from the TSS showed that the rates in 5'UTR are about 2% lower than the rates in introns, but the ratios of $r_{C\to T}/r_{G\to A}$ and of $r_{A\to G}/r_{T\to C}$ were the same in introns and 5'UTR. Similar observations were made in the analysis of 3'UTR sequences.

The excess of $A \to G$ over $T \to C$ was detected in 3'UTR sequence that are located in the one kbp long region upstream to the 3'end of the genes (Figure A.2). As in 5'UTR sequences, the substitution rates in 3'UTR sequences are lower by about 20% compared to intronic regions (see in Figure 3.1), but the ratio of $r_{A\to G}/r_{T\to C}$ was the same for both (Figure A.3). The lower substitution rates in 3'UTR and 5'UTR sequences indicates that they these regions are under negative selection.

Four fold degenerate sites Amino acids coded by 4 codons which differ only in their third position are called four fold degenerated . The third nucleotide in such FFD codons is called FFD site. In order to build a set of such sites, for each Ensemble gene we chose one transcript that overlaps a Refseq transcript. Since the number of FFD sites varies along the transcripts, we used windows of different lengths, which contain about 100,000 FFD sites. Such criteria left us with 3 windows in the 5 kbp regions downstream to the transcription starts; hence, we extended the region of analysis to 20 kbp past the TSS. The estimation of substitution in each gene was done by a Maximum Likelihood Method.

We found that for FFD sites the $A \to G$ rates are higher than $T \to C$ rates along the whole transcript (Figure A.4). The localized excess of $C \to T$ over $G \to A$ substitutions was also found in FFD sites located proximal to the TSS. However, in contrast to introns, this bias in the substitution rates is not statistically significant due to the low amount of sequence data on FFD sites.

Repetitive elements We also checked the impact of repetitive elements on substitution rates by estimating the substitution in repetitive regions whose annotation was retrieved from Repeatmasker and in the remaining non-repetitive sequences. The

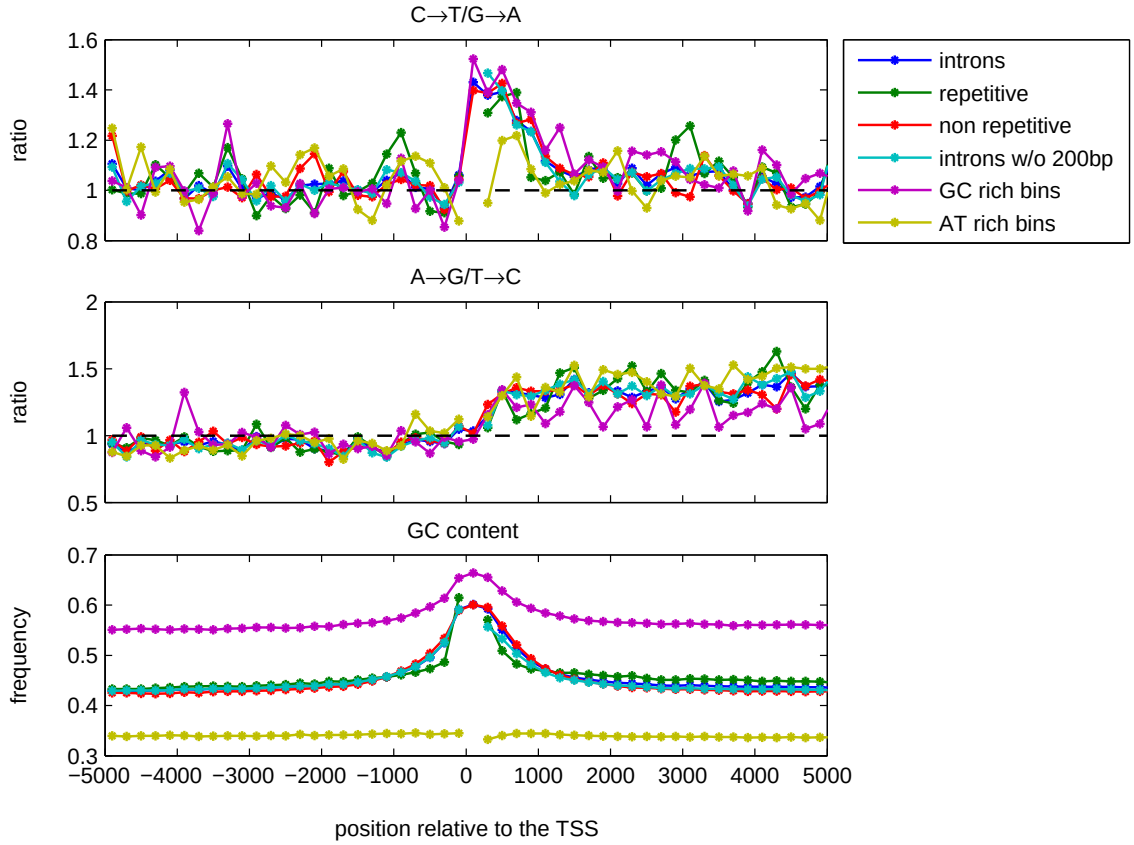

Figure 4.1: Ratios between complementary transition rates and the GC content plotted against distance from the 5'end of genes calculated in 200 bp long windows along the non-template strand, combined information from all genes and presented by six different genomic contexts. "intronic": genes that were used in Figure 3.1. "introns w/o 200 bp": the 200bp in introns' edges were excluded; " GC rich windows" and "AT rich windows": DNA sequences with GC content of above 50% and below 41%, respectively. Windows that contained less than 100 kbp long DNA sequences were omitted.

UTRs and FFD sites are assumed to evolve under stronger constraints than introns. On the other hand, repetitive elements are assumed to have less functional constraints. Indeed, the substitution rates in repetitive elements (Figure A.6) are higher than in non-repetitive regions (Figure A.7). Yet, the ratios between complementary transition rates were the same in repetitive and non-repetitive regions (Figure 4.1).

The asymmetries are not due to functional elements in intron edges The existence of local and global asymmetries in different parts of the transcripts, which are under different level of evolutionary constraints, implies that the asymmetries are most likely invoked by a bias in the molecular mutational processes, as a result of selection acting on functional elements common to all different parts of the transcript, including repeats. Candidates for such functional elements that are common to introns, UTRs

and FFDs would be splicing elements. These elements are usually found in 30 bp exons at the edges of introns and exons [20; 66]. Since previous works suggest that splicing elements might be part of the 200 bp ends of introns [92; 149], we re-analyzed our data, using sequence data after excluding the 200 bp at each end of the intron. The removal of these elements did not affect the profile of the ratios of complementary substitution frequencies (Figure 4.1); hence, it seems that the bias is not due to elements that are located in the edges of introns. Nevertheless, there are several works that suggest that the first introns are enriched with splicing motifs [151]. Even though the first introns are usually long, and removing them leaves us with little sequence data for analysis, we could observe similar substitution patterns in introns which are not first introns (Figure A.5).

4.4 The impact of CpG islands on strand asymmetries in transcribed regions

Strand asymmetries are correlated with transcription and transcription initiation in embryonic stem cells Our results, so far, suggest that the substitution rates are shaped by mutational molecular mechanisms. The asymmetry $A \rightarrow G$ vs. $T \rightarrow C$, was suggested to be the result of strand specificity of TCR and a bias in misinsertion of $A \rightarrow G$ over $T \rightarrow C$ during replication which is not attributed to difference between transcribed and non-transcribed strands [60]. In a similar fashion, TCR and biases in misinsertion rates between complementary transversions can lead to the asymmetries that are observed in Figure 3.1, i.e. biases in misinsertions of $C \rightarrow G$ over $G \rightarrow C$, $G \rightarrow T$ over $C \rightarrow A$ and $A \rightarrow T$ over $T \rightarrow A$. However, the fact that TCR acts on the whole transcript rules out its role in the restriction of the $C \rightarrow T$ over $G \rightarrow A$ bias to the first 1 kbp. Therefore we suggest that other mutational mechanisms are responsible for this bias. It is known that there are several mutagenic processes that target ssDNA, which is a by-product of RNA polymerase activity. Hence, higher frequencies of ssDNA in this region could lead to higher transition rates of $C \rightarrow T$ in the beginning of a transcript, rather than in the rest of the transcript.

Recent works show that RNA polymerase II (pol II) activity is not homogenously distributed along transcripts [77; 107]; moreover, in many genes there are marks of transcription initiation but not of complete elongation. In about 80% of genes in embryonic stem cells (ESC), there is an initiation of transcription even though just 50% of the genes are fully transcribed [61]. Guenther et al. [61] performed a genome wide mapping of histone modifications, that mark transcription initiation and elongation. Their results reveal that many "inactive" (non expressed) genes harbour histone marks associated with transcription initiation at the vicinity of their 5'ends.

We used this recent dataset, in which genes were tested for their initiation and expression states in embryonic stem cells. In particular, we used Table S3 and Table S5 in the supplementary material of their paper. Table S3 is a list of genes with the chromosomal location of their TSS and the signal of different histone modifications. The histone modification H3K4me3, is a marker for transcription initiation. We used a threshold of above two in the signal of H3K4me3 in order to determine if a gene experienced initiation. In addition, we used the chromosomal position of the TSS in order to associate an appropriate Ensembl gene. The expression status of a gene was retrieved from Table S5 in that paper. This table provides a list of genes and their expression status in ESC. We divided the genes in this dataset into 4 groups according to their expression status in ESC (denoted by exp+/-), and their initiation status (init +/-). We calculated the substitution rates in the 5 kbp upstream and downstream regions of the 5'end of genes in three groups, init+exp+, init+exp-, and init-exp- (the number of genes in the last group init-exp+ was too small for analysis).

The results show that the local excess of $C \to T$ over $G \to A$ is strongest in genes that are classified as init+exp+, a weaker bias (half as strong as for the init+exp+ group) was found in the group of init+exp- genes. In the set of init-exp- genes, the local asymmetry is actually absent (Figure 4.2). Similar behavior is also observed for the global asymmetry (the excess of $A \to G$ over $T \to C$). Hence, we concluded that initiation of transcription (in ESC) is correlated with formation of the local and global asymmetries, although these asymmetries are weaker in init+exp- than in init+exp+ genes.

The local asymmetry is correlated to GC content and distance form the TSS Both the distance from the TSS and the GC content are correlated with the local asymmetry. The excess of $C \to T$ over $G \to A$ is limited to the first 1 kbp region at the start of the transcript, a region that is also extremely enriched in G+C nucleotides (Figure 4.2). It is not clear if the local asymmetry is correlated with the distance from the TSS or if these factors are correlated indirectly through a third factor of G+C density. In order to isolate the impact of the following two parameters: GC content and distance from the TSS, on the mutation spectrum, for each window in a certain length from the TSS we built collections of two types of sequences, GC-rich (above 50%) or GC-poor (below 40%). In GC-rich sequences, the ratios of $r_{C\to T}/r_{G\to A}$ and of $r_{A\to G}/r_{T\to C}$ were similar to the ratios in all genes (compare Figures 4.2 and 4.1). The similarity between the ratios at the 2 kbp downstream is expected because the enrichment of GC in these regions. The lack of the bias in $C \to T$ over $G \to A$ in GC-rich windows that are located further downstream from the TSS implies that this bias in the first 1 kbp of the transcript is not just due to the GC content (Figure 4.2). However, in GC-rich windows, the ratio between $A \to G$ over $T \to C$ is lower than the overall ratio and fluctuates between 1 and 1.2 (Figure 4.2); this can indicate that the GC-rich sequences that are located several kbp from the 5'end of genes are subject to different molecular mechanisms than GC-poor sequences. On the other hand, in GC-poor windows, the bias between $C \to T$ and $G \to A$ is absent at the immediate 1 kbp

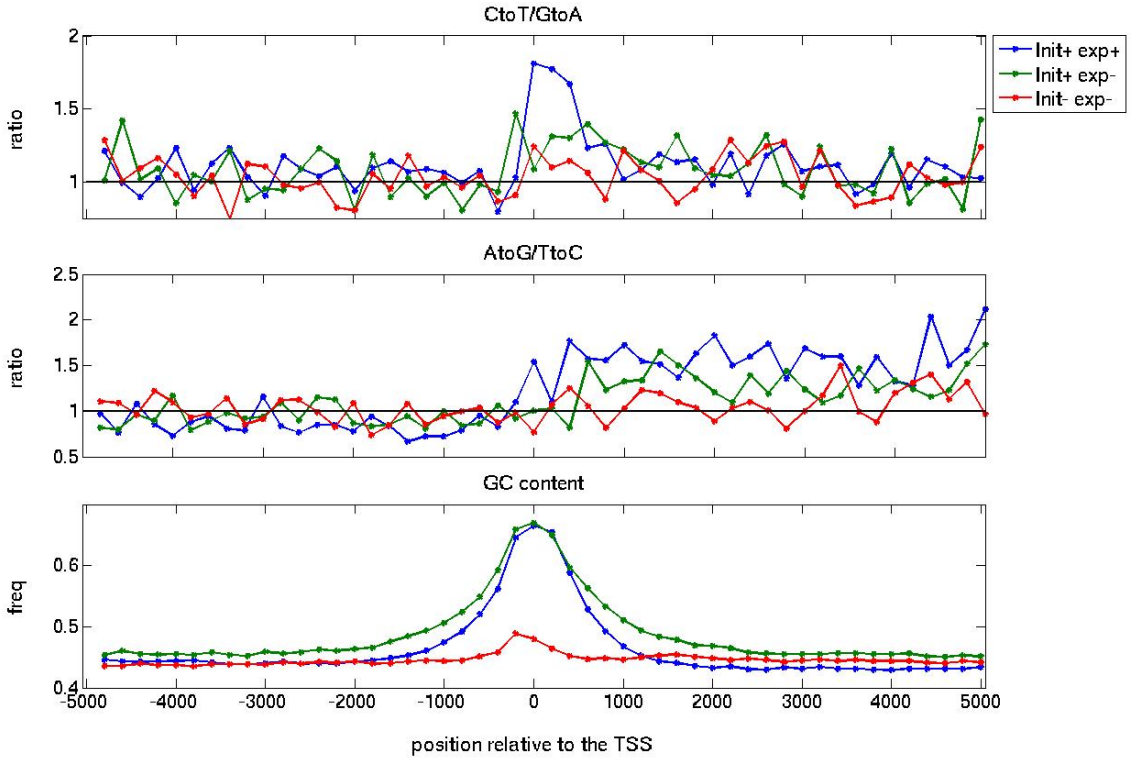

Figure 4.2: Correlation between strand asymmetry and transcription status of genes in embryonic stem cells (ESC). The ratios between complementary transition rates and the GC content are calculated in three gene classes [61]: genes that experienced initiation and transcription (exp+init+); genes that experienced initiation but not complete transcription (exp-init+); genes that experienced initiation but not complete transcription (exp-init-). The estimation of substitution frequencies has been performed using the nontemplate strand.

downstream to the TSS, but in the same regions $A \to G$ exceeds $T \to C$ (Figure 4.2). Therefore, it seems that both genomic parameters, the distance from the TSS together with the GC density, are correlated with the localized asymmetry. Another interesting observation is the dependency of CpG transition rates on the local GC content [36]. At distances of more than 2000 bp from the TSS, the CpG loss rates in GC-poor windows were more than 25% higher than in GC-rich windows (Figure A.8). The CpG loss rates in AT-rich sequences were 10 times higher than the rates in GC-rich sequences near the TSS, whereas the $C \to T$ transition rates in non-CpG sites in AT-rich sequences in the first kbp of the transcript were lower than in GC-rich sequences (Figure A.9).

Since both sets of genes, exp+init+ and exp+init-, have on average a high GC content in promoter regions but different levels of strand asymmetries, we suggest that GC content is indirectly correlated with the asymmetry level due to higher transcription activity of GC rich promoters in germline cells. Clearly, a limitation of this analysis

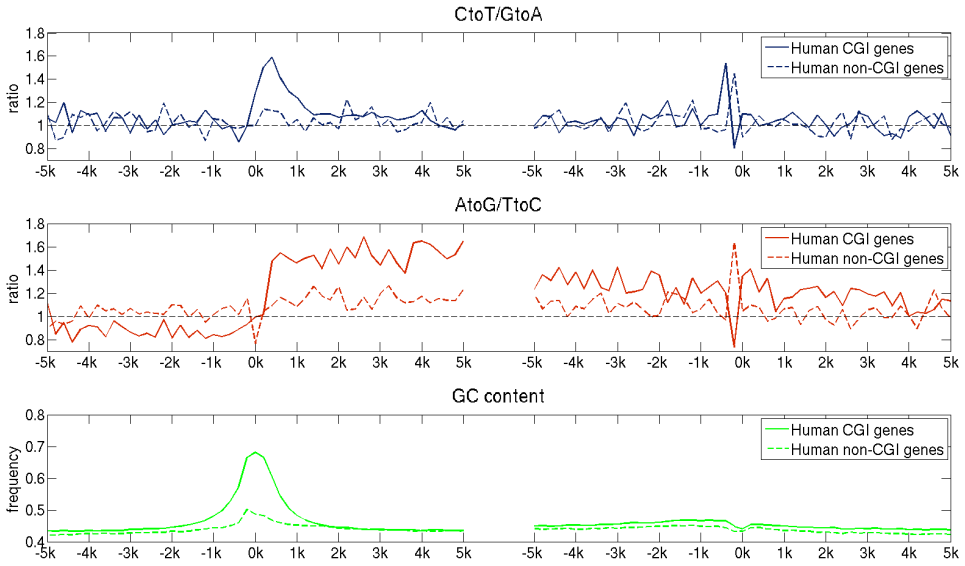

Figure 4.3: Ratios of complementary transition frequencies and GC content along human genes. Each of the three panels consists of two sections: The left section is centered on the TSSs (left 0k) and refers to 5 kbp of intergenic region upstream to the TSS and extends towards 5 kbp of intronic region downstream to the TSS. The right section on a panel is similarly centered on the 3' ends (right 0k), while the analyzed regions extend from 5kbp of intronic regions (upstream to the 3'end) to 5kbp of intergenic regions downstream to the genes 3' ends. The ratios are plotted against the distance from the 5' and 3' ends of genes. Ratios are calculated along the non-template strand from pooled 200 bp windows of genes annotated for the reference species in each taxon. For CGI-genes the ratios are presented by thick lines, for nonCGI-genes ratios as thin lines.

is that, although it was carried out in genes that are expressed in ESC, we cannot be sure that they are also expressed in germline cells, where mutations have to occur in order to be passed on to the next generation.

The localized strand asymmetry appears only in genes with CpG island promoters As we mentioned above init+ genes differ for init- in the GC content in the vicinity of the the TSS. Higher GC content of init+ genes implies that they are associated with CpG islands. Therefore, we compared the strand asymmetries CGI-genes with the ones in nonCGI-genes (see Chapter 2). The results of this analysis show that the global and local asymmetries are significantly more pronounced in CGI-genes than in nonCGI-genes (Figure 4.3). Actually, the local strand asymmetry is found only in CGI-genes (Figure 4.3).

The localized asymmetry is limited to CpG islands In order to further investigate the role of CGIs in shaping the strand asymmetries along introns, we analyzed the strand asymmetries across CGIs.

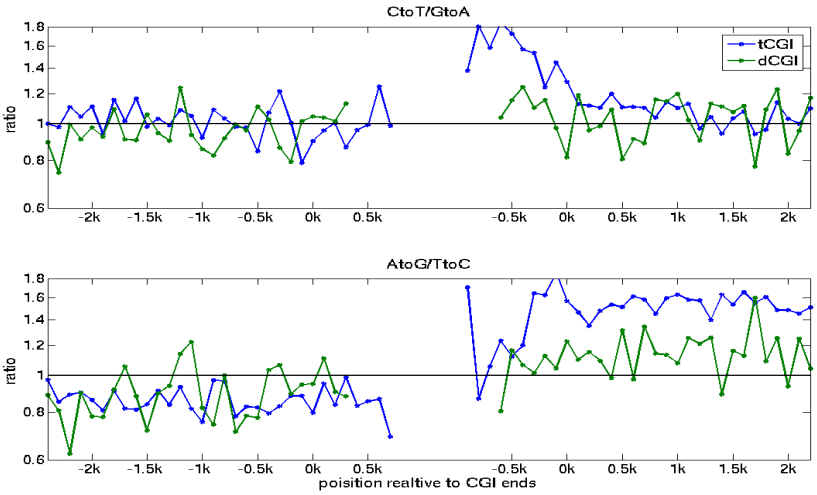

Figure 4.4: Ratios of complementary transition rates are plotted against distance from the 5end (left 0k) and 3end (right 0k) of CGIs calculated in 100 bp long windows. The analyzed regions are non-exonic. The ratios that are presented between the left and the right 0k are calculated within the CGIs. The data points between the left and the right 0k are calculated within the CGIs.

The localized asymmetry is found in the transcribed part of CGI and decays outside of it. We do not observe asymmetry between $C \to T$ and $G \to A$ IN the tCGI at the upstream intergenic part nor in dCGI. This implies that a process that couples transcription and CGI is involved in the formation of the localized strand asymmetry. The ratio $r_{C\to T}/r_{G\to A}$ is still found to be higher than one downstream to the 3'end of genes. This indicates that the definition of CGI boundaries is not accurate.

The ratio of $A \to G$ over $T \to C$ rates is lower than 1.2 within CGI, while it is above 1.4 in intronic regions which are found immediately downstream to the 3' end of the tCGI. The global assymetry starts from 200bp downstream to the TSS (see Figure 3.1).

