

# **Patterns of Two Types of Overlapping Genes in Five Mammalian Genomes**

Chaitanya Ramesh Sanna

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University  
in partial fulfillment of the requirements for the degree of

Master of Science  
In  
Computer Science

Liqing Zhang - Chair  
Lenwood S. Heath  
Zhijian Tu

July 28, 2006  
Blacksburg, Virginia

Keywords: overlapping genes, same strand, different strand, 5'-UTR, 3'-UTR

Copyright 2006, Chaitanya R. Sanna

# **Patterns of Two Types of Overlapping Genes in Five Mammalian Genomes**

Chaitanya Ramesh Sanna

## **ABSTRACT**

Increasing evidence suggests that overlapping genes is a common phenomenon in eukaryotic genomes too and are not restricted to prokaryotes alone. Here we determined overlapping genes in a set of orthologous genes in the genomes of human, chimp, mouse, rat, and dog and contrasted the patterns of overlapping between two principal types of overlapping genes, the same-strand-overlapping genes and different-strand-overlapping genes. The two types of overlapping genes are compared with respect to their frequencies, overlap lengths, region of overlap, and conservation of overlap in five species. Our results suggest the following: different-strand-overlaps are more common, both types show different patterns with respect to overlap lengths and regions of overlap, different-strand-overlapping genes are more evolutionarily conserved, and 3'-UTR evolution plays an important role in transitions between non-overlapping genes and overlapping genes.

The thesis also presents a review of related work in terms of history, origin, types, biological significance of overlapping genes, human diseases associated with them, and their comparison in prokaryotes and eukaryotes.

# Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Liqing Zhang for her persistent guidance, patience, and motivation. I am indebted to her for the valuable direction and ideas she always gave when faced with a problem.

I am very thankful to Dr. Lenwood Heath and Dr. Zhijian Tu for serving on my committee.

I am grateful to the Department of Computer Science, Virginia Tech, for supporting me through my graduate study.

I would also like to thank all my friends and roommates, especially Prasant and Prakriti for always being there for me.

Special thanks go to my parents, Ramesh and Dr. Indira Reddy, my sister Divya, and the rest of my family for their constant love and support.

# Table of Contents

<b>INTRODUCTION</b> .....	<b>1</b>
<b>BACKGROUND</b> .....	<b>4</b>
2.1 DEFINITIONS .....	4
2.2 CONCEPTS.....	6
<b>RELATED WORK</b> .....	<b>8</b>
3.1 HISTORY OF OVERLAPPING GENES .....	8
3.2 ORIGIN OF OVERLAPPING GENES.....	9
3.3 TYPES OF OVERLAPPING GENES .....	10
3.4 HUMAN DISEASES THAT INVOLVE OVERLAPPING GENES .....	11
3.5 BIOLOGICAL SIGNIFICANCE .....	12
3.6 PROKARYOTES VERSUS EUKARYOTES .....	13
<b>MATERIALS AND METHODS</b> .....	<b>14</b>
4.1 DATA COLLECTION.....	14
4.2 METHODS .....	14
<b>RESULTS AND DISCUSSION</b> .....	<b>17</b>
5.1 CLASSIFICATION AND FREQUENCY OF THE TYPES OF OVERLAPS.....	17
5.2 THE LENGTHS OF OVERLAPPING REGIONS .....	19
5.3 THE BIRTH/DEATH OF OVERLAPPING GENES .....	32
<b>CONCLUSIONS</b> .....	<b>38</b>
<b>FUTURE WORK</b> .....	<b>39</b>
7.1 ORIGIN.....	39
7.2 BIOLOGICAL SIGNIFICANCE .....	39
7.3 IN PLANTS .....	40
7.4 GENOME-WIDE STUDY .....	40
<b>APPENDIX A: FIGURES IN CHIMP, MOUSE, AND RAT</b> .....	<b>49</b>
<b>VITA</b> .....	<b>61</b>

# List of Figures

Figure 1. 1 Types of overlapping genes.....	1
Figure 2. 1 Structure of a eukaryotic gene.....	4
Figure 2. 2 Synthesis of cDNA.....	5
Figure 4. 1 Illustration of determining the causes for the transition between non-overlapping and overlapping genes .....	16
Figure 5. 1 Overlapping lengths in same strand and different strand overlapping gene pairs of 5 species.....	21
Figure 5. 2 The Cumulative distribution of overlapping lengths.....	22
Figure 5. 3 The distribution of the proportion of the overlapping lengths with respect to the sizes of the short genes, long genes, and the regions spanned by both genes.....	26
Figure 5. 4 The distribution of the ratios of the lengths of short vs. long genes for the same and different strand overlaps in comparison with the one for the neighboring and non-overlapping genes in Human and Dog.....	31
Figure 5. 5 Conservation of overlapping relationships in the five species .....	34
Figure A. 1 The distribution of the proportion of the overlapping lengths with respect to the sizes of the short genes, long genes, and the regions spanned by both genes.....	54
Figure A. 2 The distribution of the ratios of the lengths of short vs. long genes for the same and different strand overlaps in comparison with the one for the neighboring and non-overlapping genes. ....	60

## List of Tables

Table 5. 1 Statistics of overlapping genes in the 11197 orthologous genes .....	18
Table 5. 2 Classification of overlapping gene pairs based on the location of overlaps ....	18
Table 5. 3 Causes for transition between non-overlapping and overlapping genes for same strand embedded gene pairs .....	36

# List of Abbreviations

DNA – DeoxyRiboNucleic Acid

mRNA - Messenger RiboNucleic Acid

UTR – Untranslated Region

cDNA – Complementary DNA

EST – Expressed sequence tags

BLAST – Basic Local Alignment Search Tool

BLASTP – Protein-protein BLAST

bps – Base pairs

kb or kbps – Kilo Base pairs

# Chapter 1

## Introduction

During the golden age of molecular biology, the period roughly between 1950 and 1970 [Kolalta, 1977], and until overlapping genes were discovered, the scientific community believed that genes occurred on chromosomes in a fashion similar to beads that are sewn through a string [Portin, 1993], completely dismissing the possibility of overlap. Nevertheless, genes do not occur at regular intervals along the DNA of an organism. Genes can sometimes occur densely in a particular region; they can be sparsely located or can even overlap with one another. Overlapping genes are known to be abundant in viruses, mitochondria, bacterial chromosomes and plasmids [Chisholm and Johnson, 2004]. Recent studies have shown that the phenomenon of overlapping genes is not confined to prokaryotes alone and that there are a considerable number of overlapping genes in eukaryotes. A few examples are human [Bristow et al., 1993; Cooper et al., 1998; Kennerson et al., 1997; Veeramachaneni et al. 2004], mouse [Batshake et al., 1996], rat [Adelman et al., 1987], fish [Makalowskaa et al., 2005], and flies [Misener et al., 2000; Spencer et al., 1986]. Section 3.1 gives a more detailed description of the history.

Figure 1.1 illustrates two principal types of overlapping genes that are discussed in the remainder of the thesis. The arrows in the figure point to the overlapping gene pair.

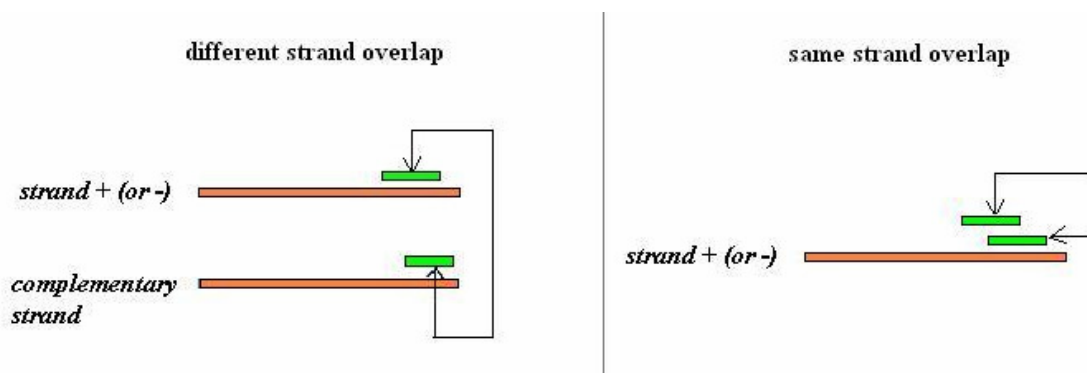


Figure 1. 1 Types of overlapping genes

In the first type of overlapping genes, the gene pair is transcribed from the same strand of the chromosome. The second type of overlapping genes consists of gene pairs in which each gene is transcribed from a different strand. However, most recent large-scale analyses have been restricted to different-strand-overlapping genes. This is mainly because different-strand overlapping genes form potential sense and antisense genes that can be important for regulation of gene expression at levels such as transcription, mRNA processing, splicing, and translation [Boi et al., 2004; Alfano et al., 2005]. There has not been much research done on same-strand-overlapping genes. However, same-strand-overlapping genes are also functionally important and are reported in a few cases [Cawthon et al., 1991; Williams et al., 2005].

It is important to have a broad comparison between the two principal types of overlaps. Here we compiled a list of genes that have orthologs in five species, human, chimp, mouse, rat, and dog, and used overlapping genes in this “orthologous-gene” set to address the following questions:

1. What are the relative proportions of same and different-strand-overlapping genes?
2. Is there any distinct difference between these two types of overlaps? We addressed this question mainly in terms of the sizes of overlapping regions.
3. Is there any difference in the lengths of the overlapping regions between the two types of overlaps?
4. Is there any constraint on the degree of overlap and how does this differ between the two types of overlaps?
5. Finally, taking advantage of this “orthologous-gene” set, we examined the evolutionary conservation of overlapping relationships and causes for transition between non-overlapping and overlapping genes, especially for same-strand-overlapping genes.

The thesis is comprised of seven chapters. Chapter 2 provides a few definitions of biological terms and explanation of concepts that are used throughout the thesis. Chapter 3 presents some aspects of work done so far in relation to overlapping genes and discusses their history, origin, biological significance, types, human diseases that involve

overlapping genes, and their occurrence in prokaryotes compared to eukaryotes. Chapter 4 discusses materials and methods used in the study. Chapter 5 explains the results and simultaneously presents a discussion regarding the same. Chapter 6 highlights major conclusions drawn from the study. Chapter 7 proposes some future directions in relation to advancing research regarding overlapping genes. Bibliography is followed by Appendix A that includes a few more figures that are referred in Chapter 5.

## Chapter 2

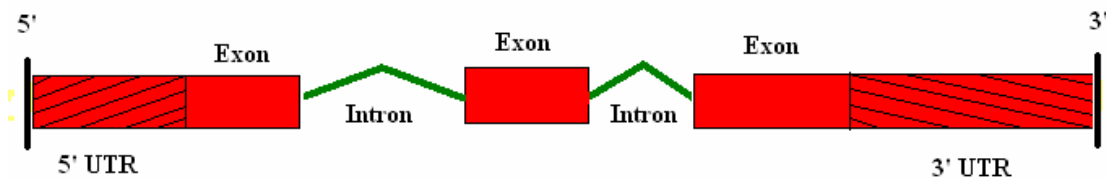
### Background

This chapter provides some definitions of biological terms (section 2.1) and concepts (section 2.2) that will be frequently used in the remainder of the thesis.

#### 2.1 Definitions

##### Gene

Inherited traits are determined by the elements of heredity that are transmitted from parents to offspring in reproduction; these elements of heredity are called *genes* [Hartl and Jones, 2002]. Gregor Mendel in 1866 was the first to discover existence of genes and rules that governed their transmission from one generation to the next. The term gene however, was coined in 1909 by Wilhelm Johannsen, a Danish botanist. Genes occur on chromosomes and are nothing but segments of DNA. The general structure of a eukaryotic gene is illustrated in figure 2.1.



**Figure 2. 1 Structure of a eukaryotic gene**

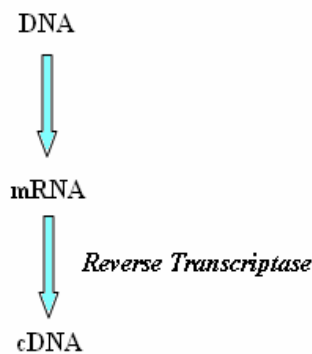
The 5' and 3' mark the direction of transcription (process that involves the making of mRNA from DNA) where the gene is usually read from the 5' end to the 3' end. There are two untranslated regions that are the non-coding regions of the gene and are called 5'UTR and 3' UTR depending on whether they are present at the 5' or 3' end. The coding region is also called the Open Reading Frame (ORF). Both non-coding and coding regions are transcribed into RNA during transcription, after which the introns are spliced out. Then translation occurs to generate the protein.

## DNA

In 1869, Friedrich Miescher discovered a new type of weak acid, abundant in the nuclei of white blood cells that turned out to be the chemical substance genes are made of. Miescher's weak acid is now called *deoxyribonucleic acid*, or *DNA* [Hartl and Jones, 2002]. DNA consists of a sequence of nucleotides of the following bases: Adenine (A), Thymine (T), Guanine (G), and Cytosine (C).

## cDNA

*cDNA* is synthesized from mRNA with the help of an enzyme called reverse transcriptase. This enzyme acts on the single stranded mRNA, producing its complementary DNA. This is based on the pairing rules where RNA base pairs (A, U, G, C) pair with their DNA complements (T, A, C, G).



**Figure 2. 2 Synthesis of cDNA**

## mRNA

*mRNA* or *Messenger RNA* carries information from DNA during transcription (in the nucleus) to protein synthesis sites (in the cytoplasm) to undergo translation that finally results in a gene product.

## EST

An *expressed sequence tag* or *EST* is a short sequence of a transcribed protein-coding or non-protein coding DNA piece with a length of about 500 to 800 base pairs. It is known to be helpful in gene discovery and sequence determination.

**BLAST** [Altschul et al., 1990]

*BLAST* is the acronym for *Basic Local Alignment Search Tool*. The tool aids in comparing biological sequences such as amino acid sequences or DNA sequences. A query sequence is compared with a library in order to identify whether any library sequences resemble the query sequence above a certain threshold. It finds regions of local similarity between sequences. Therefore, the tool is popularly used to infer evolutionary and functional associations between sequences. It is also used to identify members of gene families.

### **BLASTP**

*BLASTP* is a specific kind of BLAST program called the *Protein-protein BLAST*. This program deals with protein sequences as opposed to DNA sequences. In particular, the program accepts a protein query as input and outputs the protein sequences that it resembles the most after comparing it with the protein database specified by the user.

### **Unique Best Reciprocal Hit (UBRH)**

When a query gene translation has an unambiguous 'best' hit to a target translation, and that particular target translation has an unambiguous best hit back to the starting query translation, that gene translation pair is labeled a UBRH orthologous prediction.

## **2.2 Concepts**

### **Overlapping genes**

Every gene has a start and end co-ordinate given in terms of base pairs. In this thesis, two genes were determined to be overlapping based on their start and end co-ordinates. For example: - let start1 and end1 be the start and end co-ordinates of gene1; and start2 and end2 be the start and end co-ordinates of gene2 respectively. Then gene1 and gene2 are determined to be overlapping any one of the following conditions is satisfied:

- $start2 \geq start1$  and  $start2 < end1$
- $start1 \geq start2$  and  $start1 < end2$

Therefore genes could either partially overlap or sometimes one gene can even be completely embedded inside the other gene.

### **Alternative Splicing**

Alternative splicing can be defined as the process in which a pre-mRNA can lead to different mRNAs by including some exons and excluding others each time. This kind of splicing in turn results in producing different proteins.

## Chapter 3

### Related Work

This chapter presents some aspects of the work done so far in relation to overlapping genes.

#### 3.1 History of overlapping genes

The period between 1950 and 1970 was considered as the golden age of molecular biology when overlapping genes were considered an exception to the universal rule of nonoverlap [Kolata, 1977]. Because this assumption was so firmly ingrained in the scientific community, the first evidence of overlapping genes was readily dismissed as an artifact. Around 1973, Alan Weiner and Klaus Weber revealed that the translation of two genes of Q $\beta$ , an RNA virus that infects *Escherichia coli*, is initiated from a common site on the viral genome [Kolata, 1977]. Around the same time that overlapping genes were revealed in Q $\beta$ , researchers were also suspecting that they were present in  $\Phi$ X174 which is a small DNA virus that infects *E. coli* [Kolata, 1977]. So the first ever report of overlapping genes that was accepted by the scientific community was in  $\Phi$ X174 and G4 bacteriophages [Barrell, 1976; Shaw, 1978]. This was followed by quite a lot of research in other viral and prokaryotic genomes, which led to the discovery of more overlapping genes.

As availability of sequenced genomes increased, researchers began to discover that overlapping genes occurred in eukaryotic genomes too. But it took nearly a decade following the discovery of overlap in  $\Phi$ X174 to report similar occurrences in eukaryotes like fruit fly [Spencer et al., 1986; Henikoff et al., 1986; Misener and Walker, 2000; Misra et al., 2002; Yanicostas and Lepesant, 1990; Schultz and Butler, 1989; Schultz et al., 1989; Wong et al., 1987; Schulze et al., 2005] and mouse [Williams and Fried, 1986; Shendure and Church, 2002; Kiyosawa et al., 2003; Batshake et al., 1996] and of course in human [Bristow et al., 1993; Cooper et al., 1998; Kennereon et al., 1997; Veeramachaneni et al. 2004; Shendure and Church, 2002; Lehner et al., 2002; Fahey et

al., 2002; Yelin et al., 2003]. The following few years showed more reports of overlapping genes in other species like chicken [Farrell and Lukens, 1995; Zuniga Mejia Borja et al., 1993], frog [Kimelman and Kirschner, 1989], rat [Adelman et al., 1987], fish [Makalowskaa et al., 2005], and yeast [Malavasic and Elder, 1990; Peterson and Myers, 1993]. There are reports of overlapping genes in plants like Arabidopsis [Quesada et al., 1999], and rice [Osato et al., 2003].

### **3.2 Origin of overlapping genes**

Some molecular mechanisms involved in creation of a new individual gene are exon shuffling (ectopic recombination of exons and domains from distinct genes), gene duplication, retroposition (new gene duplicates are created in new genomic positions by reverse transcription or other processes), mobile elements or transposable elements (sequence is directly recruited by host genes), lateral gene transfer (a gene is laterally or horizontally transmitted among organisms), gene fusion (two adjacent genes fuse into a single gene), gene fission (single gene splits into two genes), and de novo origination (a coding region originates from a previously non-coding genomic region) [Long et al., 2003].

