

# **Statistical Analyses and Markov Modeling of Duplication in Genome Evolution**

By

Yi (Joey) Zhou

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Biology

New York University

May 2005

---

Bhubaneswar Mishra

© Yi (Joey) Zhou

All Rights Reserved, 2005

## **Acknowledgements**

This dissertation would not have been possible without the help and support from many people to whom I am greatly indebted.

First of all, I thank my advisor, Bud Mishra for his support, encouragement and collaboration. It is Bud who introduced me to the computational world, and guided me into the exciting interdisciplinary field of computational biology. He has taught me the value of rigor and depth in scientific research, and has demonstrated a truly great devotion to science.

I also thank everyone in the NYU/Courant Bioinformatics Group. They have been my best mentors, collaborators and friends, and have filled every day of my thesis work with care, joy and excitement. Special thanks to Nadia Ugel, Raoul-Sam Daruwala, Toto Paxia, Archisman Rudra, Marco Antoniotti, Fang Cheng, Bing Sun, Marina Spivak, Ofer Gill, Matthias Heymann, and many more. I will cherish every day that I have spent with them.

I benefited much from David Fitch, who shared with me his deep biological insights, and constantly encouraged me in the study of evolutionary

biology. Without him, this thesis would not be possible. I would like to express my gratitude to Suse Broyde, Daniel Tranchina, and Brian Dynlacht for serving on my thesis committee and for providing valuable discussions and suggestions, and special thanks to Tan Ignatius for helping me during my job hunting. I would also like to thank Todd Holmes. I have enjoyed the two years I have worked in his lab and learned many valuable experimental techniques from him.

I am greatly indebted to Lexing Ying for his care, support and encouragement. With his love, all the difficult moments become much more endurable.

Finally, I would like to thank my family. They have selflessly provided the best in the world to me, and I owe to them everything I am now.

## **Abstract**

Genome evolution, especially duplications, was studied using a computational approach. The motivation of the thesis work comes from the “evolution by gene duplication” theory proposed by Susumu Ohno in 1970’s, which postulates that duplication is one of the main forces in driving genome evolution and creating genome complexity. The research described in this dissertation investigates the duplication process systematically by analyzing whole-genome data. In particular, it studies the molecular mechanisms of the segmental duplications in mammalian genomes; the influence of duplications and other evolutionary processes on the genome statistical structures; and the measurement of phylogenetic distances between genomes based on the number of duplications and other evolutionary events. During the process, we have developed computational methods and mathematical models that take into account the nature of the data and incorporate the dynamics of the evolutionary processes.

Using a Markov model of the segmental duplication process in the mammalian genomes, we found that about 12% of these recent segmental duplications were caused by recombination mediated by the recent active interspersed repeats in the mammalian genomes. In addition, the physical instabilities in the DNA

sequence may also affect the process by introducing “fragile” sites in the genomes. A “rich gets richer” dynamics of the duplication process is suggested by the results of the analysis on the copy number distributions of the segmental duplications as well as other genomic components. Based on these observations, we propose a parsimonious genome evolution model, which includes three elementary processes: *substitution*, *duplication* and *deletion*. Using this model as our prior, we further developed a novel alignment-independent method that estimates the genomic evolutionary distance based on their word copy number variations. The phylogenomic distance measured using our method reflects the total number of substitution, duplication and deletion events since the divergence of the two genomes. Combined with conventional phylogenomic methods, we can study the modulation of the three different evolutionary processes in different lineages.

# Table of Contents

<b>Acknowledgements</b> .....	<b>iii</b>
<b>Abstract</b> .....	<b>v</b>
<b>List of Figures</b> .....	<b>xii</b>
<b>List of Tables</b> .....	<b>xvi</b>
<b>List of Abbreviations</b> .....	<b>xviii</b>
<b>List of Appendices</b> .....	<b>xix</b>
<b><u>Chapter 1.</u> Introduction</b> .....	<b>1</b>
1.1 Background .....	1
Duplication in Genome Evolution .....	3
Phylogenetic Methods .....	12
Markov Chain.....	16
Parameter Estimation.....	18
1.2 Rationale .....	21
1.3 Problem Statement .....	23
1.4 Contributions and Thesis Organization .....	25
<b><u>Chapter 2.</u> Mechanisms of Segmental Duplications in Mammalian Genomes</b> .....	<b>30</b>

2.1 Introduction and Related Work.....	30
2.2 Methods .....	34
Sequence Preparation .....	34
Mer Analyses.....	35
Repeat Analysis.....	36
Markov Model of Duplication.....	37
Model Parameter Estimation .....	39
Model Evaluation.....	46
Stability and Flexibility Computation.....	47
2.3 Mer Analysis on the Duplication Flanking Regions.....	48
2.4 Repeat Analysis on the Duplication Flanking Regions .....	51
2.5 Error Analysis on the Current Mapping of Segmental Duplications in Human Genome.....	55
2.6 Markov Model of the Duplication Process.....	62
2.7 Physical Instability in the Duplication Flanking Regions .....	75
2.8 Summary.....	80
<b><u>Chapter 3. Statistical Structure of the Genomes.....</u></b>	<b>84</b>
3.1 Introduction and Related Work.....	84
3.2 Analysis on the Distribution of the Segmental Duplications in Human Genome .....	87
3.3 Analysis on the Distribution of the Protein Domain Family Sizes .....	90



3.4 Analysis on the Distribution of the Mer Frequencies .....	93
3.5 The Effect of Repeats and Selection on the Mer Distributions .....	97
3.6 Summary .....	100
<b><u>Chapter 4.</u> A Polya’s Urn Model for Genome Evolution .....</b>	<b>102</b>
4.1 Introduction and Related Work.....	102
4.2 Polya’s Urn Model for Genome Evolution .....	104
4.3 A Simple Model for Non-overlapping Mers: The Parsimony of the Genome Evolution Model.....	106
4.4 Summary .....	121
<b><u>Chapter 5.</u> A Novel Alignment-Independent Method for Estimating Phylogenomic Distances.....</b>	<b>123</b>
5.1 Introduction and Related Work.....	123
5.2 A Refined Model for Genome Evolution .....	128
5.2 Rationale .....	130
5.3 Methods .....	133
Projection of Genome Evolution onto Mer Copy Number Evolution.....	133
Mer Copy Number Evolution Parameter Estimation .....	144
Estimation of Genome Evolutionary Distance.....	147
Method Summary.....	148

5.4 Method Verification .....	150
Simulation Data .....	150
Real Genomic Data .....	159
5.5 Study on the Modulation of Substitution and Indel Events .....	162
5.6 Summary .....	171
<b><u>Chapter 6. Summary</u> .....</b>	<b>174</b>
6.1 Summary .....	174
Mechanisms of the Recent Segmental Duplications in Mammalian Genomes .....	174
Analysis and Modeling of the Duplication Effect on the Statistical Structure of the Genomes .....	176
Alignment-Independent Phylogeny Method .....	177
6.2 Future Work .....	179
Further Examination of the Segmental Duplication Dynamics .....	179
Incorporating Heterogeneous Mutation Rate in the Alignment-Independent Phylogeny Method .....	180
Bayesian Framework for the Alignment-Independent Phylogeny Method .....	181
Generalization and Potential Application of the Alignment-Independent Phylogeny Method .....	181
<b>Appendices .....</b>	<b>183</b>
Appendix A. Choice of Duplication Divergence Interval Size .....	183

Appendix B. Empirical Simplification of Mer Copy Number Evolution	185
<b>Bibliography</b>	<b>189</b>