4.5 Strand asymmetries are found in other mammals

Similar strand asymmetries are found along the mammalian tree We tested whether the substitution asymmetries are a unique feature of the human genome or also found in other species. In mammals, we have found that the ratio $r_{A\to T}/r_{T\to C}$ along the 5kb region downstream to the TSS and 5kbp upstream to the 3'end is larger than 1 and relatively constant along the analyzed transcribed region. In addition to this global pattern there is a localized pattern, where $r_{C\to T}/r_{G\to A}$ is greater than 1 only along the first 1-2kb of transcripts, while the ratio was close to 1 outside this

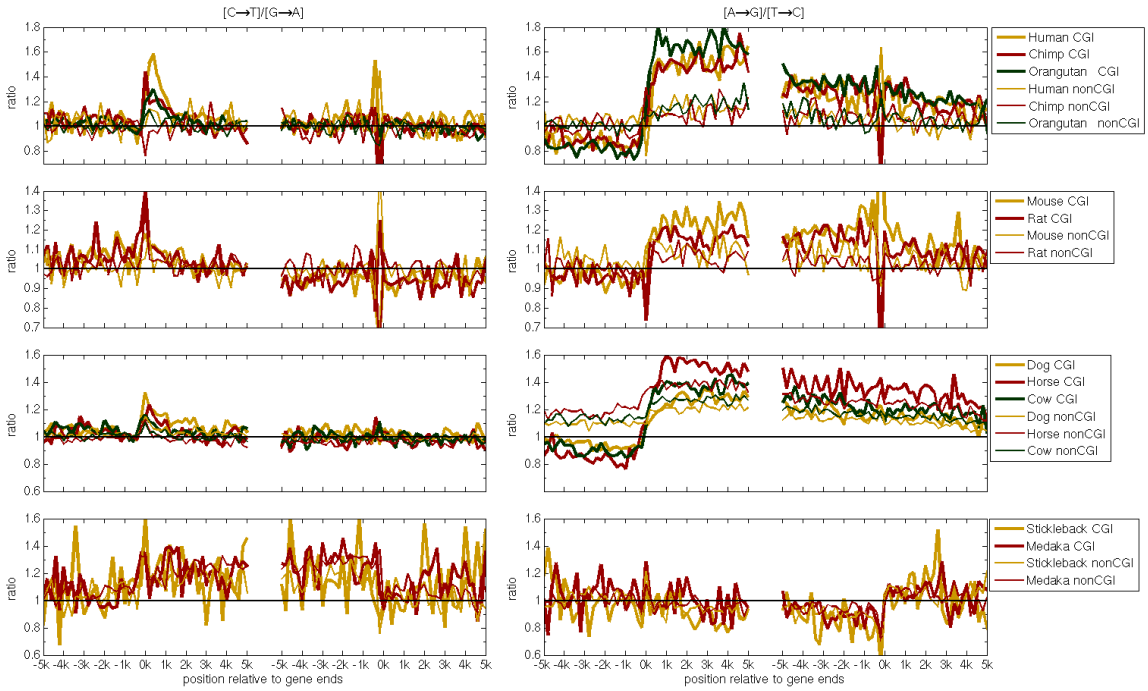

Figure 4.5: Ratios of complementary transition frequencies across vertebrates. There are four rows of panels and two columns of panels. The rows correspond to the four taxa analyzed in this study and the columns to the two types of ratios between complementary transition frequencies. Each panel consists of two sections: The left section is centered on the TSSs (left 0k) and refers to 5 kbp of intergenic region upstream to the TSS and extends towards 5 kbp of intronic region downstream to the TSS. The right section on a panel is similarly centered on the 3' ends (right 0k), while the analyzed regions extend from 5kbp of intronic regions (upstream to the 3'end) to 5kbp of intergenic regions downstream to the genes 3' ends. The ratios are plotted against the distance from the 5' and 3' ends of genes. Ratios are calculated along the non-template strand from pooled 200 bp windows of genes annotated for the reference species in each taxon. For CGI-genes the ratios are presented by thick lines, for nonCGI-genes as thin lines.

region. Similarly, the global $r_{A\to G}/r_{T\to C}$ and the local $r_{C\to T}/r_{G\to A}$ asymmetries are present in all mammalian genes. Yet the ratio $r_{A\to G}/r_{T\to C}$ declines towards the 3'ends of transcripts (Figure 4.5). And as in human, the global asymmetries extend into the intergenic regions downstream to genes. Across mammals the global and local asymmetries are significantly more pronounced in CGI-genes than in nonCGI-genes. This implies that similar CGI-linked processes in mammals are the cause of more pronounced patterns in CGI-genes (Figure 4.5).

In addition to the breaking of strand symmetry of transition rates, 3 out of 4 of the transversion rate pairs are not equal to each other (Figures B.3 and B.4). The ratios $r_{G\to T}/r_{C\to A}$, $r_{C\to G}/r_{G\to C}$ and $r_{A\to T}/r_{T\to A}$ are greater than one in intronic regions of all

mammalian species that were analyzed in this study. These three asymmetries begin at the 5'end of genes and extend beyond the 3'end of genes (Figure 4.5). These results suggest that strand asymmetries along mammalian transcripts have been shaped by similar forces during evolution.

Strand asymmetry in fish Strand asymmetries are also found in introns of non-mammalian species (Figure 4.6), but their directions are different in most cases. In stickleback and medaka introns, the $r_{A\to G}/r_{T\to C}$ ratio is smaller than 1, opposite to the asymmetry in mammals (Figure 4.5). Also the ratio $r_{C\to T}/r_{G\to A}$ is greater than 1, similar to the first 1-2 kbp downstream the TSS of genes in mammals, but it is not restricted to the vicinity of the TSS in fish genes. In contrast to mammals, there is no difference in the level of asymmetry between CGI- and nonCGI- genes in fish.

4.6 Regional patterns of nucleotide composition

The current single nucleotide substitution rates lead to TA and GC skews in mammals As was mentioned in Chapter 1, TA skew ($S_{TA} = ([T]-[A])/([T]+[A])$) and GC skew $S_{GC} = ([G]-[C])/([G]+[C])$ are observed in human introns and have been suggested to be a result of a bias in substitution rates. Over a long period of time, biases in substitution rates should accumulate and lead to skews in the base composition of complementary DNA strands [153]. Measurements of the TA skew and the GC skew [1] have been shown to be different between species both in respect to their location relative to the TSS and their levels of intensity [1; 149; 49]. It is interesting to see whether the observed biases of intronic substitution rates can lead to similar skews as seen in current genomes. We have found that for all vertebrates this is indeed the case; the direction of skews in intronic regions agrees with the current ones in CGI-genes (Figure 4.6) and in nonCGI-genes (Figure B.5). Hence, the observed substitution rates can indeed build the current skews in the genome. For nonCGI-genes the degree of the stationary skews is similar to the current one in mammals. However, in intronic regions of primate CGI-genes, the stationary TA and GC skews are greater than the current ones. This might be due to the incompleteness of our genome evolution model, which does not include other mutational processes, such as insertion and deletions, that might have being acting against the influence of substitutions on base composition around and within genes. Alternatively, it might be that the mutational force that leads to skews has became stronger during primate evolution.

Spikes of nucleotide skews at gene ends Another characteristic of both the TA and GC skews are spikes at boundaries of genes (Figures 4.6 and B.5). At the 5' end of genes there is a local increase in both skews. For example in primates and rodents the GC skew is above 0.1 while in the rest of the transcript the skew is below 0.06 (Figures 4.6 and B.5). However at the 3' end the GC skew is negative in all

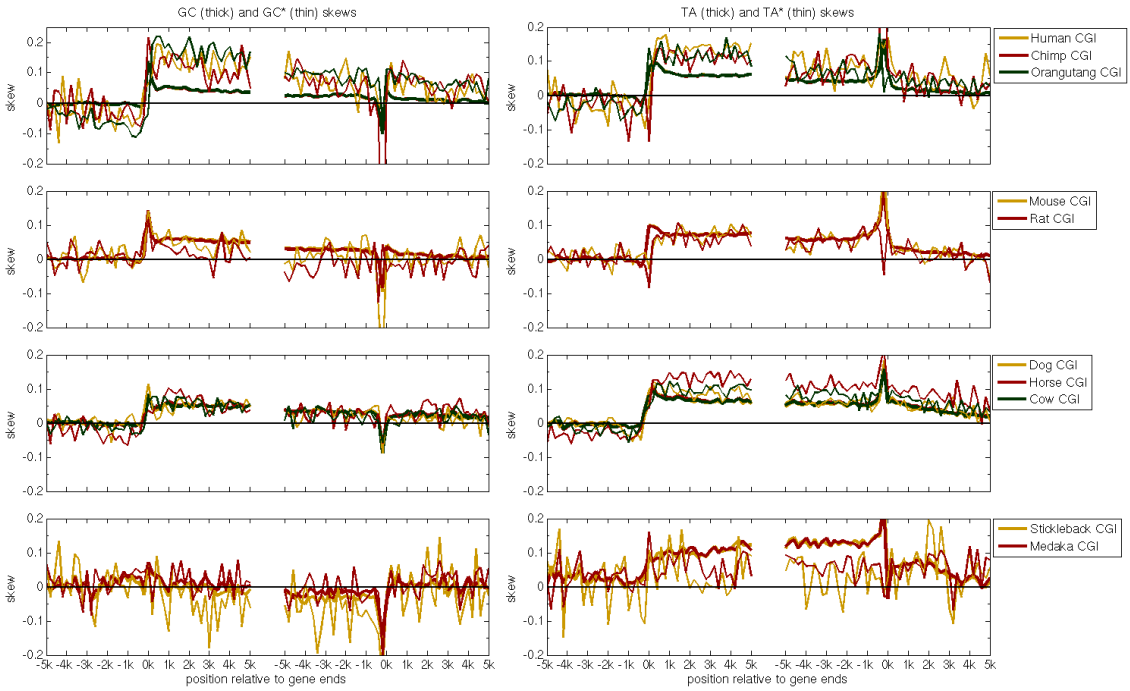

Figure 4.6: Current and stationary TA(*) and GC(*) skews along CGI-genes and their flanks. Current stationary skews are plotted with thicker lines. The skews are plotted against distance from the 5' and 3' ends of genes and are calculated along the non-template strand from pooled 200 bp windows of genes annotated for the reference species in each clade.

species except in primates, where we do not see a GC skew. Therefore the GC skew is opposite directions at both ends of genes. TA skews are positive along genes and are stronger near gene boundaries in all species, while being strongest at the 3'end. The increase in skews in gene boundaries, in particular at the 5' ends, is also found in the stationary nucleotide composition in most of species computed solely from the nucleotide substitution rates (Figures 4.6 and B.5). This suggests that substitutional biases acting over evolutionary time scales have generated these spikes in the skews.

4.7 Possible mutational processes that generate localized asymmetries in genes

Is it selection or mutational bias that has generated the substitution asymmetries in mammalian transcripts? The localized strand asymmetries are detected in different parts of the transcript and as we saw are found in all mammalian genomes that we analyzed. This indicates that the substitution force that has gener-

ated asymmetries acts on all part of the transcripts and is conserved over time. Zhang et al. [164] propose that the action of selection on splicing enhancers shaped the strand asymmetries along transcripts, while Green et al. [60] suggest that these are biases in repair process (see below). However, surveys of somatic mutations in two cell lines derived from tumor [120; 119] suggest that strand asymmetries are a consequence of mutational activity. In the two cell lines there was an excess of $r_{A\to G}$, $r_{C\to G}$, $r_{G\to T}$ and $r_{A\to T}$ over their base complementary substitution rates. Since these rates were determined using several thousands of somatic mutation instances it is more likely that bias in mutagenesis or in repair rather than selection has generated the strand asymmetries in the transcribed regions of the cancer cell lines.

Molecular mechanisms that can lead to localization of mutations Which mechanisms are responsible for the localization of cytosine deamination in the non-transcribed strand near the TSS? One possible process coupled to transcription is the activity of the Activation Induced (Cytosine) Deamination (AID) enzyme as part of the somatic hypermutation (SHM) pathway during B-cell development [112]. This enzyme induces both single nucleotide mutations and rearrangement at the immunoglobulin gene loci. It promotes cytosine deamination in ssDNA during transcription, but just at 1-2 kbp downstream to the TSS [112]. Very recently AID has been suggested to play a role in de-methylation in sperm at early stages of embryogensis by promoting deamination of methylated cytosine into uracils. Since CpG islands are known to have lower methylation levels in germ cells it is tempting to propose the AID can actively promote CGI de-methylation. During this hypothetical de-methylation, ssDNA can be formed.

AID activity is thought to be focused on so-called AID hot spots [118]. There is growing evidence that AID is expressed in germ cells, and it has even been reported to be express in the nucleus, which leads to the conjecture that AID can target ssDNA in these cells [137]. The probability that a cytosine will be targeted by AID is affected by its flanking sequence; AID targets hot spots in the form of RGYW/WRCY, where R is purine, Y is pyrimidine and W is A or T [10; 118; 132]. It has been reported that 50% of the mutations mediated by AID happened in target sites which are just 25% of all possible sequences

We propose AID can form the localized mutation, since AID mediates mutations ($C \to T$) on the first 1 kbp downstream to TSS regions of its target genes similar to the region we observed for the local asymmetry. In order to test this hypothesis, we divided the cytosine and guanine in each window into two sets, ones that are within hot spots, and the complementary ones not in hot spots and not part of CpG. The substitution rates were calculated by parsimony analysis, which attempts to attach the shortest evolutionary path from the ancestral sequence. In this analysis, we took into account only sequences where the sequences in rhesus and chimp were the same. Hence by parsimonious we assumed that ancestral sequence of human-chimp-rhesus was the one in chimp and rhesus, and the ancestral sequence was mutated in human

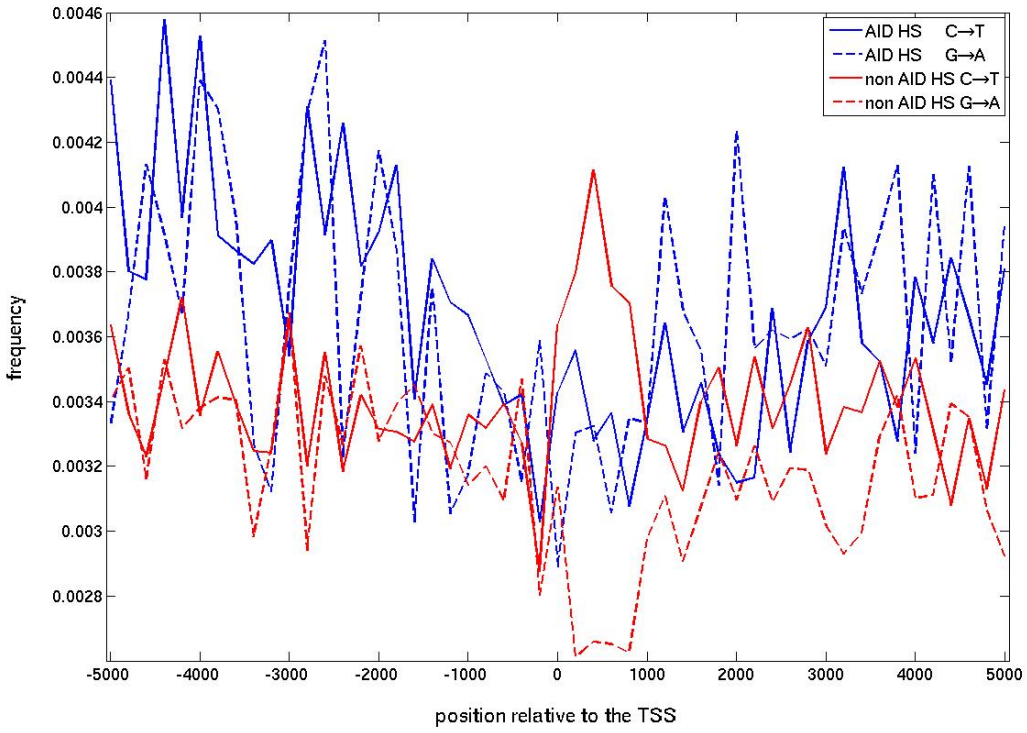

Figure 4.7: Transition frequencies of cytosine and guanine in/off AID hot spots (HS).

after the eventual split between human and chimp. Comparing the mutation patterns in these sets revealed that asymmetry does not appear in hot spots but rather in the non-hot-spots sites (Figure 4.7). Therefore, we could not substantiate the involvement of AID in the generation of the localized asymmetry on a genome wide scale.

However, there are other DNA deaminases that target ssDNA which have different sequence specificities. For example, APOBEC3F has a preference to CCN or TCN (where is N is an arbitrary base) [10]. Mutations in these motifs are higher in our data set than in the complementary motifs (data not shown), however APOBEC3F has been shown to deaminate ssDNA just in the cytoplasm [134]. Other DNA deaminases unfortunately have not yet been analyzed so extensively as AID.

Even though we could not directly associate DNA modifiers like AID with the local asymmetry, the similarity between the localization of the strand asymmetry in substitution rates and the restrictions of SHM might be the result of similar mechanisms. Possible mechanisms, which can elevate mutagenesis, are the ones that lead to formation of a ssDNA of the non transcribed strand. A ssDNA structure has been proposed to occur in variable (V-) regions during SHM in order to provide the substrate for AID, which mutates only ssDNA. Furthermore, the nucleotides in ssDNA are subjected to higher spontaneous DNA damage than nucleotides in dsDNA. Cytosines in ssDNA are

more prone to deamination than in dsDNA [47; 12]. Hence, if a localized ssDNA of the non transcribed strand is formed near the TSS, it might explain the higher rate of $C \to T$ transitions on this strand. Interestingly, the observed excess of $C \to T$ over $G \to A$ in bacterial genomes has been also suggested to be mediated by ssDNA on the non transcribed strand [43]. Since the average length of a bacterial gene in 1-2kb the induced range of the asymmetry along bacterial genes is similar to the asymmetry pattern we observed for human genes [43]. Hence, we propose that localized conformation of ssDNA near the TSS can explain the observed strand asymmetry

There are a couple of mechanisms that have been suggested to invoke ssDNA in V-regions. One of these mechanisms is the formation of RNA/DNA hybrids (R-loops) on the template strand in the first 1-2 kbp of the variable (V-) regions of immunoglobulin genes [163; 68]. R-loops are formed on the transcribed strand, leaving the non-transcribed strand in ssDNA formation at a higher frequency than its complementary strand [87; 133]. Using a biophysical model [19] we calculated the differences between the free energy levels of RNA/DNA hybrids and DNA/DNA (i.e. dsDNA) structures at various distances from the TSS for genes in our dataset. According to this model, for most genes the RNA/DNA hybrid in the first 200 bp downstream of the TSS is more stable than DNA/DNA conformation (Figure 4.8). Moreover, the average difference between the energies of these two structures is peaked at the immediate 200bp long region downstream to the TSS (Figure 4.8). The predicted higher stability can be attributed to higher GC content and the GC skew near the TSS of human genes, since GpG di-nucleotides are the main contributors to the energetic difference between RNA/DNA hybrids and DNA/DNA structures [19]. It is important to note that until now it is not fully understood what the conditions for R-loops formation are. A couple of recent studies have suggested that R-loops are initiated from 50 bp long regions, which contain the motif GGGGCTGGGG and comprise at least 50% of Gs [67; 136]. We found about 300 genes that contain such regions in their first 1 kbp. Estimating substitution rates in 1 kbp windows along these genes indeed shows a higher degree of asymmetry (data not shown), but the small number of genes is not enough to establish a significant association. A recent study suggests that, on top of the higher stability of RNA/DNA compared to the DNA/DNA structure, a capping enzyme can promote formation of transcriptional R-loops in vitro [74]. Since the formation of a cap is a necessary process at early stages of transcription, this finding implies that R-loops are indeed found at higher frequencies near the TSS than in the rest of the transcript.

Beside R-loops non-B DNA conformations, called G-quadruplexes (G4), might also be formed near the TSS of human genes on the non transcribed strand [33]. The formation of a G4 structure in the non template strand and in parallel the formation of an R-loop on the template strand is often called G-loops and has also been observed in different situations [34]. It is not known when and where such structures are formed, but several DNA sequence motifs have been suggested to have higher probability to form G-quadruplexes [161]. Recent studies reveal an enrichment of such motifs in promoter regions and in the first 500 bp of human genes [33]. Therefore, these G4

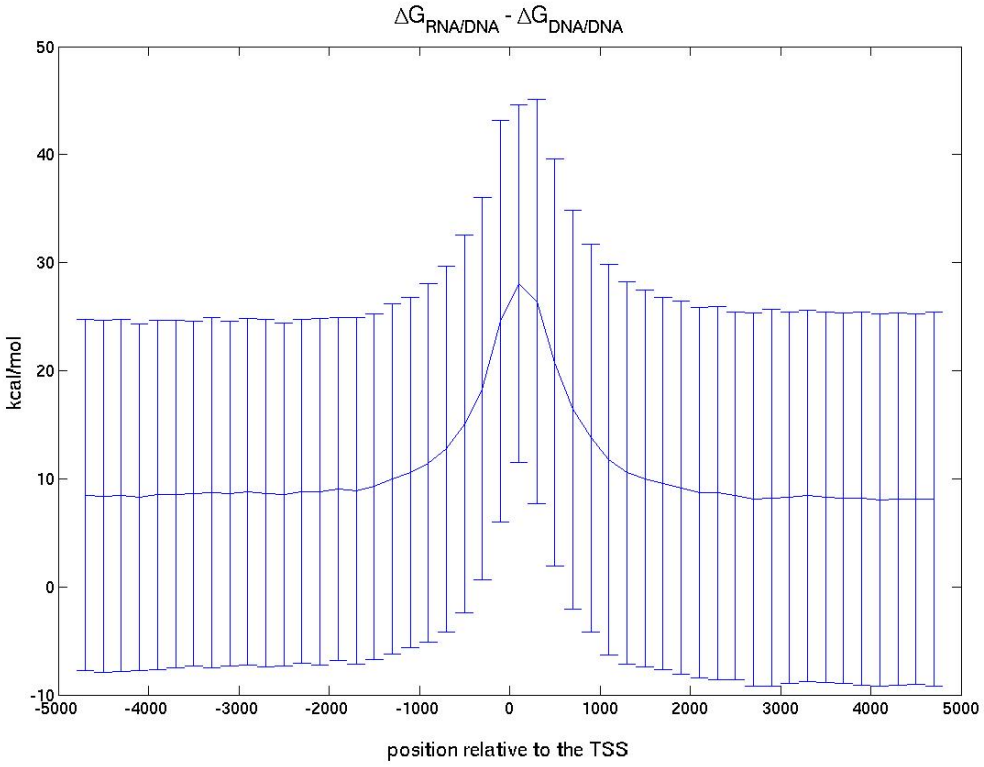

Figure 4.8: The mean (and the variation) of the energetic gap between the free energy of RNA/DNA and of DNA/DNA hybrids.

conformations have the potential to create a gradient of ssDNA of the non-template strand along human transcripts.

In summary, we propose the following model for the generation of the localized substitution bias. We assume that the non-transcribed strand at the start of genes is in ssDNA formation at a higher frequency than in regions further downstream. There are several mechanisms that may induce a localized formation of ssDNA of the transcribed strand. These mechanisms include either the higher occupation time of pol II near the TSS [104] or the formation of G/R-loops. As a consequence of these mechanisms, the transcribed strand near the TSS is protected either by RNA pol II complex or by the RNA/DNA hybrid, whereas the non-transcribed strand is left in ssDNA formation, which is prone to higher rates of cytosine deamination. This is due to spontaneous chemical processes [47] or due to the enzymatic activity of DNA deaminases like AID or APOBEC3 [112; 134]. These processes eventually can invoke the observed higher rate of $C \rightarrow T$ transition on the non-transcribed strand.

4.8 Mutational processes that may generate global asymmetries

The global strand asymmetry might also share similarity with SHM The excess of $A \to G$ over $T \to C$ substitutions in mammals is proposed [60] to be a byproduct of transcription coupled repair (TCR) (see Chapter 1). TCR is activated when RNA polymerase II stalls due to DNA lesions [63]. Green et al. [60] suggested that lesions that halt RNA pol II can be base mismatches that are formed during replication. There are two replication errors contributing to substitutions of A in G on the non-transcribed strand. The first error is misincorporation of Gs at template Ts when the non-transcribed strand is copied. The second error is misincorporation of Cs at template As during the replication of the transcribed strand; these errors result in G-T and A-C (non transcribed-transcribed) mismatches, which become the substitution-mutations $A \to G$ on the non-transcribed strand when they are repaired into G-C base pairs. In a similar fashion, a substitution $T \to C$ on the non-transcribed strand is a result of misincorporations of Cs at template As and of Gs at template Ts, which results in C-A and T-G mismatches. When the base pair mismatches are repaired via the TCR pathway, the non-transcribed strand serves as a template for the correction of the transcribed strand. Therefore, the balance between the rates of $A \to G$ and $T \to C$ is determined by the balance between misincorporations and their repair. Since the miss-incorporation rate of purines into the strand being copied is higher than the rate of pyrimidines, the rate $r_{A\to G}$ is higher than of $r_{T\to C}$ on the non-transcribed strand.

The above model has been criticized by different groups mainly because mismatches in the DNA are not expected to halt transcription and therefore should not initiate TCR. An alternative model, that we suggest, is that mismatches are introduced by a removal of DNA lesions that block transcription via error-prone nucleotide-excision repair (NER) mechanisms. According to this, the DNA lesion is excised together with a surrounding DNA sequence and the resulting gap [114] is filled by a low fidelity DNA polymerase, such as polβ [18]. As a consequence the newly synthesized DNA is prone to accumulate mutations and to form DNA mismatches. Unfortunately, the in-vivo error spectra of DNA polymerases are not known and therefore we can not determine which DNA polymerases might be the best candidates to explain the observed biases without further studies.