The origin of overlapping genes is explained by two popular hypotheses, overprinting [Keese and Gibbs, 1992; Makalowska et al., 2005] and adoption of signals from neighboring gene locus [Makalowska et al., 2005]. Overprinting is defined as the process of generating new genes from preexisting nucleotide sequences. Overprinting is a widespread phenomenon in viruses too [Pavesi, 2006]. Grasse [1977] who coined the term “overprinting” first discussed the prospect of producing novel genes from already existing nucleotide sequences. Another theory explains overlap as the result of two (or more) genes brought together during the phenomenon of chromosomal rearrangement [Shintani et al., 1999]. According to [Clark et al 2001] gene overlap is due to reduction or even complete elimination of intergenic regions that may be caused by a mutational bias towards deletion [Fukuda et al., 2003]. Yet another theory attributes the usage of alternative polyadenylation sites as being the reason for overlap as in the case of Mink and Chrne genes [Dan et al., 2002]. A study among 50 bacterial species has shown that

mutations at the 3' upstream end of a gene were a more common reason of overlapping genes among closely related species than among species that are not very closely related [Fukuda et al., 2003]. The same study also shows that mechanisms that could lead to the formation of overlapping genes would probably depend on actual structures of the gene pairs involved [Fukuda et al., 2003].

### 3.3 Types of overlapping genes

Overlapping genes are categorized into several types. Makalowska [2005] divided overlapping genes into the following categories based on the direction of transcripts and regions shared by overlapping genes: (a) Genes on the same strand and sharing the same locus, but coding for different proteins (b) Genes that share only the promoter region (c) Nested genes (d) Embedded gene (e) Genes on opposite strands with overlapping locus but no overlap in the exonic regions (f) Tail-to-tail overlap in the exonic region (g) Head-to-head overlap that involves 3'UTRs and coding sequence.

In some studies overlapping genes were classified into three major categories depending on the direction of transcripts: (a) Convergent ( $\rightarrow\leftarrow$ ), (b) Unidirectional ( $\rightarrow\rightarrow$  or  $\leftarrow\leftarrow$ ), and (c) Divergent ( $\leftarrow\rightarrow$ ) [Chang and Chang; Fukuda et al., 2003]. Another study classified overlapping genes into two types: (a) Tandem ( $\rightarrow\rightarrow$ ) and (b) Antiparallel ( $\rightarrow\leftarrow$  or  $\leftarrow\rightarrow$ ) [Johnson and Chisholm, 2004]. A more detailed description and discussion regarding the types and arrangements of overlapping genes can be found in [Boi et al., 2004].

Although there are several ways to categorize overlapping genes, we only considered the direction of transcripts as a factor to decide the types of overlapping genes in this thesis. Therefore overlapping genes were categorized into same-strand and different-strand overlapping genes based on only one factor, i.e., the direction of transcripts.

In addition, although the exact numbers for each of the types of overlapping genes vary in different studies, in [Fukuda et al., 2003] they mention that divergent overlapping genes were rare when compared to the other two types. In another study that involved

about 50 bacterial species, although the number of convergent and divergent gene pairs was the same, the fraction of divergent overlapping genes was found to be lower than that of convergent overlapping genes [Fukuda et al., 2003]. Another study has shown that unidirectional overlapping genes are most common in bacterial genomes [Fukuda et al., 2003; Eyre-Walker, 1995].

### **3.4 Human diseases that involve overlapping genes**

Research has shown that the etiology of many human diseases have been traced to overlapping genes. In fact, this number seems to be on the rise with the progress of time. Some examples of human diseases include the following: In patients diagnosed with endometriosis, the proliferation of endometrial cells is attributed to the reduced expression of the exon 1B isoform of the basic fibroblast growth factor (bFGF) antisense transcript [Mihalich et al., 2003]; A disturbance of the interaction between XLas and ALEX, which are structurally unrelated mammalian proteins translated from alternative overlapping reading frames of a single transcript, leads to abnormal human phenotypes, including mental retardation and growth deficiency [Nekrutenko et al., 2005]; The last exon of CYP21 gene overlaps with Tenascin X, which plays a crucial role in the regulation of collagen deposition by dermal fibroblasts and contributes to the human Ehlers-Danlos syndrome [Mao et al., 2002]; SNURF-SNRPN sense/UBE3A antisense transcription units are involved in Prader-Willi and Angelman syndromes [Runte et al., 2001]; the arthropathy-camptodactyly syndrome involves the PRG4 gene overlapping at 3' end with TPR gene [Marcelino et al., 1999]; an inherited form of  $\alpha$ -thalassemia is caused by the silencing of the  $\alpha$ -globin gene through the generation of an aberrant cis-NAT [Tufarelli et al., 2003]. The human F8A gene is nested within FACTOR VIII gene [Levinson et al., 1990] and its gene product HAP40 is known to be involved in Huntington's disease [Peters and Ross, 2001]. In addition, a study by Karlin et al [2002] concluded that the majority of overlapping genes groups (defined as- when exons of one gene are contained within the introns of another) encoding significantly many amino acid long runs are potentially associated with disease.

### 3.5 Biological Significance

Several theories have been proposed in order to attempt and explain the biological significance of the occurrence of overlapping genes in various genomes. This section discusses some popular theories that best explain the biological significance of overlapping genes.

The first theory suggests that the occurrence of overlapping genes is a mechanism to compress maximum amount of information into genes [Krakauer, 2000; Normark et al., 1983]. According to [Spencer et al., 1986] numerous cases of overlapping genes in viruses and bacteria demonstrate that parsimonious exploitation of coding capability of DNA is fairly common amongst prokaryotes. Since overlapping genes share a common region of nucleotide sequence, they tend to make the genome more compact while still encoding for different polypeptides. In this way, new peptides can be produced without increasing the size of the genome. Hence, this theory suggests that genome compactness is a major factor for the occurrence of overlapping genes. However, another recent study has shown that overlapping genes cannot actually be considered as a direct mechanism to reduce genome size [Johnson and Chisholm, 2006].

Another theory suggests that overlapping genes are a mechanism to regulate gene expression via translational coupling [Krakauer, 2000]. This theory of transcriptional or functional coupling was also suggested by another study [Inokuchi et al., 2000]. A considerable amount of research has been conducted on different strand overlapping genes commonly referred to as natural sense-antisense transcripts (NATs) and their significance in genomes. In [Lavorgna et al., 2004] the authors report how NATs regulate gene expression in eukaryotic genomes and identify transcriptional interference, RNA masking and dsRNA-dependent mechanisms as the three general mechanisms by which gene expression is regulated by antisense transcription.

Other studies also suggest the role of overlapping genes in X-inactivation [Lee et al 1999], RNA interference [Billy et al., 2001], alternative splicing [Munroe and Lazar, 1991], and RNA editing [Kumar and Carmichael, 1997].

### **3.6 Prokaryotes versus Eukaryotes**

One major difference between overlapping genes in prokaryotes and eukaryotes is related to frequency of same strand and different strand overlapping genes. Studies have shown that in microbial genomes, majority of overlapping genes occur on the same strand when compared to different strand overlapping genes [Johnson and Chisholm, 2006]. This result is consistent with earlier studies that showed that same strand overlapping genes are more common in prokaryotes [Fukuda et al., 1999, 2003; Eyre-Walker, 1995]. Earlier when overlapping genes were considered quite rare in eukaryotic genomes, studies showed that anti-parallel overlapping of genes was also a rare occurrence [Dan, 2002]. However, as time progressed research has proven quite the contrary with respect to the frequency of different strand overlapping genes in eukaryotic genomes. More recent studies about eukaryotic genomes showed that different strand overlapping genes are found to be more common when compared to same strand overlapping genes [Veeramachaneni, 2004; Yelin et al., 2003].

In addition, overlaps are known to be more common in prokaryotes when compared to eukaryotes because they are rapidly evolving genomes. According to [Krakauer, 2000] overlap is more prevalent in genomes that evolve rapidly with high mutations rates, like in viruses, bacteria, and mitochondria.

## Chapter 4

# Materials and Methods

### 4.1 Data Collection

The dataset was obtained from <http://www.ensembl.org> using the online data-mining tool Biomart. Because the study is concerned with overlapping genes not due to alternative splicing and also in comparing the evolution of overlapping genes in mammals, protein-coding genes that are presumably orthologous in human, chimp, mouse, rat, and dog were used. ENSEMBL contains all-against-all BLASTP results for all these species and the pair wise species comparisons. A gene was added to the list of putative orthologous genes only if the gene in human has a unique best reciprocal hit to a gene in each of the other four species. With this criterion, we identified 11197 putative orthologous genes in the five species. These genes were then checked for possible overlaps using the gene annotation in ENSEMBL. According to the annotated start and end positions of the genes, two genes are considered to overlap with each other if they share a region longer than 20 base pairs on the chromosomes. The overlapping pairs of genes were then classified into same-strand overlap if both genes are on the same strand or into different-strand overlap if both genes are each on different strands.

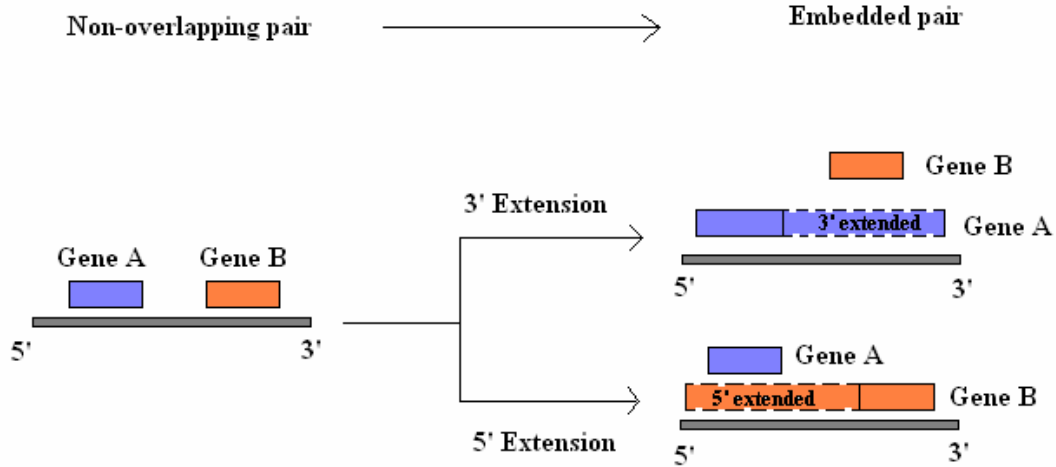
### 4.2 Methods

A great variation in the extent of overlap between two genes was observed. The extreme is when one gene is completely embedded inside another. The type of evolutionary constraint is expected to be different depending on size and location of overlapping regions. To investigate the extent of overlap in these mammalian genomes, we measured the percentage of the overlap regions with respect to the lengths of the shorter gene, the longer gene, and the total region spanned by both genes.

The orthologous gene set provides a unique opportunity to address evolutionary conservation of overlapping genes. The question asked is: whether the overlapping relationship is evolutionarily conserved across these species. In other words, do overlapping genes in one species have the same overlapping relationship in other species as well? Is there any difference between the two types of overlaps? For all the overlapping pairs in each genome, we determined whether the corresponding orthologous gene pairs in the other species are overlapping as well. Three scenarios can occur: the two orthologous genes in a second species also overlap, do not overlap but are on the same chromosome, or are on different chromosomes.

The analysis of evolutionary conservation of overlapping genes shows that there is little conservation of overlapping relationships, especially for same-strand-overlapping genes. This raises an important question about the evolutionary origin of overlapping genes. What causes the fast transition between non-overlapping and overlapping genes? Theoretically, two neighboring but non-overlapping genes can become overlapping through 5' change, i.e. capturing and employing an upstream alternative start signal, 3' change, i.e. capturing and employing a downstream alternative termination signal, or a combination of both mechanisms.

However, it is hard to differentiate computationally whether the transition is due to 5' change or 3' change alone or the combined result of 5' and 3' changes. Fortunately, we can take advantage of the fact that most overlapping genes on the same strand are embedded forms and decide for these cases whether the transition is caused by the 5' or 3' end extension. This idea is illustrated in Figure 4.1 where two genes A and B are non-overlapping at first, with gene A occurring before gene B if read in the direction of 5' to 3'. This non-overlapping pair becomes overlapping in two possible ways. Firstly, the 3' end of gene A can extend by capturing and employing a downstream alternative termination signal leading to gene B being embedded in gene A. An alternative scenario could be one in which the 5' end of gene B extends by capturing and employing an upstream alternative start signal leading to gene A being embedded in gene B.



**Figure 4. 1 Illustration of determining the causes for the transition between non-overlapping and overlapping genes**

The rationale is that for embedded gene pairs that are on the same strand, we look at corresponding orthologous gene pairs in other species that do not overlap but are on the same strand and chromosome and compare their gene orders with the one in the embedded pair. If the gene orders are the same, the cause of the transition is 3' end extension, otherwise 5' end extension. Using this idea, we were able to determine the causes of transition for 56, 70, 74, 204, and 332 cases in human, chimp, mouse, rat, and dog, respectively.

## Chapter 5

# Results and Discussion

### 5.1 Classification and Frequency of the types of overlaps

The statistics of the classification and frequency of the two types of overlapping genes in the five genomes are shown in table 5.1. Of the 11197 orthologous genes, there are altogether 564, 585, 674, 853, and 971 pairs of overlapping genes in human, chimp, mouse, rat, and dog respectively. Because some large genes overlap with multiple genes, there are a total of 982, 1007, 1121, 1201, and 1277 unique genes involved in overlap in human, chimp, mouse, rat, and dog respectively. The proportions of overlapping genes in the five genomes range from about 8.8% to 11.4%. The total amount of overlapping genes in each genome is probably higher than these estimates because we only considered genes that have orthologs in all the genomes that only formed a subset of the total genes that could be present in the genome.

Among the overlapping genes, we found that different-strand-overlapping genes are more common when compared to same-strand-overlapping genes for all genomes (Table 5.1). In chimp and mouse, percentages of different-strand overlapping genes are nearly four times that of same-strand-overlapping genes. In human, the ratio is even higher: 85% vs. 15%. The contrast becomes less distinct in both rat (60% vs. 40%) and dog (56% vs. 44%). It is interesting to note that this pattern is quite opposite from what is observed in prokaryotes. Previous studies of overlapping genes in 198 microbial genomes show that 84% of the overlapping genes are on the same strand with the remaining on different strands [Chisholm and Johnson, 2004]. This clearly shows that majority of overlapping genes in prokaryotes are same strand overlapping genes. This suggests that the significance of the two types of overlapping genes in the genome is different between prokaryotes and eukaryotes. However, the implication of this difference between the two kingdoms of prokaryotes and eukaryotes is still an unanswered question.

**Table 5. 1 Statistics of overlapping genes in the 11197 orthologous genes**

Species	No. of Overlapping genes	Overlapping genes (%)	No. of Overlapping pairs	No. of Same strand overlapping pairs	Same strand overlapping pairs (%)	No. of Different strand overlapping pairs	Different strand overlapping pairs (%)
Human	982	8.77%	564	83	14.72%	481	85.28%
Chimp	1007	8.99%	585	116	19.83%	469	80.17%
Mouse	1121	10.01%	674	139	20.62%	535	79.38%
Rat	1201	10.73%	853	340	39.86%	513	60.14%
Dog	1277	11.40%	971	427	43.97%	544	56.02%

Depending on locations of overlapping regions, overlapping genes can be further divided into different categories. Different-strand-overlapping gene categories include 5' overlapping genes (where the 5' regions of both the genes overlap), 3' overlapping genes (where the 3' regions of both the genes overlap), and embedded genes (where one gene is completely contained in the other gene). Same-strand-overlapping gene categories include 5'-3' overlapping genes (i.e., the 5' of a gene overlaps with the 3' of another gene) and embedded genes. The classification of these types is shown in Table 5.2.

**Table 5. 2 Classification of overlapping gene pairs based on the location of overlaps**

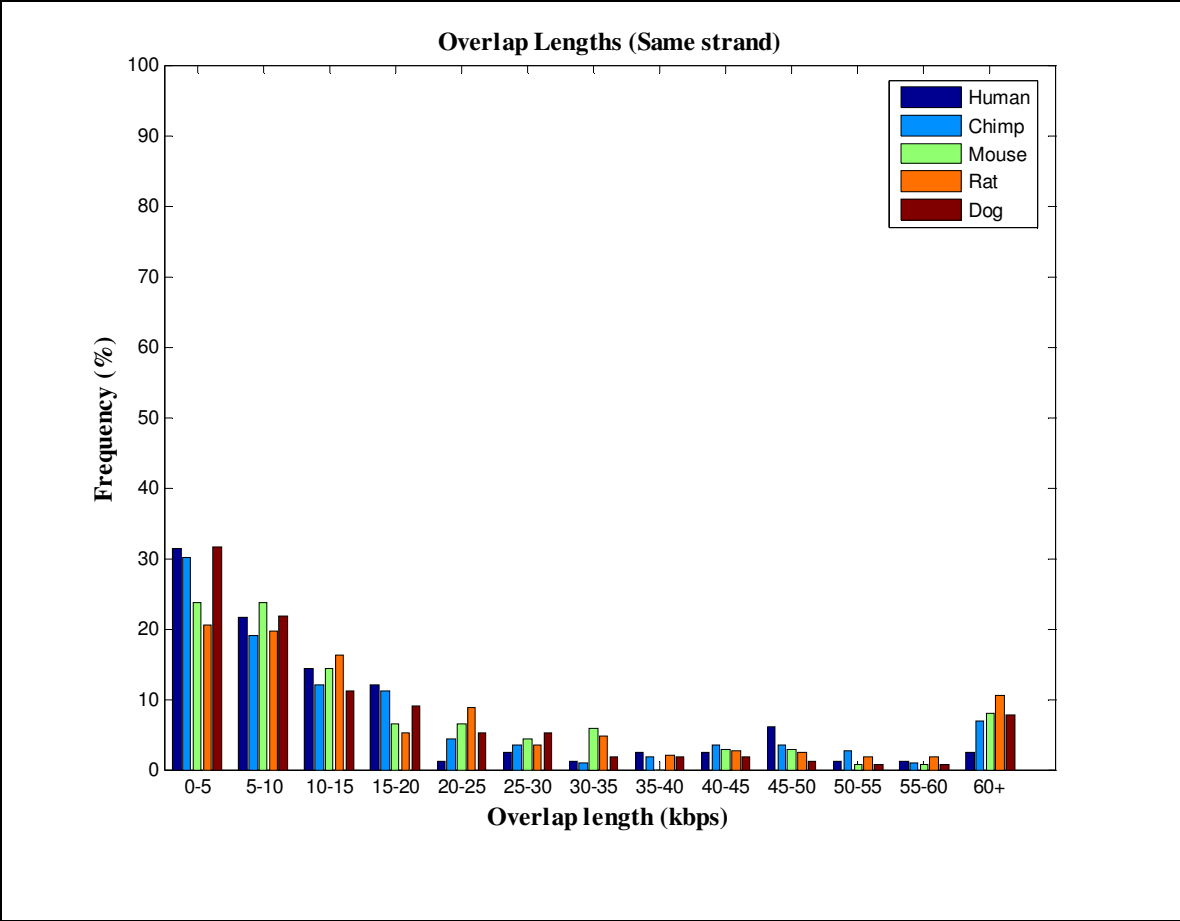
	Same strand		Different strand		
	5' - 3'	Embedded	5'	3'	Embedded
Human	22 (26.5%)	61 (73.5%)	95 (19.8%)	246 (51.1%)	140 (29.1%)
Chimp	40 (34.5%)	76 (65.5%)	95 (20.3%)	223 (47.5%)	151 (32.2%)
Mouse	46 (33.1%)	93 (66.9%)	76 (14.2%)	282 (52.7%)	177 (33.1%)
Rat	91 (26.8%)	249 (73.2%)	79 (15.4%)	125 (24.4%)	309 (60.2%)
Dog	72 (16.9%)	355 (83.1%)	31 (5.7%)	73 (13.4%)	440 (80.9%)