## List of Figures

Figure 1. The procedures of mer analysis in the segmental duplication flanking sequences. ....	35
Figure 2. The average frequencies of the 6-mer ‘AAAAAA’ in the flanking sequences at different physical locations relative to the duplication breakpoint. ....	48
Figure 3. The significantly over-represented 5 and 6-mers in the duplication flanking sequences. ....	50
Figure 4. The appearance frequencies of various subfamilies of repeats detected in the duplication flanking regions in the human (hg16 and hg15), mouse (mm3) and rat (rn3) genomes. ....	53
Figure 5. The figure represents schematically the definition of <i>gap</i> , <i>shift</i> and <i>internal indel</i> . ....	56
Figure 6. Analyses on the <i>gaps</i> between the matching homologous repeats in the flanking regions and the mapped duplication boundaries. ....	57
Figure 7. Analyses on the <i>shifts</i> between the positions of the matching homologous repeats in the duplication flanking regions. ....	59
Figure 8. A schematic display of our mathematical model formulating the changes in the distribution of flanking region pairs over different states as a	

Markov process over evolution time. ....	68
Figure 9. The fitting of the model to the distribution of Alu and L1 repeats in the duplication flanking regions in the human genome (hg16 shown here) and Mouse, Rat genomes.....	71
Figure 10. The helix stability and the DNA flexibility in the repeat-less flanking sequences in the human (hg16 shown here) and mouse genomes.....	76
Figure 11. The schematic representation of the propagating effects of duplication during recent human genome evolution.. ....	82
Figure 12. The distribution of the potential duplication 'hot-spots' on the human genome.....	88
Figure 13. Protein domain family size distribution in some genomes examined. ....	92
Figure 14. Mer-frequency distribution in some genomes examined. ....	94
Figure 15. Amino acid mer-frequency distribution in <i>E. coli K12</i> and <i>H. pylori</i> .. ....	95
Figure 16. The 6-mer frequency distribution of the resulting sequence of a simple "evolution by duplication" simulation. ....	97

Figure 17. The 7-mer copy number distribution in <i>S. cerevisiae</i> before and after the removal of repetitive elements from the analysis. ....	98
Figure 18. The relation between the total copy number of each 7-mer in <i>E. coli</i> K12 and <i>P. abyssi</i> and the number of copies that have experienced changes when compared to the genomes of their close relatives <i>E. coli</i> O157 and <i>P. horikoshii</i> , respectively.....	100
Figure 19. The three processes occurring during graph evolution: <i>deletion</i> , <i>duplication</i> , and <i>substitution</i> .....	109
Figure 20. Mer-frequency distribution in some genomes examined.. ....	112
Figure 21. Comparison of models in fitting of the distributions of different mer sizes in <i>Escherichia coli</i> K12 genome.....	113
Figure 22. The model is also successfully applied to the distribution of 3-amino acid mers computed from all <i>Escherichia coli</i> K12 proteins. ....	115
Figure 23. The refined Polya's Urn model for genome evolution. ....	130
Figure 24. The evolutionary events on the genomic sequence level can be projected onto each individual mer as the changes in their copy numbers.. ..	130
Figure 25. The summerization of the development of our method.....	131

Figure 26. The relation between estimated $p_{-1}(m)$ and the mer size $m$ according to the random sequence model.....	139
Figure 27. Notations for the copy numbers of the submers and subsequences in a length- $m$ mer. ....	140
Figure 28. The design of our method.....	148
Figure 29. The scheme for <i>in silico</i> evolution simulation.....	150
Figure 30. Evaluation of the method without consideration of sequence correlation using <i>in silico</i> simulated data.. ....	153
Figure 31. Evaluation of the method with consideration of sequence correlation using <i>in silico</i> simulated data.....	155
Figure 32. The empirical distribution of the ratio among external branches computed from the simulated quartet datasets where equal external branch lengths are expected.....	170

## List of Tables

Table 1. Summerization of the model parameters. ....	40
Table 2. All possible transitions between different states of the duplication flanking regions in a short evolution period $\Delta t$ . ....	65
Table 3. The contribution of repeat recombination, estimated by the model from the datasets in different regions.. ....	73
Table 4. The enrichment of the “fragile” sites in the repeat-less duplication flanking sequences in different mammalian genomes.....	79
Table 5. Graph model parameters ( $\mu_s, \mu_r / \mu_d$ ) fitted to the mer-frequency distribution data (6 to 8-mer for prokaryotic genomes and 8 to 10-mer for eukaryotic genomes) from the whole genome analysis.....	117
Table 6. The sensitivity of our method evaluated using the <i>in silico</i> evolved quartet datasets.....	158
Table 7. Comparison of our methods with the alignment results for closely related sequences without rearrangements. ....	160



Table 8. The phylogeny trees inferred from the orthologous sequences of four mammalian species.....	162
Table 9. The two different scenarios of differential modulation of the substitution and indel events.....	169
Table 10. The probability of getting an MLE (assuming iid Bernoulli random variable) estimation $\hat{p}$ outside a specific error bound $\delta$ of the true probability $p$ , <i>i.e.</i> $\hat{p} < p(1-\delta)$ or $\hat{p} > p(1+\delta)$ , in a sample of size $N$ ....	184

## List of Abbreviations

Mer: short words of DNA or protein sequences, i.e. oligo-nucleotides or oligo-peptides of a particular length. Therefore, for mers of length  $m$ , there are totally  $4^m$  DNA mers and  $20^m$  protein mers.

ML: Maximum Likelihood methods.

MP: Maximum Parsimony methods.

SINE: Short Interspersed Nucleotide Elements.

LINE: Long Interspersed Nucleotide Elements.

L1: Long interspersed nucleotide element 1.

bp: base pair.

$\Delta G$ : Gibbs free energy.

Array-CGH: array-based Comparative Genomic Hybridization.

## List of Appendices

[Appendix A. Choice of Duplication Divergence Interval Size](#).....183

[Appendix B. Empirical Simplification of Mer Copy Number Evolution](#).....185

# Chapter 1

## Introduction

### 1.1 Background

Genome evolution is the underlying process that ultimately determines the structure, regulation and variation of the other processes in a biological system, and is the key to understanding the development and diversity of biological traits. It is a complex process that involves various molecular mechanisms such as point mutations, insertions, deletions, duplications transpositions, translocations, inversions and recombinations, and is further affected by selective constraints, effective population size, and various environmental factors. In the genomic era, the research on genome evolution has moved from a gene-centric view to a global perspective at genome-wide scale. The construction of large-scale cellular networks, including regulatory, metabolism and protein-protein interactions, has allowed the investigation on the relation between the evolutionary rates of a gene and its position in the network [176], bringing evolutionary studies to a system level. Recently, there has been great progress in both tool development and biological insights on various problems of genome evolution that were not accessible before.

Comparative genomics has played an essential role in the recent development of the genome evolution field. Most of the current genome evolution research requires tools to compare a genome in its entirety to itself or to other genomes. For example, well-conserved inter-genome regions hint at a selection advantage, and intra-genome duplicated regions suggest interesting evolutionary dynamics. Currently available sequence alignment tools have incorporated many innovations to greatly advance very detailed comparative genomics studies. Most of local alignment tools use exact or inexact k-mers as homology seeds for local alignment extension, such as BLAST [4], PatternHunter [110], CHAOS [29], BLASTZ [150][149], BLAT [93] and PASH [89]. The global alignment tools, such as MUMmer [40], AVID [25] and LAGAN [28] are built on top of the local alignment tools, creating the global alignment by chaining the significant local alignments and applying the local alignment tool iteratively in the gaps. Some tools even incorporate the statistical structure of the genome, to reach higher sensitivity and specificity (e.g., mapping out wobble positions, as in WABA [94]). Most of these algorithms are applicable to whole-genome-scale comparisons, and can safely detect homology levels higher than 80%.