In a recent paper, Steele reviewed the data on mutations that result from SHM along introns of genes that are part of the immunoglublin locus [142]. Steele found that in this locus, on the non-transcribed strand, the mutation rates of $A \to G$, $A \to T$ and $A \to C$ excess their reverse complement mutation rates. We also observe an excess of $r_{A\to G}$ over $r_{T\to C}$ and an excess of $r_{A\to T}$ over $r_{T\to A}$ along human introns (Figure 3.1). But on the other hand, we find no or only a slight excess of $r_{A\to C}$ over

$r_{T\to G}$ (Figure 3.1). These similarities imply that SHM and TCR employ similar but not identical sets of proteins.

How are strand biases introduced in SHM? During SHM, the induction of AID leads to the deamination of cytosines into uracils on the non-transcribed strand of immunoglobulin genes [98]. The U:G mismatches are assumed to invoke mutations in two phases [111]. The first phase, the handling of U:G mismatches leads to transitions or transversions of C:G base pairs. If one strand is primarily targeted by AID, then the rates of $C \to T$, $C \to G$ and $C \to A$ will exceed the rates of the reverse complement strand. The second phase, mutations can occur in the sequence surrounding of the U:G mismatch mainly in A:T base pairs [142]. This has been suggested to be due to the removal of parts of the surrounding sequence containing an abasic site that is created when the uracil is excised by uracil-DNA glyocsylase (UNG). The resulting gap is then filled by the error prone DNA polη [126] promoting mutations of the type $A \to T$. As a consequence, the non-transcribed strand of immunoglobulin genes accumulates transversions in $A \to T$ over $T \to A$, as we also find in mammalian introns. Even though the above mentioned pathways can explain part of the global asymmetries along introns, there are two arguments against these hypotheses. First, in SHM the non-transcribed strand is often damaged and is also the one that is actively repaired [98], while in TCR, the transcribed strand is subject to repair processes [63]. Second, the bias of $r_{A\to G}$ over $r_{T\to C}$ can not be explained by the traditional models for SHM.

RNA mediated repair model for TCR An intriguing model, which has been recently suggested for SHM, provides a mechanism that can explain the observed global asymmetries, including the excess of $r_{A\to G}$ over $r_{T\to C}$ on the non-transcribed strand, via a repair of the transcribed strand. Higher mutation rates in As on the non-transcribed strand of immunoglobulin genes have been postulated to be caused by repair of the transcribed strand, via a combination of reverse transcription and RNA editing [46]. According to this model, a DNA strand that contains abasic sites and/or uracil is first transcribed into RNA. This RNA then forms a secondary structure with hairpins. Adenosines in the RNA hairpins are known targets for the RNA editing enzyme adenosine deaminase (ADAR) and are converted to inosines [46]. In the next step, reverse transcription of such an edited RNA results in a DNA sequence that is inserted into the DNA strand instead of the damaged fragment. Compared to the original transcribed sequence the newly synthesized transcribed-strand accumulates several mutations $T \to C$, $T \to A$ and $T \to G$ [46] that appear on the non-transcribed strand as $A \to G$, $A \to T$ and $A \to C$. By our results for introns, the substitution frequencies of $A \to G$ and $A \to T$ exceed their reverse complement substitution frequencies, while $r_{A\to C}$ is only slightly exceeding $r_{T\to G}$ (Figure 3.1). This suggests that such a repair mechanism can introduce the global strand asymmetries along genes.

5 Asymmetries in intergenic regions originate in CpG Islands

CpG islands (CGIs) are suggested to be origins of bidirectional replication (OBRs) and origins of antisense transcription. Since both replication and transcription processes can lead to strand asymmetry, such an association should leave a trace in the intergenic regions of the genome in the form of strand asymmetries in mutational rate. Substitution analysis around CGI reveals bidirectional asymmetries between complementary substitution rates that are similar to the one that is found around the OBR in bacteria. These asymmetries may be induced due to differences in the replication of the leading and lagging strand and the bidirectionality of substitution implies that a significant number of CGIs overlaps OBRs.

5.1 Strand asymmetries in intergenic regions in the vicinity of genes

Opposite strand asymmetries upstream to the 5'end of mammalian genes In upstream regions to the 5'end human genes, there is an opposite bias in the $A \to G$ vs. $T \to C$ substitution frequencies compared to those in 5' downstream regions (Figure 3.1). These asymmetries in upstream regions are found in CGI-genes but not in nonCGI genes (Figure 4.5). An excess of $r_{A\to G}$ over $r_{T\to C}$ is observed in the upstream regions of genes in other mammalian species. Furthermore, the ratio $r_{A\to G}/r_{T\to C}$ and $r_{C\to G}/r_{G\to C}$ is lower than one in the upstream regions of all mammalian CGI-genes, while in nonCGI-genes the rates are either equal to one as seen in primates and rodents or greater than one as in laurasiatheria. For $r_{G\to A}/r_{C\to A}$ and $r_{A\to T}/r_{T\to A}$ the ratio is close to 1 in the 5' flanking regions of the genes in most mammalian species (Figures B.3 and B.4).

Skews in upstream regions In the upstream regions of primate and laurasiatheria CGI-genes one might expect to find skews since there are strand asymmetries in substitution rates in these regions (Figure 4.6). However, the TA and GC skews are close to 0 in the 5' intergenic flanking regions of CGI-genes in both taxa. This might imply that even though we observe that two out of six of the single-nucleotide substitution rates are strand asymmetric, it is not enough to induce non-zero skews. Indeed, in

laurasiatheria, the stationary skews are absent, which indicates that other mutational processes compensate the impact of strand specific nucleotide mutations. But the stationary TA and GC skews in primates suggest that the strand asymmetries are able to induce strand asymmetries in nucleotide distribution in these regions (Figure 4.6).

5.2 Long range strand asymmetries around CpG islands

Strand asymmetries are found in intergenic regions which are located upstream to the TSS of CGI related genes The reversal of the asymmetry in the upstream regions of CGIs might not come as a surprise since many of CGIs are found to be bidirectional promoters and origins of bidirectional replication. Therefore, it is possible that reversal of the asymmetry in upstream regions of genes reflects an extensive antisense transcription [116] and in particular originating from CGI. Therefore, we wondered what is the range of the asymmetries in intergenic regions around CGIs. To investigate this, we derived the profile of 18 nucleotide substitutions rates around CGI (see Chapter 2). Then in order to measure the deviation of substitutions from strand symmetric case we calculated the ratios between pairs of complementary transition rates i.e. the ratio of the rates of $C \to T$ over $G \to A$ and the ratio of rates of $A \to G$ over $T \to C$ (Figure 5.1). In the intergenic regions upstream to tCGI and dCGI, the ratio $r_{A\to G}/r_{T\to C}$ is less than 1 (the mean and standard deviation of the ratio in the first 10 windows upstream to dCGIs is 0.934 +/- 0.02), whereas, in intergenic downstream regions it is greater than 1 (mean and standard deviation in the first 10 windows downstream is 1.09 +/- 0.023). In addition, on both sides of the CGI (5' and 3') the $r_{A\to G}/r_{T\to C}$ ratio is constant along several hundreds of kbp. These properties of the ratio profile indicate that there is a bidirectional asymmetry around CGIs which has the CGI as its origin and that the mutational pressure which leads to the asymmetry is constant over long distances. In contrast, the ratio of $r_{C\to T}/r_{G\to A}$ fluctuates around 1 along the flanking intergenic regions of CGI, an indication that the strand symmetry between these rates holds around CGIs. Among the four pairs of complementary transversion rates, three are symmetric and only the symmetry between the rates of $C \to G$ to $G \to C$ is broken (Figures C.2 and C.3).

We also performed the substitution analysis around the middle point between two consecutive CGIs. On a genome-wide scale, there is a change in the direction of strand asymmetries at the mid-point (Figure 5.2). The ratio $r_{A\to G}/r_{T\to C}$ is greater than 1 in the region upstream of the mid-point whereas, in the downstream region the ratio is lower than 1. The change in the direction of the asymmetry is smooth (Figure 5.2) around the mid-point, in contrast to the sharp change in the $r_{A\to G}/r_{T\to C}$ ratio at the CGIs (Figure 5.1). The analysis around the mid-point rules out the scenario that the ratio profiles in Figure 5.1 can be found around any random positions in the

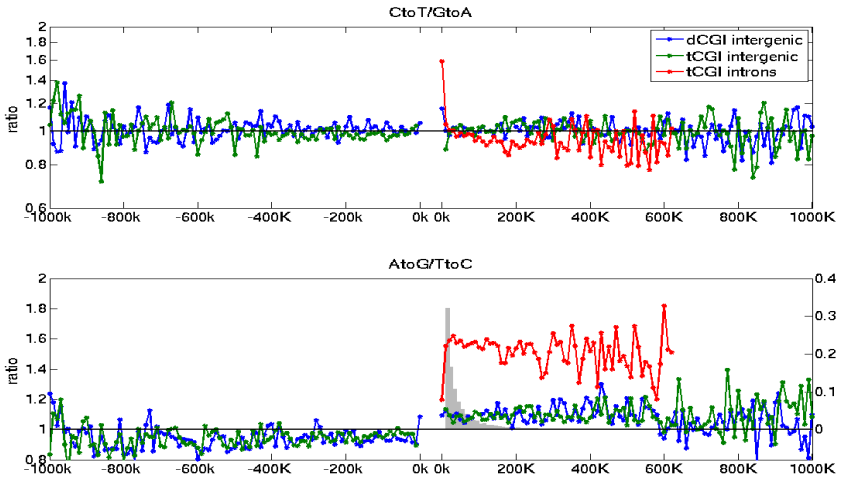

Figure 5.1: Ratios between complementary substitution rates in intergenic (blue, green) and intronic (red) regions. The ratios are plotted against the distance from the 5' end (left 0 k) and 3' end (right 0 k) of CGIs calculated in 10-kbp long windows. For dCGI and tCGIs, the analyzed sequences are intergenic (see corresponding blue and green bars in Figure 2.2) and are taken from the reference strand as it is described in Figure 2.2. The analyzed intronic sequences are of genes that their TSS is located within tCGI (see red bars in Figure 2.2). The ratios in these regions are computed using the nontemplate strand of a gene (see Figure 2.2); intronic sequences are only available for the 3' side (left to the gap) analysis. The ratios at 0 k are calculated within the CGIs, for details, see Figure 4.4. A shaded histogram of gene lengths in the human genome is presented at the bottom panel demonstrating that strand asymmetries between $A \to G$ versus $T \to C$ extend over distances larger than of a typical length of a gene.

genome; and therefore it establishes the role of CGIs as origins of the bidirectional asymmetry.

The strand bias is quantitatively similar in tCGIs and dCGIs As we saw in Chapter 4, transcribed regions (excluding exons) show a similar qualitative asymmetry, i.e. $r_{A\to G}/r_{T\to C} > 1$. A similar excess of $A \to G$ over $T \to C$ in introns has been suggested to be a result of transcription coupled repair (TCR). If only transcription induces the asymmetries in intergenic regions around transcripts, then one would expect a higher level of strand bias around tCGI than around dCGIs. However, the ratio profile is quantitatively identical around these two classes of CGIs. This observation is a strong indication that the transcription of genes is not the cause of long range asymmetries (a point which will be further addressed below).

The ratio $r_{A\to G}/r_{T\to C}$ is greater in introns than in intergenic regions To check whether transcription has an additional impact on the ratios between complementary substitution rates, we also performed a sliding window analysis along intronic regions of genes whose TSS is inside of a tCGI and we denote them as CGI-genes (see

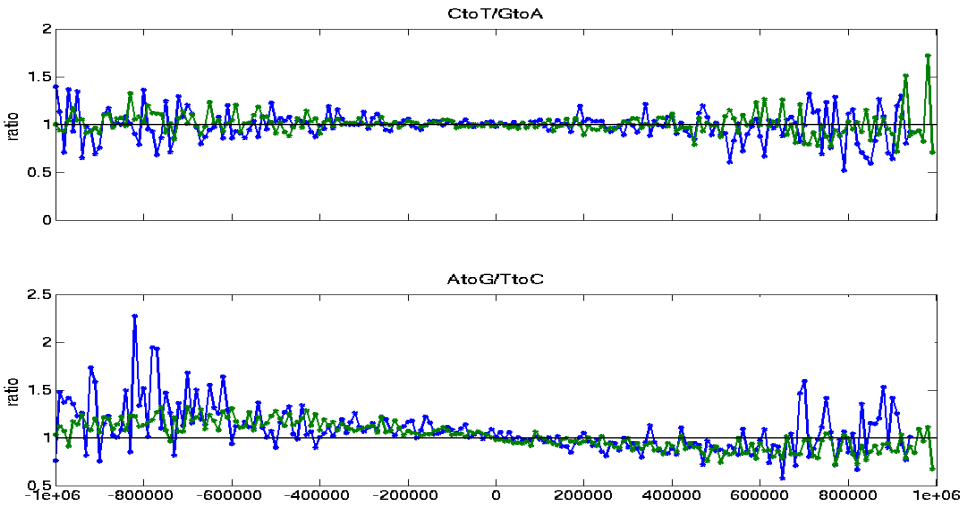

Figure 5.2: Ratios of complementary transitions rates in intergenic regions plotted against the distance from the middle point (0k) between consecutive CGIs calculated in 10 kbp long windows. All analyzed regions are interegenic. The middle point is between a CGI and its 5' adjacent CGI. The results for two classes of CGI are presents: tCGI (blue) and dCGI (green).

Figure 2.2 and Chapter 2). For these introns, the ratios were calculated in windows of 10 kbp starting from the TSS up to 1Mbps downstream to the tCGI (Figure C.4). We found out that in introns the ratio of $A \to G$ to $T \to C$ rates is 1.6 compared to only 1.1 in intergenic regions at a similar distance from the 3'end of tCGIs (Figure 5.1). Hence, the greater excess of $A \to G$ over $T \to C$ substitutions along the non-template strand of CGI-genes compared to intergenic regions suggests that the transcription process has a primary impact on the asymmetries in introns. From the fact that the ratio $r_{A\to G}/r_{T\to C}$ is relatively constant along introns (Figure 5.1) one can conclude that these transcription associated mutational forces are not dependent on the distance from CGIs. Note that by our definition (see Figure 2.2 and Chapter 2), at any given bin the intergenic regions are pooled out of the 3' intergenic flanking regions of CGI-genes that are shorter than the ones whose sequences are taken for the analysis of the intronic regions.

In order to further test whether the transcription process is the primary force that shapes the asymmetries in introns we measured the ratios in intronic regions pooled from genes whose TSS is not inside a CGI (denoted as nonCGI-genes). We divided the set of nonCGI-genes into two classes according to the direction of their transcription relative to the closest CGI, which can be either dCGI or tCGI. The first class is composed of genes that are transcribed outwards from the closest CGI. The second class called inwards genes and it contains genes that are transcribed towards the closest CGI. We found that the ratio $r_{A\to G}/r_{T\to C}$ is about 1.4 on the non-template strand of both classes of genes (Figure 5.3). Therefore for some genes the direction of the

asymmetries is opposite to the one in the intergenic regions. In introns of genes that are transcribed outwards from the CGIs the direction of the asymmetry is the same as in intergenic regions, (Figure 5.3), whereas, introns of genes that are transcribed towards CGIs have asymmetries opposite to those found in the average intergenic regions at similar distances from the CGI (see Figure 5.3 and Figure 5.1). Hence, the direction of the strand asymmetries along genes is determined by the direction of transcription.

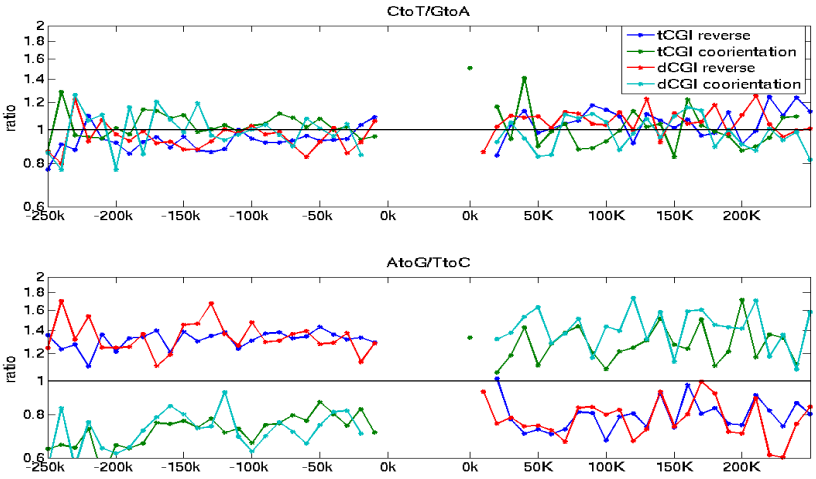

Figure 5.3: Ratios of complementary transition rates in intronic regions of inward and outward genes relative to tCGIs and dCGIs. All rates are calculated using the reference strand as defined in the Chapter 2.

Low $r_{A\to G}/r_{T\to C}$ ratio in introns that overlap tCGIs In intronic regions that overlap tCGIs the ratio $r_{A\to G}/r_{T\to C}$ is about 1.1 (this is the first bin in the analysis of 3' intronic regions in Figure 5.1 and more detailed analysis can be found in Figure 4.4), while in intronic regions downstream to the tCGI, the ratio fluctuates between 1.5 and 1.7 (Figure 5.1). Moreover, this ratio in introns that overlap tCGIs is similar to the one in intergenic regions (Figure 5.1). Therefore, the asymmetry level within tCGI is not dependent on transcription; this implies that TCR is not active in the transcribed parts of CGIs or that its impact on substitution pattern is obscured by other mutational processes.

Additional asymmetry between $C \to T$ and $G \to A$ is found in intronic regions In contrast to intergenic regions, we observed that in intronic regions the symmetry is additionally broken between $C \to T$ and $G \to A$ rates. This symmetry is broken in two different ways and on different scales. The first asymmetry is an excess of $C \to T$ over $G \to A$ that is restricted to intronic regions which overlap CGI (Figures 5.1 and 4.4). This asymmetry is localized at the first 2 kbp downstream to the TSS (Figure 4.5). An opposite and weaker bias is found in regions that are located 50-

500 kbp downstream to the CGIs (Figure 5.1). This bias has been previously observed by Green et al. [60] and has been suggested to be a result of TCR.

5.3 Mechanisms that can generate strand asymmetries in intergenic regions

Possible mechanisms that can generate bidirectional strand asymmetries What processes can generate these two patterns? How are these mutational processes associated with CGI? Replication can form strand asymmetries since it copies differently the two DNA strands. The bidirectional strand asymmetries in intergenic regions can be evidence that CGIs are origins of bidirectional replication. The association of CGIs with OBR [4; 55] has been suggested more than one decade ago and recently further established in a genome wide study [17]. Similar patterns of $r_{A\to G}/r_{T\to C}$, i.e. bidirectional asymmetries, have been associated with origin of replication in bacterial and mitochondrial genomes (see Chapter 1). Therefore, we suggest that the asymmetries around CGIs and the change in the direction of the asymmetry in the CGIs (Figure 5.1) are due to initiation of bidirectional replication in these regions.

The range of the asymmetries is several hundreds kbps which is similar to the average length of a replicon and one order of magnitude longer than the average size of transcripts (Figure 5.1). In case only transcription would invoke these asymmetries, CGIs would be origins of unknown transcripts of length of hundreds of kbps transcribed in both directions from CGIs. In particular, such transcripts should originate not only from tCGIs but also from dCGIs. Recent articles suggest that over 90% of the genome is transcribed, therefore most of annotated intergenic regions are transcribed at least once [146]. If such unknown transcripts induced this asymmetry their orientation would be biased relative to the CGIs, i.e. transcription would tend to occur in outward direction from the CGIs. The estimated rates for introns and for intergenic regions suggest that a surplus of at least 17 % of the nucleotides in the intergenic regions around CGIs should be transcribed in outward orientation relative to CGIs; however there is not sufficient data at this time to check this prediction.

Are all CGIs associated with OBRs? Cadoret et al. [17], have found that 99 out of 506 CGIs in the ENCODE regions serve as OBRs in a specific cell line. It is possible that different CGIs are origins of replication in different cell types, since the factors that link between CGIs and OBRs might be cell specific. The substitution rates that we measured in the present study are the result of mutations that occurred in or during the production of human germ cells. Up to now, data on OBRs in these cell types is not available.

Strand asymmetries are due to a difference in DNA polymerases error spectra The association of OBR with strand asymmetries in the human genome

implies that on the leading strand, which serves as the template of a new lagging strand, the ratios $r_{A\to G}/r_{T\to C}$ and $r_{C\to G}/r_{G\to C}$ are greater than 1 (Figures 5.1, C.2 and C.3) as in many bacterial genomes [131]. Also in human mitochondrial DNA (mtDNA), one observes that along the so called heavy strand, the rate of $A \to G$ substitutions is higher than of $T \to C$ substitution [40].

There are two main mechanisms that can invoke a ratio $r_{A\to G}/r_{T\to C}$ greater than 1 on the leading strand (and on the heavy strand of mtDNA). In bacteria (and for mtDNA), it has been suggested that adenine on the leading (heavy) strand is subject to higher deamination rates since the leading (heavy) strand is found in ssDNA conformation and adenine is less stable in this conformation than when it is base-paired to T in dsDNA conformation. However, the exposure of ssDNA should also induce even stronger asymmetry between $C \to T$ and $G \to A$, since the rates of cytosine deamination in ssDNA are 140 higher than in dsDNA [47]. Indeed, along the heavy strand of the mitochondrial genome one observed a stronger excess of the rates of $C \to T$ over $G \to A$ than of $A \to G$ over $T \to C$ [40]. We find that the ratio $r_{C\to T}/r_{G\to A}$ is close to 1 (Figure 5.1) and therefore it is unlikely that the exposure of the leading strand is the main cause for the excess of $A \to G$ over $T \to C$ on this strand (even though it is possible that the repair mechanism which handles cytosine deamination is more efficient in human than in bacteria). This suggests that the mechanism that has introduced asymmetry along the human mitochondrial DNA is different than the one that has generated the asymmetry in the nuclear DNA.

Another potential source of the asymmetry between the leading and lagging strand has been found in yeast. A recent study [84] demonstrated that the lagging and leading strands are synthesized by two different DNA polymerases, polδ and polϵ, respectively. Assuming that two different polymerases are also used in human cells we suggest that the observed asymmetries are due to a difference between the error spectra of these two DNA polymerases. During synthesis of the nascent DNA there are 12 possible single base-base mismatches that can occur. Since there are 2 DNA polymerases participating in DNA replication there are 24 potential misincorporation rates. Each of the 12 single-base mutation rates is a combination of two errors that are caused by polδ and polϵ. The replication errors that contribute to substitution of A by G on the leading strand, are misincorporation of G opposite template T by polϵ. and of C opposite to template A by polδ that result in G-T and A-C (leading-lagging) mismatches. These mismatches become substitution-mutations $A \to G$ when they are repaired into G-C base pair. In a similar manner, substitution of T by C on the leading strand is a result of misincorporation by polδ of C opposite to A on the template strand and of and G opposite to T template on the tempalte strand by polδ, which results in C-A and T-G mismatches. Unfortunately, the error rate during replication of these two polymerase, polδ and polϵ are currently only known for yeast [147] but not for the human polymerases. Therefore, we can not examine this model but rather we predict that the rates that lead to $A \to G$ on the leading strand are higher than the error rates that lead to $T \to C$.