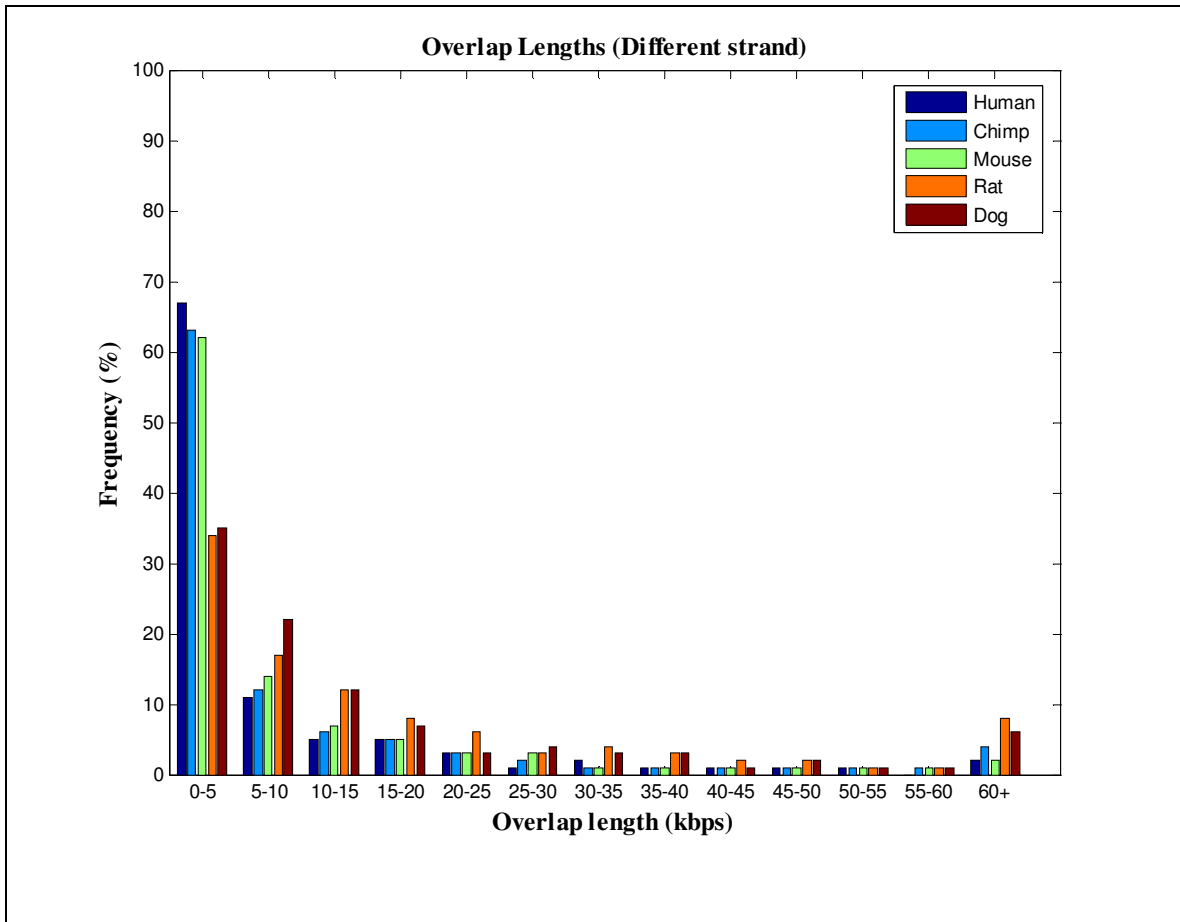
The majority of same-strand-overlapping genes, ranging from about 66% to 83% in the five species, are embedded forms. The different-strand-overlapping genes show different patterns among species: in human, chimp, and mouse, 3' overlapping genes are the most common type (about 50%) with the embedded form less common (about 31%) and 5' overlapping genes the least common (about 18%); in rat and dog, the most common form is the embedded form, followed by 3' and 5' overlapping genes.

There are only a few estimates on the proportions of the three types of different-strand-overlapping genes. For example, Veeramachaneni et al. [2004] found that of a sample of 774 different-strand-overlapping genes in human, about 54% are 3' overlaps, about 30% 5' overlaps, and about 16% the embedded overlaps. A similar pattern is also observed in mouse for a sample of 542 different-strand-overlapping genes with about 54% being 3' overlaps, about 37% 5' overlaps, and about 9% the embedded form. Shendure and Church [2002] reported that about 72% of a sample of 185 different-strand-overlapping genes in human and mouse are 3' overlaps, about 22% embedded forms, and only about 6% 5' overlaps. Two other studies from Lehner et al. [2002] and Yelin et al. [2003] also examined frequencies of the types of different-strand overlaps. These studies and the current one are consistent in that for both human and mouse, the most prevalent type of different-strand overlaps is 3' overlap, but they differ greatly in the exact proportions of 3' overlap. Furthermore, there is no consensus as to which one of the two remaining types (5' and embedded form) is more common. Unfortunately, there are no estimates available for frequencies of various types of the same-strand overlaps that we can compare with.

## **5.2 The lengths of overlapping regions**

The overlap distance between each pair was also calculated for the same strand and different strand overlapping data set. This result is plotted and illustrated in figure 5.1.



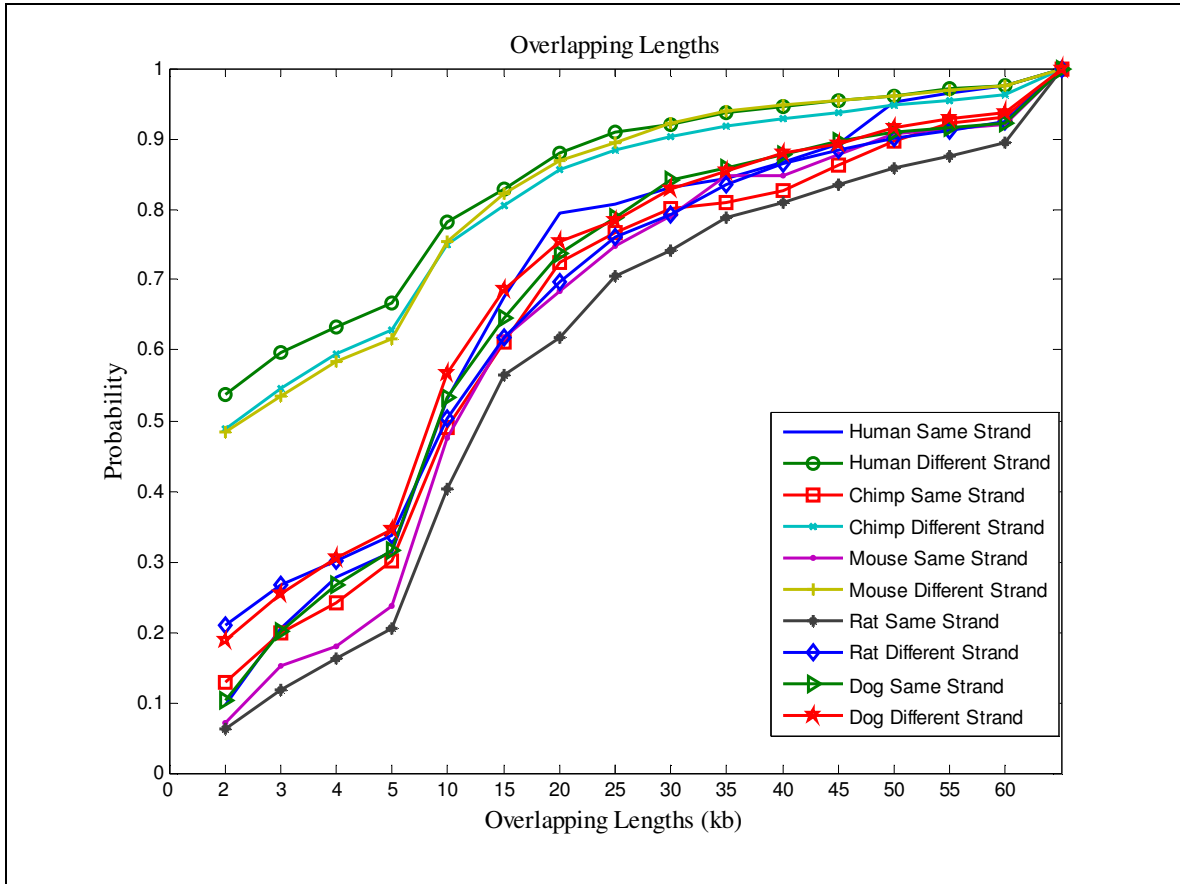


**Figure 5. 1 Overlapping lengths in same strand and different strand overlapping gene pairs of 5 species**

The above figure clearly illustrates that not many same-strand-overlapping gene pairs have overlapping lengths within 5000 bps as compared to different-strand-overlapping gene pairs. About 60% of gene pairs in all the five species have overlapping lengths within 15000 bps.

The figure illustrates that most of the different-strand-overlapping gene pairs have overlapping lengths of less than 5000 base pairs. About 60% of gene pairs in human, chimp, and mouse have overlapping lengths within 5000 bps. About 30% of gene pairs in rat and dog have overlapping lengths within 5000 bps. Frequency of overlapping lengths greater than 5000 bps is less than 20% in all the five genomes.

The cumulative distribution of the overlapping lengths reveals an interesting dichotomy as shown in Figure 5.2.

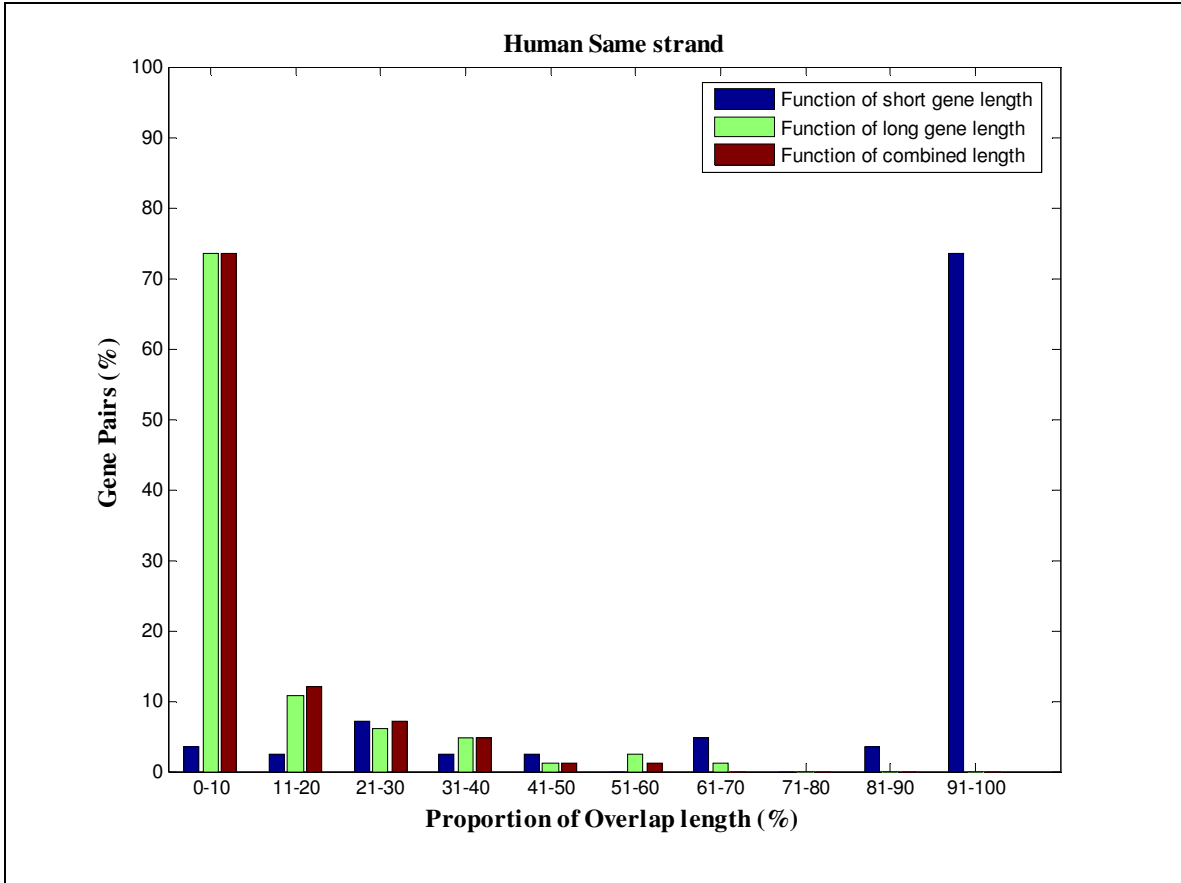


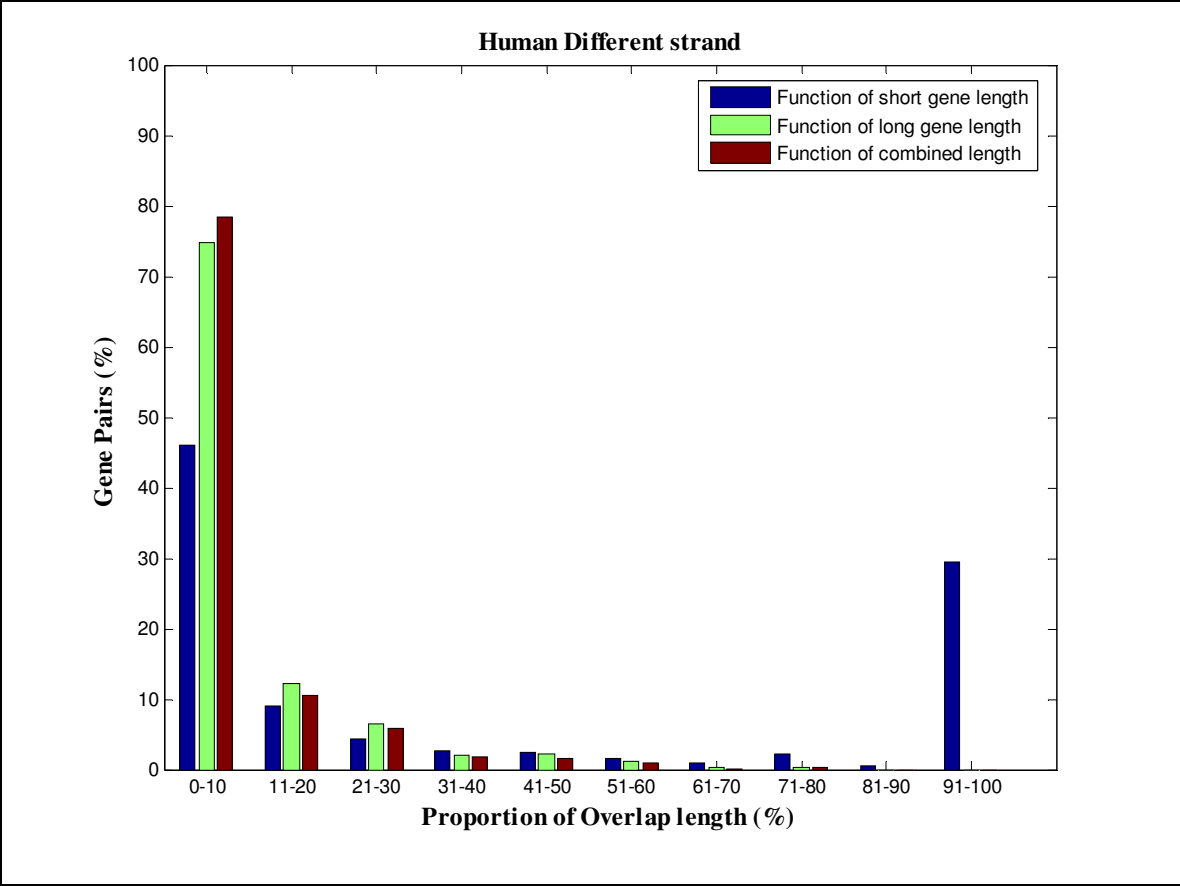
**Figure 5. 2 The Cumulative distribution of overlapping lengths**

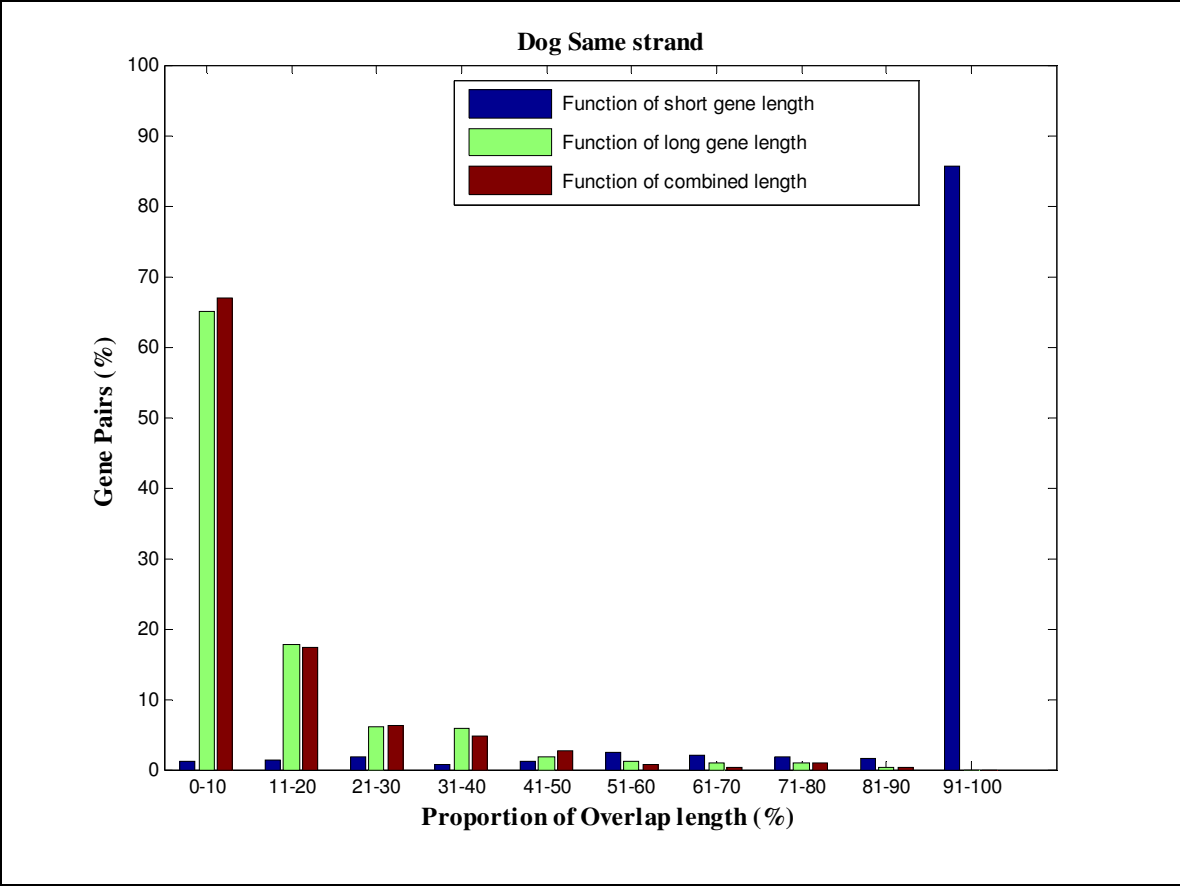
For human, chimp, and mouse, the overlapping lengths of different-strand-overlapping genes tend to be smaller than those of same-strand-overlapping genes, with about 50% of the overlapping lengths shorter than 2000 bps, in contrast to only about 10% for the same-strand-overlapping genes. For rat and dog, the distributions of the overlapping lengths are similar for the two types of overlapping genes.