The availability of whole-genome-scale alignment results from comparison

studies either between different genomes or against the same genome has motivated the recent studies on many complex and interesting evolution problems. For example, based on the sequence comparison between different genomes, on a gross scale, one can detect the structure and distribution of the synteny groups and infer the possible mechanisms for large scale genome rearrangements [57][182][133]. On a finer scale, the mutational pattern has been extensively studied and co-variation has been observed among the rates of different mutational events in different regions of the genome [157][74][188][140]. Furthermore, the conserved non-coding sequences that are potentially functional with different lineage-specificities [45][112][39][180][43] can be identified using phylogenetic foot-printing [72][22][130] among distantly related genomes and phylogenetic shadowing [23] among closely related sequences. Combining the sequence analysis in a single genome and its close relatives, the rate of the (retro)transposition over the evolutionary history of the genome can be inferred [57][182][183] [105].

### **Duplication in Genome Evolution**

One of the important topics in genome evolution has been the study of duplication process. In 1970, S. Ohno [124] proposed “evolution by gene duplication”, first suggesting the essential role of the duplication process

in genome evolution. As the results from large scale sequencing and experimental effort as well as comparative genomics become available, Ohno's theory, expanded to a more general view including duplications of both gene and non-genic regions, has gained much attention and progress. On the gene level, large scale detection of paralogous genes in various genomes [141][103][109] led to the analysis of the age, scale and functional category of the duplication genes. The ubiquitousness of the gene duplication phenomenon and the variation in the duplication pattern has led to deep appreciation of the complexity of the duplication process [69][55][32][155][77][1]. The rates of gene duplication and deletion have been examined in different genomes, and were found to be on a similar scale as the substitution rate [109].

The fate of the duplicated genes is also of great interest to biologists. Ohno's theory [124] argues that after gene duplication, one of the duplicated copy preserves the original function while the selection pressure on the other copy is relaxed, allowing it to accumulate various mutations. The mutational copy eventually becomes a pseudogene by loss-of-function, or by chance give rise to an advantageous gene with a gained new function. This theory was later referred to as the "mutation during nonfunctionalization" (MDN) model by A. Hughes [81]. Under such a model, the population genetic theory predicts that a

duplicate gene is much more likely to experience loss-of-function in typical situations than gaining a new function, suggesting a low retention rate of the duplicated genes [178]. However, many [81][175] have criticized the MDN theory based on the observations of negative and positive selection in the duplicated gene pairs and the high retention rate of duplicated genes in tetraploid fish lineages and *Xenopus laevis* [171][121]. A. Hughes proposed “gene sharing” as an alternative theory [81]. In his theory, the singleton genes first gain multiple functions and go through a period of gene sharing (one gene performing multiple functions). The following gene duplication then allows each daughter gene to specialize one of the functions of the ancestral gene. Under similar assumptions, A. Force and M. Lynch proposed the duplication-degeneration-complementation (DDC) model [54]. Similar to Hughes’ model, it suggests that, after duplication, the two gene copies acquire complementary loss-of-function mutations in independent sub-functions, such that both genes are required to produce the full complement of functions of the single ancestral gene. Population genetic theory [177] predicts that when duplicated genes are preserved by sub-functionalization, it potentially extends the time period during which both genes are exposed to natural selection, thereby enhancing the chance of gaining rare beneficial mutations to novel functions. It also partially releases the selection pressure on both copies by reducing their pleiotropic constraints, allowing further



fine-tuning on the specific subfunctions. Both the gene sharing and the DDC model have found support from individual experimental data, such as the *Hox* genes and the *nodal* genes in zebrafish (reviewed in [136]), and the great retention rate of the duplicated genes in tetraploid fish lineages and *Xenopus laevis* [171][121].

Several groups [109][98] have conducted large scale experiments in various genomes to examine the mutation rates in duplicated genes of different ages, and confirmed the temporary relaxation of the selection pressure right after the duplication occurs. However, not so infrequently, the two duplicated genes evolve “asymmetrically” at expression and/or sequence level [42][176][36], i.e. one duplicated gene or part of that gene has gone through a significantly different divergence rate or selective pressure from the other duplicate copy when compared to their out-group ortholog. To explain such diversification in the duplicated genes, more specific models have been proposed. For example, Nowak [123] tried to explain the retention of the functionally redundant genes in a population by the balance between the fitness provided by the redundancy and the variance in the mutational rate of the duplicated genes during germline or somatic development. Gibson and Spring [58] proposed that duplications of multi-domain proteins may be preserved by purifying selection

because deleterious mutations in the duplicated copy can cause a dominant-negative phenotype by incorporating a mutation protein into the protein complex and disrupting its normal function.

All the above models describe the duplicated gene fixation as a surviving process from neutral or negative selection. There are also evidences that duplicated genes can be fixed by positive selection. For instance, by producing more of the same protein, duplicated genes can be retained in the genome through dosage compensation effect [96]. In other cases, the duplicated genes can go through Darwinian positive selection and provide functional or structural variation, contributing to the adaptive evolution of the organism [80][191].

To understand the role of the duplicated genes in the evolution of the biological system, research has been conducted to study their expression pattern, genetic dispensability, and their positions in the cellular networks [176]. It has been found that duplicated genes diverge faster at the expression level than at the protein coding level, indicating that transcription evolution is much faster than protein sequence evolution [70]. From a large scale knock-out experiment in yeast, it was found that if a gene has a duplicate in the genome; its knockout has less effect on the fitness of the organism compared to the knockouts of

the singleton genes [71]. Such experimental results confirmed the long-held speculation that the duplicated genes provide robustness to the genome. Interestingly, the genetically dispensable duplicated genes tend to have a medium level of expression correlation, instead of being highly or anti-correlated [88]. These observations are consistent with the sub-functionalization model discussed earlier, where duplicated genes can be fixed in the genome by becoming complementary either at the expression level (expect to be anti-correlated) or at the protein function level (expect to be correlated). To examine the duplication process on a system level, the duplicated genes have also been characterized in various cellular networks, and were found to be more constrained to the part of the network with more specific functions [31][17][114]. For example, in *C. elegans*, most of the duplicated developmental genes are in the late development stage instead of the early stage in which genetic changes tend to be more fatal [31].

The availability of genome data and sequence analysis tools also shed much light on the otherwise inexplicable possible whole genome duplication events in several evolutionary lineages at different evolutionary times, particularly in early vertebrates [48][82][117], *Arabidopsis* [62][173], and *Saccharomyces* [184][152]. The sequencing of a related species, *Kluyveromyces waltii*, that diverged

from *Saccharomyces cerevisiae* before the duplication event and the comparative study on the gene orders and copy numbers provided the most convincing evidence for a whole genome duplication in *S. cerevisiae* followed by a massive gene loss [92]. Although whole genome duplication theory is favored in the cases of early vertebrates and *Arabidopsis*, more convincing evidence are still lacking and may also depend on similar comparative studies as those performed on yeast.

Although whole genome duplication events can bring large impact on genome evolution, more often, duplications occur at a scale much smaller than the whole genome. The duplicated segments do not necessarily cover a functional gene unit, and in fact may not carry any coding regions at all. Therefore, one needs to expand the gene-centric view in studies of the duplication process to a genome-scale view.