How are ORIs mechanistically linked to CpG islands? It is not yet known how ORIs are determined in mammalian species. It is even less understood why replication is initiated from CGIs [2]. Since in CGIs are enriched both in set of TSSs and ORIs, it is tempting to suggest that the same factors that are associated with transcription initiation are involved in replication initiation [3] such as: low methylation levels in CGI [4; 127], particular histone modifications [93] and transcription factor binding sites [17]. It has also been suggested that CGIs harbor the binding site for Origin Recognition Complex (ORC), which is essential for replication initiation [75]. It is further possible that transcription factors that are usually associated with transcription can also promote replication, e.g. c-Jun or c-Fos [106]. Transcription *per se* has been shown not to be necessary for the use of CGIs as ORI but transcription can impact the timing of replication, since replication is initiated first from transcribed regions [56].

Surplus of nearly 17% of unknown transcripits in outgoing direction from CGI can generate asymmetries We have already suggested that the asymmetries within intergenic regions might be caused by replication. An alternative model is that unknown transcription around CGI induces these asymmetries. The fact that the vast majority of the genome is transcribed [53] may suggest that transcription and TCR are active on a genome wide level and could over time have generated the bidirectional asymmetry around CGIs. However, in this case the transcriptional activity has to be biased to transcribe the strand in an outward orientation relative to the CGIs.

In order to quantify the amount of transcription that is needed to generate the observed asymmetry in intergenic regions, we assumed that in intergenic regions downstream to the CGI, a fraction p of the sequence is evolved according to an asymmetric model and (1-p) fraction is evolved according to a symmetric model, in which complementary substitution rates are equal to each other. Under our model the relation between the frequency of substitution of α in β in intergenic *(i)* regions, $r^i_{\alpha\to\beta}$ and the frequencies by asymmetric (as) and symmetric (s) models, $r^s_{\alpha\to\beta}$ and $r^{as}_{\alpha\to\beta}$, respectively, is described by:

$$r^i_{\alpha\to\beta} = pr^{as}_{\alpha\to\beta} + (1-p)r^s_{\alpha\to\beta} \tag{5.1}$$

The fraction p can be different for complementary bases since the frequency of such bases are, in general, not equal to each other in regions that evolved under asymmetric model, for example, there is an excess of Ts over As in intronic regions (Figure C.5). However, even in regions that have been (so far) known to be subject to the strongest asymmetric mutational forces in the genome, i.e. introns, the bias is not significantly high (less than 10% surplus of Ts over As in introns). Therefore, we approximate the A and T content to be the same in sequences, which evolved either according to symmetric or asymmetric model, and by that we can reduce the number of parameters in our model.

Therefore, the substitution rates $A \to G$ and $T \to C$ in intergenic regions are described by:

$$r^i_{A\to G} = pr^{as}_{A\to G} + (1-p)r^s_{A\to G} \tag{5.2}$$

and

$$r^i_{T\to C} = pr^{as}_{T\to C} + (1-p)r^s_{T\to C} \tag{5.3}$$

Notice that $r^s_{A\to G} = r^s_{T\to C}$ because of the definition of the complementary symmetric model. From the above two equations we derive that:

$$p = \frac{r^i_{A\to G} - r^i_{T\to C}}{r^{as}_{A\to G} - r^{as}_{T\to C}} \tag{5.4}$$

To get the rates of $A \to G$ and $T \to C$ in intergenic regions we averaged the rates that are given in Figure C.2 over 10 consecutive windows downstream to the 3' end of dCGI. To get the rates of the asymmetric model we averaged the substitution rates in intronic regions (Figure C.4) over 10 consecutive windows downstream to the 3' end of tCGI:

$$r^i_{A\to G} = 0.0033 \pm 0.0001, \quad r^{as}_{A\to G} = 0.0034 \pm 0.0001 \tag{5.5}$$

$$r^i_{T\to C} = 0.0031 \pm 0.0001, \quad r^{as}_{T\to C} = 0.0022 \pm 0.0001 \tag{5.6}$$

Using these values we found that at least $p = 0.167$ fraction of the DNA evolved under the asymmetric model, i.e. there is surplus of about 17% in transcription proceeding outward to the CGI, compared to transcription towards the CGI. Here, the parameters we used for the asymmetric model are estimated in transcribed regions of known genes. It is likely that the transcription levels of unknown genes are lower than of known genes and therefore the mutational bias should be also higher. Hence, 17% surplus of outward expression is a conservative lower bound of bidirectional transcription from a given CGI and it is reasonable to assume a higher excess.

Implication on genomic architecture: co-orientation of transcription and replication The association of CGI with replication and transcription suggests that these processes are coupled to each other through CGIs. Such association has several evolutionary consequences. CO-occurrence of ORI and TSSs implies that genes with CGI promoter are transcribed from the leading strand in human. It has been suggested before that in human, transcription is coordinated with the replication fork progress [69] similar to some bacterial species [130; 129]. The co-orientation of transcription and replication is suggested to reduce head-on collisions between DNA and RNA polymerases in bacteria [130]. On the other hand, there is an ongoing discussion whether most human genes are coded on the leading strand or not [69; 110].

At the vast majority of CGIs that overlap TSS there is a paused RNA polymerases at any given time [61] and therefore an absence of transcription does not exclude the possibility that the RNA polymerase and other parts of transcriptional machinery attract the replication machinery. Moreover, recent research suggests that in mammalian CpG islands there is an overproduction of short RNA even in the absence of a full transcript production [138]. Such short RNA are transcribed both from the sense and antisense strand, where the sense transcripts accumulate downstream to the TSS and the antisense are primarily found upstream to the TSS [138]. Similar phenomena have been discovered for replication initiation from CpG islands, where overproduction of short DNA sequences that overlap CGI have been detected during S phase of the cell cycle [55]. It is possible that transcription of several dozens bps of long non-coding RNAs is part of initiation of replication in particular of the leading strand.

6 Weak to strong bias in promoters and CpG islands

As was discussed in Chapter 1, animal genomes are rich in weak bases (A and T) and poor in strong bases (G and C). One of the mutational forces that lead to this enrichment is an excess of $S \to W$ substitutions (and mutations) over $W \to S$ at a genome wide level. However, there are DNA loci which are enriched in GC nucleotides. Among these regions are CGIs and regions surrounding the 5'end of genes (which often coincide with CGIs). One can predict that the ratio $r_{W\to S}/r_{S\to W}$ peaks in CGIs and at the TSS of mammalian genes. In this section we analyzed the profiles of $r_{W\to S}/r_{S\to W}$ and stationary GC content (GC) around and along genes and CGIs. We found that recombination played a role in the recent evolution of CGIs in human. A subclass of CGIs, the ones which are further than 10 kbp from any annotated gene, are subject to mutational forces that have increased the GC content in these subclass. In promoter of dog and stickleback we observe high $r_{W\to S}/r_{S\to W}$ ratio as in distal CGI we suggest that recombination via biased gene conversion is responsible for this increase.*

6.1 Correlation of $r_{W\to S}/r_{S\to W}$ ratio with crossover rates in human CpG islands

An excess of $W \to S$ over $S \to W$ is found within dCGI We found out that the ratios are different between tCGI and dCGI. In the last group the ratio of $r_{W\to S}/r_{S\to W}$ is about 2, whereas in tCGIs the ratio of $r_{W\to S}/r_{S\to W}$ is about 1 (Figure 6.1). In the immediate flanking regions of both types of CGIs this ratio is lower than 1 (Figure 6.1). Hence, the $r_{W\to S}/r_{S\to W}$ profile implies that a GC enriching substitution pattern is a feature of the CGIs but the level of this enrichment is higher in tCGIs.

In the next step, we used the estimated substitution rates to calculate the stationary GC content, denoted by GC*, i.e., the GC content that is acquired after a long time assuming that the substitution rates do not change [37]. GC* in dCGI (about 0.6, Figure 6.1) is higher than GC* in tCGI (about 0.42, Figure 6.1). The GC content in both classes of CGI is lower at equilibrium than the current GC frequency of 0.65 in both types of CGIs (Figure 6.1).

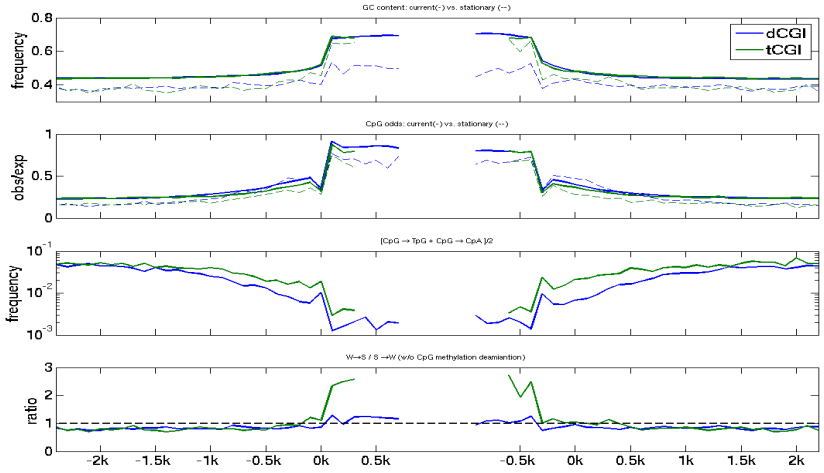

Figure 6.1: GC content, CpG odds, CpG methylation-deamination rates, $r_{W\to S}/r_{S\to W}$ ratio and total substitution rates in the proximity of 5 and 3 ends (left and right 0k, respectively) of dCGI and tCGI. In the top two panels, the current values of GC content and CpG odds (continuous lines) are compared with the corresponding stationary quantities (dashed lines). In the middle panel, $r_{CpG\to CpA+CpG\to TpG}/2$ (continuous) are compared with $r_{CpG\to CpT+CpG\to ApG}/2$ (dashed). The data points between the left and the right 0k are calculated within the CGIs.

The recombination rate is correlated with the $r_{W\to S}/r_{S\to W}$ ratio within CGIs
The primary process that is known to increase the $r_{W\to S}/r_{S\to W}$ ratio is recombination through GC-biased gene conversion (BGC) [50]. In regions where a heteroduplex is formed mismatches can occur. It has been suggested that repair of mismatches between S bases (GC) and W bases (AT) will be corrected preferentially into S bases [103]. Therefore, we further subdivided each class of CGIs according to the estimated recombination rates in every CGI. At the first stage, CGIs that overlap with recombination hotspot were grouped into a subclass called hotspots. At the second stage, the remaining CGIs were subdivided into further 3 subclasses according to their recombination rates: below 0.2 cM/Mb, 0.2-0.8 cM/Mb, above 0.8 cM/Mb. Indeed, there is a positive correlation between the ratio $r_{W\to S}/r_{S\to W}$ and the recombination rate in tCGI and in dCGI (Figure 6.2), this correlation exists also for CGI within genes (gCGIs see Chapter 2) and proximal CGI (pCGIs) that are CGIs at distance less of 10 kbp from a gene (Figure 6.2). In addition, the $r_{W\to S}/r_{S\to W}$ ratios in dCGIs are greater than the one in tCGI at any recombination level; this indicates that recombination is not the only factor that impact the $r_{W\to S}/r_{S\to W}$ ratio but also the absence or presence of TSS on CGI contributes to the bias of $W\to S$ over $S\to W$.

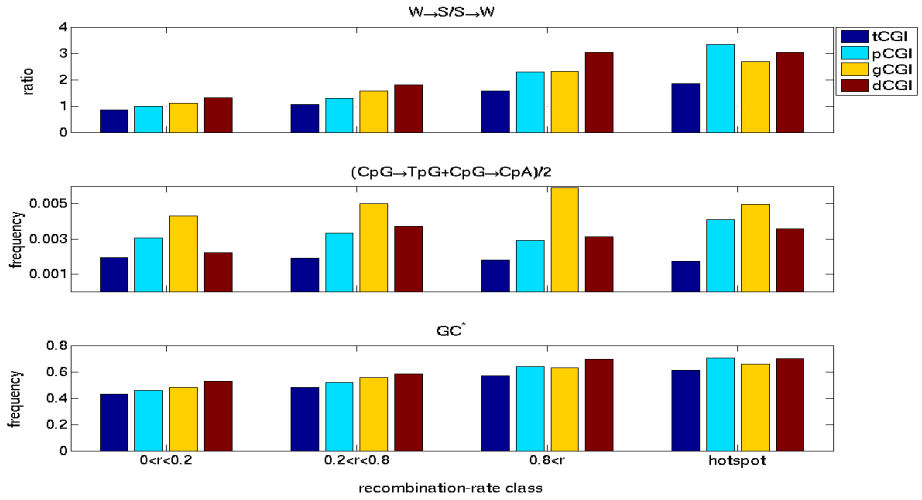

Figure 6.2: Dependence of $r_{W\to S}/r_{S\to W}$ ratio, CpG deamination frequencies, and the stationary GC content from the recombination rate for four classes of CGIs (see Chapter 2). The CGIs in the four (t-,p-,g-,d-) CGI classes are subdivided according to four recombination rate ranges ($0 \leq r < 0.2$, $0.2 \leq r < 0.8$, $0.8 \leq r$, hot spots), which are denoted on the horizontal axis.

6.2 Substitution signature of recombination in vertebrate promoters

The profile of GC stationary content along genes and their flanks Since mammalian promoters are associated with CGIs, the GC content near the TSS is higher by about 50% than the genome wide GC content [165]. We used the estimated substitution rates in order to calculate the stationary GC content (denoted by GC*) along genes (Figures 6.3 and B.6).

All mammals examined, except for the dog, show a clear peak in GC content at the TSS, while in the GC* profile the peak is lower indicating the loss of CpG islands. This trend is most pronounced in primate and rat CGI-genes, where the GC* profile is almost flat along the regions examined (Figure 6.3). This result is surprising, since CpG islands are usually believed to play a regulatory role in the transcription of genes [4] and therefore, one expects the current GC peaks to be preserved also in GC* profiles. Interestingly, given the close relation to human and chimpanzee, the orangutan genome shows a high rate of strong to weak substitutions leading to significantly lower GC content throughout the examined regions. In cow and horse genomes the GC* near the TSS is about 50%, which is higher than the value of 43% GC* content in distal regions (Figure 6.3). Also in the mouse genome, the expected GC* is, at 60%, well above 50% but still lower than the current GC content near the TSS. However, most surprisingly, in dogs, the current GC content is about 60% at the

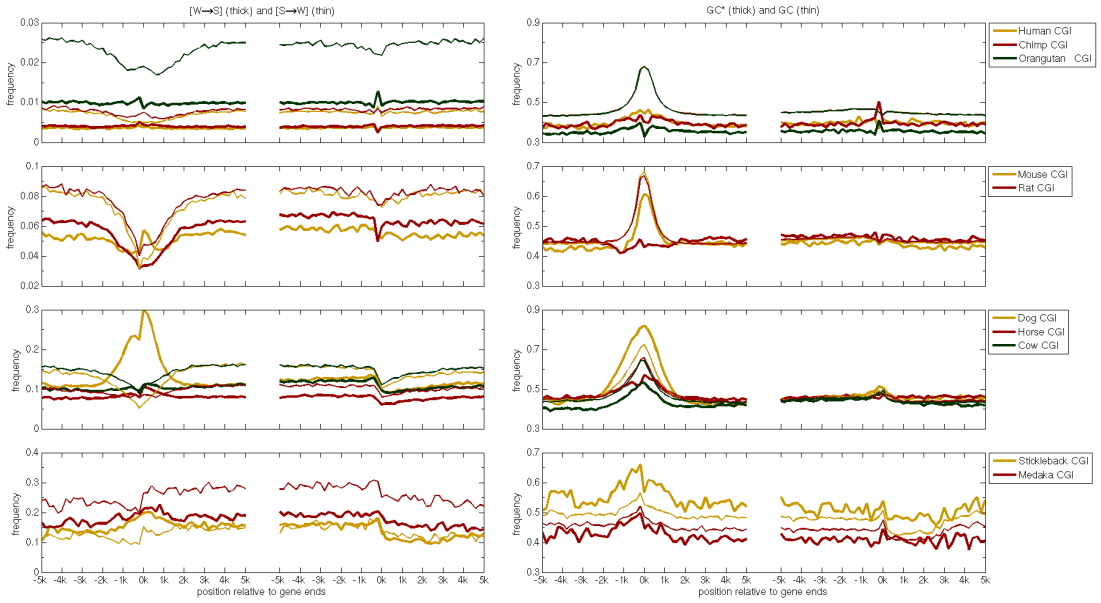

Figure 6.3: The weak to strong bias along CGI-genes and their flanks. The frequencies of $W \to S$ (thick line), $S \to W$ (thin line), the stationary GC content (GC*, thick) and the GC content (thin) are plotted against distance from the 5'end and 3'end of genes and calculated along the non-template strand from pooled 200 bp windows of genes annotated for the reference species in each taxon. Only the results for CGI-genes are presented.

TSS, while the GC* value is about 70%. So dog is the only mammalian species where the recent nucleotide substitution rates lead to an increase of GC content around the TSS.

In stickleback, the GC* content is higher than the current GC content along the regions we analyzed (Figure 6.3). The GC content at the 5'end of CGI-genes is slightly greater than 50%. However, the GC* is above 60% and hence stickleback promoters are expected to become richer in GC content than mammalian promoters. Also in the intergenic regions, GC* is higher than in other species; it is around 50% at the upstream intergenic regions and about 48% in the regions downstream to the TSS (Figure 6.3).

The ratio $r_{W\to S}/r_{S\to W}$ increases near the TSS of vertebrates genes The GC* content reflects the balance between the rates $r_{W\to S}$ and $r_{S\to W}$. Therefore, it is not surprising that the ratio $r_{W\to S}/r_{S\to W}$ increases near the TSS of vertebrate genes. In particular this is true for CGI-genes in mammals and stickleback. In most specie the modification of cytosine into uracil occurs at higher rates than other single base mutations; therefore, at the genome wide level the ratio $r_{W\to S}/r_{S\to W}$ is lower than 1 in animal genomes [70]. Since the rate of CpG methylation-deamination ($S \to W$ mutation) is significantly lower near mammalian TSSs than in the rest of the genome, it might induce the peak in the $r_{W\to S}/r_{S\to W}$ ratio profile near the TSS. However, even

after excluding the CpG methylation-deamination rates in the calculation of the ratio $r_{W\to S}/r_{S\to W}$ (using only the single-base substitution rates) one observes an increase of this ratio near the TSS of mammalian CGI-transcripts. The ratio is greater than 1 in the vicinity of the TSS of laurasiatheria and mouse genes. In stickleback in particular, the ratio is greater than 1 along all analyzed regions. These results imply that in non-CpG sites in vertebrates, the mutational pressures favor strong bases near the TSS.

6.3 BGC as a putative mechanism to increase GC content

Possible higher rates of biased gene conversion in CGI GC- biased gene conversion, a by-product of recombination, is considered to be the primary mutational process that increases the ratio $r_{W\to S}/r_{S\to W}$ in the genome [70; 103; 50; 37] (see also Chapter 1). Indeed, the $r_{W\to S}/r_{S\to W}$ ratio is correlated with the estimated recombination rates in the different CGI categories (Figure 6.2). However, while, the ratio $r_{W\to S}/r_{S\to W}$ is greater in CGIs than in their flanks, the average recombination rate is the same in CGIs and in their immediate flanking regions (Fig. 6.1). This implies that recombination can not solely explain the profile of $r_{W\to S}/r_{S\to W}$. We suggest two explanations that can account for this difference. First, the recombination rates that we use in this study are actually an estimation of the rate of crossing over events and they miss the non-crossing over events. Therefore, it is possible that the rates of recombination are locally higher within CGI regions but crossing over is suppressed due to purifying selection. Second, the $W \to S$ vs. $S \to W$ bias can be shaped by positive selection for high GC content in CGIs due to their role is gene regulation [143]. The GC rich sequence of CGIs is associated with an open chromatin structure that allows the binding of RNA polymerase and transcription factors to the DNA [4]. The above explain the preference for high stationary GC content in tCGI. The stronger bias of $W \to S$ vs. $S \to W$ in dCGI might point to distinct mutational mechanisms that shape the GC content in tCGI and dCGI.

Suggested model for CGI evolution We suggest that the gain of GC bases by BGC is a transient state. All CGIs have first emerged in the genome by BGC activity (which also increases the total substitution rates); but once a CGI becomes functional in the context of nearby genes (tCGI) it also becomes constrained and has lower mutation rates that preserves its GC content. Indeed, there is a difference in the total substitution rates between the two classes; within dCGI the rate is 40% greater than in tCGI (Figure 6.1). Interestingly, the total substitution rate is higher in both tCGI and dCGI than in their flanking regions suggesting that these regions have undergone a period of positive selection for high GC content or BGC (Figure 6.1).

Recombination rates might shaped GC content in vertebrate promoters The mutational pattern of BGC is observed near TSSs of CGI-genes in dog, cow, horse, stickleback and medaka (Figure 6.3). In rodents, the picture is slightly different. In contrast to other mammals, at the 2kb regions centered at the TSS, the substitution frequencies are lower than in regions that are found at a distance of 1-5kbp from the 5'ends. However, the drop in rates implies that negative selection near TSSs is stronger for mouse genes than in the rest of mammalian species. Within the 2kbp long regions centered in mouse CGI-gene TSSs, there is an increase (and a decrease) in $r_{W\to S}$ ($r_{S\to W}$) (Figure 6.3). In sum, the signature of BGC mutations rates is found near the TSS in all mammals and fish. However, these substitution patterns can be also formed due to a positive selection on the local GC content near the TSS. A pair of substitutions, which does not impact the GC content but has been found to be positively correlated with recombination rates is $G:C \to C:G$ [36]. In dog genes, these rates also increase in the vicinity of the 5'end of genes, in particular in CGI-genes (Figure B.9). Therefore, we suggest that BGC have shaped the patterns of $r_{W\to S}/r_{S\to W}$ ratios in the vicinity of the TSS in all species.

Average specie-chromosome length correlates with $r_{W\to S}/r_{S\to W}$ Previous studies have revealed that $r_{W\to S}/r_{S\to W}$ in dogs is significantly higher than in any other mammal. It has been speculated that this is due to the shorter length of the dog chromosomes, which is suggestive for higher recombination rates along the chromosome [70]. A support for this theory is that chicken micro-chromosomes, compared to macro-chromosomes, have a higher GC content, a higher observed/expected ratio of CpG, and a higher density of CpG islands [102]. This observation was recently extended to vertebrate genomes, in which the three genomic features are anticorrelated with the chromosomal length [62]. Therefore, two species that have a similar population size but different average chromosomal length would have different $r_{W\to S}/r_{S\to W}$ ratios. Dog, which has the shortest chromosomes among the mammals studied here, is expected to have the highest recombination rates.

In a similar fashion, one can explain the difference between stickleback and medaka. The average medaka chromosome is about 29 Mb long, 1.3 times longer than that of stickleback (22 Mb). If recombination rates are roughly reversely proportional to the chromosome length in stickleback, then the recombination rates should be about 1.3 times higher than in medaka. The value of $r_{W\to S}/r_{S\to W}$ in non-CpG sites in stickleback is about 1.3 compared to the ratio of 0.85 in medaka, that is, 50% greater. This difference can also explain the higher density of CGIs in the stickleback (150 CGI/Mb) genome compared to the medaka genome (37 CGI/Mb) [62].