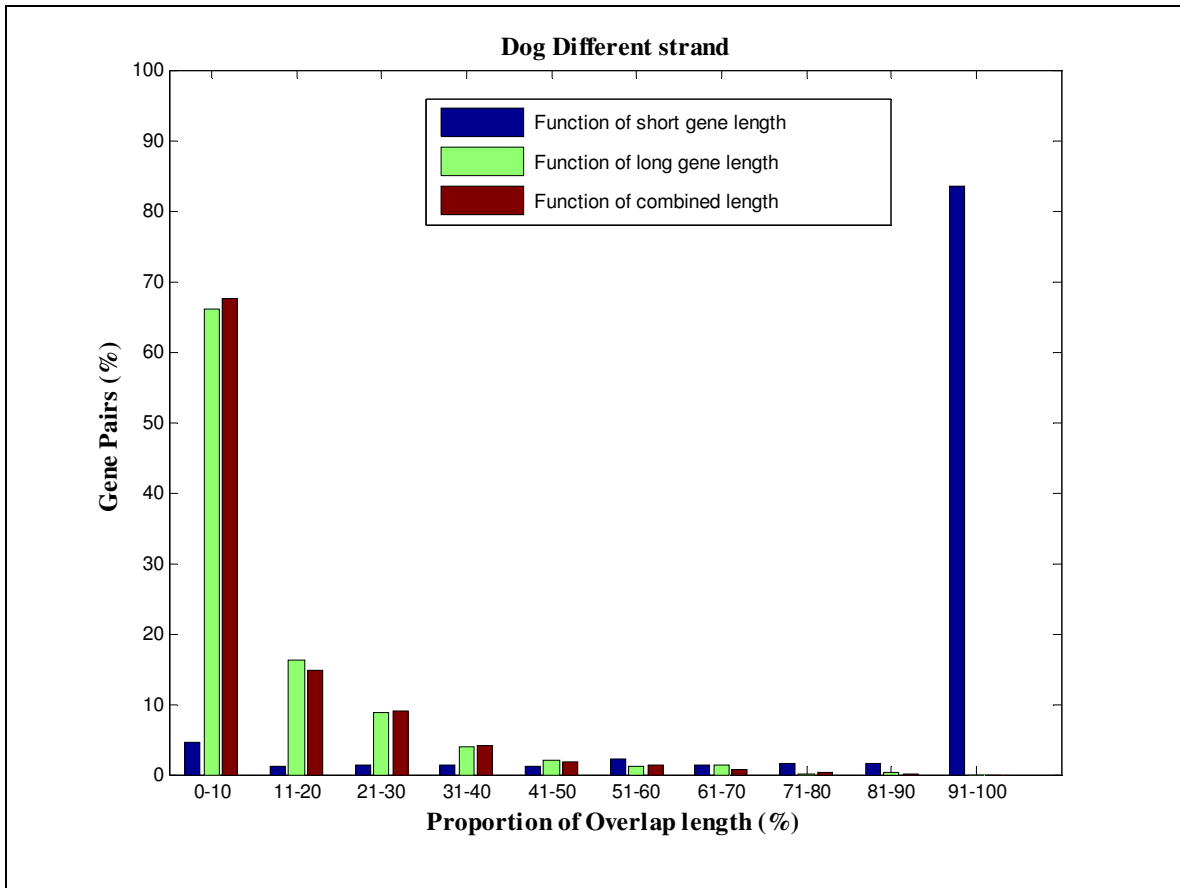
The lengths of overlapping regions exhibit great variation. However, how does this variation compare to the variation in gene lengths? Is there any difference between the two types of overlaps? Therefore three degrees of overlap are computed; the percentages of the overlapping lengths as a function of the lengths of the short gene (i.e.

the shorter one in the overlapping pair), the long gene (i.e. the longer one in the overlapping pair), and the entire region spanned by both genes. The distributions of the three degrees of overlap in human and dog were shown in Figure 5.3 as examples (refer appendix A.1 for chimp, mouse, and rat).









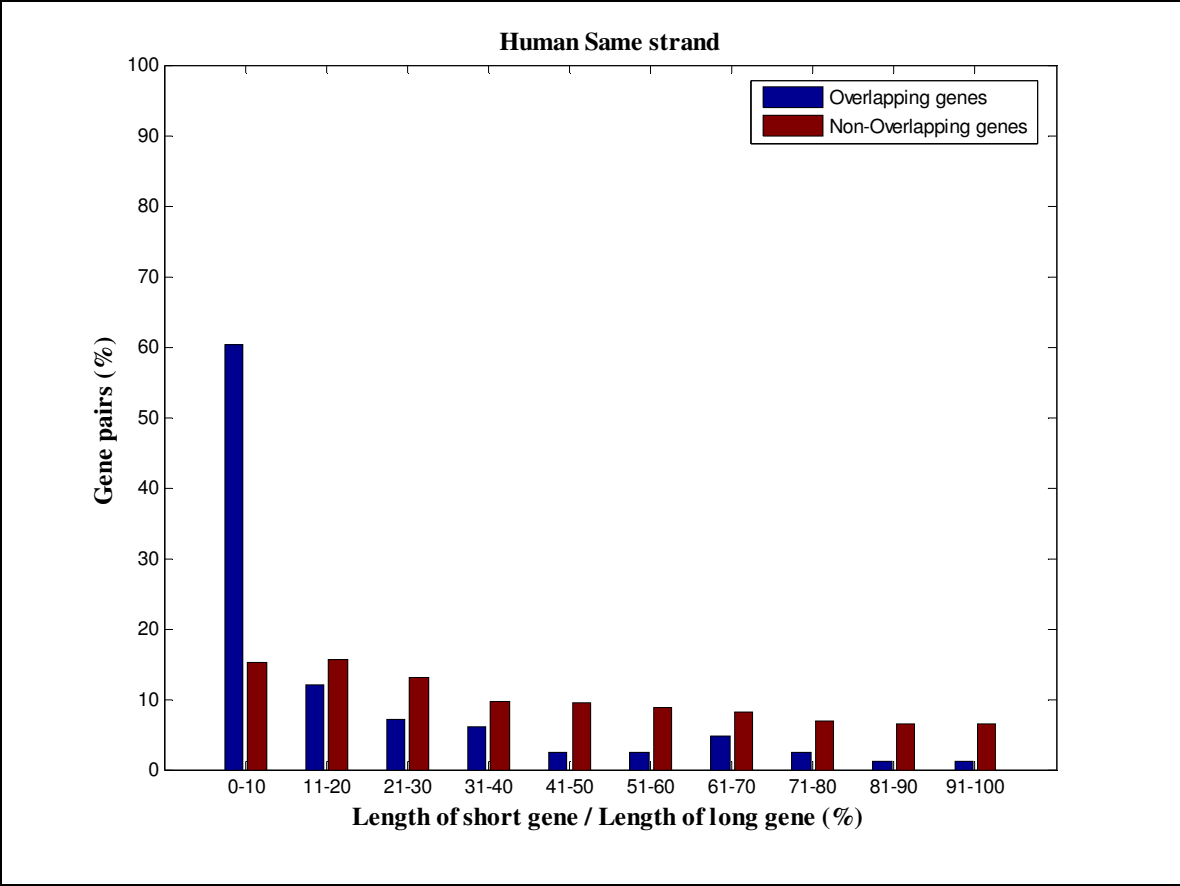
**Figure 5.3 The distribution of the proportion of the overlapping lengths with respect to the sizes of the short genes, long genes, and the regions spanned by both genes.**

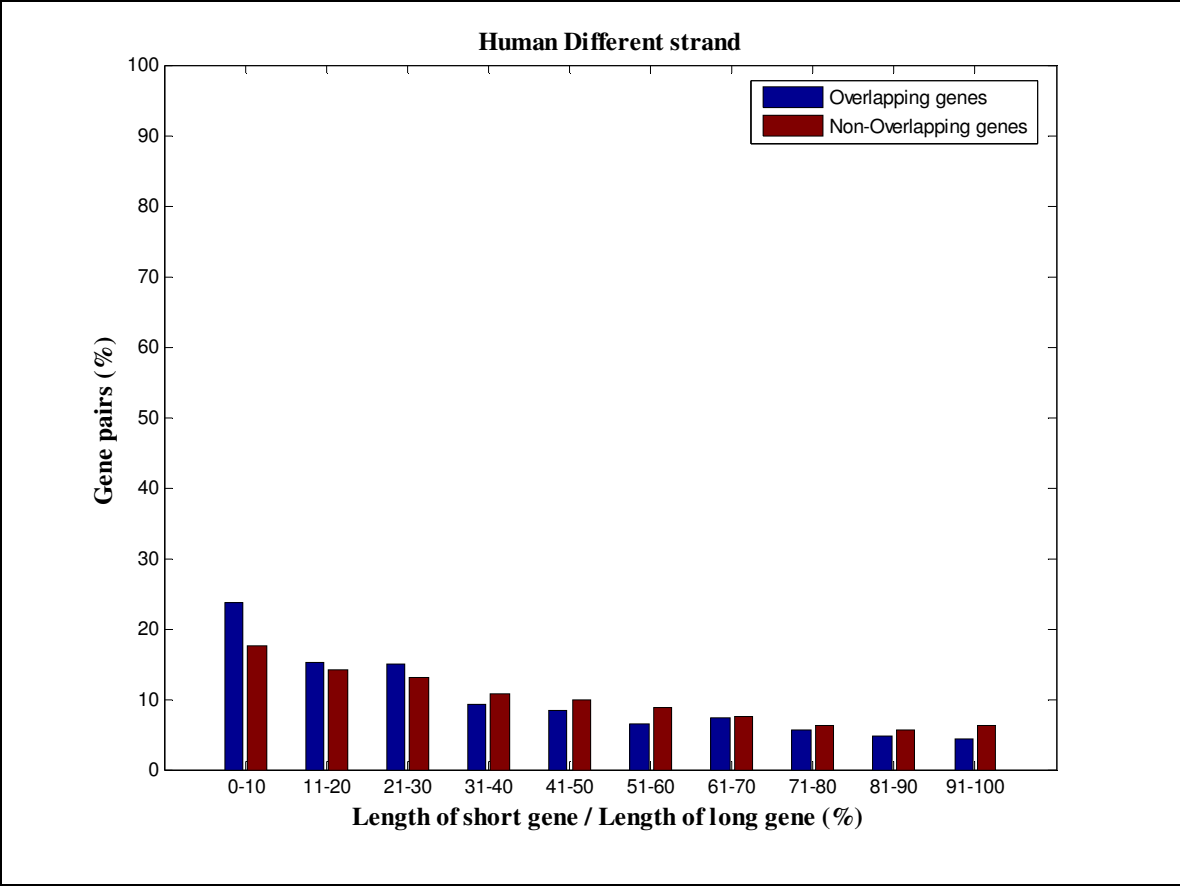
For the same-strand overlaps, all species show similar distributions for the three degrees of overlap (Figure 5.3 and appendix A.1). For the majority of overlapping genes, overlapping regions account for 90-100% of short genes, consistent with the observation that the majority of same-strand overlaps are embedded forms (Table 5.2). Furthermore, for the majority of overlapping genes, overlapping regions account for less than 10% of the long genes. The distribution of overlapping lengths as a function of the lengths of combined gene regions follows closely to that for long genes, suggesting that for the majority of same-strand-overlapping genes, lengths of the short genes are negligible when compared to long genes.

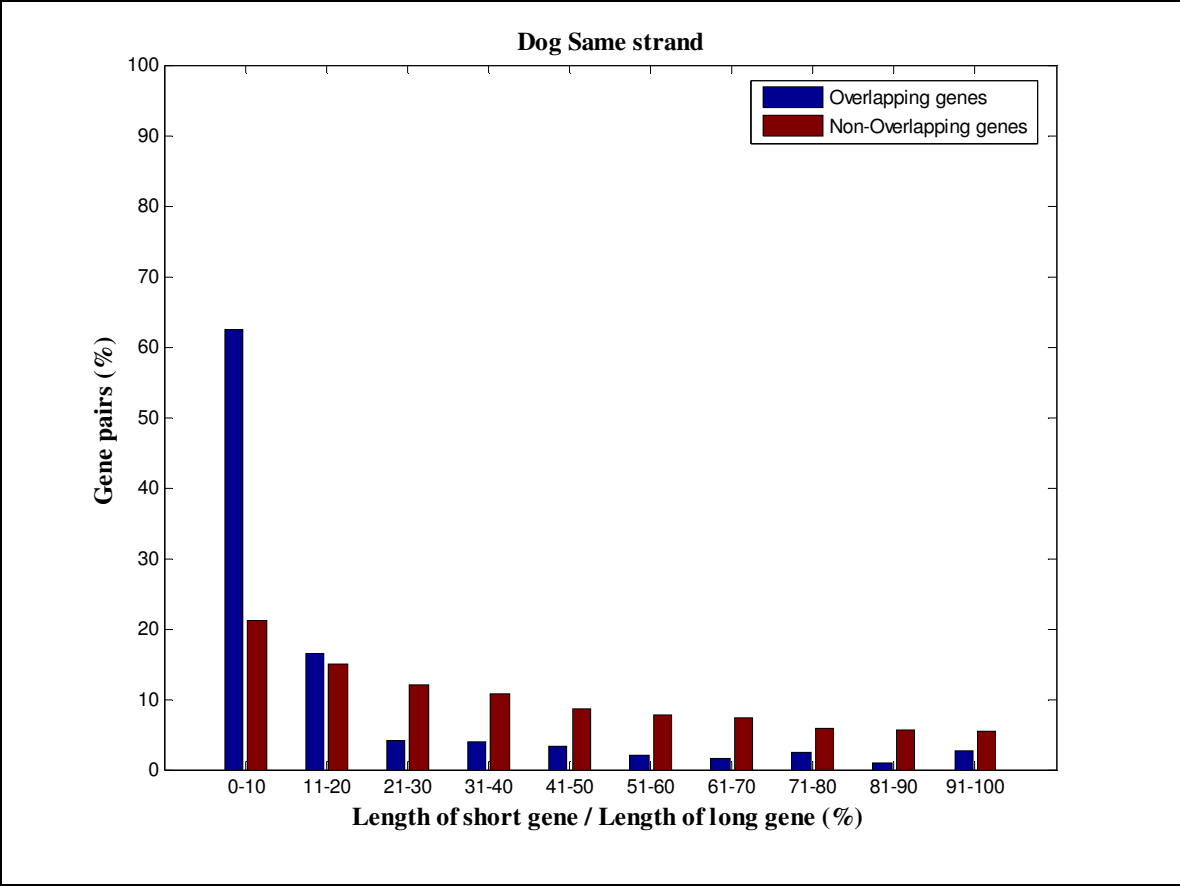
For different-strand-overlapping genes, we observed another dichotomy with human, chimp, and mouse showing one pattern, and rat and dog another (Figure 5.3 and

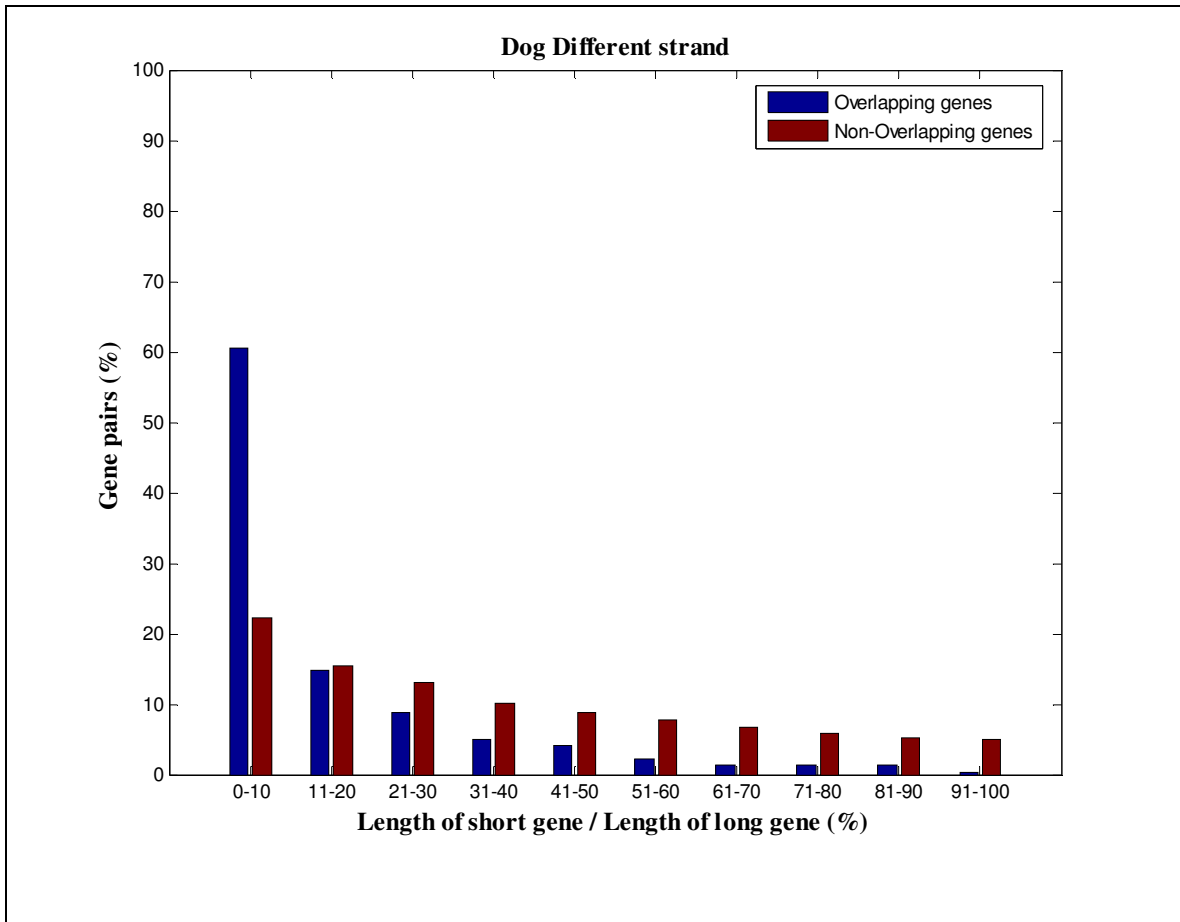
appendix A.1). In human, chimp, and mouse, distribution of overlapping lengths as a function of the short gene lengths shows two peaks. One peak contains about 37-46% of the cases with the overlapping regions accounting for less than 10% of the lengths of the short genes; the other contains about 30-34% of the cases with the overlapping regions accounting for 90-100% of the lengths of the short genes. In rat and dog, the distribution of the overlapping lengths as a function of the short gene lengths is similar to that for the same-strand-overlapping genes, consistent with the observation that the majority of the overlapping genes on different strands are embedded forms in these species (Table 5.2). For all species, the distributions of the overlapping lengths as a function of the lengths of the long genes and the regions spanned by overlapping genes are very similar.