Recently, large segmental duplications have been detected and cataloged in various mammalian genomes. These duplications, which happened quite recently (30-60Mya), covered both coding and non-coding regions and include both intra-chromosomal and inter-chromosomal events [13][34][12][35][170]. They are distributed in the genome in a clustered manner, mostly around pericentromeric and subtelomeric regions, and have been suggested to have contributed to

the evolutionary dynamics of the mammalian genomes. Different studies have confirmed the significant association between segmental duplications and syntenic breakpoints [11][6], indicating their role in large genomic rearrangement events. Additionally, many of the duplicated segments in the human genome have been found to be involved in further rearrangements, some leading to genetic diseases [51][85]. The genic contents of the segmental duplications suggest that they may also play a role in adaptive evolution and a domain accretion process [144].

Using in-laboratory evolution experiments and various new experimental techniques, such as array-CGH (Comparative Genomic Hybridization), various genome rearrangements, including segmental duplications, have been traced on a genomic scale in a time series, from which the exact sequence and onset time of the events can be recorded. For example, adaptive segmental duplications have been observed during the in-laboratory evolution of *E. coli* [142] and *S. cerevisiae* [46] strains. The mutational spectrum has also been studied in *C. elegans* strains evolved under a regime in which effects of selection were greatly reduced relative to genetic drift [41]. During the micro-evolution of cancer development, a genome goes through a large amount of rearrangements, in the forms of duplications, deletions and translocations (reviewed by

[101][14][2][154]), and cause the copy numbers of different genomic regions to fluctuate considerably [106][107][2][134]. Similar techniques are now being applied to study these processes in cancer cells.

The availability of the genomic sequence has also greatly facilitated the studies on the molecular mechanisms of the duplication process. Repeat elements, especially transposable elements were found to play an important role. A famous case of repeat's involvement in gene duplication is the duplication of the  $\gamma$ -globin gene by unequal crossover mediated by L1 long interspersed repetitive elements (LINE) in an early ancestor of simian primates [53]. More recently, Alu, a short interspersed nucleotide element (SINE) in the primate genomes, were found to be actively involved in various chromosome rearrangements, including duplications, deletions and translocations, by creating recombination hotspots in both genetic diseases, such as tumor, and genomic polymorphisms in the normal population (reviewed in [97]). Detailed breakpoint flanking sequence analyses in the in-laboratory evolved *E. coli* [142] and *S. cerevisiae* [46] strains showed that the large genomic evolutionary events were mostly caused by the homologous recombination or transposition of the mobile elements (insertion sequences, or transposable elements and their relics). However, duplications can also be caused by repeat-independent mechanisms. For example, the presence of left-

handed helical Z-DNA structure can induce recombination events by altering chromatin organization [158]. Double strand breakage followed by non-homologous end joining (NHEJ) can also lead to gene amplification [44].

### **Phylogenetic Methods**

Another important area of genome evolution is to understand and estimate the evolutionary relations and distances between different genomes. There has been significant progress in the field of molecular phylogeny recently. According to the types of data used, the current phylogeny methods can be roughly divided into three classes: those that are based on sequence alignments, gene orders, or sequence compositions. Depending on the approaches used for evolutionary tree inference, there are, again, roughly three classes: maximum parsimony (MP), maximum likelihood (ML) and distance-based approaches.

The sequence alignment based methods have been advanced quite rapidly and are now easily accessible through various well-implemented program packages, such as PAML [189], PAUP [165], and PHYLIP [52]. The one-parameter Jukes-Cantor model [87] represents the first step towards modeling the substitutions between two aligned sequences. Since then, methods that contain more and more parameters have been invented to incorporate the complexity in the

sequence evolutionary process – from Kimura’s two-parameter model [95] that incorporates the difference between transition and transversion rate, to the HKY model that considers base composition bias [75], to the REV model that contains a set of eight parameters for all the possible reversible mutations between different bases and base compositions, and to the various complicated non-stationary models (reviewed in [66] and [7]). To account for the region-specific mutational rates, the substitution rate variation along the sequences has been further incorporated into the above models, mostly as a Gamma function (general review in [122]). Recently, codon biases have also been incorporated when dealing with sequences from coding regions [60][120]. Furthermore, instead of building separate trees for each different sequence, methods have been developed to infer phylogeny relationships from concatenated multiple sequences [59]. Although the complexity of the methods make them more faithful to the biology, it also leads to a significant increase in the parameter space, which results in larger variances in the computed results. It has been shown that for most data, the simple models often give more robust results [122].

In spite of a large body of literature, most of the sequence alignment-based methods focus only on the substitutions, and indels are often ignored or treated as separate characters from the alignment positions [91]. For example, assuming



there is no parallel or reverse evolution; one can treat the transposons in the genomes as individual characters, and use their insertion or deletion events to infer phylogenetic distance [21]. However, indel events (duplications and deletions) other than transposon insertions/deletions occur at very different scales and rates and also contribute significantly to the genome evolutionary process [63][132]. In addition, the non-parallel non-reversible evolution assumptions can be violated, i.e. the gaps in the sequence alignments may be generated by multiple events [15], and therefore should not always be treated as a single indivisible character. However, methods that can properly incorporate the indels into the evolutionary distance remain to be developed.

Another class of phylogeny methods is based on gene order. Such methods can be applied to species that have diverged a long time ago, since they only look at the order of the orthologous genes and not the specific sequences. However, there are two inherent difficulties with such approaches. First, the computational expense grows exponentially with the number of genes if one wishes to model the rearrangements (transpositions and inversions) properly and to find *the* correct answer. Second, events such as gene duplication, deletion and inter-chromosomal translocations are difficult to deal with in those methods. To simplify the computational task, Sankoff *et al.* [146][145] first proposed

the “breakpoint analysis” to compute the most parsimonious rearrangement distance between two sequences. Later, Prezner [73][9] developed an efficient polynomial-time algorithm to compute the reversal and transpositional distance between the gene order data. More recently, programs such as GRAPPA [119][167] and GRIMM [168] have further increased the efficiency of the breakpoint distance method and even extended it to a maximum likelihood (ML) approach. However, methods from this class are only applicable to small genomes (a few hundred genes) [167], and cannot handle sequences in which there are too many gene duplication or deletion events.

The third class of methods does not require the sequential information of the sequences under comparison, but only relies on the composition of the sequences. According to the scale of the composition unit, methods have been developed based on the contents of genes [65][78], protein domains or short nucleotide (amino acid) mers [137][162]. The gene based methods [65][78] have been mainly used to conjecture the contents of the common ancestral genome of divergent species [159]. The domain-based method [187] has been applied to infer the phylogenetic relation of the completed genomes. Although not directly relied upon for sequence alignments, both the gene-based and domain-based methods heavily depend on genome annotations, whose accuracy varies

with the underlying model and the amount of knowledge we have about the organism. The mer-based methods are the only class that does not require extensive pre-analysis on the sequence, therefore are more resistant to the errors in the sequencing or assembly data. The mer-based methods [137][162] have been used to reconstruct the phylogeny relationships of bacteria or phages. However, so far this class of methods is mostly empirical-based without theoretical explanations.

## **Markov Chain**

A better understanding of genome evolution lies in a deeper comprehension of the dynamics of its mechanisms: A limited number of evolutionary mechanisms with simple dynamics, through repetitions and interactions, can lead to an unlimited number of complex evolutionary paths. The increasing number of genome-wide datasets available today provides an unprecedented opportunity to study genome evolutionary mechanisms in more details, especially in a quantitative manner. In this thesis, I am interested in studying the duplication and other events in genome evolution by modeling them as Markov chains.