Recombination rates are higher near the TSS due to adaptation Compared to cow and horse, in dogs genes the ratio of $r_{W\to S}$ over $r_{S\to W}$ is only higher near the TSS. Higher recombination rates along dog chromosomes can not solely explain the peak in $r_{W\to S}/r_{S\to W}$ near the TSS. Interestingly, dog and stickleback, which have the highest $r_{W\to S}/r_{S\to W}$ ratios, are the species that have experienced the greatest number

of adaptation events or bottlenecks among the studied species [22; 59]. The evolution of regulatory elements is often considered to be the prime driver of adaptation. Hence, it is possible that recombination rates are higher at the TSS in both dog and stickleback genes, due to the selection of newly recombinant promoter regions. The recombination rates along the genome are heterogeneous and determined by unknown factors; the rates themselves have been constantly changing during evolution. The recombination maps in human and chimpanzee reveal a low number of overlapping hotspots [123; 159]; the reason for that is not clear, but it is possible that selection can shape these rates across the genome during evolution. For example, in species like stickleback, which constantly adapt to new environments, recombination rates are expected to be higher near the TSS, since crossing over events can increase the fitness of the specie in a new environment. Another scenario is that recombination rates do not differ significantly between species, but the fixation rate of $W \rightarrow S$ substitutions increases near the TSS due to positive selection on GC rich motifs [165]. However, the fixation dynamic of these GC-alleles is similar to the one is caused by BGC [109] and therefore these scenarios are indistinguishable by sequence comparison [50].

Are CpG islands vanishing? In this study, a CpG island is defined as DNA sequence longer than 400bp that fulfills two conditions; (1) the GC content is above 50%; (2) the CpG observed/expected ratio is greater than 0.6 (see Chapter 2). However, the value of the stationary GC content in tCGIs (GC*=0.41) suggests that these regions are predicted to lose their CGI properties over a long period of time, while dCGI with (GC*=0.58) will keep them. This result is counter intuitive since tCGI are assumed to play a major role in gene regulation [4] and one would expect that substitution rates would maintain the GC content in the vicinity of TSS. It is possible that the mutational forces that increases the GC content and build up the CGIs are more active in these regions only prior to time of the association of CGI with transcription. When CGI harbor TSS and becomes a dominate regulator of gene transcription, the mutational forces that induces mutations might have deleterious impact. An alternative explanation might be that other mutational forces such as insertion and deletion can lead to an increase of GC content that counter balance the impact of nucleotide substitution rates.

7 Summary

The availability of mammalian genomes and their corresponding alignments together with high quality genome annotation enables us to gain insights into differences in mutational processes in different contexts along human chromosomes. In particular, one can address the question of substitution signatures that are associated with different cellular processes. We study the impact of transcription on substitution patterns in the vicinity of the 5' end and 3' end of genes. Also, an analysis of substitution patterns within and around CpG islands, which are mammalian sequence features, was presented. The analysis reveals rich and (to some extent) unexpected patterns of mutational patterns that are associated with transcription processes, CpG islands, or both.

There are three transcription-associated substitution patterns that have been observed in human, of which two are related to CpG islands. The first is a sharp decline in the deamination rate of methylated CpG dinucleotides, which is observed in the vicinity of the 5' end of genes due to abundance of CpG islands in these regions that are subject to lower methylation levels compared to CpG dinucleotides elsewhere in the genome. The second is a strand asymmetry in complementary substitution rates, which extends from the 5' end to 1 kbp downstream from the 3' end, associated with transcription-coupled repair. The third is a localized strand asymmetry, an excess of $C \to T$ over $G \to A$ substitutions in the nontemplate strand confined to the first 2 kbp downstream of the 5' end of genes at CpG islands. This pattern might be induced by a higher exposure of the nontemplate strand near the 5' end of genes that in turn leads to a higher cytosine deamination rate. This type of substitution asymmetry is similar to the one that is observed as a consequence of the somatic hypermutation pathway. It might be that various proteins that are active during somatic hypermutation, such as a DNA mutator, Activation Induced cytidine Deaminase (AID), which solely targets single-stranded DNA, are also active in mammalian germline cells. The necessary ssDNA conformation can be induced by R-loops or G4 structures, which preferentially occur at the 5' ends of genes.

The transcription-associated substitution patterns are not unique to human and can also be found in other mammalian species, such as chimpanzee, orangutan, mouse, rat, horse, cow and dog. Fish also have strand asymmetry patterns in introns, but these asymmetries are different to those in mammals, pointing out that transcription-associated repair or mutagenesis processes have been evolving in the vertebrate lineage.

Strand-specific substitution processes exist also in intergenic regions. CpG islands are origins of bidirectional strand asymmetries that extend over hundreds of thousands of base pairs. These asymmetries can be induced by a DNA replication process which has CpG islands as its initiation sites. Alternatively, these asymmetries in intergenic regions can be the signature of unknown transcripts, such as very long non-coding RNAs. In intergenic regions downstream of genes, there are strand asymmetries that are similar to the ones in introns, implying that RNA polymerase continues to transcribe regions even further than the 3' ends of genes.

In this thesis, we also study the ratio of $W \rightarrow S$ over $S \rightarrow W$ substitution frequencies. The genome-wide ratio is lower than 1 in mammals and in all animals that were sequenced so far. But in human CpG islands that are far from any annotated gene, the ratio is higher than 1. Previous studies showed that this ratio is positively correlated with crossing-over rates. When CpG islands are divided according to the crossing-over rates, the ratio increases with crossing-over rates. Therefore, we suggest that recombination, possibly via associated processes, such as biased gene conversion, plays a driving force in creating CpG islands. In CpG islands that overlap TSSs of human genes and of most mammalian genes there is an opposite bias, i.e. an excess of $S \rightarrow W$ substitution rate over $W \rightarrow S$, a bias that can lead to the loss of the CpG islands from TSS regions in the long term. In striking contrast, in dog and stickleback we observed an increase in GC content in the long term and we suggest that this is an evidence for the role of recombination in shaping the promoter regions of these species.

The GC content at CpG islands has been known as genomic punctuation marks in the landscape of CpG methylation-deamination rates. Therefore, results of this thesis imply that mammalian CpG islands are much more boundary elements for multiple mutation processes, especially the ones that establish strand asymmetries. This, we speculate, is a mutational trace of transcription and replication that genome-wide tend to initiate at CpG islands.

Bibliography

[1] S. Aerts, G. Thijs, M. Dabrowski, Y. Moreau, and B. De Moor. Comprehensive analysis of the base composition around the transcription start site in metazoa. *BMC Genomics*, 5(1):34, 2004.

[2] M. I. Aladjem. Replication in context: dynamic regulation of dna replication patterns in metazoans. *Nat Rev Genet*, 8(8):588–600, 2007.

[3] F. Antequera. Structure, function and evolution of cpg island promoters. *Cell Mol Life Sci*, 60:1647 – 1658, 2003.

[4] F. Antequera and A. Bird. Cpg islands as genomic footprints of promoters that are associated with replication origins. *Current Biology*, 9(17):R661–R667, 1999.

[5] P. F. Arndt, C. B. Burge, and T. Hwa. Dna sequence evolution with neighbor-dependent mutation. *J Comput Biol*, 10(3-4):313–322, 2003.

[6] P. F. Arndt and T. Hwa. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, 21(10):2322–2328, 2005.

[7] P. F. Arndt, T. Hwa, and D. A. Petrov. Substantial regional variation in substitution rates in the human genome: importance of gc content, gene density, and telomere-specific effects. *J Mol Evol*, 60(6):748–763, 2005.

[8] P. F. Arndt, D. A. Petrov, and T. Hwa. Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. *Mol Biol Evol*, 20(11):1887–1896, 2003.

[9] N. Arnheim and P. Calabrese. Understanding what determines the frequency and pattern of human germline mutations. *Nat Rev Genet*, 10(7):478–488, 2009.

[10] R. C. L. Beale, S. K. Petersen-Mahrt, I. N. Watt, R. S. Harris, C. Rada, and M. S. Neuberger. Comparison of the differential context-dependence of dna deamination by apobec enzymes: correlation with mutation spectra in vivo. *J Mol Biol*, 337(3):585–596, 2004.

[11] K. Bebenek and T. A. Kunkel. Analyzing fidelity of dna polymerases. *Methods Enzymol*, 262:217–232, 1995.

[12] A. Beletskii and A. S. Bhagwat. Transcription-induced mutations: increase in c to t mutations in the nontranscribed strand during transcription in escherichia coli. *Proc Natl Acad Sci U S A*, 93(24):13919 – 13924, 1996.

[13] A. P. Bird. Dna methylation and the frequency of cpg in animal dna. *Nucleic Acids Res*, 8(7):1499 – 1504, 1980.

[14] A. P. Bird. Cpg islands as gene markers in the vertebrate nucleus. *Trends Genet*, 3:342 – 347, 1987.

[15] C. Bock and T. Lengauer. Computational epigenetics. *Bioinformatics*, 24(1):1–10, 2008.

[16] V. A. Bohr, C. A. Smith, D. S. Okumoto, and P. C. Hanawalt. Dna repair in an active gene: removal of pyrimidine dimers from the dhfr gene of cho cells is much more efficient than in the genome overall. *Cell*, 40(2):359–369, Feb 1985.

[17] J. C. Cadoret, F. Meisch, V. Hassan-Zadeh, I. Luyten, C. Guillet, L. Duret, H. Quesneville, and M. Prioleau. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *PNAS*, 105(41):15837–15842, 2008.

[18] Y. Canitrot, M. Frechet, L. Servant, C. Cazaux, and J. S. Hoffmann. Overexpression of dna polymerase beta: a genomic instability enhancer process. *FASEB J*, 13(9):1107–1111, 1999.

[19] E. Carlon and T. Heim. Thermodynamics of rna/dna hybridization in high-density oligonucleotide microarrays. *Phys. A: Stat. Mech. Appl.*, 362:433–449, 2006.

[20] J.-V. Chamary and L. D. Hurst. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol Biol Evol*, 21(6):1014–1023, Jun 2004.

[21] W.-H. Chen, J. de Meaux, and M. J. Lercher. Co-expression of neighbouring genes in arabidopsis: separating chromatin effects from direct interactions. *BMC Genomics*, 11(1):178, 2010.

[22] P. F. Colosimo, K. E. Hosemann, S. Balabhadra, G. Villarreal, M. Dickson, J. Grimwood, J. Schmutz, R. M. Myers, D. Schluter, and D. M. Kingsley. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, 307(5717):1928–1933, 2005.

[23] I. H. G. S. Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[24] M. G. S. Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[25] R. M. G. S. A. Consortium. Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, 316(5822):222–234, 2007.

[26] M. S. Cooke, M. D. Evans, M. Dizdaroglu, and J. Lunec. Oxidative dna damage: mechanisms, mutation, and disease. *FASEB J*, 17(10):1195–1214, 2003.

[27] G. M. Cooper, D. A. Nickerson, and E. E. Eichler. Mutational and selective effects on copy-number variants in the human genome. *Nat Genet*, 2007.

[28] E. C. Cox and C. Yanofsky. Altered base ratios in the dna of an escherichia coli mutator strain. *Proc Natl Acad Sci U S A*, 58(5):1895–1902, 1967.

[29] CSAC. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005.

[30] J. G. De Boer. *Encyclopedia of life sciences*, chapter Mutations and the genetic code. John Wiley & Sons, Ltd: Chichester, 2001.

[31] D. R. Denver, P. C. Dolan, L. J. Wilhelm, W. Sung, J. I. Lucas-Lled, D. K. Howe, S. C. Lewis, K. Okamoto, W. K. Thomas, M. Lynch, and C. F. Baer. A genome-wide view of caenorhabditis elegans base-substitution mutation processes. *Proc Natl Acad Sci U S A*, 106(38):16310–16314, 2009.

[32] D. R. Denver, K. Morris, M. Lynch, L. L. Vassilieva, and W. K. Thomas. High direct estimate of the mutation rate in the mitochondrial genome of caenorhabditis elegans. *Science*, 289(5488):2342–2344, 2000.

[33] Z. Du, Y. Zhao, and N. Li. Genome-wide analysis reveals regulatory role of g4 dna in gene transcription. *Genome Research*, 18(2):233–241, 2008.

[34] M. L. Duquette, P. Handa, J. A. Vincent, A. F. Taylor, and N. Maizels. Intracellular transcription of g-rich dnas induces formation of g-loops, novel structures containing g4 dna. *Genes & Development*, 18(13):1618–1629, 2004.

[35] L. Duret. Mutation patterns in the human genome: more variable than expected. *PLoS Biol*, 7(2):e1000028, 2009.

[36] L. Duret and P. F. Arndt. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*, 4(5):e1000071, 2008.

[37] L. Duret and P. F. Arndt. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genetics*, 4(5):e1000071, 2008.

[38] M. Dye and N. Proudfoot. Multiple transcript cleavage precedes polymerase release in termination by rna polymerase ii. *Cell*, 105:668–669, 2001.

[39] W. Enard, A. Fassbender, F. Model, P. Adorjn, S. Pbo, and A. Olek. Differences in dna methylation patterns between humans and chimpanzees. *Curr Biol*, 14(4):R148–R149, Feb 2004.

[40] J. J. Faith and D. D. Pollock. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics*, 165(2):735–745, 2003.

[41] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

[42] P. Flicek, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. P. Hubbard, A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal, and S. Searle. Ensembl 2008. *Nucl. Acids Res.*, 36:D707–714, 2008.

[43] M. P. Francino and H. Ochman. Deamination as the basis of strand-asymmetric evolution in transcribed escherichia coli sequences. *Mol Biol Evol*, 18(6):1147–1150, 2001.

[44] A. C. Frank and J. R. Lobry. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, 238:65 – 77, 1999.

[45] S. A. Frank and M. A. Nowak. Cell biology: Developmental predisposition to cancer. *Nature*, 422(6931):494, 2003.

[46] A. Franklin, P. J. Milburn, R. V. Blanden, and E. J. Steele. Human dna polymerase-eta, an a-t mutator in somatic hypermutation of rearranged immunoglobulin genes, is a reverse transcriptase. *Immunol Cell Biol*, 82(2):219–225, 2004.

[47] L. Frederico, T. Kunkel, and B. Shaw. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry*, 29(10):2532–2537, 1990.

[48] E. C. Freidberg, G. C. Walker, W. Siede, R. D. Wood, R. A. Schultz, and T. Ellenberger. *DNA Repair and Mutagenesis*. ASM press, 2006.

[49] S. Fujimori, T. Washio, and M. Tomita. Gc-compositional strand bias around transcription start sites in plants and fungi. *BMC Genomics*, 6(1):26, 2005.

[50] N. Galtier and L. Duret. Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution. *Trends in Genetics*, 23(6):273–277, 2007.

[51] N. Galtier, G. Piganeau, D. Mouchiroud, and L. Duret. Gc-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, 159(2):907–911, 2001.

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

[53] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. What is a gene, post-encode? history and updated definition. *Genome Research*, 17(6):669–681, 2007.

[54] R. A. Gibbs, G. M. Weinstock, M. L. Metzker, D. M. Muzny, E. J. Sodergren, S. Scherer, G. Scott, D. Steffen, K. C. Worley, P. E. Burch, G. Okwuonu, S. Hines, L. Lewis, C. DeRamo, O. Delgado, S. Dugan-Rocha, G. Miner, M. Morgan, A. Hawes, R. Gill, Celera, R. A. Holt, M. D. Adams, P. G. Amanatides, H. Baden-Tillson, M. Barnstead, S. Chin, C. A. Evans, S. Ferriera, and C. Fosler. Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428:493 – 521, 2004.

[55] M. Gomez and F. Antequera. Overreplication of short dna regions during s phase in human cells. *Genes & Development*, 22(3):375–385, 2008.

[56] M. Gomez and N. Brockdorff. Heterochromatin on the inactive x chromosome delays replication timing without affecting origin usage. *PNAS*, 101(18):6923–6928, 2004.

[57] D. Graur. *Encyclopedia of life sciences*, chapter Mutational change in evolution. John Wiley & Sons, Ltd: Chichester, 2008.

[58] D. Graur and W.-H. Li. *Fundamentals of Molecular Evolution*. Sinauer, 2000.

[59] M. M. Gray, J. M. Granka, C. D. Bustamante, N. B. Sutter, A. R. Boyko, L. Zhu, E. A. Ostrander, and R. K. Wayne. Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics*, 181(4):1493–1505, Apr 2009.

[60] P. Green, B. Ewing, W. Miller, P. Thomas, and E. Green. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet*, 33(4):514 – 517, 2003.

[61] M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, 2007.

[62] L. Han, B. Su, W.-H. Li, and Z. Zhao. Cpg island density and its correlations with genomic features in mammalian genomes. *Genome Biol*, 9(5):R79, 2008.

[63] P. C. Hanawalt and G. Spivak. Transcription-coupled dna repair: two decades of progress and surprises. *Nat Rev Mol Cell Biol*, 9(12):958–970, 2008.

[64] HapMap. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, 2007.

[65] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J Mol Evol*, 22(2):160–174, 1985.

[66] M. M. Hoffman and E. Birney. Estimating the neutral rate of nucleotide substitution using introns. *Mol Biol Evol*, 24(2):522–531, 2007.

[67] F.-T. Huang, K. Yu, B. Balter, E. Selsing, Z. Oruc, A. Khamlichi, C.-L. Hsieh, and M. Lieber. Sequence dependence of chromosomal r-loops at the immunoglobulin heavy-chain s-class switch region. *Mol. Cell. Biol.*, 27:5921–5932, 2007.

[68] F.-T. Huang, K. Yu, C.-L. Hsieh, and M. Lieber. Downstream boundary of chromosomal r-loops at murine switch regions: Implications for the mechanism of class switch recombination. *Proc. Natl. Acad. Sci.*, 103:5030–5035, 2006.

[69] M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton Carafa, A. Arneodo, and C. Thermes. Human gene organization driven by the coordination of replication and transcription. *Genome Research*, 17(9):1278–1285, 2007.

[70] D. G. Hwang and P. Green. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(39):13994–14001, 2004.

[71] K. J. Impellizzeri, B. Anderson, and P. M. Burgers. The spectrum of spontaneous mutations in a saccharomyces cerevisiae uracil-dna-glycosylase mutant limits the function of this enzyme to cytosine deamination repair. *J Bacteriol*, 173(21):6807–6810, 1991.

[72] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nat. Genet*, 33:245–254, 2003.

[73] T. Jukes and C. Cantor. *Mammalian protein metabolism*, chapter Evolution of protein molecules, pages 21–123. Academic Press, New York, 1969.

[74] S. Kaneko, C. Chu, A. J. Shatkin, and J. L. Manley. Human capping enzyme promotes formation of transcriptional r loops in vitro. *Proceedings of the National Academy of Sciences*, 104(45):17620–17625, 2007.

[75] C. Keller, E.-M. Ladenburger, M. Kremer, and R. Knippers. The origin recognition complex marks a replication origin in the human top1 gene promoter. *J. Biol. Chem.*, 277(35):31430–31440, 2002.

[76] P. Khaitovich, B. Muetzel, X. She, M. Lachmann, I. Hellmann, J. Dietzsch, S. Steigele, H.-H. Do, G. Weiss, W. Enard, F. Heissig, T. Arendt, K. Nieselt-Struwe, E. E. Eichler, and S. Pbo. Regional patterns of gene expression in human and chimpanzee brains. *Genome Res*, 14(8):1462–1473, 2004.

[77] T. Kim, L. Barrera, M. Zheng, C. Qu, M. Singer, T. Richmond, Y. Wu, R. Green, and B. Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436:876880, 2005.

[78] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

[79] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120, 1980.

[80] A. S. Kondrashov. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Hum Mutat*, 21(1):12–27, Jan 2003.

[81] M. Krawczak, J. Reiss, and D. N. Cooper. The mutational spectrum of single base-pair substitutions in mrna splice junctions of human genes: causes and consequences. *Hum Genet*, 90(1-2):41–54, 1992.

[82] R. M. Kuhn, D. Karolchik, A. S. Zweig, T. Wang, K. E. Smith, K. R. Rosenbloom, B. Rhead, B. J. Raney, A. Pohl, M. Pheasant, L. Meyer, F. Hsu, A. S. Hinrichs, R. A. Harte, B. Giardine, P. Fujita, M. Diekhans, T. Dreszer, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent. The ucsc genome browser database: update 2009. *Nucleic Acids Res*, 37(Database issue):D755–D761, 2009.

[83] T. A. Kunkel and K. Bebenek. Dna replication fidelity. *Annu Rev Biochem*, 69:497–529, 2000.

[84] T. A. Kunkel and P. M. Burgers. Dividing the workload at a eukaryotic replication fork. *Trends in Cell Biology*, 18(11):521–527, 2008.

[85] M. J. Lercher, J.-V. Chamary, and L. D. Hurst. Genomic regionality in rates of evolution is not explained by clustering of genes of comparable expression profile. *Genome Res*, 14(6):1002–1013, 2004.

[86] M. J. Lercher, E. J. B. Williams, and L. D. Hurst. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol*, 18:2032–2039, 2001.

[87] L. Li, K. M. Murphy, U. Kanevets, and L. J. Reha-Krantz. Sensitivity to phosphonoacetic acid: a new phenotype to probe dna polymerase [delta] in saccharomyces cerevisiae. *Genetics*, 170:569–580, 2005. 10.1534/genetics.104.040295.

[88] T. Lindahl and B. Nyberg. Rate of depurination of native deoxyribonucleic acid. *Biochemistry*, 11(19):3610–3618, 1972.

[89] T. Lindahl and B. Nyberg. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry*, 13(16):3405–3410, 1974.

[90] J. Lobry. Properties of a general model of dna evolution under no-strand-bias conditions. *J Mol Evol*, 40(3):326–30, 1995.

[91] J. R. Lobry. Asymmetric substitution patterns in the two dna strands of bacteria. *Mol Biol Evol*, 13(5):660–665, 1996.

[92] E. Louie, J. Ott, and J. Majewski. Nucleotide frequency variation across human genes. *Genome Res*, 13(12):2594 – 601, 2003.

[93] I. Lucas, A. Palakodeti, Y. Jiang, D. Young, N. Jiang, A. Fernald, and M. Le Beau. High-throughput mapping of origins of replication in human cells. *EMBO reports*, 8(8):770–777, 2007.

[94] S. E. Luria and M. Delbrck. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491–511, 1943.

[95] M. Lynch. *The Origins of Genome Architecture*. Sinauer Associates, 2007.

[96] M. Lynch. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A*, 107(3):961–968, 2010.

[97] M. Lynch, W. Sung, K. Morris, N. Coffey, C. R. Landry, E. B. Dopman, W. J. Dickinson, K. Okamoto, S. Kulkarni, D. L. Hartl, and W. K. Thomas. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A*, 105(27):9272–9277, 2008.

[98] N. Maizels. Immunoglobulin gene diversification. *Annual Review of Genetics*, 39(1):23, 2005.

[99] J. Majewski. Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am J Hum Genet*, 73(3):688 – 692, 2003.

[100] J. Majewski and J. Ott. Distribution and characterization of regulatory elements in the human genome. *Genome Res*, 12(12):1827 – 36, 2002.

[101] S. D. McCulloch and T. A. Kunkel. The fidelity of dna synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res*, 18(1):148–161, 2008.

[102] H. A. McQueen, J. Fantes, S. H. Cross, V. H. Clark, A. L. Archibald, and A. P. Bird. Cpg islands of chicken are concentrated on microchromosomes. *Nat Genet*, 12(3):321–324, 1996.

[103] J. Meunier and L. Duret. Recombination drives the evolution of gc-content in the human genome. *Mol Biol Evol*, 21(6):984–990, 2004.

[104] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448:553 – 560, 2007.

[105] J. Mrazek and S. Karlin. Strand compositional asymmetry in bacterial and large viral genomes. *PNAS*, 95(7):3720–3725, 1998.

[106] Y. Murakami, M. Satake, Y. Yamaguchi-Iwai, M. Sakai, M. Muramatsu, and Y. Ito. The nuclear protooncogenes c-jun and c-fos as regulators of dna replication. *PNAS*, 88(9):3947–3951, 1991.