The observation of the overlapping lengths with respect to the short and long genes led us to think that there might be a huge difference between the lengths of the short vs. long genes. To address this issue, we calculated the ratios of the lengths of the short genes vs. the long genes for all the overlapping gene pairs for the two types of the overlapping genes and compared them with the corresponding genome pattern for each type. The genome pattern was obtained by calculating the ratio of the lengths of the neighboring but not overlapping genes (again, short vs. long genes) either on the same or different strand in the sample of 11197 genes. Figure 5.4 shows the distributions in human and dog (refer appendix A.2 for chimp, mouse, and rat).









**Figure 5. 4 The distribution of the ratios of the lengths of short vs. long genes for the same and different strand overlaps in comparison with the one for the neighboring and non-overlapping genes in Human and Dog.**

In human, chimp, and mouse, for the same-strand-overlapping genes, the ratios of lengths of short vs. long genes are mostly less than 10%, as compared to the rather uniform distribution of the ratios for the genome pattern. In the case of different-strand-overlapping genes, the distribution of the overlapping genes is generally consistent with that of the neighboring but non-overlapping genes with a small peak in the bin of 0-10%.

In rat and dog, the patterns are similar for same-strand and different-strand overlaps: compared to genes that are neighboring but non-overlapping, the distribution of the ratios of the lengths of the short genes vs. the long genes is highly skewed towards the bin of 0-10%.

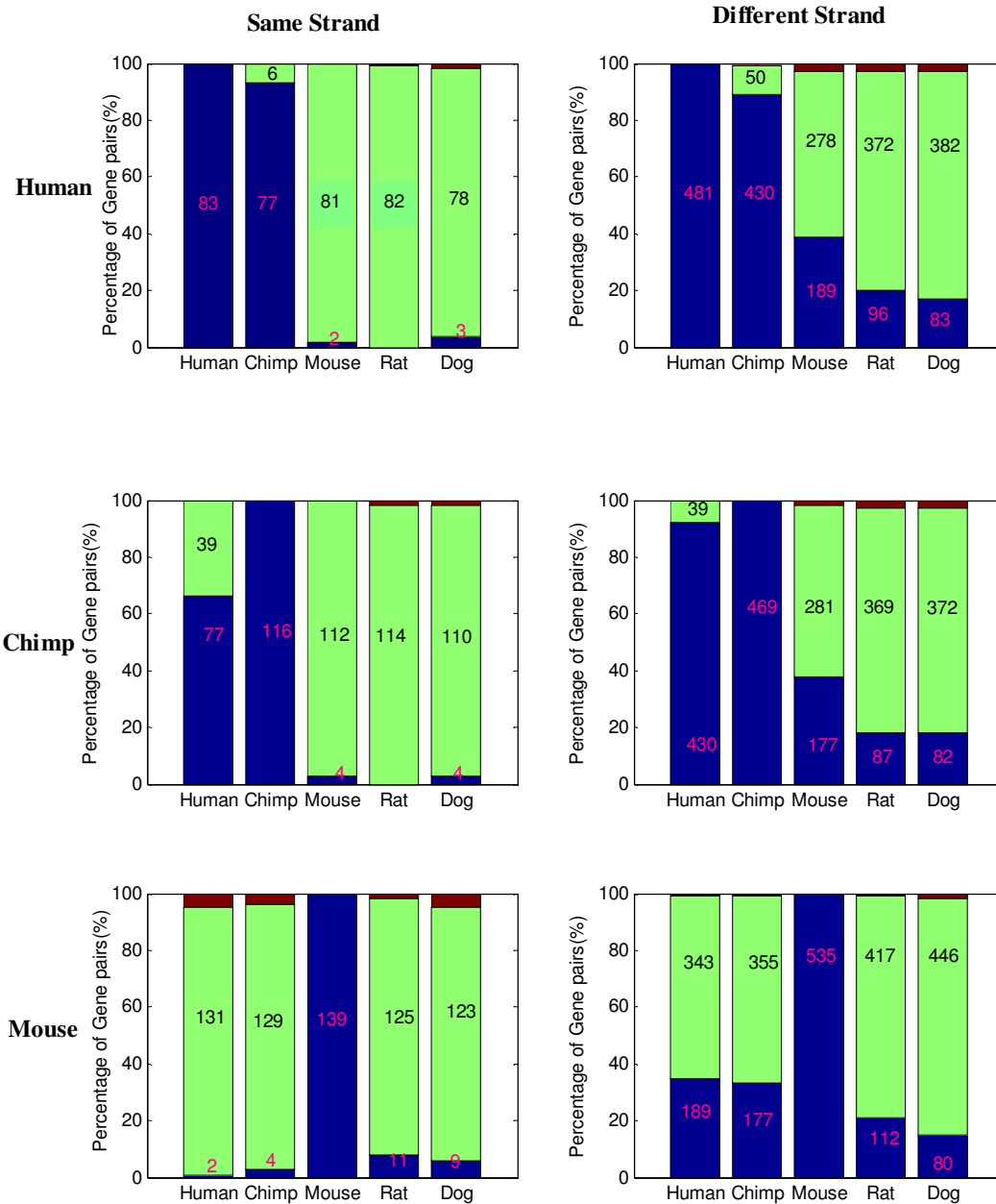
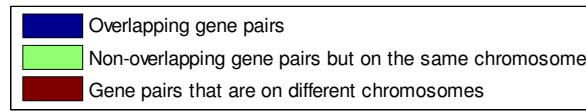
Because various statistics on overlapping genes in the current study is largely consistent with previous studies for human and mouse, the patterns shown here are most likely real for the two species. In rat and dog, most of the different-strand-overlapping genes are embedded forms, similar to the same-strand-overlapping genes, thus showing that these two types of overlapping genes exhibit similar patterns. Given the fact that gene annotation in rat and dog has low quality due to the rather limited large-scale data on full-length cDNAs and ESTs and the fact that mouse and rat diverged only about 40 million years ago and comparison of the two genomes show no systematic changes between them [Rat Genome Sequencing Consortium, 2004], we believe that the patterns observed in rat and dog, especially for different-strand-overlapping genes, is an artifact of gene annotation in the two species.

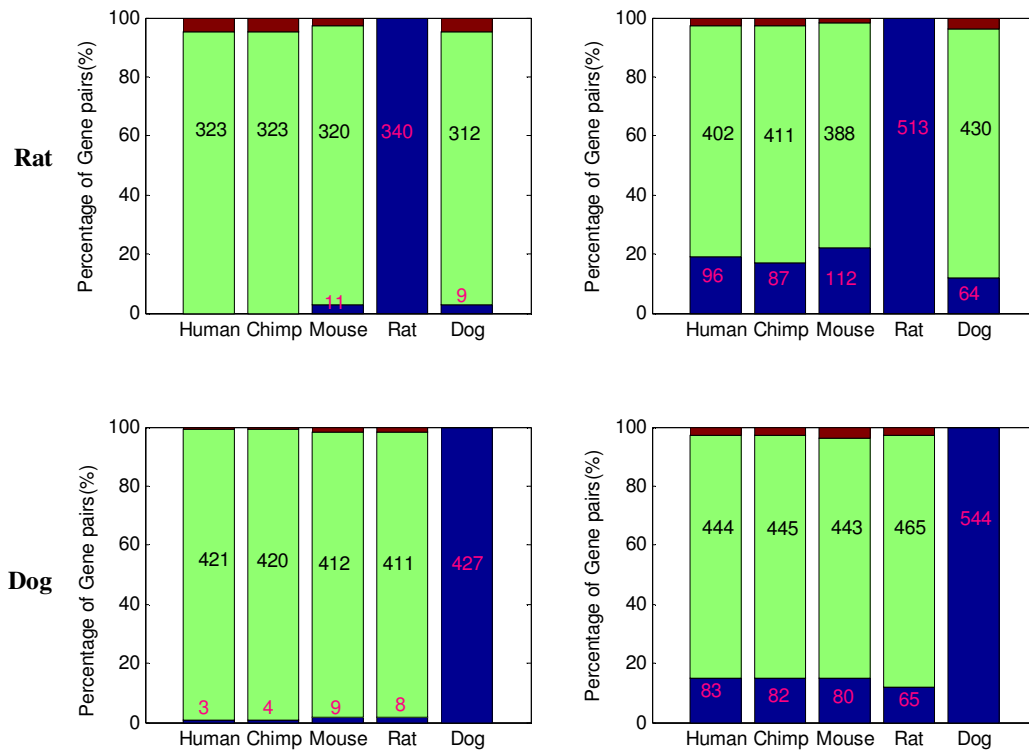
### **5.3 The birth/death of overlapping genes**

We looked to see if overlapping gene pair relationships in one species were conserved across the other four species with the hope of gaining some kind of insight into probable causes of birth and death among overlapping genes.

So for each set (same-strand and different-strand) of overlapping gene pairs in each species, we checked their corresponding orthologous gene pairs in the other four species and got a count of the following: number of orthologous gene pairs that overlap, number of orthologous gene pairs that do not overlap but occur on the same chromosome and the number of orthologous gene pairs that do not overlap because they do not even occur on the same chromosome.

The results are illustrated in figure 5.4, which shows the percentage of overlapping genes (blue), percentage of non-overlapping genes that occur atleast on the same chromosome (green), and percentage of genes that occur on different chromosomes altogether (maroon). For all the overlapping genes of the reference species (solid blue bar), the other four bars show the percentages of overlapping gene pairs, non-overlapping gene pairs that are at least present on the same chromosome, and gene pairs that occur on different chromosomes in the rest of the four species.





**Figure 5. 5 Conservation of overlapping relationships in the five species**

We observed that for the same-strand-overlapping genes, mostly the orthologous genes of overlapping gene pairs in one species seldom overlap in another species, showing little conservation across different species except human and chimp due to their recent species split. The percentages of conserved overlaps range from 0%-8% for the same-strand-overlapping genes, excluding the human and chimp comparison. In contrast, the percentages of conserved overlaps range from 12%-41% for the different-strand-overlapping genes. Therefore, overlaps of opposite strands tend to be more evolutionarily conserved than those of the same strand. Similar observation was made in a comparison of overlapping genes in human and fugu where about 23.3% of the consecutive relationships were found to be conserved between human and fugu whereas only 13.5% of the same strand gene pairs were found conserved between the two organisms [Dahary et al., 2005].

Interestingly, although the orthologous gene pairs are no longer overlapping in the other species, they remain mostly on the same chromosome (Table 5.3). In fact, in 94%-100% of the cases, the genes either show conserved overlapping relationships or are non-overlapping but on the same chromosome for all the species.

Veeramachaneni et al. [2004] examined 255 cases of different strand overlaps where both overlapping genes in human have mouse orthologous genes and found that in 95 cases (37%), genes were also overlapping in mouse, in 150 cases (59%), genes do not overlap but appear to be on the same chromosome in mouse, and the remaining are on different chromosomes in mouse. Hence, in about 96% cases, mouse shows a conserved chromosome arrangement as those of human genes, similar to our results of 97% on human and mouse comparisons (Table 5.3). Similarly, when using mouse as the reference species, they found that of the 240 cases of overlapping genes in mouse, 39% also overlap in human and about 99% have conserved chromosomal arrangement, consistent with the corresponding percentages (35% and 99%) obtained in the current study (Table 5.3).

The observation that two genes are overlapping in one species but non-overlapping and neighboring on the same chromosome in other species indicates the highly dynamic evolution of either 5'-UTRs or 3'-UTRs. There are two possible scenarios explaining the observation: one is that two ancestral genes did not overlap originally and became overlapping later; the other is that two ancestral genes overlapped initially and became non-overlapping later. We cannot decide which scenario is applicable to the genes in the current study because there are no estimates on the evolutionary rates of birth and death of 5' and 3'-UTRs.

Nonetheless, we were able to computationally determine causes for transitions between non-overlapping and overlapping for the embedded-same-strand-overlapping genes. All five species show a similar pattern: the majority of the transitions (about 57%-63%) were due to 3' end extension and the remaining due to 5' end extension as shown in Table 5.3.

**Table 5. 3 Causes for transition between non-overlapping and overlapping genes for same strand embedded gene pairs**

Species	Total no. of embedded gene pairs	5' extension	3' extension
Human	56	22 (39.3%)	34 (60.7%)
Chimp	70	26 (37.1%)	44 (62.9%)
Mouse	74	28 (37.8%)	46 (62.2%)
Rat	204	81 (39.7%)	123 (60.3%)
Dog	332	142 (42.8%)	190 (57.2%)

Several previous studies examined the origin of overlapping genes in eukaryotes. For example, Shintani et al. [1999] studied the evolutionary origin of TCP1 and ACAT2 that are overlapping in their 3'-UTR on different strands and suggested that the overlap arose during transitions from therapsid reptiles to mammals and has been retained for more than two hundred million years. They proposed that the two genes were brought together and became overlapping during the chromosomal rearrangement. Dan et al. [2002] examined the origin of the overlapping genes of Mink and Chrne and found that the two genes overlap in some mammals but not others owing to different usages of alternative polyadenylation sites.

However, apart from the studies that examined individual cases of overlapping genes in eukaryotes, there have not been any systematic investigations of the causes for the transitions between overlapping and non-overlapping genes, especially for the same strand overlapping genes. Here, our results on 736 pairs of overlapping genes in the four mammalian genomes indicate that 3'-UTR evolution plays a predominant role in the transition. This seems to be consistent with the finding that at least half of all human genes encode multiple transcripts with alternative 3' termini [Iseli et al., 2002]. The higher frequency of transition contributed by 3'-UTR changes than 5'-UTR changes implies that it is easier to capture and employ a downstream alternative termination signal than an upstream alternative start signal. However, this could simply be due to the bias in 3'-UTR annotations.

With increasing empirical evidence for gene structures, we might discover that many genes have multiple alternative 5' or 3'-UTRs for transcription and become overlapping with each other in certain transcribed forms and it is likely that overlapping genes are a rule rather than exception in many eukaryotic genomes.

## Chapter 6

### Conclusions

It can confidently be concluded that overlapping genes are a common phenomenon among eukaryotic genomes because of increasing evidence that support their occurrence. Interestingly, there are also cases in which some large genes overlap with multiple genes (section 5.1). With respect to frequencies of occurrence of the two types of overlapping genes, different strand overlapping genes are more common than same strand overlapping genes in eukaryotic genomes (section 5.1). The most prevalent type of overlap that occurs in different strand overlapping genes is when their 3' UTR regions overlap, and in same strand overlapping genes is when one gene is embedded in the other.

In human, chimp, and mouse, the overlapping lengths of different-strand-overlapping genes tend to be smaller than those of the same-strand-overlapping genes. For rat and dog, the distributions of overlapping lengths are similar for the two types of overlapping genes. The percentages of overlapping lengths with respect to the short gene, long gene, and the region spanned by both the regions show similar distributions for same-strand-overlapping genes in all species. The percentages of overlapping lengths with respect to the short gene, long gene, and the region spanned by both the regions for different-strand-overlapping genes shows a dichotomy with human, chimp, and mouse showing one pattern and rat and dog another.

Different-strand overlaps tend to be more evolutionarily conserved when compared to same-strand-overlapping genes. Although orthologous gene pairs may not always overlap in the other species, they at least remain on the same chromosome in majority of the cases. The transitions from non-overlapping to overlapping are mostly caused due to 3' extensions in all the five species.

## **Chapter 7**

### **Future Work**

#### **7.1 Origin**

So far, we have seen several gene overlap cases identified in both prokaryotic and eukaryotic genomes. Considerable number of studies reported the occurrence of overlapping genes but very few of them actually claim to have identified the origin of overlapping genes. For example a phylogenetic approach was used to clarify the origin of overlapping genes in viruses like tymoviruses, luteoviruses, lentiviruses and paramyxoviruses [Keese and Gibbs, 1992; Jordan et al., 2000]. Another recent study that investigated the origin and evolution of overlapping genes was in bacteriophages that belonged to the family Microviridae [Pavesi, 2006]. However, there has not been much research done to establish the exact causes or origins of such overlaps in large scales especially in eukaryotic genomes. Hence, this could be one of the future directions.

#### **7.2 Biological Significance**

Even though there have been reports that studied specific pairs of overlapping genes [Nekrutenko and Wadhawan, 2005; Knee et al., 1994; Bristow et al., 1993] in detail and explained their biological significance, there is still one basic question that is unanswered with respect to overlapping gene pairs in general: is there an exact biological significance of this occurrence that explains its significance in all genomes? Of course, some studies have shown that there exists a regulatory relationship between genes involved in sense-antisense overlap. However, it is still unclear whether the presence of regulatory mechanisms is important and applicable to all overlapping gene pairs in general. Studies have shown that antisense RNA may alter gene regulation leading to various pathologies [Vanhee-Brossollet and Vaquero, 1998]. Hence, this provides another

avenue for more research to question the biological significance of the occurrence of overlapping genes in various genomes and comprehend the pathological situations.

### **7.3 In Plants**

There have been many reports as we have seen regarding overlapping genes in prokaryotes and mammals, but not much is known about their occurrence in plants [Terryn and Rouze, 2000]. A few cases of overlapping genes have been reported in the model organism *Arabidopsis* [Quesada et al., 1999; Ito et al., 1997; Glover et al., 1998; Wang et al., 2005], maize [Schmitz and Theres, 1992; Joanin et al., 1997; Ansaldi et al., 2000] and rice [Osato et al., 2003,]. However, they have not been studied with the same depth used in studies of bacterial and viral genomes. For example, though a lot studies have been conducted reporting human diseases that involve overlapping genes, there are no reports of plant diseases that may have involved overlapping genes. Thus, there is plenty of room for active research regarding overlapping genes among plants.