Markov chain is a useful tool to model statistical and random behaviors in physical and biological sciences. Given a finite or countable set of states  $E$ , a

Markov chain is, by definition, a stochastic process (i.e. a set of random variables)  $x_t$  such that its future depends only on the immediate past [49][172]. More precisely, for transition probability  $P$ , we have the following condition:

$$P(x_t = j | x_0 = i_0, \dots, x_{t-1} = i_{t-1}) = P(x_t = j | x_{t-1} = i_{t-1})$$

where  $j, i_0, \dots, i_{t-1}$  are states from  $E$ . Often we specify the initial distribution  $\mu$

where

$$P(x_0 = j) = \mu_j$$

and denote the one step transition matrix.

$$P(x_t = j | x_{t-1} = i_{t-1}) \text{ as } P_{ij}.$$

The  $n$  step transition matrix is defined to be  $P^n$ .

A Markov chain is irreducible if for any pair  $i, j$ , there exists an integer  $n$  such that  $P^n_{ij} > 0$ . Intuitively, this means that any two states of  $E$  communicate with each other eventually. A Markov chain is recurrent if the process starting from any state  $j$  will return to  $j$  at a later time with probability 1. A Markov chain is positive recurrent if the expected time of returning to any state  $j$  is finite.

For a given state  $i$  in  $E$ , let  $D_i = \{n | P^n_{ii} > 0\}$ . The greatest common divisor  $d_i$  of  $D_i$  is called the period of state  $i$ . If the chain is recurrent, the period is the

same for all the states. If  $d_i$  is equal to 1 for all  $i$ , then the chain is called aperiodic. A distribution  $\boldsymbol{\pi}$  is called the stationary distribution of a Markov chain, if

$$P\boldsymbol{\pi} = \boldsymbol{\pi}, \text{ and } \sum_j \pi_j = 1.$$

Suppose the Markov chain is irreducible, positive recurrent and aperiodic, then there exists a stationary distribution  $\boldsymbol{\pi}$ . Moreover, for any  $i$ , we have  $P_{ij}^n \rightarrow \pi_j$  in the limit as  $n$  goes to infinity.

However, most of the processes in evolutionary biology belong to the group of non-stationary Markov processes, in which the transition matrix changes over time, or depends on the current state, such that  $P(\mathbf{x}_t = \mathbf{j} | \mathbf{x}_{t-1} = \mathbf{i})$  is a function of  $t$  and  $\mathbf{i}$ .

### **Parameter Estimation**

In our models of genome evolution, various parameters need to be estimated by fitting the model to the data. The procedure of parameter estimation is important because the optimal parameters computed this way may reflect the underlying dynamics of the biological processes.

The problem of parameter estimation is to estimate an unknown probability density function based on observed data [143]. More precisely, given independent observations  $\{x_1, x_2, \dots, x_n\}$ , assuming that the probability density function  $f(x|\theta)$  is parameterized by a variable  $\theta$ , we want to find  $\theta$  which best describes the data  $x$ . A simple example is where  $f$  is equal to Gaussian distribution with unit variance, the parameter  $\theta$  which controls the mean, and  $\{x_1, x_2, \dots, x_n\}$  are the independent samples generated by this Gaussian distribution. In general, there are two ways for parameter estimation: the frequentist view and the Bayesian view.

The frequentist view does not assume any prior knowledge about  $\theta$ . We define the likelihood function to be

$$L(\theta) = f(x_1, x_2, \dots, x_n | \theta) = f(x_1 | \theta) \dots f(x_n | \theta).$$

Then we simply look for

$$\theta_{MLE} = \text{Arg max}_{\theta} f(x_1, x_2, \dots, x_n | \theta)$$

This method is called Maximum likelihood estimation (MLE) and the result  $\theta_{MLE}$  is called a maximum likelihood estimator.

In the Bayesian setting, we assume prior knowledge about how likely each

$f(\mathbf{x}|\boldsymbol{\theta})$  is relative to the true distribution. We quantify this knowledge as a prior distribution with prior density function  $p(\boldsymbol{\theta})$  on the parameter  $\boldsymbol{\theta}$  space. Given the data  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ , the posterior density function  $p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$  which quantifies our belief about the true  $\boldsymbol{\theta}$  after we observe the data is given by:

$$p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \frac{f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

To estimate  $\boldsymbol{\theta}$  in the Bayesian framework, one can choose  $\boldsymbol{\theta}$  that maximizes  $p(\boldsymbol{\theta}|\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ . Such an estimator of  $\boldsymbol{\theta}$  is called the maximum a posteriori (MAP) estimator. Other useful estimators include the posterior mean estimate and posterior median estimator.

When the prior knowledge is available and dependable, the Bayesian method is usually more accurate than the frequentist method in practice. However, when the prior knowledge is not available, the frequentist method is the reasonable choice. We like to point out that when the prior density function  $p(\boldsymbol{\theta})$  is uniform over  $\boldsymbol{\theta}$ , the MLE is the same as the MAP.

In the previous discussion, we fixed the model  $f$  of the distributions which generate the data. However, in practical problems, the choice of model space is often not unique. Based on the models we use, we obviously obtain different

distribution functions. An over-simplified model often gives a poor fit to the observed data, while a complicated model with many degrees of freedoms is often unstable and uninformative. An important principle of statistics and learning theory is called Occam's Razor, which says that one should not increase, beyond what is necessary, the number of entities required to explain anything, therefore the simplest model which fits the data reasonably well should be favored. We practice this principle throughout the thesis by using the most parsimonious model to explain complex sequence evolutionary problems. Although the models we use may be an over-simplification of the biological reality, by avoiding over-fitting, the parsimonious models can help uncover the most essential features of the underlying process.

## **1.2 Rationale**

Susumu Ohno first proposed the “*evolution by gene duplication*” theory in 1970's [124]. As I have discussed in the previous section, it has recently attracted a renewed attention. It is now widely believed that duplication is one of the key processes that create robustness, plasticity and novelty during genome evolution [71]. Furthermore, the duplication process may facilitate speciation events by random silencing of the duplicated genes or by mediating large genome



rearrangements [108][37][185][169]. Therefore, a necessary and important step towards a full understanding of genome evolution is to address such questions as how duplications happen, what effect they have on genome structures, and how they are modulated in association of other evolutionary processes. One practical use for such work is in cancer biology. The causal mechanisms for copy number fluctuations in tumor genomes may share some commonality with the mechanisms that caused the recent segmental duplications. The answer to the questions related to the duplication process may thus have implications for related medical research.

With the availability of whole-genome data, a great deal of work has been done at the genomic level, aimed at understanding the duplication process and its effect on genomic structures (as I have surveyed in the Background). However, many questions still remain open. I describe three of them further: First, although the recent segmental duplications in the mammalian genomes have been mapped [13][34][12][35][170], their mechanism is still unclear. Second, several models [186][79][20][174][19] have been proposed to study the effect of duplications on the genomic structure. However, most of these models are built to reflect the changes on higher-level cellular information, *i.e.* protein interactions [19], gene families [186], and expression clusters [20], which are hard to interpret using

basic evolutionary mechanisms. Third, despite the great advances in the development of phylogenomic methods, we still lack a method that can incorporate the effect of duplication and deletion (indels) events with the commonly used substitution estimation in measuring the genomic evolutionary distances. The absence of such distance measurement greatly hinders research on the modulation of different sequence evolutionary processes.

In this thesis, I present my work to understand and quantify the mechanisms of segmental duplications in the mammalian genomes using Markov models, and develop novel and efficient phylogenomic approaches incorporating duplication and deletion events based on our parsimonious genome evolution model.