[107] G. Muse, D. Gilchrist, S. Nechaev, R. Shah, J. Parker, S. Grissom, J. Zeitlinger, and K. Adelman. Rna polymerase is poised for activation across the genome. *Nat. Genet.*, 39:1507–1511, 2007.

[108] M. W. Nachman and S. L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, Sep 2000.

[109] T. Nagylaki. Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences of the United States of America*, 80(20):6278–6281, 1983.

[110] A. Necsulea, C. Guillet, J.-C. Cadoret, M.-N. Prioleau, and L. Duret. The relationship between dna replication and human genome organization. *Mol Biol Evol*, pages 303–308, 2009.

[111] J. M. D. Noia and M. S. Neuberger. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem*, 76:1–22, 2007.

[112] V. Odegard and D. Schatz. Targeting of somatic hypermutation. *Nat. Rev. Immunol.*, 6:573–583, 2006.

[113] S. H. Orkin, S. E. Antonarakis, and H. H. Kazazian. Base substitution at position -88 in a beta-thalassemic globin gene. further evidence for the role of distal promoter element acaccc. *J Biol Chem*, 259(14):8679–8681, 1984.

[114] W. P. Osheroff, H. K. Jung, W. A. Beard, S. H. Wilson, and T. A. Kunkel. The fidelity of dna polymerase beta during distributive and processive dna synthesis. *J Biol Chem*, 274(6):3642–3650, 1999.

[115] A. Palleja, E. Guzman, S. Garcia-Vallve, and A. Romeu. In silico prediction of the origin of replication among bacteria: A case study of bacteroides thetaiotaomicron. *OMICS: A Journal of Integrative Biology*, 12(3):201–210, 2008.

[116] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach, and A. Soldatov. Transcriptome analysis by strand-specific sequencing of complementary dna. *Nucleic Acids Res*, 37(18):e123, 2009.

[117] B. Paten, J. Herrero, K. Beal, S. Fitzgerald, and E. Birney. Enredo and pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, 18(11):1814–1828, 2008.

[118] P. Pham, R. Bransteitter, J. Petruska, and M. Goodman. Processive aid-catalysed cytosine deamination on single-stranded dna simulates somatic hypermutation. *Nature*, 424:103–107, 2003.

[119] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M.-L. Lin, G. R. Ordez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter, L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196, 2010.

[120] E. D. Pleasance, P. J. Stephens, S. O'Meara, D. J. McBride, A. Meynert, D. Jones, M.-L. Lin, D. Beare, K. W. Lau, C. Greenman, I. Varela, S. Nik-Zainal, H. R. Davies, G. R. Ordoez, L. J. Mudie, C. Latimer, S. Edkins, L. Stebbings, L. Chen, M. Jia, C. Leroy, J. Marshall, A. Menzies, A. Butler, J. W. Teague, J. Mangion, Y. A. Sun, S. F. McLaughlin, H. E. Peckham, E. F. Tsung, G. L. Costa, C. C. Lee, J. D. Minna, A. Gazdar, E. Birney, M. D. Rhodes, K. J. McKernan, M. R. Stratton, P. A. Futreal, and P. J. Campbell. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 463(7278):184–190, 2010.

[121] P. Polak and P. F. Arndt. Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Research*, 18(8):1216–1223, 2008.

[122] P. Polak and P. F. Arndt. Long-range bidirectional strand asymmetries originate at cpg islands in the human genome. *Genome Biol Evol*, 2009:189–197, 2009.

[123] S. E. Ptak, D. A. Hinds, K. Koehler, B. Nickel, N. Patil, D. G. Ballinger, M. Przeworski, K. A. Frazer, and S. Paabo. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet*, 37:429 – 434, 2005.

[124] S. E. Ptak, D. A. Hinds, K. Koehler, B. Nickel, N. Patil, D. G. Ballinger, M. Przeworski, K. A. Frazer, and S. Pbo. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet*, 37(4):429–434, 2005.

[125] H.-Q. Qu, S. Lawrence, F. Guo, J. Majewski, and C. Polychronakos. Strand bias in complementary single-nucleotide polymorphisms of transcribed human sequences: Evidence for functional effects of synonymous polymorphisms. *BMC Genomics*, 7:213, 2006.

[126] C. Rada, J. M. D. Noia, and M. S. Neuberger. Mismatch recognition and uracil excision provide complementary paths to both ig switching and the a/t-focused phase of somatic mutation. *Mol Cell*, 16(2):163–171, 2004.

[127] T. Rein, T. Kobayashi, M. Malott, M. Leffak, and M. L. DePamphilis. Dna methylation at mammalian replication origins. *J. Biol. Chem.*, 274(36):25792–25800, 1999.

[128] B. Rhead, D. Karolchik, R. M. Kuhn, A. S. Hinrichs, A. S. Zweig, P. A. Fujita, M. Diekhans, K. E. Smith, K. R. Rosenbloom, B. J. Raney, A. Pohl, M. Pheasant, L. R. Meyer, K. Learned, F. Hsu, J. Hillman-Jackson, R. A. Harte, B. Giardine, T. R. Dreszer, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent. The ucsc genome browser database: update 2010. *Nucleic Acids Res*, 38(Database issue):D613–D619, 2010.

[129] E. P. C. Rocha. The organization of the bacterial genome. *Annual Review of Genetics*, 42(1):211–233, 2008.

[130] E. P. C. Rocha and A. Danchin. Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat Genet*, 34(4):377–378, 2003.

[131] E. P. C. Rocha, M. Touchon, and E. J. Feil. Similar compositional biases are caused by very different mutational effects. *Genome Research*, 16(12):1537–1547, 2006.

[132] I. B. Rogozin and Y. I. Pavlov. The cytidine deaminase aid exhibits similar functional properties in yeast and mammals. *Mol Immunol*, 43(9):1481–1484, Mar 2006.

[133] D. Ronai, M. Iglesias-Ussel, M. Fan, Z. Li, A. Martin, and M. Scharff. Detection of chromatin-associated single-stranded dna in regions targeted for somatic hypermutation. *J. Environ. Monit.*, 204:181–190, 2007.

[134] B. Rosenberg and F. Papavasiliou. Beyond shm and csr: Aid and related cytidine deaminases in the host response to viral infection. *Adv. Immunol*, 94:215–244, 2007.

[135] K. Rothkamm, I. Krger, L. H. Thompson, and M. Lobrich. Pathways of dna double-strand break repair during the mammalian cell cycle. *Mol Cell Biol*, 23(16):5706–5715, 2003.

[136] D. Roy, K. Yu, and M. R. Lieber. Mechanism of r-loop formation at immunoglobulin class switch sequences. *Mol Cell Biol*, 28(1):50–60, 2008.

[137] S. Schreck, M. Buettner, E. Kremmer, M. Bogdan, H. Herbst, and G. Niedobitek. Activation-induced cytidine deaminase (aid) is expressed in normal spermatogenesis but only infrequently in testicular germ cell tumours. *J Pathol*, 210(1):26–31, 2006.

[138] A. C. Seila, J. M. Calabrese, S. S. Levine, G. W. Yeo, P. B. Rahl, R. A. Flynn, R. A. Young, and P. A. Sharp. Divergent transcription from active promoters. *Science*, 322(5909):1849–1851, 2008.

[139] J. C. Shen, W. M. Rideout, and P. A. Jones. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded dna. *Nucleic Acids Res*, 22(6):972–976, Mar 1994.

[140] C. C. A. Spencer, P. Deloukas, S. Hunt, J. Mullikin, S. Myers, B. Silverman, P. Donnelly, D. Bentley, and G. McVean. The influence of recombination on human genetic diversity. *PLoS Genet*, 2(9):e148, 2006.

[141] F. Squartini and P. F. Arndt. Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Mol Biol Evol*, 25(12):2525–2535, 2008.

[142] E. J. Steele. Mechanism of somatic hypermutation: critical analysis of strand biased mutation signatures at a:t and g:c base pairs. *Mol Immunol*, 46(3):305–320, 2009.

[143] G. Strathdee, A. Sim, and R. Brown. Control of gene expression by cpg island methylation in normal cells. *Biochem. Soc. Trans.*, 32(Pt 6):913–915, 2004.

[144] N. Sueoka. Intrastrand parity rules of dna base composition and usage biases of synonymous codons. *J Mol Evol*, 40(3):318–325, 1995.

[145] D. Takai and P. A. Jones. Comprehensive analysis of cpg islands in human chromosomes 21 and 22. *Proc Natl Acad Sci USA*, 99:3740 – 3745, 2002.

[146] The-ENCODE-Project-Consortium. Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

[147] D. Thomas, J. Roberts, R. Sabatino, T. Myers, C. Tan, K. Downey, A. So, R. Bambara, and T. Kunkel. Fidelity of mammalian dna replication and replicative dna polymerases. *Biochemistry*, 30(51):11751–9., 1991.

[148] M. Touchon, A. Arneodo, Y. d'Aubenton Carafa, and C. Thermes. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Res*, 32(17):4969–4978, 2004.

[149] M. Touchon, A. Arneodo, Y. d'Aubenton Carafa, and C. Thermes. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucl. Acids Res.*, 32(17):4969–4978, 2004.

[150] M. Touchon, S. Nicolay, A. Arneodo, Y. dAubenton Carafa, and C. Thermes. Transcription-coupled ta and gc strand asymmetries in the human genome. *FEBS Lett*, 555:579–582, 2003.

[151] M. Touchon, S. Nicolay, A. Arneodo, Y. d'Aubenton Carafa, and C. Thermes. Transcription-coupled ta and gc strand asymmetries in the human genome. *FEBS Lett*, 555(3):579–582, 2003.

[152] M. Touchon, S. Nicolay, B. Audit, E.-B. B. of Brodie, Y. d'Aubenton Carafa, A. Arneodo, and C. Thermes. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc Natl Acad Sci U S A*, 102(28):9836–9841, 2005.

[153] M. Touchon and E. P. Rocha. From gc skews to wavelets: A gentle guide to the analysis of compositional asymmetries in genomic data. *Biochimie*, 90(4):648–659, 2008.

[154] M. Wan, S. S. Lee, X. Zhang, I. Houwink-Manville, H. R. Song, R. E. Amir, S. Budden, S. Naidu, J. L. Pereira, I. F. Lo, H. Y. Zoghbi, N. C. Schanen, and U. Francke. Rett syndrome and beyond: recurrent spontaneous and familial mecp2 mutations at cpg hotspots. *Am J Hum Genet*, 65(6):1520–1529, Dec 1999.

[155] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[156] M. Weber, I. Hellmann, M. B. Stadler, L. Ramos, S. Paabo, M. Rebhan, and D. Schubeler. Distribution, silencing potential and evolutionary impact of promoter dna methylation in the human genome. *Nat Genet*, 39:457 – 466, 2007.

[157] M. T. Webster, N. G. Smith, M. J. Lercher, and H. Ellegren. Gene expression, synteny, and local similarity in human noncoding mutation rates. *Mol Biol Evol*, 21(10):1820–1830, 2004.

[158] M. T. Webster and N. G. C. Smith. Fixation biases affecting human snps. *Trends Genet*, 20(3):122–126, 2004.

[159] W. Winckler, S. R. Myers, D. J. Richter, R. C. Onofrio, G. J. McDonald, R. E. Bontrop, G. A. McVean, S. B. Gabriel, D. Reich, P. Donnelly, and D. Altshuler. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308:107 – 111, 2005.

[160] K. H. Wolfe, M. Gouy, Y. W. Yang, P. M. Sharp, and W. H. Li. Date of the monocot-dicot divergence estimated from chloroplast dna sequence data. *Proc Natl Acad Sci U S A*, 86(16):6201–6205, Aug 1989.

[161] V. K. Yadav, J. K. Abraham, P. Mani, R. Kulshrestha, and S. Chowdhury. Quadbase: genome-wide database of g4 dna occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucl Acids Res*, 36(suppl):381–385, 2008.

[162] L. Y. Yampolsky and A. Stolzfuz. *Encyclopedia of life sciences*, chapter Mutational biases. John Wiley & Sons, Ltd: Chichester, 2008.

[163] F. H. Yu, V. Yarov-Yarovoy, G. A. Gutman, and W. A. Catterall. Overview of molecular relationships in the voltage-gated ion channel superfamily. *Pharmacol. Rev.*, 57:387–395, 2005.

[164] C. Zhang, W.-H. Li, A. R. Krainer, and M. Q. Zhang. Rna landscape of evolution for optimal exon and intron discrimination. *Proc Natl Acad Sci U S A*, 105(15):5797–5802, 2008.

[165] L. Zhang, S. Kasif, C. R. Cantor, and N. E. Broude. Gc/at-content spikes as genomic punctuation marks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(48):16855–16860, 2004.

A Appendix A

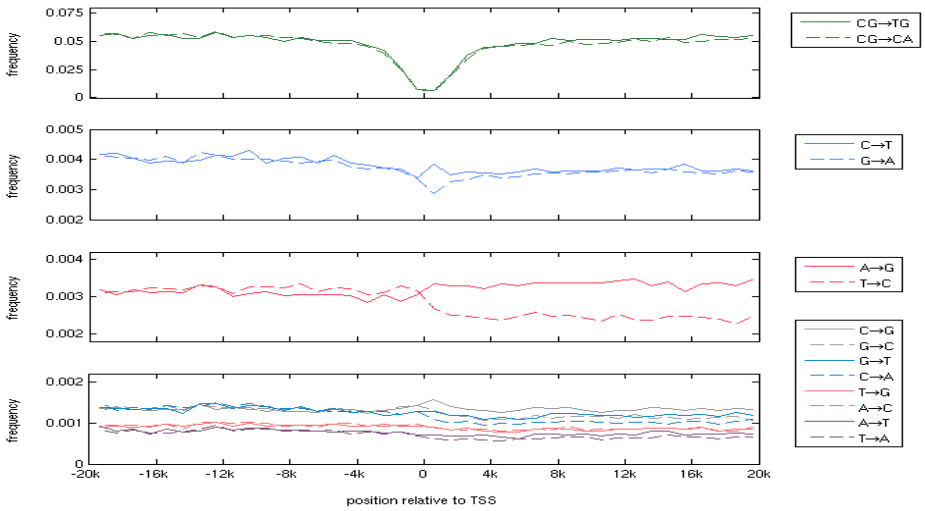

Figure A.1: Substitution rates in introns and in intergenic regions in a 20 kbp long regions centered on the 5'end of genes. The plots show the estimated twelve single nucleotide substitution rates, as well as, the CpG deamination rates in non-overlapping 1000 bp long windows.

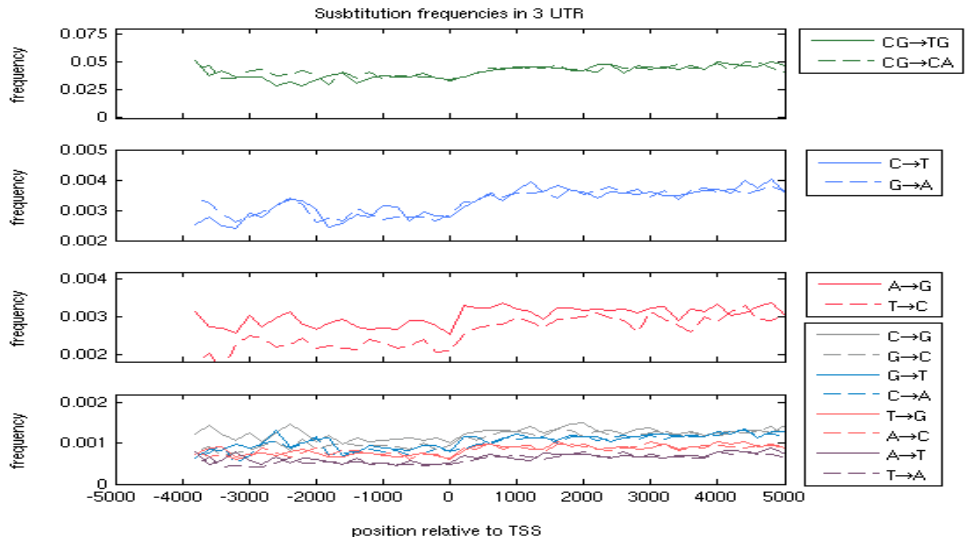

Figure A.2: Substitution rates in 3UTR and 3 downstream intergenic region of human genes. Each point in a plot is the estimated substitution rate in a window of length 200 bp that is located at certain distance from the 3end.

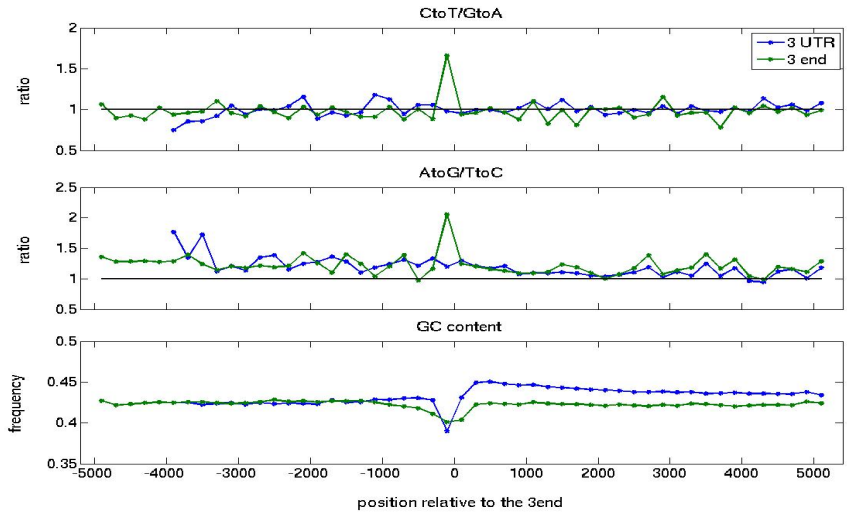

Figure A.3: The ratios between complementary transition rates plotted against the distance from the 3'end of genes. The ratios in 3'UTR and introns were calculated using the estimated substitution rates in Fig. A.2 and Fig. 4.1, respectively.

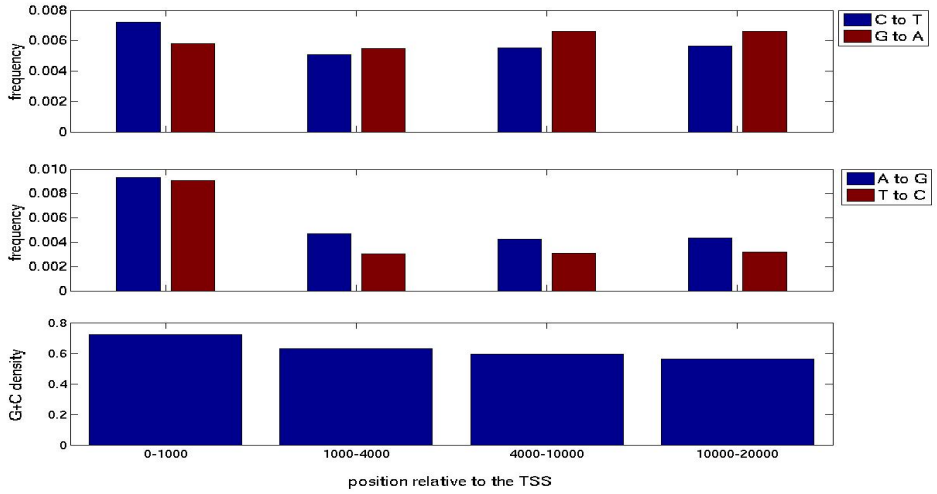

Figure A.4: Transition rates and GC content in FFDs in non-CpG sites and in windows with varying lengths.

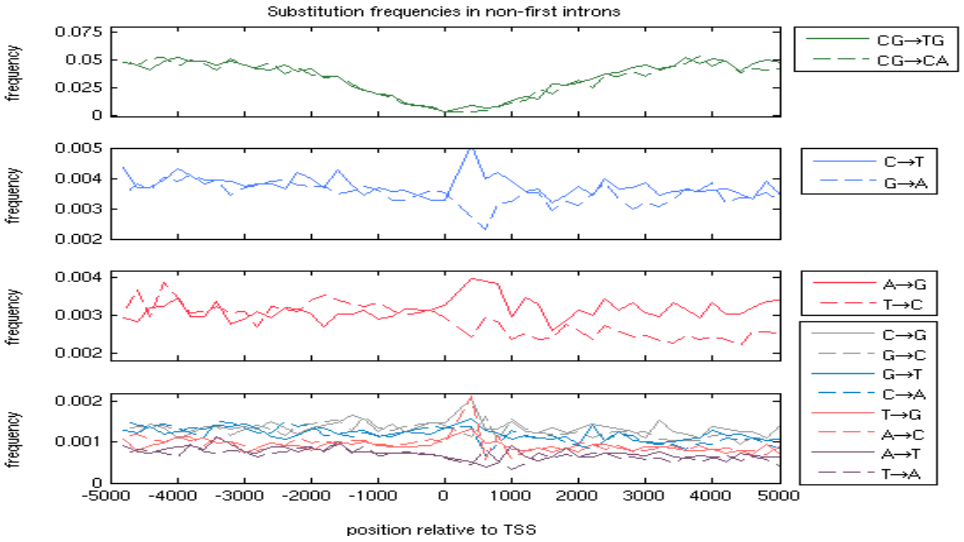

Figure A.5: Substitution rates in DNA sequences in a 10 kbp long region centered on the 5'end after excluding first introns in addition to exons.

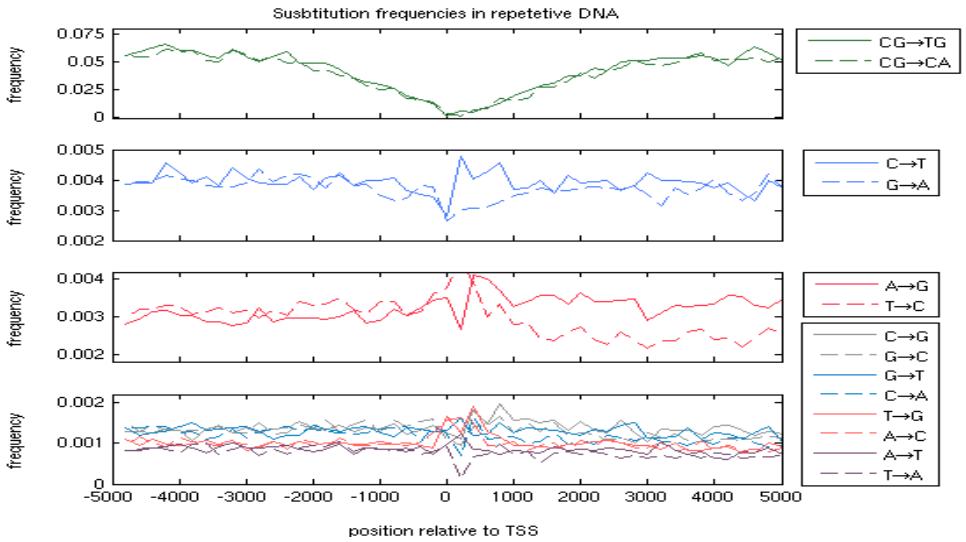

Figure A.6: Substitution rates in repetitive DNA sequences in a 10 kbp region centered on the 5'end of genes.