### **7.4 Genome-wide study**

A somewhat comprehensive report of overlapping genes is available for higher eukaryotes [Boi et al., 2004], bacterial genomes [Fukuda et al., 2003], and vertebrate genomes [Makalowska et al., 2005]. However, such reviews are missing for plants or flies. The main reason for this is that genome-wide studies for most plants have been restricted to one each in *Arabidopsis* [Wang et al., 2005] and rice [Osato et al., 2003]. In addition, both these studies looked at antisense transcripts thereby leaving out any overlapping genes that might occur on the same strand. Hence, though there have been reports discussing the individual overlap cases in some plants and fruit flies, a more comprehensive genome-wide study is required in order to learn more about their origin, evolution, biological significance etc.

# Bibliography

- Adelman JP, Bond CT, Douglass J, Herbert E (1987) Two mammalian genes transcribed from opposite strands of the same DNA locus, *Science* 235, 1514-1517.
- Alfano G, Vitiello C, Caccioppoli C, Caramico T, Carola A, Szego MJ, McInnes RR, Auricchio A, Banfi S (2005) Natural Antisense transcripts associated with genes involved in eye development. *Human Molecular Genetics* 14 (7), 913-923.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *J Mol Biol* 215, 403-410.
- Ansaldi R et al (2000) Multiple S gene family members including natural antisense transcripts are differentially expressed during development of maize flowers. *J. Biol. Chem.* 275, 24146-24155.
- Barrell BG, Air GM, Hutchison CA (1976) Overlapping genes in bacteriophage phiX174. *Nature* 264, 34-41.
- Batshake B, Sundelin J. (1996) The mouse genes for the EP1 prostanoid receptor and the PKN protein kinase overlap. *Biochem. Biophys. Res. Commun* 227, 70-76.
- Billy E, Brondani V, Zhang H, Muller U, Filipowicz W (2001) Specific interference with gene expression induced by long, double stranded RNA in mouse embryonal teratocarcinoma cell lines. *Proc. Natl. Acad Sci USA* 98, 14428-14433.
- Boi S, Solda G, Tenchini ML (2004) Shedding Light on the Dark Side of the Genome: Overlapping Genes in Higher Eukaryotes. *Current Genomics* 5, 509-524.
- Bristow J, Tee MK, Gitelman SE, Mellon SH, Miller WL (1993) Tenascin-X: A Novel extracellular matrix protein encoded by the human XB gene overlapping P450c21B, *Cell Biology* 122, 265-278.

- Cawthon et al. (1991) cDNA sequence and genomic structure of EV12B, a gene lying within an intron of the neurofibromatosis type 1 gene. *Genomics* 446-60.
- Chang YF, Chang CH. Identification and Characterization of Conserved Overlapping Genes in *Vibrio* Genomes.
- Chisholm W, Johnson ZI (2004) Properties of overlapping genes are conserved across microbial genomes. *Genome Research* 14, 2268-2272.
- Clark MA, Baumann L, Thao ML, Moran NA, Baumann P (2001) Degenerative minimalism in the genome of a psyllid endosymbiont. *J. Bacteriol.* 183, 1853-1861.
- Cooper et al. (1998) Divergently transcribed overlapping genes expressed in liver and kidney and located in the 11p15.5 imprinted domain. *Genomics* 49, 38-51.
- Dahary D, Stein OE, Sorek R (2005) Naturally occurring antisense: Transcriptional leakage or real overlap? *Genome Research* 15, 364.
- Dan I, Watanabe NM, Kajikawa E, Ishida T, Pandey A, Kusumi A (2002) Overlapping of MINK and CHRNE gene loci in the course of mammalian evolution. *Nucleic Acid Research* 30 (13), 2906-2910.
- Eyre-Walker A (1995) The Distance between *Escherichia coli* genes is related to gene expression levels. *J. Bacteriol.* 177, 5368-5369.
- Fahey ME, Moore TF, Higgins DG (2002) Overlapping antisense transcripts in the human genome. *Comp. Funct. Genomics* 3, 244-53.
- Farrell CM, Lukens LN (1995) Naturally occurring antisense transcripts are present in chick embryo chondrocytes simultaneously with the down-regulation of the alpha 1 (I) collagen gene. *J. Biol. Chem* 270, 3400-3408.
- Fukuda Y, Washio T, Tomita M (1999) Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Nucleic Acids Res.* 27, 1847-1853.
- Fukuda Y, Nakayama Y, Tomita M (2003) On dynamics of overlapping genes in bacterial genomes. *Gene* 323: 181-187.
- Glover et al (1998). Cloning and characterization of MS5 from *Arabidopsis*: a gene critical in male meiosis. *Plant J.* 15, 345-356.

- Grasse P. (1977) Evolution of living organisms: Evidence for a new theory of Transformation. Academic Press, New York, 297.
- Hartl DL, Jones EW (2002) Essential Genetics: A Genomics Perspective, Jones and Bartlett Publishers, Inc.
- Henikoff S, Keene MA, Fechtel K, Fristom JW (1986) Gene within a gene: nested Drosophila genes encode unrelated proteins on opposite strands DNA strands. Cell 44, 33-42.
- Inokuchi Y, Hirashima A, Sekine Y, Janosi L, Kaji A (2000) Role of ribosome recycling factor (RRF) in translational coupling. EMBO J. 19, 3788-3798.
- Iseli et al. (2002) Long-range heterogeneity at the 3' ends of human mRNAs. Genome Research 12, 1068-1074.
- Ito et al (1997). A serine/threonine protein kinase gene isolated by an in vivo binding procedure using the Arabidopsis floral homeotic gene product, AGAMOUS. Plant Cell Physiol. 38, 248-258.
- Joanin P et al (1997) Sense and antisense transcripts of the maize MuDR regulatory transposon localized by in situ hybridization. Plant Mol. Biol. 33, 23-36.
- Jordan IK, Sutter BA IV and McClure MA (2000) Molecular evolution of the paramyxoviridae and rhabdoviridae multiple-protein-encoding P gene. Mol Biol Evol 17, 75-86.
- Karlin S, Chen C, Gentles AJ, Cleary M (2002) Associations between human disease genes and overlapping gene groups and multiple amino acid runs. Proc. Natl. Acad. Sci. U.S.A. 99, 17008-17013.
- Keese PK and Gibbs A (1992) Origin of genes: "big bang" or continuous creation? Proc Natl Acad Sci USA 89, 9489-9493.
- Kennerson ML, Nassif NT, Dawkins JL, DeKroon RM, Yang JG, Nicholson GA (1997) The Charcot-Marie-Tooth binary repeat contains a gene transcribed from the opposite strand of a partially duplicated region of the COX10 gene. Genomics 46, 61-69.
- Kimelman D, Kirschner MW (1989) An antisense mRNA directs the covalent modification of the transcript encoding fibroblast growth factor in Xenopus oocytes. Cell 59, 687-696.

- Knee RS, Pitcher SE, Murphy PR (1994) Basic Fibroblast growth factor sense (FGF) and antisense (gfg) RNA transcripts are expressed in unfertilized human oocytes and in differentiated adult tissues. *Biochem. Biophys. Res. Commun.* 205, 577-583.
- Kolata GB (1977) Overlapping genes: More than Anomalies? *Science* 196 (4295), 1187-1188.
- Krakauer DC (2000) Stability and Evolution of Overlapping genes. *Evolution* 54 (3), 731- 739.
- Kumar M, Carmichael GG (1997) Nuclear antisense RNA induces extensive adenosine modifications and nuclear retentions of target transcripts. *Proc Natl Acad Sci USA* 94, 3542-3547.
- Lavorgna G, Dahary D, Lehner B, Sorek R, Sanderson CM, Casari G (2004) In search of Antisense. *Trends in Biochemical Sciences* 29 (2), 88-94.
- Lee JT, Davidow LS, Warshawsky D (1999). Tsix a gene antisense to Xist at the X-inactivation center. *Nat Genet.* 21, 400-404.
- Lehner B, Williams G, Campbell RD, Sanderson CM (2002) Antisense transcripts in the Human genome. *Trends in Genetics* 18, 63-65.
- Levinson B, Kenwick S, Lakich D, Hammonds Jr G, Gitschier J (1990) A transcribed gene in an intron of the human factor VIII gene. *Genomics* 7, 1-11.
- Long M, Betran E, Thornton K, Wang W (2003) The Origin of New Genes: Glimpses from the Young and Old. *Nature* 4, 865-875.
- Makalowski I, Lin CF, Makalowski W (2005) Overlapping genes in vertebrate genomes. *Computational Biology and Chemistry* 29, 1-12.
- Malavasic MJ, Elder RT (1990). Complementary transcripts from two genes necessary for normal meiosis in the yeast *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 10, 2809-2819.
- Mao JR, Taylor G, Dean WB, Wagner DR, Afzal V, Lotz JC, Rubin EM, Bristow J (2002) Tenascin -X deficiency mimics Ehlers-Danlos Syndrome in mice through alteration of collagen deposition. *Nat Genet.* 30, 421-425.

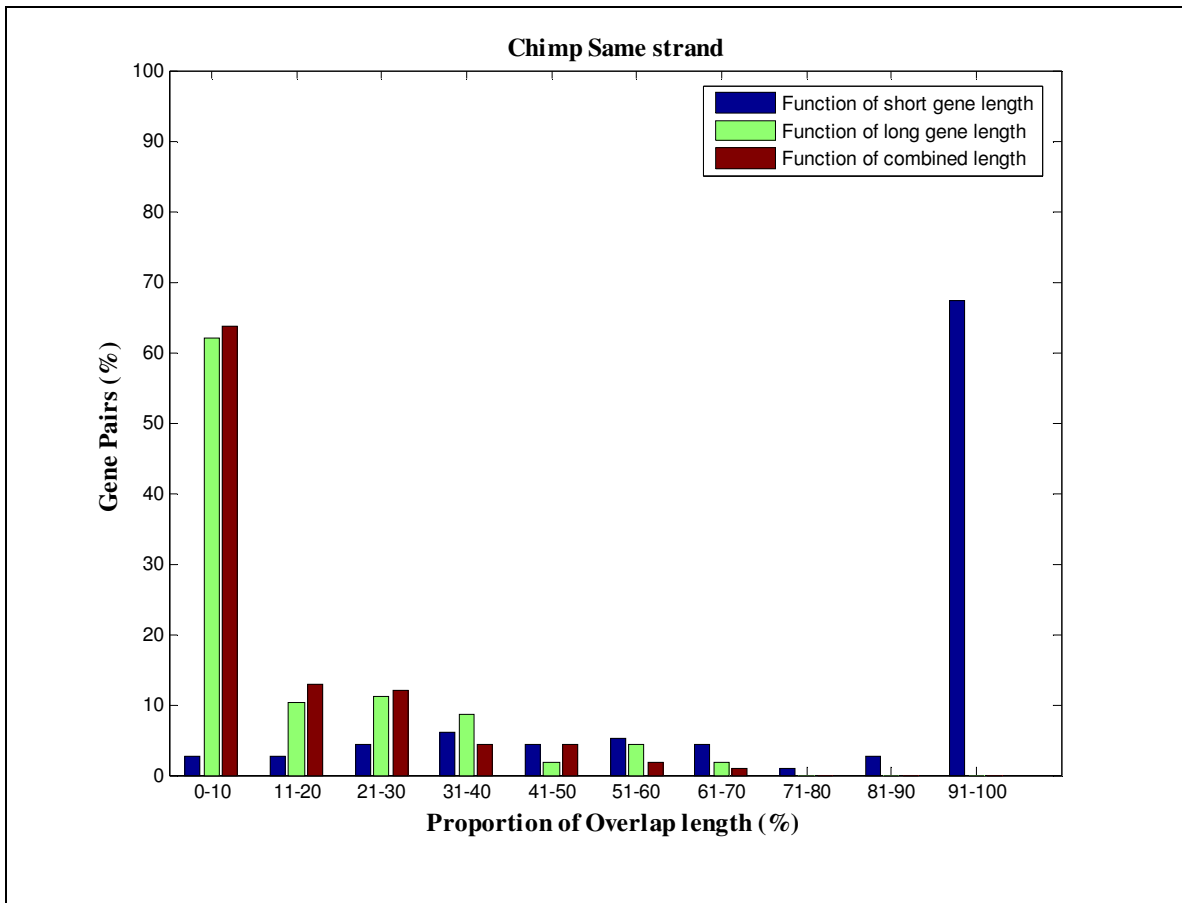
- Marcelino J, Carpten JD, Suwairi WM, Gutierrez OM, Schwartz S, Robbins C, Sood R, Makalowska I, Baxevanis A, Johnstone B, Laxer RM, Zemel L, Kim CA, Herd JK, Ihle J, Williams C, Johnson M, Raman V, Alonso LG, Brunoni D, Gerstein A, Papadopoulos N, Bahabri SA, Trent JM, Warman ML (1999) CACP encoding a secreted proteoglycan, is mutated in camptodactyly-arthropathy-coxa vara-pericarditis syndrome. *Nat Genet.* 23, 319-322.
- Mihalich A, Reina M, Mangioni S, Ponti E, Alberti L, Vigano P, Vignali M, Di Blasio AM (2003) Different basic fibroblast growth factor and fibroblast growth factor-antisense expression in eutopic endometrial stromal cells derived from women with and without endometriosis. *J. clin. Endocrinol. Metab.* 88, 2853-2859.
- Misener SR, Walker VK (2000) Extraordinarily high density of unrelated genes showing overlapping and intraintronic transcription units. *Biochim. Biophys. Acta* 1492, 269-270.
- Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kamimker JS, Millburn GH, Prochnik SE, Smith CD, Tupy JL, Whitfield EJ, Bayraktaroglu L, Berman BP, Bettencourt BR, Celniker SE, de Grey AD, Drysdale RA, Harris NL, Richter J, Russo S, Schroeder AJ, Shu SQ, Stapleton M, Yamada C, Ashburner M, Gelbart WM, Rubin GM, Lewis SE (2002) Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol.* 3, RESEARCH0083.
- Munroe SH, Lazar MA (1991) Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. *J Biol Chem.* 266, 22083-22086.
- Nekrutenko A, Wadhawan S, Goetting-Minesky P, Makova KD (2005) Oscillating Evolution of a Mammalian Locus with Overlapping Reading Frames: An XLas/ALEX Relay. *PLoS Genetics* Vol. 1, No. 2, e18 DOI: 10.1371/journal.pgen.0010018.
- Normark S, Bergstrom S, Edlund T, Grundstrom T, Jaurin B, Lindberg FP, Olsson O (1983) Overlapping genes. *Annu. Rev. Genet.* 17, 499-525.

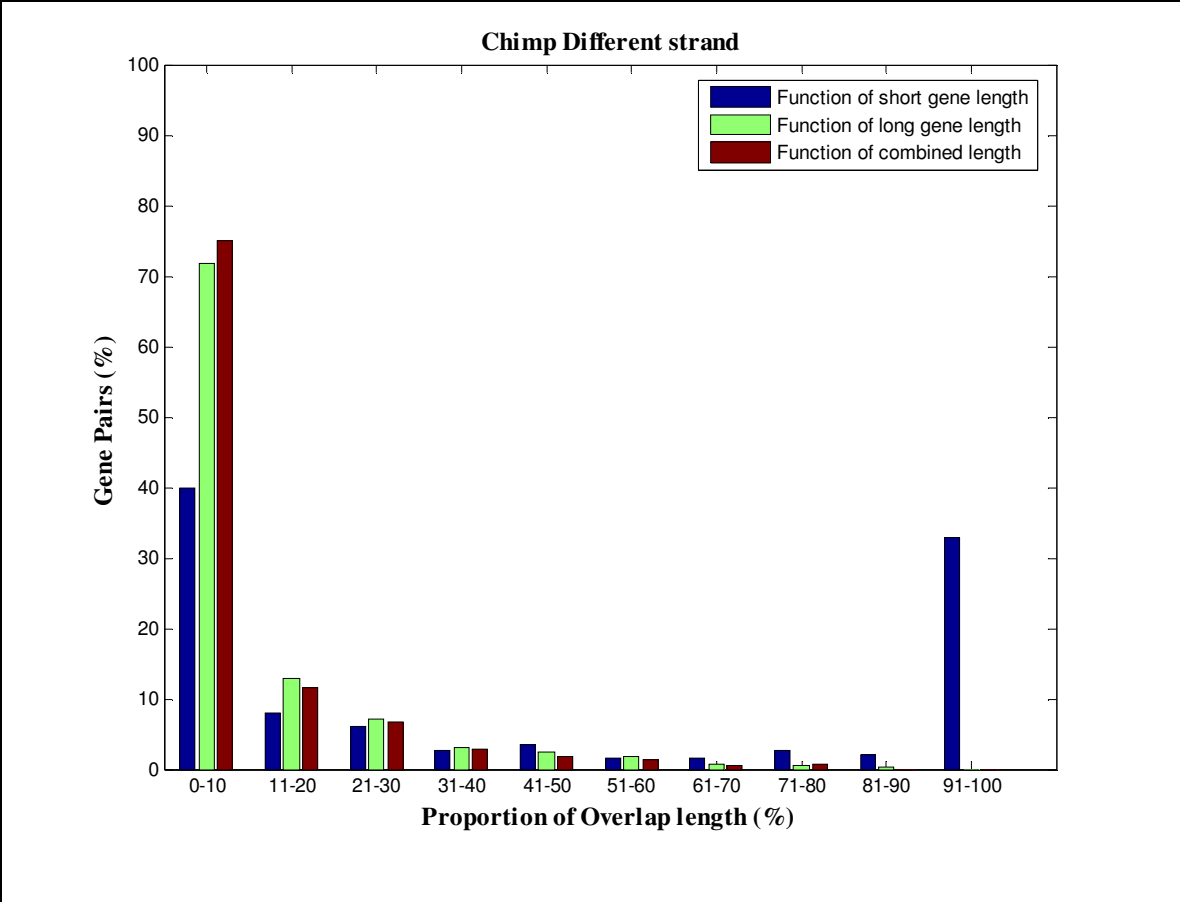
- Osato N, Yamada H, Satoh K, Ooka H, Yamamoto M, Suzuki K, Kawai J, Carninci P, Ohmoto Y, Murakami K, Matsubara K, Kikuchi S, Hayashizaki Y (2003) Antisense transcripts with rice full-length cDNAs. *Genome Biol.* 5, R5.
- Pavesi A (2006) Origin and evolution of overlapping genes in the family Microviridae. *Journal of General Virology* 87, 1013-1017.
- Peters MF, Ross CA (2001) Isolation of a 40 kDa Huntington-associated protein. *J. Biol. Chem.* 276, 3188-3194.
- Peterson JA, Myers AM (1993) Functional Analysis of mRNA 3' end formation signals in the convergent and overlapping transcription units of the *S.cerevisiae* genes RHO1 and MRP2. *Nucl. Acids Res.* 21, 5500-5508.
- Portin P (1993) The concept of the gene: short history and present status. *Q Rev. Biol.* 68, 173-223.
- Quesada V, Ponce MR, Micol JL (1999) OTC and AUL1, two convergent and overlapping genes in the nuclear genome of *Arabidopsis thaliana*. *FEBS Lett.* 461, 101-106.
- Rat Genome Sequencing Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428, 493-521.
- Runte M, Huttenhofer A, Gross S, Kiefmann M, Horsthemke B, Buiting K (2001) The IC-SNURF-SNRPN transcript serves as the host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A. *Hum. Mol. Genet.* 10, 287-2700.
- Schmitz G, theres K (1992) Structural and functional analysis of the Bz2 locus of *Zea mays*: characterization of overlapping transcripts. *Mol. Gen. Genet.* 233, 269-277.
- Schultz RA, Butler BA (1989) Overlapping genes of *Drosophila melanogaster*: organization of the z600-gonadal-Eip28/29 gene cluster. *Genes Dev* 3(2), 232-42.
- Schultz RA, Shlomchik W, Cherbas L, cherbas P (1989) Diverse Expression of overlapping genes: the *Drosophila* Eip28/29 gene and its upstream neighbors.