### **1.3 Problem Statement**

The goal of the work presented in this thesis is to study the duplication process systematically by examining the different aspects of the process: the origination of the segmental duplications; the effect of the duplication process on the genomic structure; and the role of the duplication and deletion processes in genomic evolutionary distances.

**Origination of Segmental Duplication:** It is known that tandem duplications can be caused by unequal crossing over, and small interspersed duplications can be caused by retro-transposition [50]. However, a clear delineation of mechanisms responsible for the recent large segmental duplications found in the mammalian genomes remains elusive. The excessive content of interspersed repeats in the flanking regions suggests a repeat-recombination mechanism [8][10]. However, the involvement of the fast evolutionary dynamics of the repeat elements [105] and the possible errors in the genome assembly and duplication mappings [34] make it necessary to study the problem using carefully designed mathematical models that take into account these complexities.

**Effect of Duplication on Genomic Structures:** Although existing literature [128][30][111][56] have reveals interesting statistical features on genomic and proteomic data, no systematic analysis on different scales across a wide variety of species has been performed. Since the changes on different scales are mostly rooted in the changes in the genomic sequences, a general model at the genomic level that incorporates only the basic molecular evolutionary processes can be developed to explain the observations at different scales in one unified framework. Such a model could also help us study the relative frequencies of the

duplication events in comparison to other evolutionary events in various species.

### **Role of Duplication and Deletion Events in Genomic Evolutionary Distances:**

Most of the current phylogenetic distance measurements are based on the alignment of orthologous gene sequences from different genomes (see [122] for a general introduction). These types of methods suffer from three shortcomings: 1) the methods focus mostly on the coding sequences of the genes, and ignore the changes in the non-coding regions; 2) the measurements rely on the sequence alignment algorithms which assume no rearrangements in the sequence, and whose parameters should have reflected the real rates of the evolutionary processes, but are often chosen somewhat arbitrarily in practice; 3) the distances are usually represented only by the number of substitution events, while other events are assumed to be correlated and hence neglected. To study the role of duplication and deletion events during genome evolution and how they are modulated, there is a demand for an alignment-independent genome-wide measurement of evolutionary distance that incorporates the counts of all evolutionary events.

## **1.4 Contributions and Thesis Organization**

This thesis addresses the problems in various aspects of the duplication

process, including its origination, effect, and modulation in the evolving genomes. At the same time, we have also developed novel mathematical models and statistical tools that can be used in other similar genome evolution studies.

The main contributions of the thesis are:

- We test and quantify the repeat-recombination mechanism in the recent segmental duplication process in mammalian genomes using a Markov model that incorporates the evolutionary dynamics of the involved genomic elements and the possible errors in genome assemblies and duplication mappings [194]. The model suggests that about 30% of the recent human segmental duplications were caused by a recombination-like mechanism, among which 12% were mediated by Alu. A similar picture is found in the mouse and rat genomes. Therefore, in contrast to the previous research which suggested a larger role for Alu in these duplications [10], the recent segmental duplication in the mammalian genomes is shown to be a multi-mechanism process, and a significant proportion of the duplications is caused by some repeat-independent mechanism.
- Our further analysis on the physical features of the duplication flanking sequences suggests that one of repeat-independent mechanisms of segmental duplication in mammalian genomes may be shared by genetic

instability, and is related to physical instability in the DNA sequences [194].

- We examine the statistical structure of the genomes on different scales (from small oligonucleotide mers, peptides, to protein families), and of different organisms (from bacteria, archaea, to eukaryota), and find that the distributions of genomic components on different scales are all featured by the over-representation of high-frequency elements [193]. Since no significant difference in selection pressure has been found for mers with different copy numbers, such distributions can be explained by the effect of duplication process.
- Motivated by these generic features, especially the mer frequency distributions, we propose a parsimonious model for genome evolution [193]. Our model is reminiscent of Polya's Urn model, and contains three indispensable processes: duplication, deletion and substitution. The model can explain the frequency distribution of mers of different sizes in various genomes much better than previous models, and can be easily generalized to explain the statistics of other genomic components. The parameters in the model, when fitted to the real data, represent the average activity level of the three processes over the entire evolutionary history of the genome, reflecting the effect of modulation on the activity of the

process instead of the details of individual events during evolution.

- Based on our genome evolution model, we developed a novel alignment-independent phylogenetic method based on the mer (oligonucleotides of a certain length) frequency statistics in the sequences. In our method, we measure the evolutionary distance between two genomes by the total number of duplication, deletion and substitution events occurred since the divergence of the genomes from their common ancestor. This number is estimated from the changes in the mer copy numbers between two genomes using Maximum Likelihood approach. The comparison of the phylogenetic trees constructed using our method to the trees constructed using other methods can reveal the relative influences of different evolutionary processes to the genomic evolutionary distance.

This thesis is organized as follows. Chapter 2 describes our study on the mechanisms of the recent segmental duplications in the human, mouse and rat genomes, using statistical analysis and modeling on the duplication flanking regions. Chapters 3 and 4 describe the results from our systematic analyses on the statistical structure of various genomes, and a parsimonious genome evolution model inspired by our observation in the genome structure

studies. Chapter 5 describes the novel alignment-independent phylogenomic method based on mer statistics, its verifications and applications.



## **Chapter 2**

# **Mechanisms of Segmental Duplications in Mammalian Genomes**

### **2.1 Introduction and Related Work**

In all the genomes examined so far, duplications have been found in both coding and non-coding regions, and at various scales and ages [69][55][32][155][77][1]. In particular, the mammalian genomes are filled with duplicated sequences of different sizes. In the last few years, researchers have found that 3.5~5% of the human genome [13][34], 1.2~2% of the mouse genome [12][35], and 3% of the rat genome [170] contain recent segmental duplications (genomic sequence blocks whose identity level is higher than 90% and length larger than 1kb). Those segmental duplications are mostly interspersed, and were suggested to play a role in the domain accretion in the human genome [144] the dynamic large-scale rearrangement events during genome evolution [11][6], and various genomic disorders [51][85].

Nonetheless, a clear delineation of mechanisms responsible for those recent duplications in the mammalian genomes remains elusive: Unequal

crossovers usually cause tandem duplications. For example, L1 (long interspersed repeat family 1) elements have been shown to be involved as recombination breakpoints in large tandem duplications [53]. L1 retrotransposon machinery can also cause interspersed duplications smaller than 1kb [50]. Recently, a detailed analysis on the duplication breakpoints in a specific genomic region showed that some segmental duplications may have been caused by Alu-mediated recombination events [8]. Later, Bailey's group [10] reported that a significant portion of the interspersed segmental duplications terminated within an Alu repeat. These results led to the suggestion that the primate-specific burst of Alu retro-transposition activity is the primary cause of the recent boom of segmental duplications in the human genome [10]. However, given the highly dynamic nature of the Alu repeats in the recent past [105], estimation of its contribution to the segmental duplication process could be biased if its evolutionary dynamics are not taken into consideration.

Assuming that at least some of the duplication mechanisms has left recognizable sequence signatures in the duplication flanking sequences, in order to recover all sequence-dependent mechanisms, we started by mer analyses. Such a general analysis can point out the candidate hypotheses, which are followed by further verifications by rigorous statistical tests and mathematical modeling. Our

analyses suggested that there are multiple processes involved in the duplication mechanism.