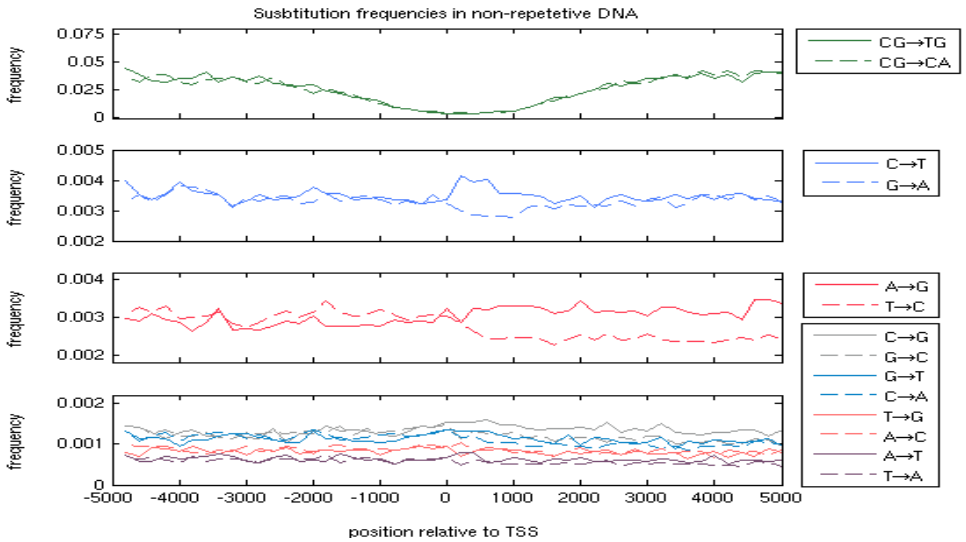

Figure A.7: Substitution rates in non-repetitive DNA sequences in a 10 kbp region centered on the 5'end of genes.

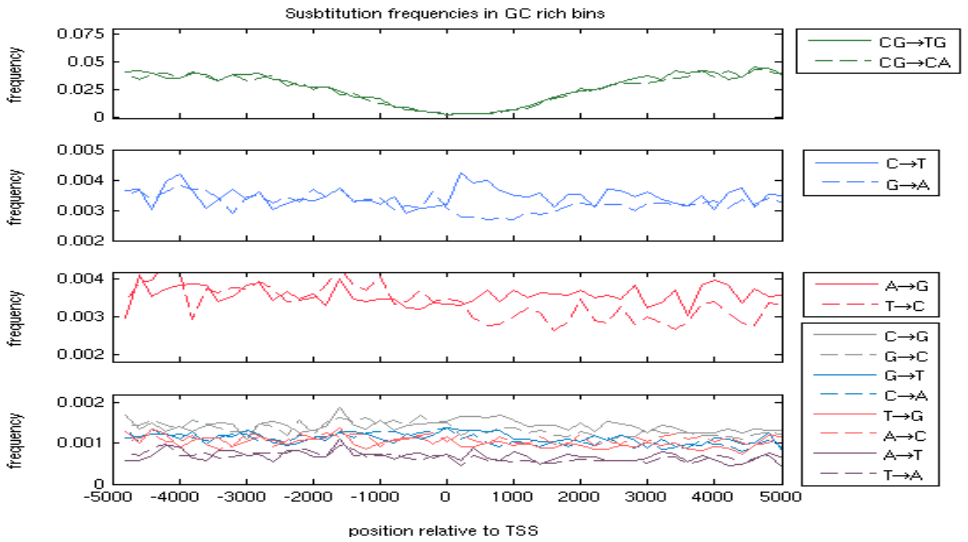

Figure A.8: Substitution rates in GC-rich pooled sequences in a 10 kbp long region centered on the 5end. In each window, just sequences with above 50% GC content were pooled out from all genes.

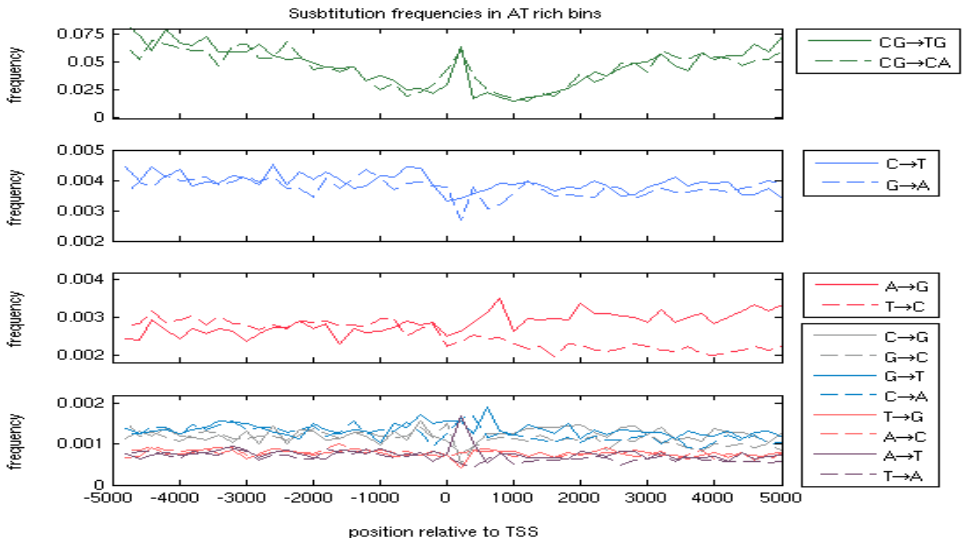

Figure A.9: Substitution rates in AT-rich (GC-poor) pooled sequences in a 10 kbp long region centered on the 5end. In each window, just sequences with less then 41% GC content were pooled out from all genes.

B Appendix B

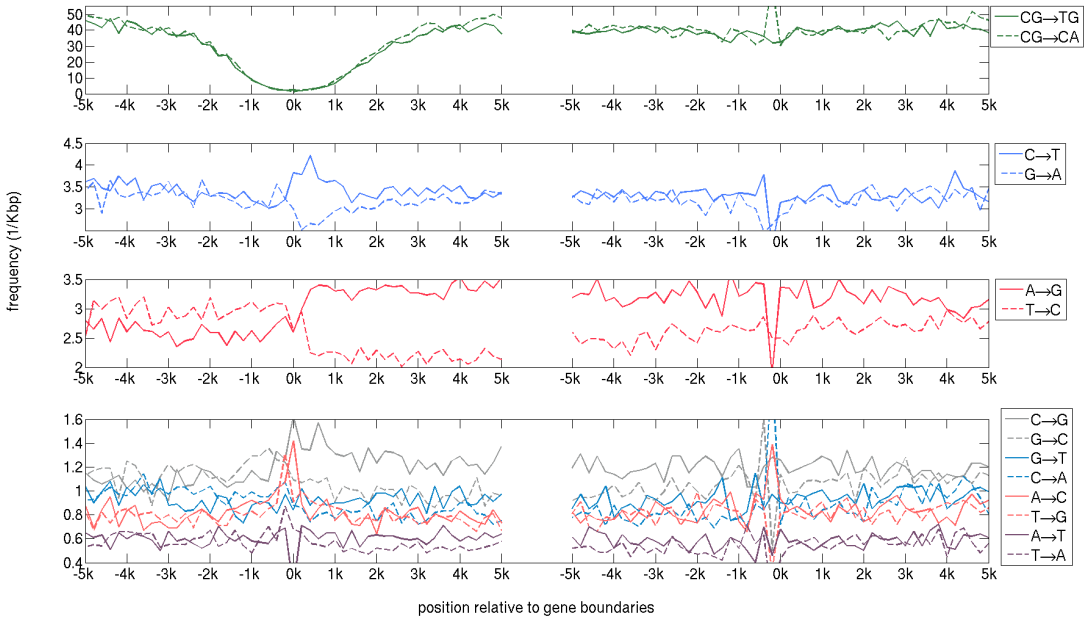

Figure B.1: Substitution rates in introns and in intergenic regions in the vicinity of 5' and 3'ends of human CGI-genes. The plots show the estimated twelve single nucleotide substitution rates and the CpG deamination rates in non-overlapping 200 bp long windows along the non template strand. The distances of the windows' centers from the 5'end or 3'end are indicated on the x- axes. The estimation of substitution frequencies has been performed using the non-template strand.

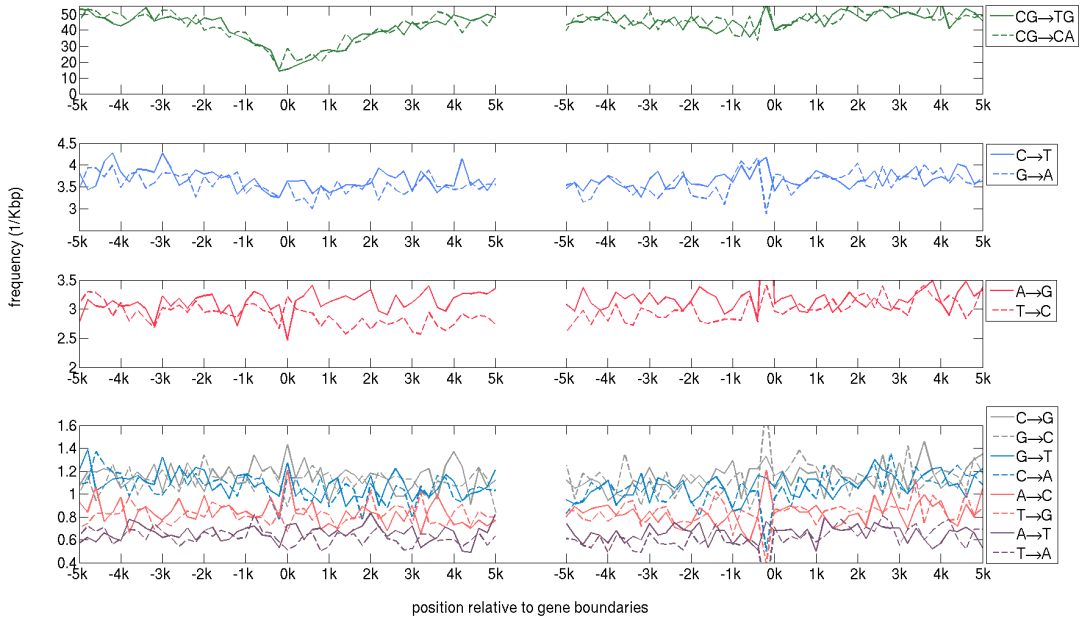

Figure B.2: Substitution rates in introns and in intergenic regions in the vicinity of 5' and 3'ends of human CGI-genes. The plots show the estimated twelve single nucleotide substitution rates and the CpG deamination rates in non-overlapping 200 bp long windows along the non template strand. The distances of the windows' centers from the 5'end or 3'end are indicated on the x- axes. The estimation of substitution frequencies has been performed using the non-template strand.

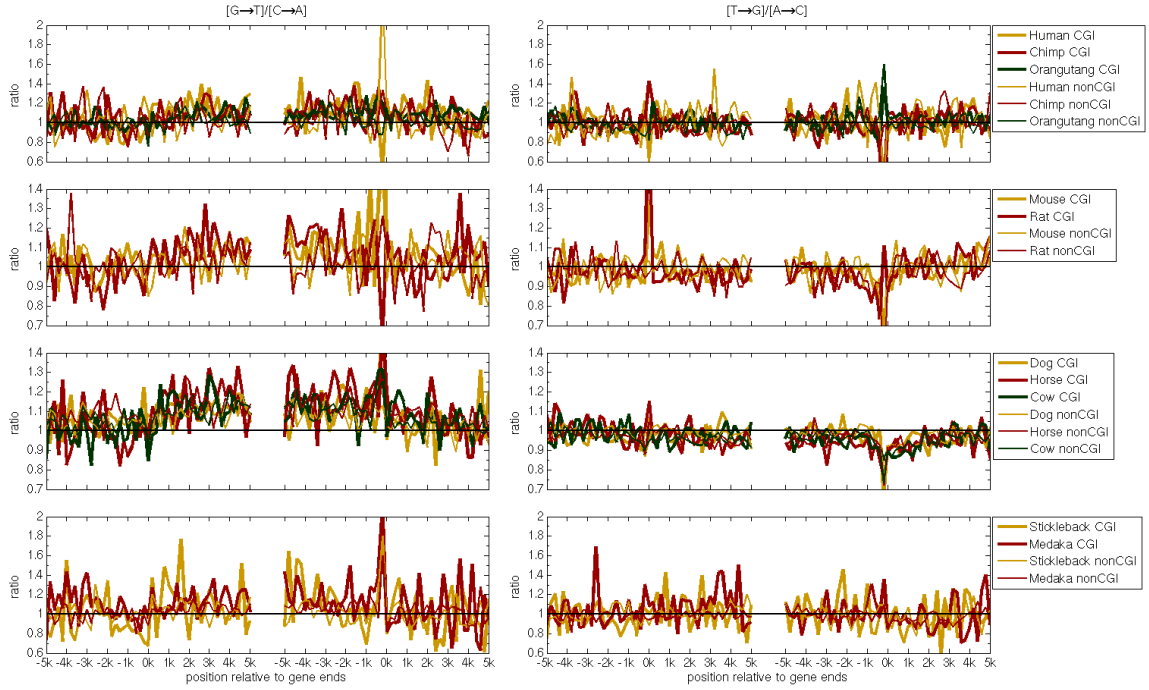

Figure B.3: Profiles of $r_{G\to T}/r_{C\to A}$ and $r_{T\to G}/r_{A\to C}$ ratios across vertebrates. The ratios are plotted against distance from the 5' and 3' ends of genes and are calculated along the non-template strand from pooled 200 bp windows of genes annotated for the reference species in each clade. For CGI-genes the ratios are presented by thicker lines. See Fig. 4.5 for further details.

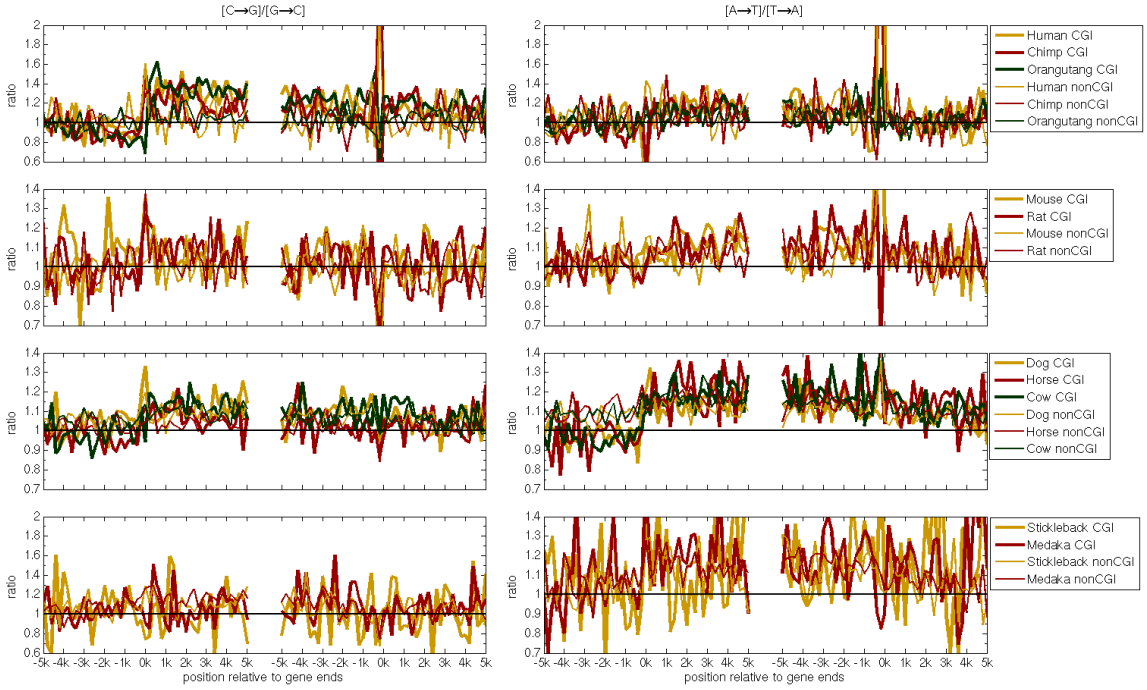

Figure B.4: Profiles of $r_{C\to G}/r_{G\to C}$ and $r_{A\to T}/r_{T\to A}$ ratios across vertebrates. The ratios are plotted against distance from the 5' and 3' ends of genes and are calculated along the non-template strand from pooled 200 bp windows of genes annotated for the reference species in each clade. For CGI-genes the ratios are presented by thicker lines. See Fig. 4.5 for further details.

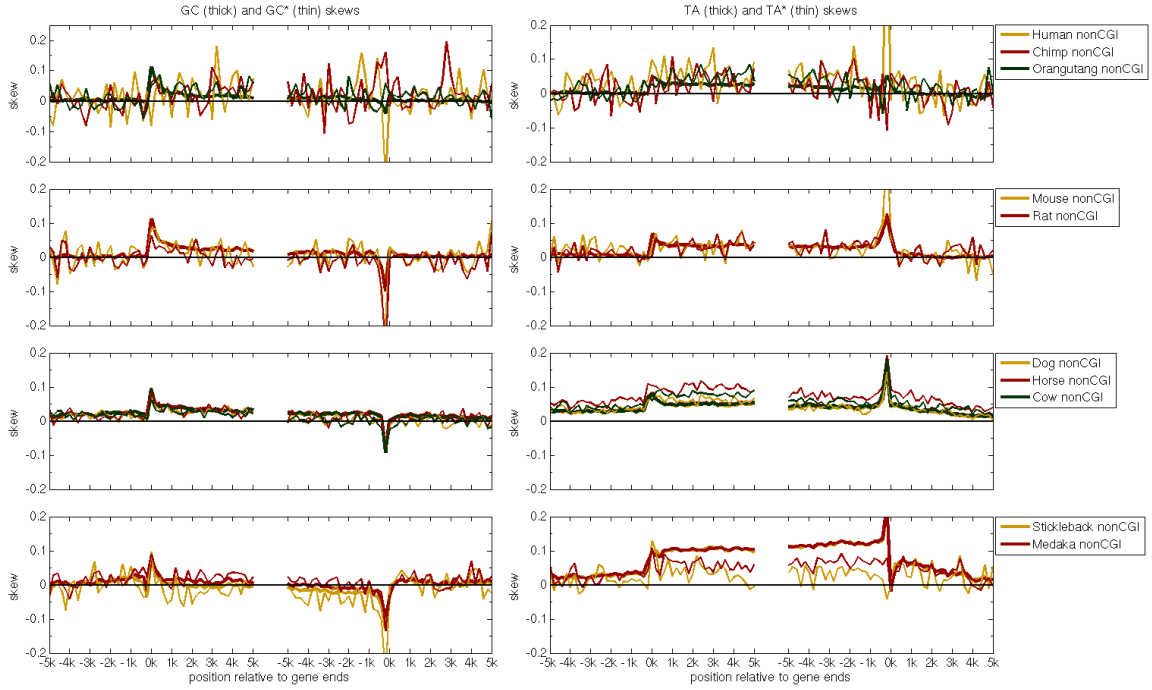

Figure B.5: The current and stationary TA(*) and GC(*) skews along nonCGI-genes and their flanks. Current stationary skews are plotted with thicker lines. The skews are plotted against distance from the 5' and 3' ends of genes and are calculated along the non-template strand from pooled 200 bp windows of genes annotated for the reference species in each clade.

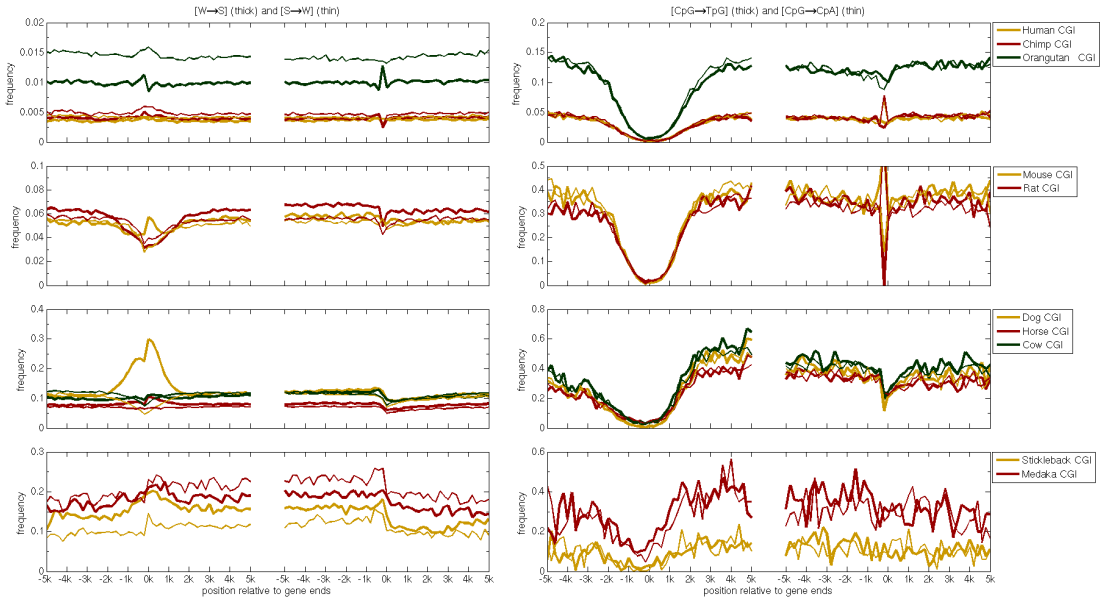

Figure B.6: The frequencies of $S \to W$, $W \to S$ and CpG methylation deamination process around the ends of CGI-genes in vertebrates. The frequencies $S \to W$ (thin lines) are calculated without the substitution in CpGs and compared to $W \to S$ frequencies (thick lines). The methylation deamination rates $CpG \to TpG$(thick) and $CpG \to CpA$ (thin) are presented in the right panels. See Fig. 4.5 for further details.

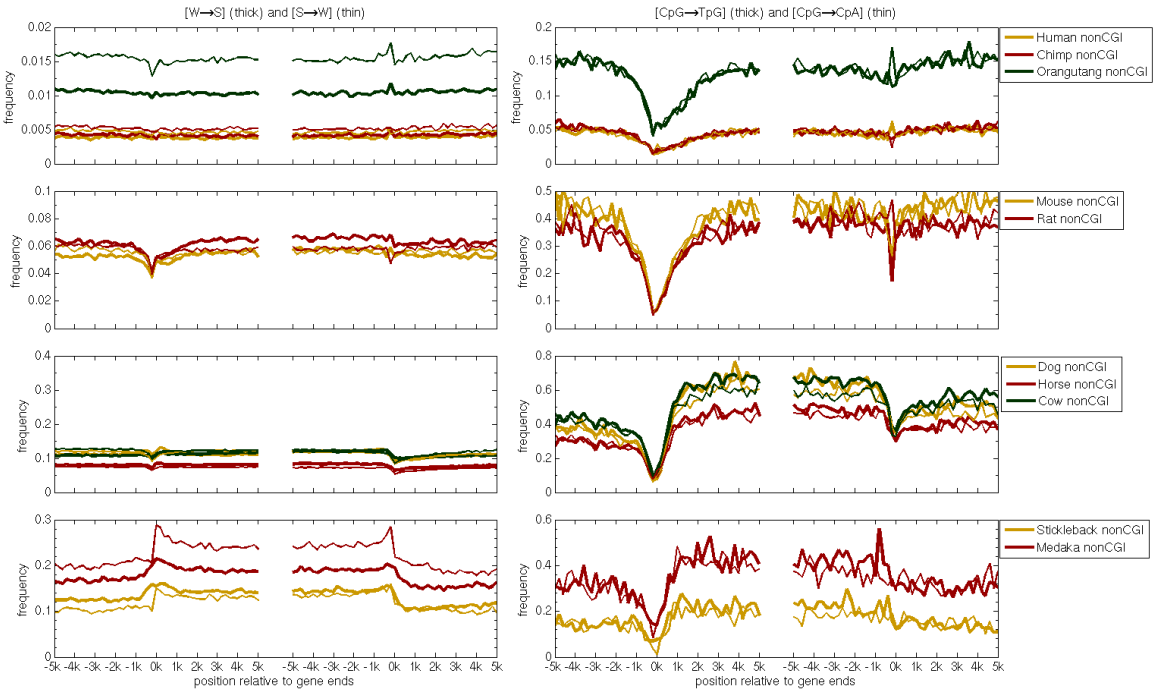

Figure B.7: The frequencies of $S \to W$, $W \to S$ and CpG methylation deamination process around the ends of nonCGI-genes in vertebrates.