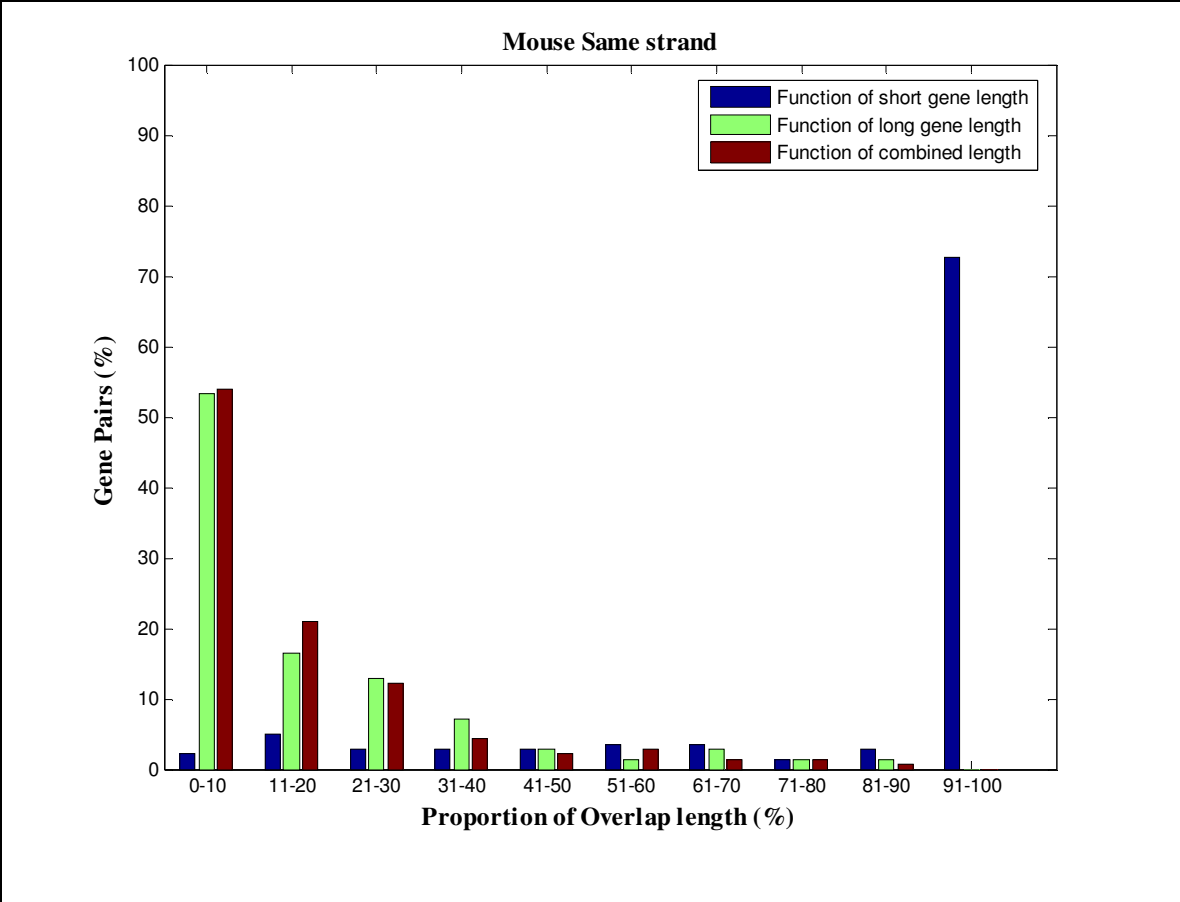
- Schulze SR, Curio-Penny B, Li Y, Imani RA, Rydberg L, Geyer PK, Wallrath LL (2005). Molecular genetic analysis of the nested *Drosophila melanogaster* Lamin C gene. *Genetics* 171, 185-196.
- Shaw DC, Walker JE, Northrop FD, Barrell BG, Godson GN, Fiddes JC, Gene K (1978) A new overlapping gene in bacteriophage G4. *Nature* 272, 510-5.
- Shendure J, Church GM (2002) Computational discovery of sense-antisense transcription in the human and mouse genomes, *Genome Biology* 3(9):research0044.1-0044.14
- Shintani S, O'hUigin C, Toyosawa S, Michalova V, Klein J (1999) Origin of gene overlap: The case of TCP1 and ACAT2, *Genetics* 152, 743-754.
- Spencer CA, Gietz. RD, Hodgetts R.B (1986) Overlapping transcription units in the dopa decarboxylase region of *Drosophila*. *Nature* 322, 279-281.
- Terryn Nancy and Rouze Pierre (2000) The sense of naturally transcribed antisense RNAs in plants. *Trends in plant science* 5(9).
- Tufarelli C, Stanley JA, Garrick D, Sharpe JA, Ayyub H, Wood WG, Higgs DR (2003) Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat. Genet.* 34, 157-165.
- Vanhee-Brossollet C, Vaquero C (1998). Do natural antisense transcripts make sense in eukaryotes? *Gene* 211, 1-9.
- Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I (2004) Mammalian Overlapping Genes: The Comparative Perspective, *Genome Research* 14, 280-286.
- Wang XJ, Gaasterland T, Chua NH (2005) Genome-wide prediction and identification of cis-natural antisense transcripts in *Arabidopsis thaliana*. *Genome Biology* 6:R30.
- Williams T, Fried M (1986) A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3' ends. *Nature* 322, 275-9.
- Williams BAP, Slamovits CH, Patron NJ, Fast NM, Keeling, PJ (2005) A High Frequency of overlapping gene expression in compacted eukaryotic genomes. *PNAS* 102, 10936-10941.

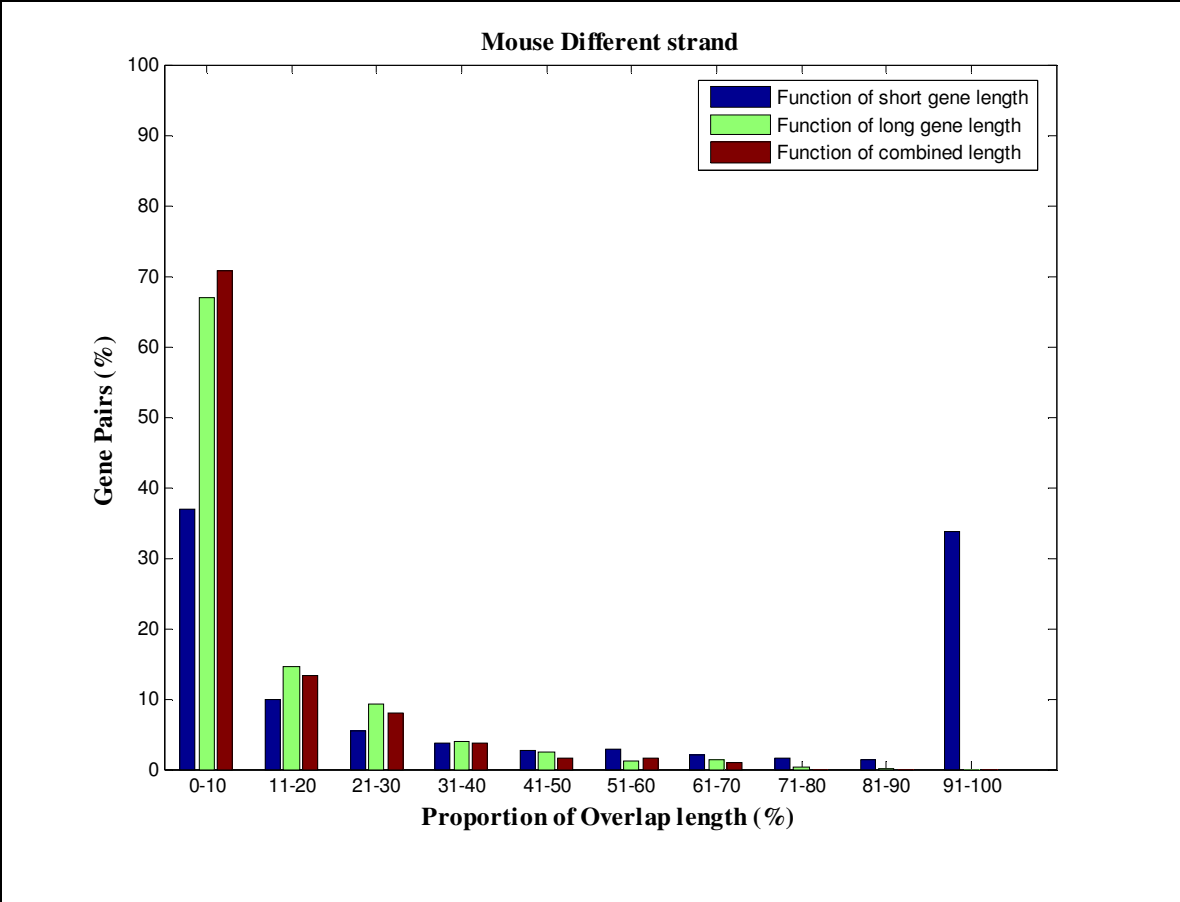
- Wong F, Yuh ZT, Schaefer EL, Roop BC, Ally AH (1987) Overlapping transcription units in the transient receptor potential locus of *Drosophila melanogaster*. *Somat Cell Mol Genet.* 13(6), 661-9.
- Yanicostas C, Lepesant JA (1990) Transcriptional and translational cis-regulatory sequences of the spermatocyte-specific *Drosophila* janusB gene are located in the 3' exonic region of the overlapping janusA gene. *Mol. Gen. Genet* 224, 450-458.
- Yelin et al. (2003) Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnology* 4, 379-86.
- Zuniga Mejia Borja A, Meijers C, Zeller R (1993) Expression of alternatively spliced bFGF first coding exons and antisense mRNAs during chicken embryogenesis. *Dev. Biol* 157, 110-118.

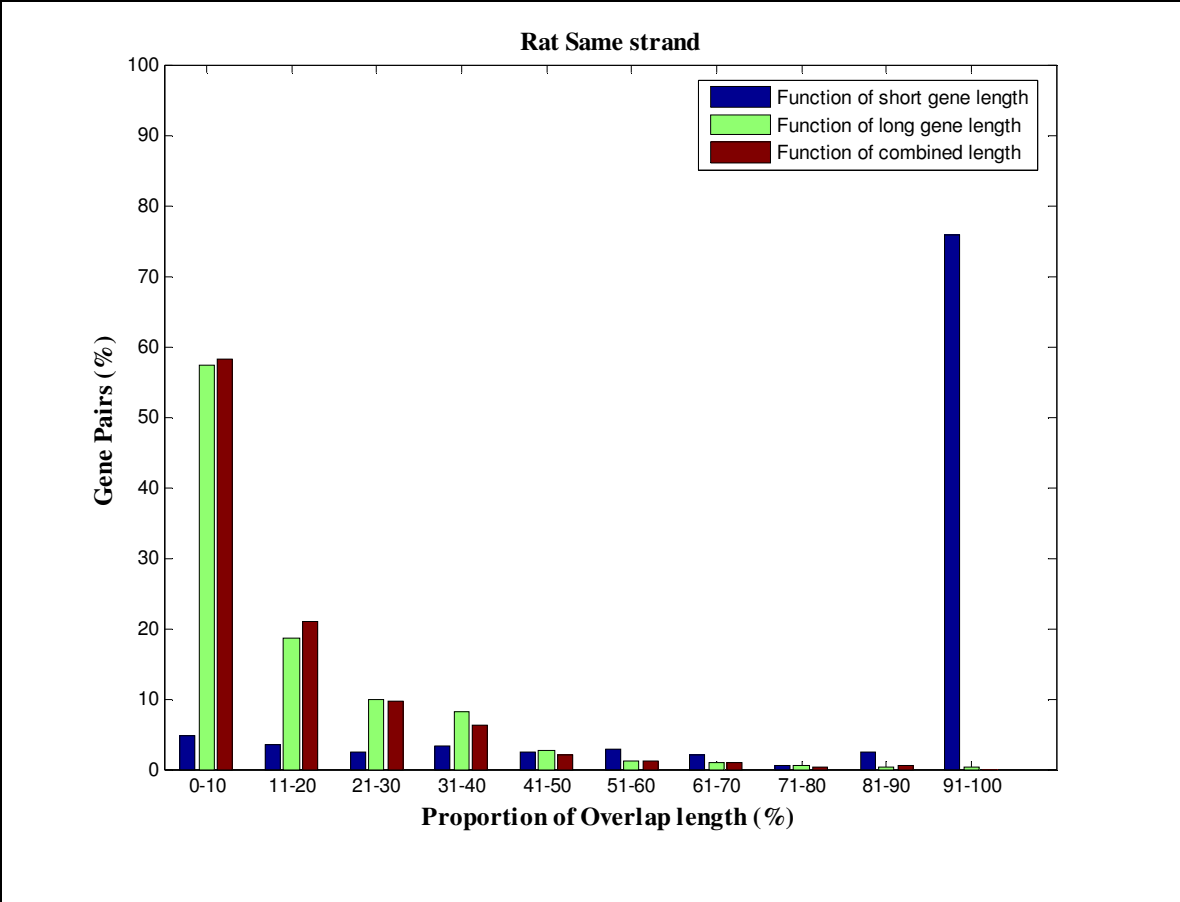
## Appendix A: Figures in Chimp, Mouse, and Rat