In our mer analysis, we found an enrichment of Alu repeats and A(T)-tract sequences in the duplication flanking regions in the human genome. Such results indicate the involvement of interspersed repeats and regions with unusual physical properties of the DNA sequences in the duplication process. Following the clues from our mer analysis results, we further analyzed the content of various repeats in the flanking regions of the recent segmental duplications in three mammalian genomes. Consistent with the results from other groups [8][10], we found an overrepresentation of the most recently active interspersed repeats (Alu in human, and L1 in rodents), especially the younger subfamilies, suggesting that the recombination between homologous repeats may contribute to segmental duplications.

To quantitatively assess the relative contribution of Alu recombination mechanism to the process of segmental duplication without a prior bias, we developed a dynamic mathematical model which formulates the evolution of the repeat distribution in the duplication flanking regions (see **Figure 4** for the definition of flanking regions) as a Markov process with the time

measured by the divergence level in the duplicated sequences since duplication [194]. The results from the model suggests that although the duplication flanking regions may have been involved in Alu recombination significantly more often than pairs of randomly selected genomic regions, Alu recombination contributes to only about 10~12% of the segmental duplications in the human genomes. Therefore, the largest fraction of duplications remains unaccounted for by just recombination between interspersed repeats as we demonstrated through our computation.

Through a more detailed analysis of the sequence physical properties, we discovered that the regions flanking duplications are enriched for sequences with low helix stability and high DNA flexibility. These physicochemical properties also characterize sequences known to be “fragile” sites [115][118] for genetic rearrangement. Thus, segmental duplications may share a mechanism linked to genetic instability.

It is worth mentioning that we are keenly aware of the possible inaccuracies in the mapping data, especially at the duplication boundaries, and have performed analysis to estimate the errors in the current mapping of the segmental duplications in the human genome. Although we could not completely

exclude the effect of such inaccuracies, in the parsimonious design of our model minimized the influence of these errors. The results, as can be seen below, are quite robust in spite of the presence of these errors.

## **2.2 Methods**

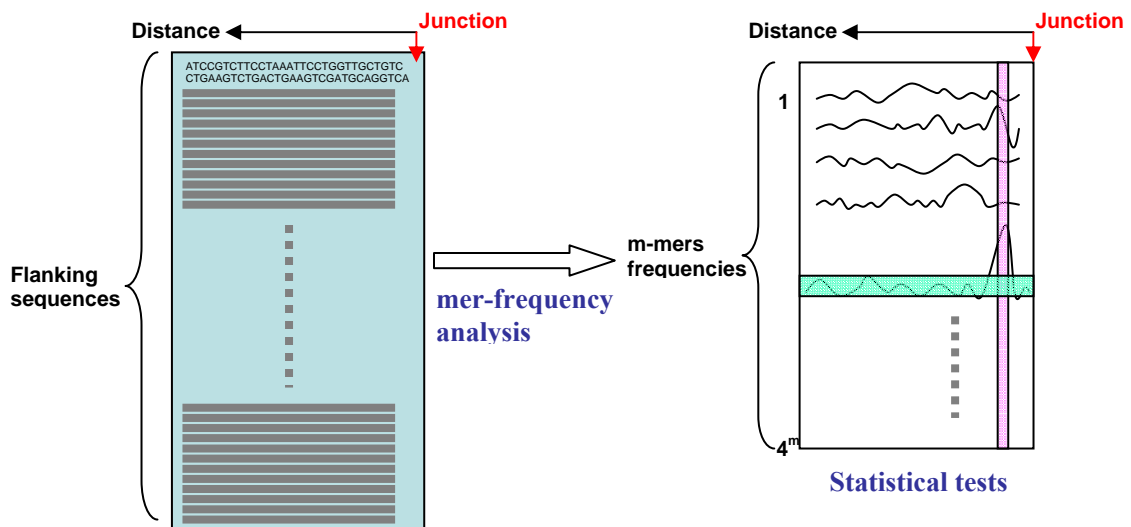
### **Sequence Preparation**

We used 4 different segmental duplication mapping datasets from 3 different mammalian genomes in our study: the July 2003 Human genome assembly (hg16) [34] [<http://chr7.ocgc.ca/humandup/>]; the April 2003 Human genome assembly (hg15) [13] [<http://genome.ucsc.edu/>]; the February 2003 Mouse genome assembly (mm3) [12]; and the June 2003 Rat genome assembly (rn3) [35]. To avoid redundancy and ambiguity, we only selected the duplication pairs that satisfy the following criteria: 1) only duplicated once; 2) cannot be included in any other duplications; 3) inter-chromosomal or at least 9kb apart; 4) longer than 6kb. A list of the filtered duplication pairs can be found at [<http://www.pnas.org/cgi/content/full/0407957102/DC1>]. Two control sequence sets are created for each dataset: One contains sequences randomly chosen from the corresponding genome assembly [<http://genome.ucsc.edu/>] to control the general genomic background noise; the other contains sequences randomly

selected from inside the duplicated regions to control the potential compositional bias in the duplicated regions.

### Mer Analyses

We calculated the frequencies of each possible 5- or 6-mer at each base-pair position relative to the breakpoints. A mer is considered over-represented at a particular position if its frequency at this position is at least 8 standard deviations higher than its mean frequency at other positions, and the absolute copy number at this position is above 20 (see Figure 1). To test the significance of the over-representation of certain mers in a region (a set of continuous positions), such as A-/T-rich mers, a two-sample *t*-test was used on the frequency values in the region of interest and the frequency values at all other positions.



**Figure 1.** The procedures of mer analysis in the segmental duplication

**flanking sequences. The flanking sequences are arranged into a file where their duplication junctions are aligned. The frequency of each  $m$ -mer ( $4^m$  in total) at each position, measured by the base pair distance to the duplication junction, is computed. For each  $m$ -mer, its frequency at each particular position is tested for significant over-representation by comparing it to the frequencies at all other positions using t-test. The significantly enriched mers as well as the position of their enrichment are recorded.**

The mers identified as over-represented at one or multiple positions were arranged according to their over-representing positions. Interestingly, most of the over-represented mers aligned well to form an Alu consensus sequence (see Figure 3). The analysis is repeated with all the Alu-containing sequences removed.

### **Repeat Analysis**

The repeats are identified according to the genome annotation database [<http://genome.ucsc.edu/>]. We consider a repeat as present in that flanking sequence, if its length is longer than a threshold (100bp in this study). For a pair of flanking regions to be identified as having a common repeat in a specific region (labeled as +/+), the repeat sequences have to be on the same side of the duplicated segments, in the same direction, and share at least 100bp of

high similarity. For Alu family, sequences from any subfamilies share high similarity [16][16]. For L1 family, however, only sequences from the same subfamily are found to be highly similar [156]. In our model, the frequency of +/- flanking region pairs in each age group is further normalized by subtracting the average frequency of repeats inside the duplicated segments, assuming that the repeats inside the duplicated region resulted from some recombination-independent mechanism and are uniformly distributed.

### **Markov Model of Duplication**

Our model tests the hypothesis that recombination between homologous repeats from a family  $X$  [e.g., Alu or L1] contributes to the recent segmental duplication processes in mammalian genomes. If some of the segmental duplication were caused by repeat recombination, these duplications should contain compatible repeat configurations (+/+) in its flanking regions right after the duplication events. On the contrary, if repeat recombination does not contribute to the segmental duplication process, the configurations of repeats in the flanking regions should be similar to those expected from randomly selected genomic regions. However, after the initial duplication events, the (+/+) repeat configuration in the flanking regions of those segmental duplications caused by repeat recombination starts to change due to mutation accumulation, forming



a specific pattern of distribution of configurations over time. Similarly, in the flanking regions of the duplications not caused by repeat recombination, the repeat configuration also changes and forms another specific pattern of distribution over time. How the repeat configuration changes over evolutionary time follows a Markov process determined by the dynamics of the repeat elements (see the transition matrix  $M_x$  below). Given enough time and assuming constant evolutionary rates, the repeat configurations in the flanking regions will reach a stationary distribution over different duplication age groups.