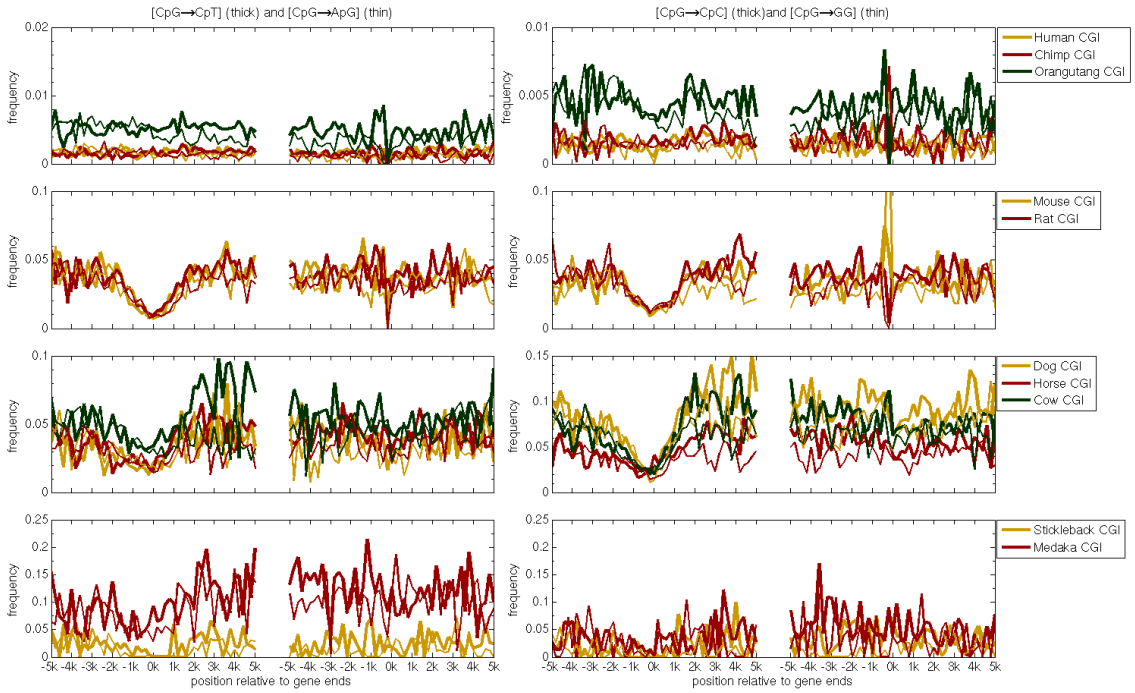

Figure B.8: The frequencies of $S \to W$, $W \to S$ and CpG methylation deamination process around the ends of CGI-genes in vertebrates.

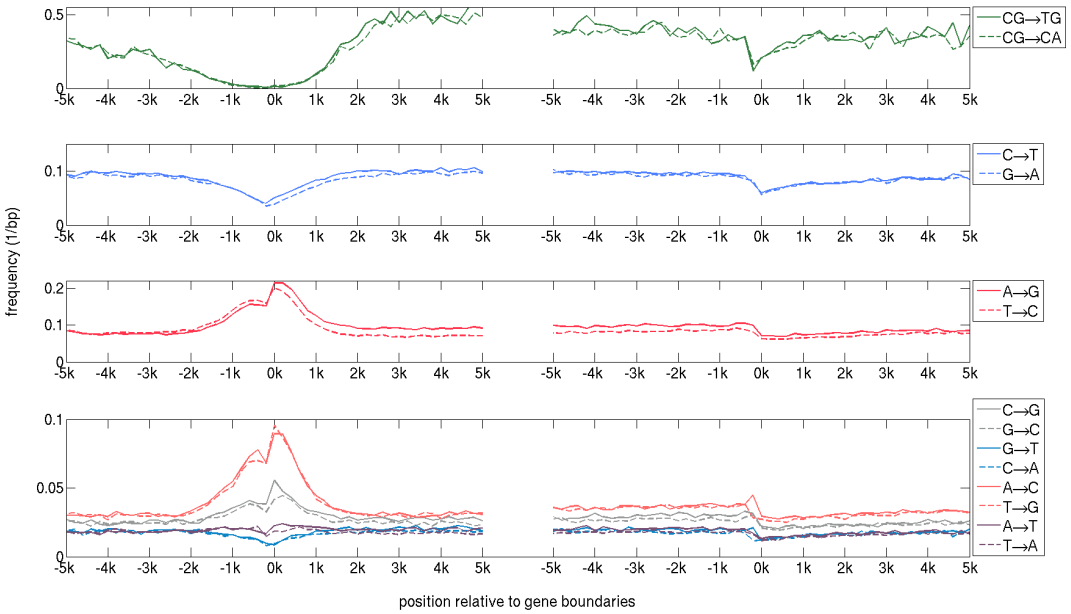

Figure B.9: Substitution rates in introns and in intergenic regions in the vicinity of 5' and 3'ends of dog CGI-genes. The plots show the estimated twelve single nucleotide substitution rates and the CpG deamination rates in non-overlapping 200 bp long windows along the non template strand. The distances of the windows' centers from the 5'end or 3'end are indicated on the x- axes. The estimation of substitution frequencies has been performed using the non-template strand.

C Appendix C

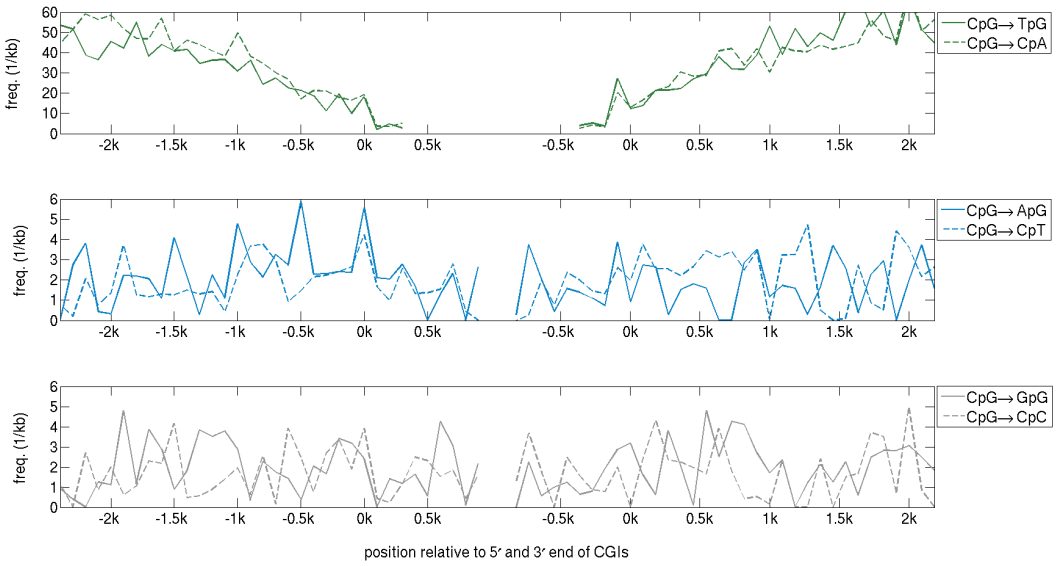

Figure C.1: CpG methylation-deamination rates in the proximity of 5 and 3 ends (left and right 0k, respectively) of dCGI. In the top two panels, the current values of GC content and CpG odds (continuous lines) are compared with the corresponding stationary quantities (dashed lines). The data points between the left and the right 0k are calculated within the CGIs.

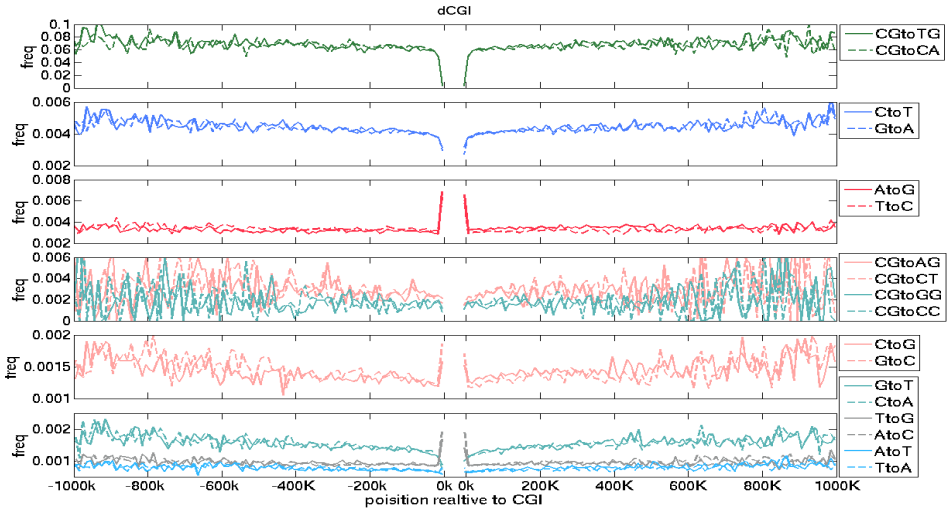

Figure C.2: Substitution frequencies within dCGIs and along their 1 Mbps long upstream and downstream intergenic regions. The plots show the estimated twelve single nucleotide substitution frequencies and 6 CpG deamination frequencies in non-overlapping 10 kbp long windows. The 5' and 3' ends of dCGIs are represented by left and right 0k, respectively. The distances of the windows' centers from the 5'end and 3' end of dCGIs are indicated by negative and positive values, respectively. The estimated of rates within dCGIs are indicated by the data point at 0ks (see Fig. C.5 for the amount for sequence in each window).

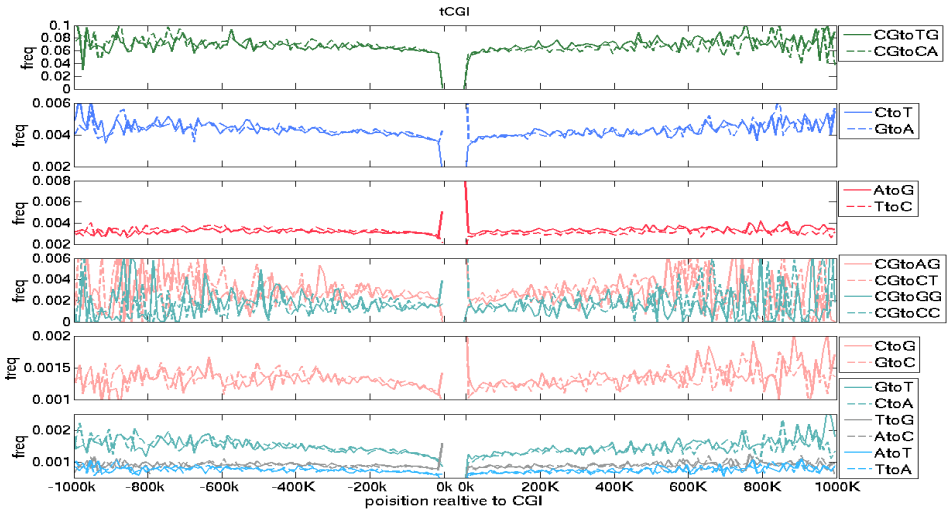

Figure C.3: Substitution rates within and around tCGIs (see Fig. C.2 for further details).

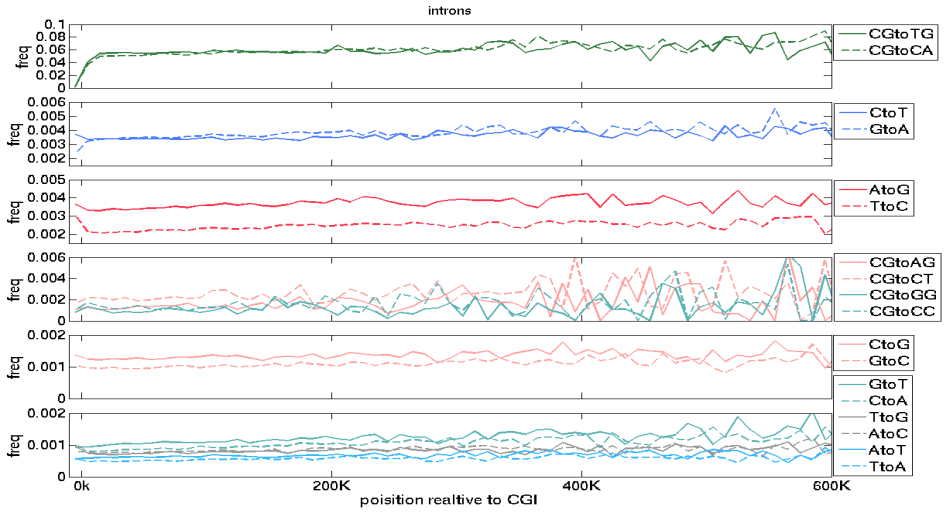

Figure C.4: Substitution rates in intronic regions of genes that have TSS inside of tCGIs. The positions are relative to the 3' end of tCGI (see Fig. C.2 for further details).

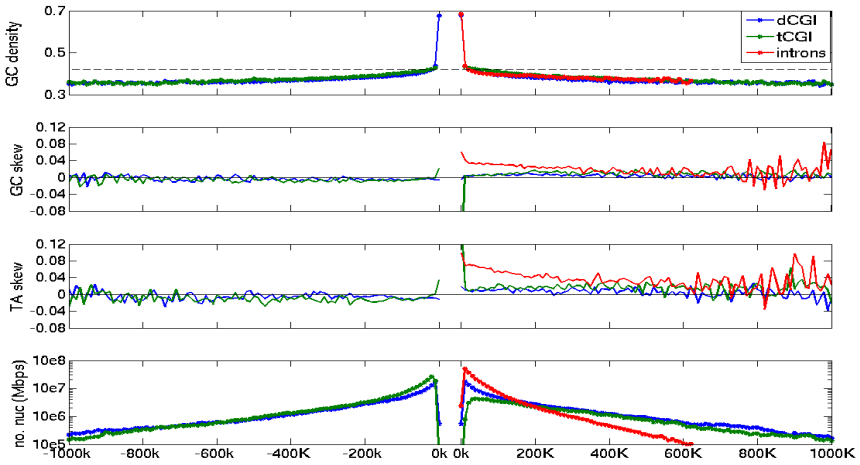

Figure C.5: Statistical features of sequences that were used for the estimation substitution rates in Figure 2 and Fig. 5.1, C.2, C.3 and C.4. TA- and GC-skews are defined in the main text.

Notation and abbreviations

DNA	deoxyribonucleic acid
A	Adenine
G	Guanine
C	Cytosine
T	Thymine
bp	base-pair
kbp	kilo base-pair
Mbp	Mega base-pair
RNA	Ribonucleic acid
ssDNA	single stranded DNA
dsDNA	double stranded DNA
SNP	single nucleotide polymorphism
α	a base
β	a base
$\alpha \to \beta$	a mutation or substitution from a base α to β
$r_{\alpha\to\beta}$	the frequency of substitution from a base α to β
W	a weak base (A or T)
S	a strong base (C or G)
$r_{W\to S}$	the frequency of substitutions from a weak base to a strong base
$r_{S\to W}$	the frequency of substitutions from a strong base to a weak base
BGC	biased gene conversion
CGI	CpG island
TCR	transcription coupled repair
$[\alpha]$	the frequency of a base α in a sequence
S_{TA}	TA skew $(=([T]-[A])/([T]+[A])$
S_{GC}	GC skew $(=([G]-[C])/([G]+[C])$
ORI	origins of replication initiation
OBR	origins of bi-directional replication initiation
TSS	transcription start site
FFD	four fold degenerate sites
$\vec{\alpha}$	nucleotide sequence $\vec{\alpha} = (\alpha_1, \ldots, \alpha_S)$
$\vec{\beta}$	nucleotide sequence $\vec{\beta} = (\beta_1, \ldots, \beta_S)$
Pr	probability

$A(t)$	random variable over the sequence space
$A_i(t)$	random variable over $\{A, C, G, T\}$ at a site i
Q	transition rate matrix
$Q^{(1)}$	context-free transition rate matrix
$Q^{(3)}$	transition rate matrix of triplets
P	transit ion matrix
$P^{(1)}$	transition matrix for context free substitution model
$P^{(3)}$	transition matrix for triplets
π	stationary distribution
$\otimes$	Kronecker tensor product
L	likelihood function
UTR	untranslated regions
tCGI	tss associated CGI
dCGI	distal CGI
gCGI	gene CGI
pCGI	proximal CGI
ESC	embryonic stem cell
AID	activation induced (cytidine) deaminase
SHM	somatic hyper-mutation
G4	G-quadruplex
R-loop	RNA/DNA hybrids
pol	polymerase
NER	nucleotide-excision repair

Zusammenfassung

Die Verfügbarkeit von Säugetiergenomen und ihrer wechselseitigen Alignments sowie zugehöriger Genomannotationen von hoher Qualität ermöglichen es uns, Einblicke in die Verschiedenheit von Mutationsprozessen in unterschiedlichen Kontexten entlang menschlicher Chromosomen zu erhalten. Insbesondere kann die Frage angegangen werden, welche Substitutionsmuster mit verschiedenen zellulären Prozessen assoziiert sind. Wir haben die Auswirkung von Transkription auf Substitutionsmuster in der Umgebung der 5'- und 3'-Enden von Genen untersucht. Zudem wird eine Analyse der Substitutionsmuster in und um CpG-Inseln vorgestellt, welche säugerspezifische Sequenzbestandteile darstellen. Die Analysen enthüllen reichhaltige und (in gewissem Maße) unerwartete Mutationsmuster, die mit Transkriptionsprozessen, CpG-Inseln oder beidem assoziiert sind.

Im Menschen wurden drei Transkriptions-assoziierte Substitutionsmuster beobachtet, von denen zwei mit CpG-Inseln in Zusammenhang stehen. Das erste Muster, eine starke Abnahme der Deaminierungsrate von methylierten CpG-Dinukleotiden, wurde im näheren Umfeld des 5'-Endes von Genen beobachtet, da die dort häufig auftretenden CpG-Inseln meist ein schwächeres Methylierungsniveau aufweisen als CpG-Dinukleotide an anderen Stellen im Genom. Das zweite Muster, eine strangspezifische Asymmetrie in komplementären Substitutionsraten, erstreckt sich vom 5'-Ende bis zu 1 kbp hinter dem 3'-Ende und ist mit Transkriptions-gekoppelter Reparatur assoziiert. Das dritte Muster wird von einer örtlich begrenzten Strangasymmetrie gebildet, einem Überschuss von $C \to T$ gegenber $G \to A$-Substitutionen im Nicht-Template-Strang, der auf die ersten 2 kbp hinter dem 5'-Ende von Genen nahe CpG-Inseln beschränkt ist. Dieses Muster könnte von einer höheren Exponiertheit des Nicht-Template-Strangs nahe dem 5'-Ende von Genen bedingt sein, welche zu einer höheren Cytosin-Deaminierungsrate führt. Diese Art von Substitutionsasymmetrie ähnelt derjenigen, die als Folge des somatischen Hypermutationsweges zu beobachten ist. Möglicherweise sind einige während der somatischen Hypermutation aktive Proteine, wie etwa der DNA-Mutator Activation Induced cytidine Deaminase (AID), welcher einzelsträngige DNA bindet, auch in Keimbahnzellen von Säugern aktiv. Die nötige ssDNA-Konformation kann von R-Loops oder G4-Strukturen induziert werden, die vorzugsweise am 5'-Ende von Genen auftreten.

Die Transkriptions-assoziierten Substitutionsmuster sind nicht auf den Menschen beschränkt und können auch in anderen Säugerspezies beobachtet werden, so etwa bei

Schimpanse, Orang-Utan, Maus, Ratte, Pferd, Rind und Hund. Fische zeigen auch Strangasymmetrie-Muster in Introns, jedoch unterscheiden sich diese Asymmetrien von denen in Säugern, was darauf hinweist, daß Transkriptions-assoziierte Reparatur beziehungsweise Mutageneseprozesse in der Wirbeltierlinie evolvierten.

Strangspezifische Substitutionsprozesse existieren auch in intergenischen Regionen. CpG-Inseln sind der Ausgangspunkt von bidirektionalen Strangasymmetrien, die sich über Hunderttausende von Basenpaaren erstrecken. Diese Asymmetrien können von DNA-Replikationsprozessen ausgelöst werden, die CpG-Inseln als Initiationsorte nutzen. Alternativ können die Asymmetrien in intergenischen Regionen Anzeichen von unbekannten Transkripten sein, wie zum Beispiel sehr langen nichtkodierenden RNAs. In intergenischen Regionen abwärts von Genen treten Strangasymmetrien auf, die denen in Introns ähneln, was darauf schließen lässt, dass die RNA-Polymerase die Transkription in Bereiche fortsetzt, die hinter dem 3'-Ende von Genen liegen.

In der vorliegenden Arbeit wird auch das Verhältnis von $W \to S$ gegenüber $S \to W$-Substitutionshäufigkeiten untersucht. Das genomweite Verhältnis ist in Säugergenomen und allen bisher sequenzierten Tiergenomen kleiner als 1. In humanen CpG-Inseln, die weitab von annotierten Genen liegen, ist das Verhältnis dagegen größer als 1. Bisherige Studien zeigten, dass dieses Verhältnis positiv mit Crossing-Over-Raten korreliert ist. Werden CpG-Inseln nach Crossing-Over-Raten aufgeteilt, so steigt das Verhältnis mit der Crossing-Over-Rate. Dies deutet darauf hin, dass Rekombination, möglicherweise über damit verbundene Prozesse wie etwa gerichtete Genkonversion, eine treibende Kraft in der Entstehung von CpG-Inseln darstellt. In CpG-Inseln, die mit den TSSs von humanen und den meisten anderen Säugergenen überlappen, gibt es eine entgegengesetzte Tendenz, nämlich einen berschuss der $S \to W$ gegenüber $W \to S$-Rate. Diese Tendenz kann auf lange Sicht zum Verlust von CpG-Inseln in TSS-Regionen führen. In einem bemerkenswerten Gegensatz dazu beobachteten wir in Hund und Stichling einen langfristigen Anstieg in GC-Gehalt und folgern, dass in diesen Spezies Rekombination eine Rolle bei der Formung der Promoterregionen spielt.

Der GC-Gehalt in CpG-Inseln ist bereits bekannt als "genomisches Satzzeichen" im genomischen Verlauf der CpG-Methylierungs-Deaminierungsraten. Daher deuten die Ergebnisse der vorliegenden Arbeit an, dass Säuger-CpG-Inseln verstärkt als Begrenzungen für mannigfache Mutationsprozesse wirken, vor allem für diejenigen, die Strangasymmetrien herausbilden. Dies, so lässt sich spekulieren, ist eine Mutationsspur von Transkription und Replikation, welche genomweit dazu neigen, an CpG-Inseln eingeleitet zu werden.

Curriculum vitae

For reasons of data protection, the curriculum vitae is not included in the online version

Erklärung zur Urheberschaft

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Paz Polak Berlin, im Juni 2010