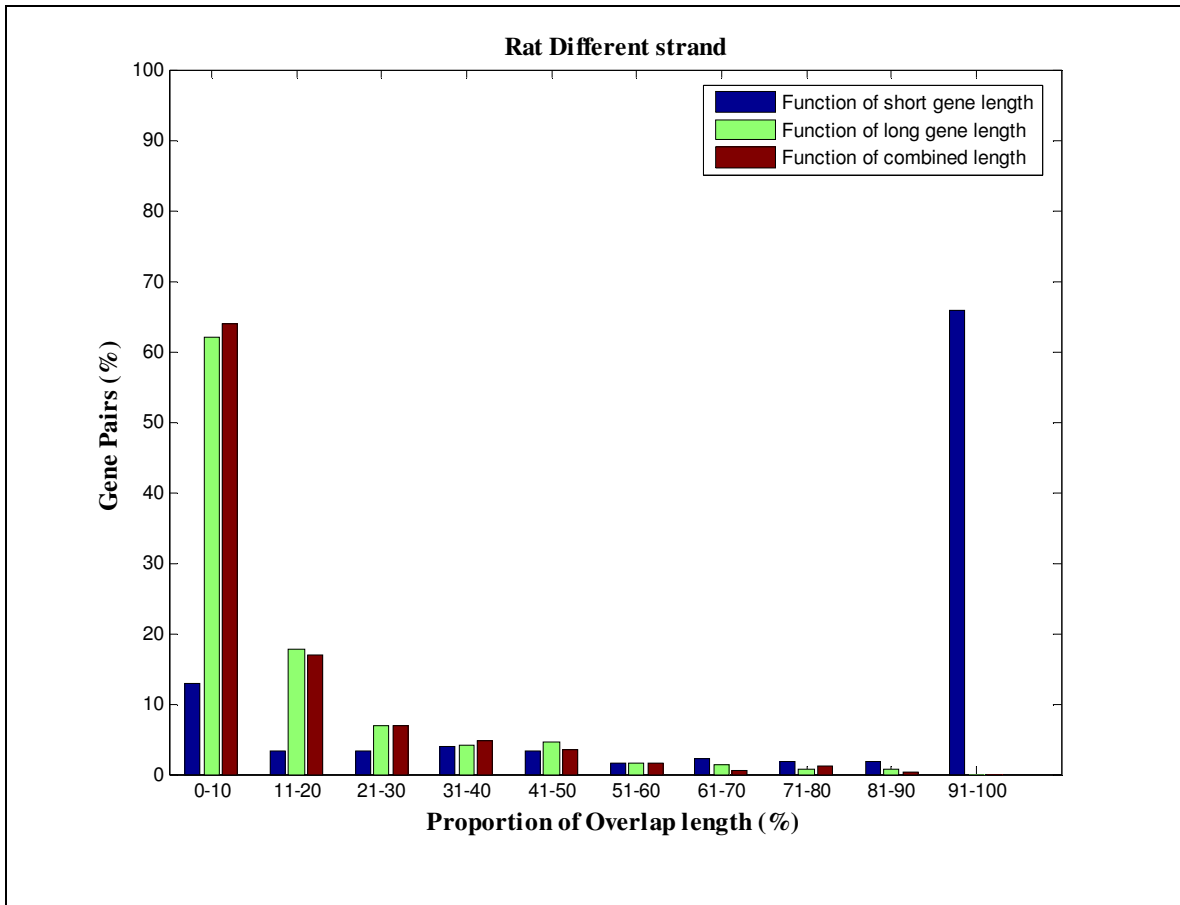




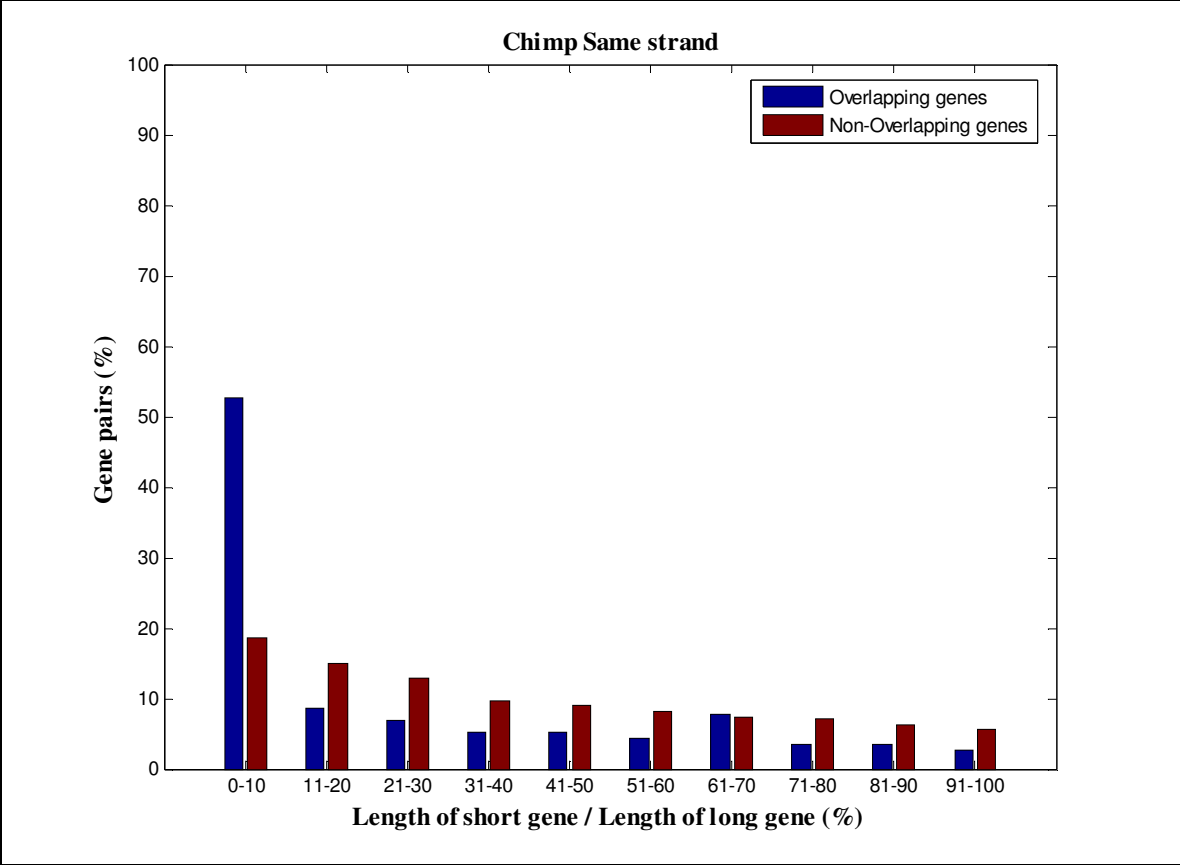


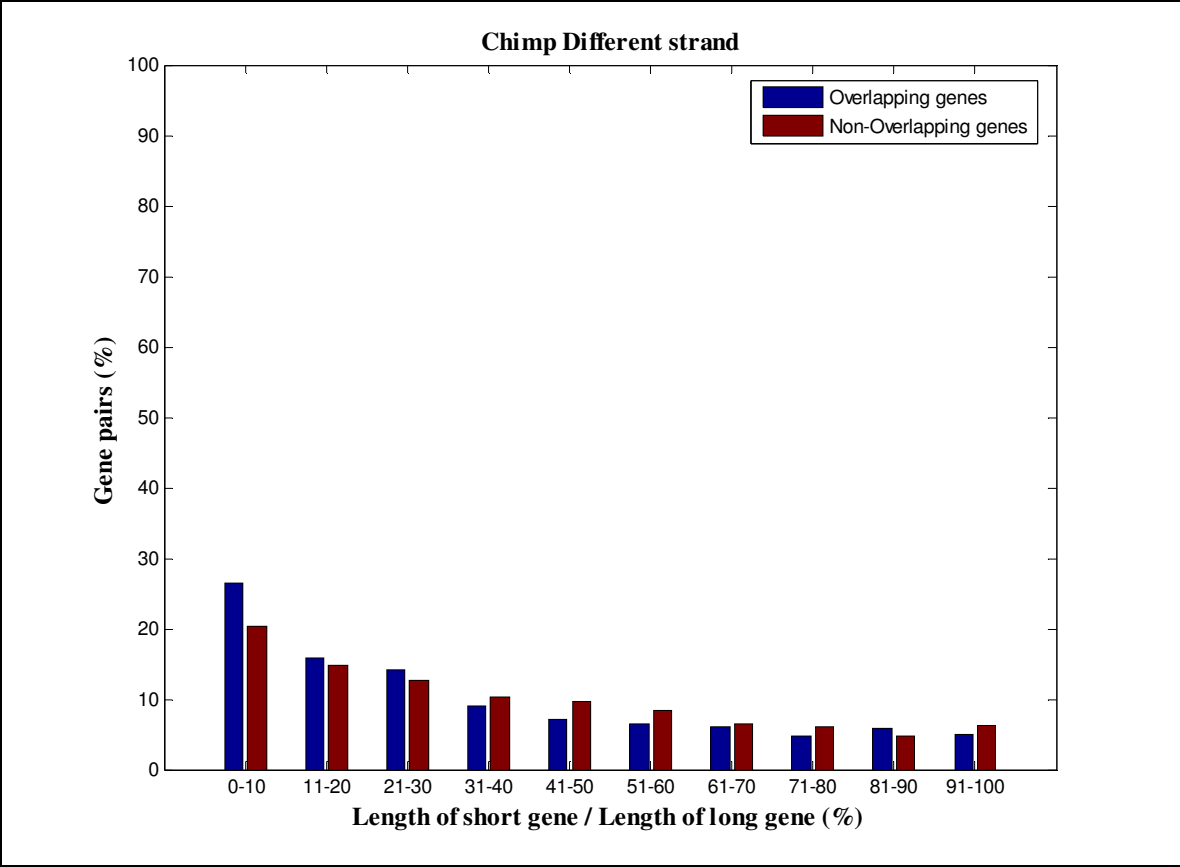


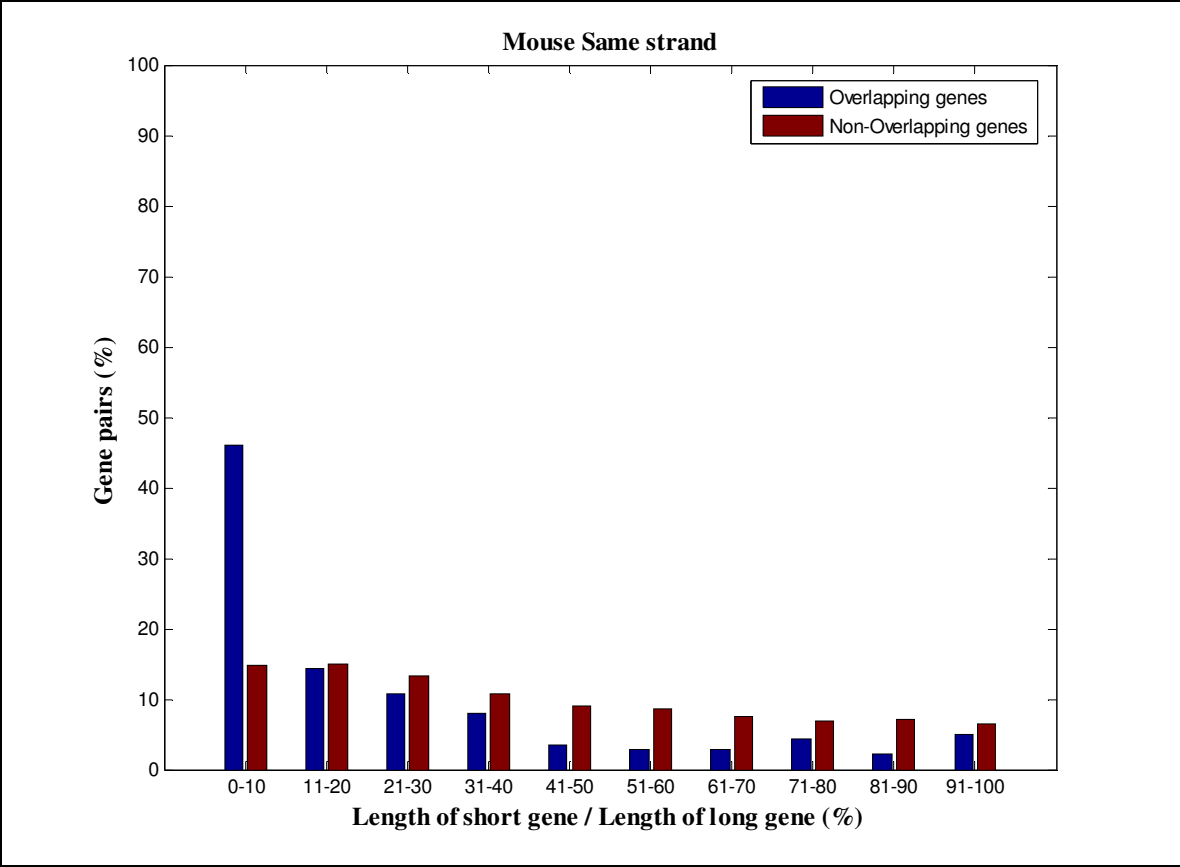


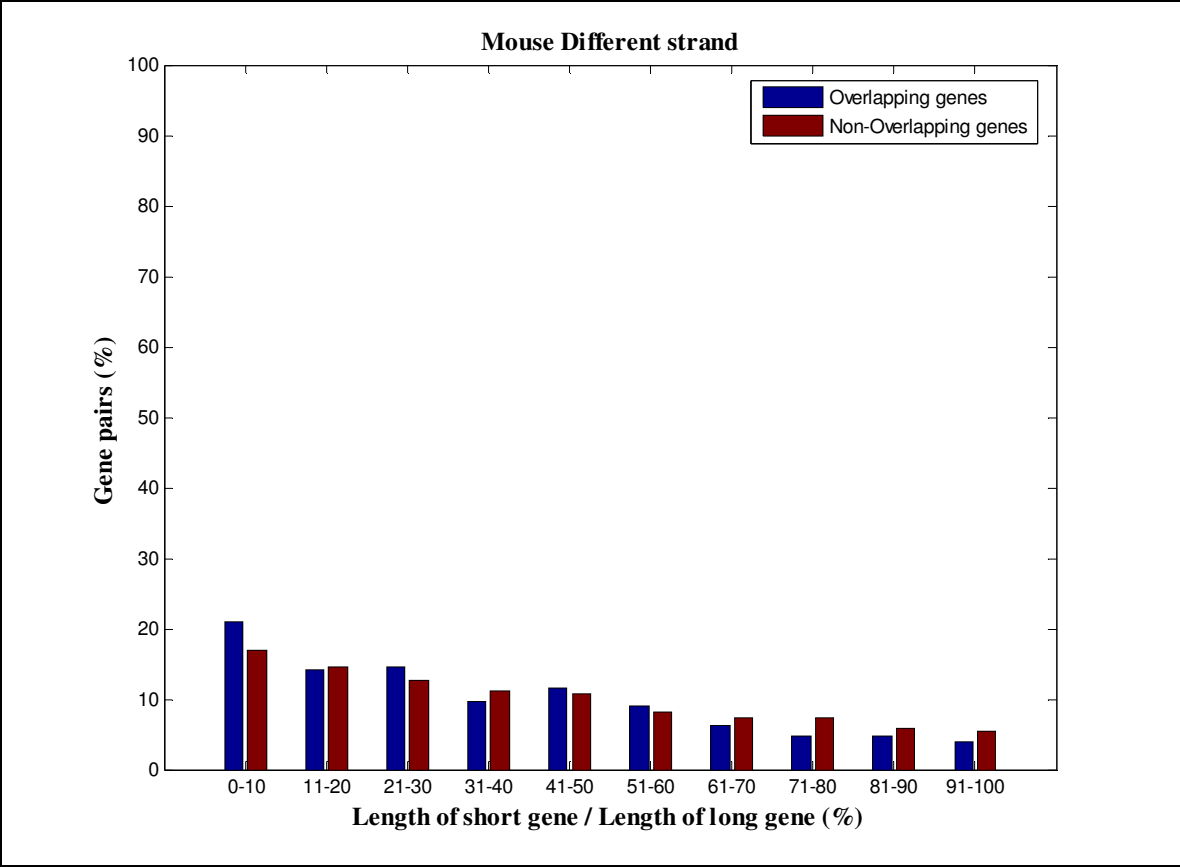


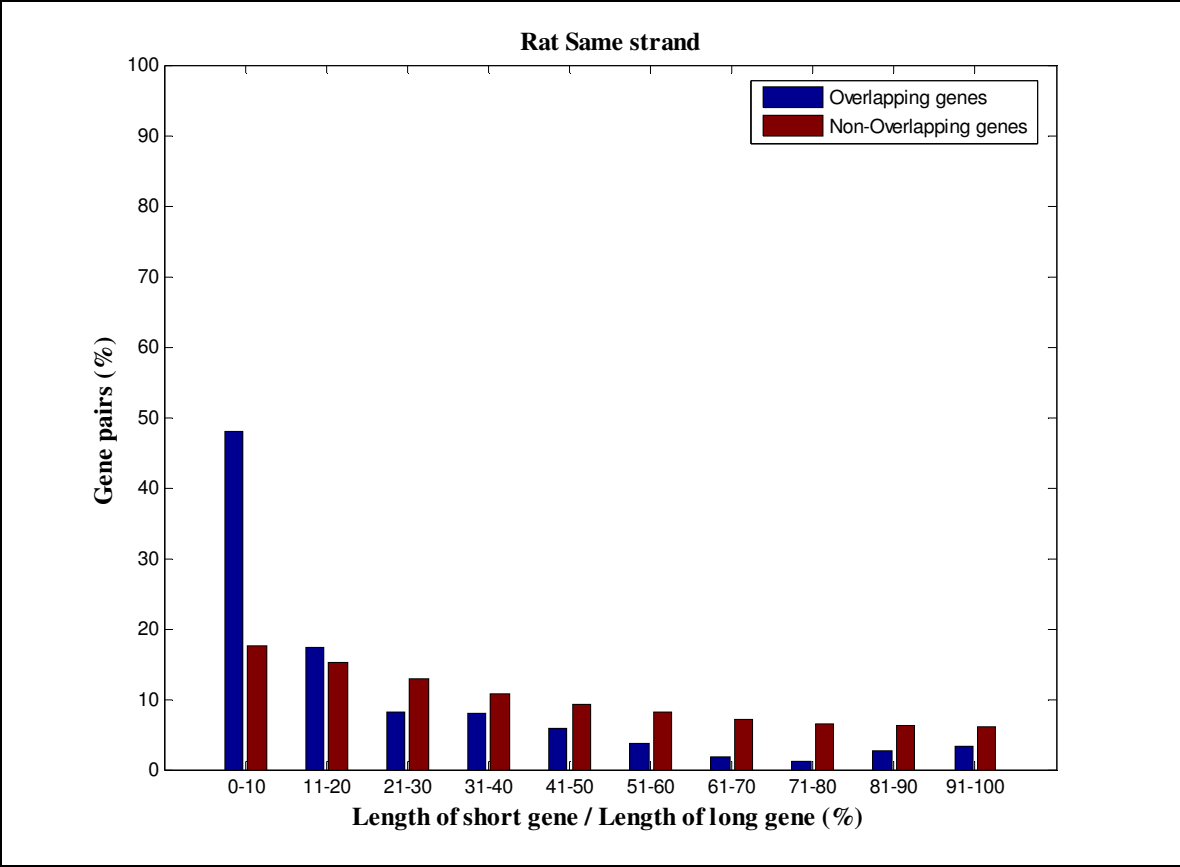
**Figure A. 1** The distribution of the proportion of the overlapping lengths with respect to the sizes of the short genes, long genes, and the regions spanned by both genes.

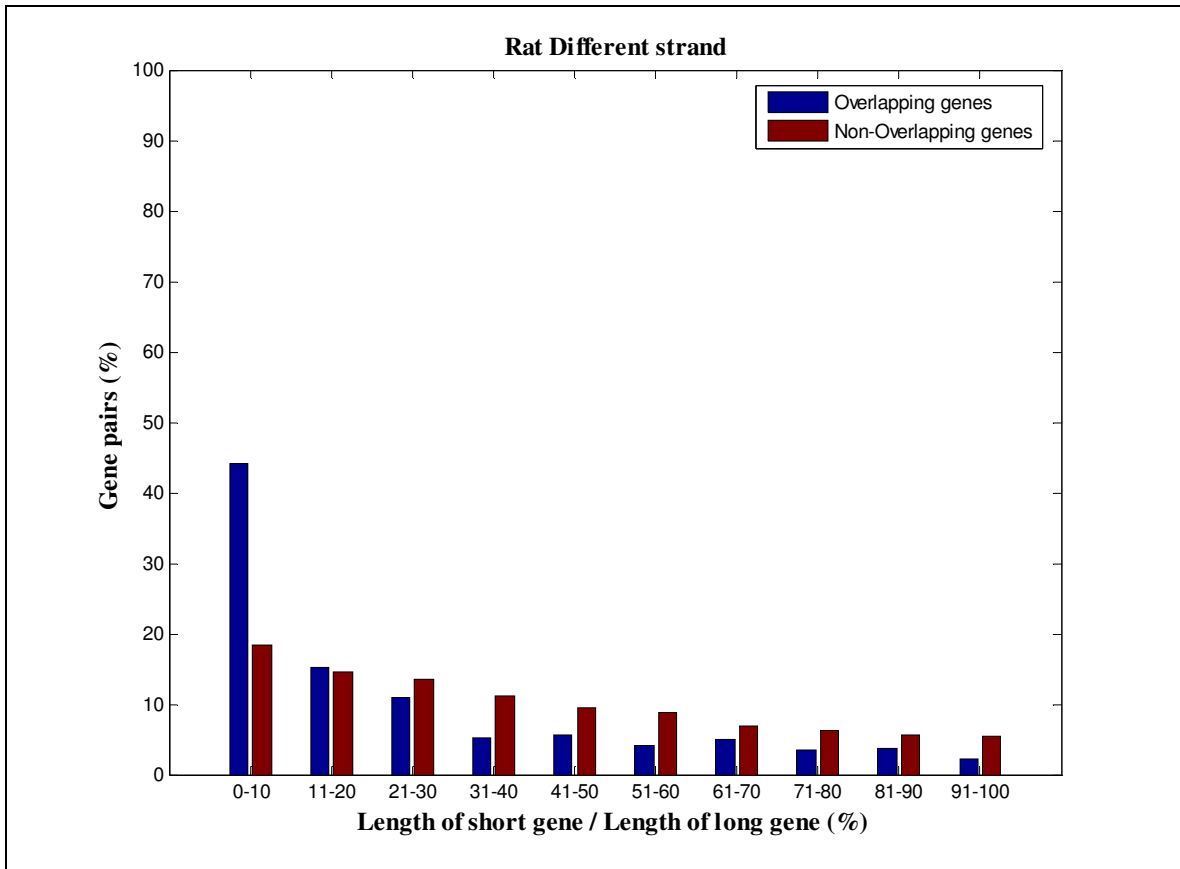












**Figure A. 2** The distribution of the ratios of the lengths of short vs. long genes for the same and different strand overlaps in comparison with the one for the neighboring and non-overlapping genes.

## **Vita**

Chaitanya R. Sanna was born in Hyderabad, India. She obtained her Bachelor of Science degree (Magna Cum Laude) in Computer Science from George Mason University, Fairfax, Virginia, in the summer of 2004. In the Fall of 2004, she began graduate studies in Computer Science at Virginia Polytechnic Institute and State University. Her research interests include bioinformatics.