However, the stationary distribution varies depending on how many duplications were caused by repeat recombination: If no duplications are caused by repeat recombination, we expect to see a stationary distribution anticipated from an initial repeat configuration in two randomly selected genomic regions. In contrast, if repeat recombination contributes significantly to the duplication process, the stationary distribution will deviate from the former case. The deviated part should follow the stationary distribution anticipated from an initial repeat configuration of (+/+), and the degree of deviation should imply the portion of the duplications caused by repeat recombination. Therefore, to test our hypothesis, we need to compare the data against the null model, in which no duplications are caused by repeat recombination. The portion of

duplications caused by repeat recombination can be estimated using our model if the data deviate significantly from the null model.

## **Model Parameter Estimation**

### *Known Parameters*

All but two of the model parameters can be derived from the existing literature. They are enumerated in Table 1. We chose a flanking region size that is large enough to minimize the effect of mapping and annotation errors (by allowing some *gaps* and *shifts*, see **Figure 5**, **Figure 6**, **Figure 7**), and yet sufficiently restrictive to distinguish the signals from the genomic background noise. To establish the most appropriate size of the flanking regions to be used in the study, we applied the model to the datasets generated from several different flanking region lengths (200bp, 500bp, 1000bp and 2000bp). The estimation of repeat recombination, measured by the  $h_I$  value (see below for definition), reaches its highest in the 500bp and 1000bp datasets, thereby, suggesting these two sizes to be optimal choices. The data presented in this dissertation use a flanking region length of 500bp.

Rates from literature	Reference	Value
$\mu_g$ : average genome mutation rate	Waterston, <i>et. al.</i>	0.14%/Myr
$\mu_{Alu}$ : average mutation rate in Alu	Kapitonov, <i>et. al.</i>	0.16%/Myr
$\mu_{L1}$ : average mutation rate in L1	Smit, <i>et. al.</i>	0.23%/Myr
$\phi_{Alu}$ : Alu amplification rate	Liu, <i>et. al.</i>	0.86cous/Mb/Myr
$\phi_{L1}$ : L1 amplification rate	Liu, <i>et. al.</i>	0.32cous/Mb/Myr
Parameters in the model	Calculation	Value
$p_{Alu}$ : Alu frequency in random regions		0.130
$p_{L1}$ : L1 frequency in random regions		0.135
$c$ : divergence interval		1.0%
$\Delta t$ : time interval		3.57Myr
$\alpha$ : rate of entering an older age group <sup>1</sup>	$\epsilon/(2 \cdot \mu_g \cdot \Delta t)$	1
$\beta_{Alu}$ : Insertion rate of Alu in 500bp in $\Delta t$	$\phi_{Alu} \cdot \Delta t \cdot (L_{Alu} + 500bp - 2L_{Thr})$	$1.73 \times 10^{-3}$
$\beta_{L1}$ : Insertion rate of L1 in 500bp in $\Delta t$	$\phi_{L1} \cdot \Delta t \cdot (L_{L1} + 500bp - 2L_{Thr})$	$6.72 \times 10^{-3}$
$\gamma_{Alu}$ : Decaying rate of Alu in $\Delta t$	$\mu_{Alu} \cdot \Delta t/30\%$	$1.78 \times 10^{-2}$
$\gamma_{L1}$ : Decaying rate of L1 in $\Delta t$	$\mu_{L1} \cdot \Delta t/30\%$	$2.56 \times 10^{-2}$
$R_{Alu}$ : Background Alu distribution	$\begin{pmatrix} (1 - p_{Alu})^2 \\ 2(1 - p_{Alu})p_{Alu} + \frac{1}{2}p_{Alu}^2 \\ \frac{1}{2}p_{Alu}^2 \end{pmatrix}$	$\begin{pmatrix} 0.757 \\ 0.235 \\ 0.009 \end{pmatrix}$
$R_{L1}$ : Background L1 distribution	$\begin{pmatrix} (1 - p_{L1})^2 \\ 2(1 - p_{L1})p_{L1} + \frac{1}{2}p_{L1}^2 \\ \frac{1}{2}p_{L1}^2 \end{pmatrix}$	$\begin{pmatrix} 0.748 \\ 0.243 \\ 0.009 \end{pmatrix}$

**Table 1.** In the top half of the table we summarize the rates, the most recent reference for each rate, and its estimated value. In the bottom half of the table, we summarize the parameters, the mathematical formula to compute each parameter, and its computed value subsequently used in the model. The genomic data with segmental duplication mapping is then used to estimate the remaining two parameters,  $h_1$  and  $f_1$ .

The model that we propose is based on certain simplifying assumptions and estimated parameters taken from the existing literature; they are enumerated below: The model assumes constant rate of duplication, mutation, and repeat amplification. The sequence divergence rate  $\alpha$  in the duplication

pairs is estimated based on the average mutation rate in the human genome  $\mu_g = 0.14\%$  per million years (Myr) [182]. We take  $\alpha = \epsilon / (2 \cdot \mu_g \cdot \Delta t)$  because of the parallel divergence in the duplication pairs. The sequence diversity intervals ( $\epsilon$ ) between different age groups is set to be 1%, so that each age group contains at least 100 samples (see Appendix A for how the interval is chosen). These choices were based on our empirical studies showing that these sample sizes were needed for us to estimate the corresponding statistics within a reasonable error bounds. A detailed analysis on error bounds based on different sample size is listed in Table 10 in Appendix A. The range of the duplication pairs divergence is set to be between 0.5% and 8.5%. The duplications with lower divergence levels are omitted to avoid assembly or mapping errors. Given the duplication divergence rate ( $\alpha$ ) and interval ( $\epsilon$ ), the time interval is selected to be  $\Delta t = \epsilon / (2 \mu_g \alpha) = 3.57$  Myr. The amplification rates for Alu and L1 in humans are estimated in [105] by comparing syntenic sequences from human, baboon (which diverged from human  $\approx 25$  Myr ago), and chimpanzees (which diverged from human  $\approx 5$  Myr ago). Using those data, we estimated their amplification rates in a flanking sequence of size 500 bp, given that the insertion can result in a detectable repeat size that is larger than a threshold ( $L_{Thr}$ , 100 bp in this dissertation). If the duplication flanking regions contain two independently inserted repeats, they are counted as in state (+/+) only when the two repeats

are positioned in certain ways (see Repeat Analysis). Therefore, the rate at which duplication pairs change from (+/-) to (+/+) state is not  $\beta$ , but should divide  $\beta$  by 2 to correct for the appropriate orientation, assuming insertion occurs with equal probability in both orientations in the genome. If we assume that when the divergence levels of Alu's and L1's reach above 30% the repeats become unrecognizable for the RepeatMasker, and the divergence level in the repeats are uniformly distributed, then  $\gamma_{\text{Alu}} \approx \mu_{\text{Alu}} \cdot \Delta t / 30\% = 1.78 \times 10^{-2}$ , and  $\gamma_{\text{L1}} \approx \mu_{\text{L1}} \cdot \Delta t / 30\% = 2.56 \times 10^{-2}$ . The background repeat distribution  $R$  is computed from the corresponding repeat frequencies in randomly selected 500bp genomic regions.

Given the parameters in Table 1, we can write down the transitional probabilities between different states within a time interval  $\Delta t$  (see Table 1). The transition matrix can be expressed as follows:

$$M_X = \begin{pmatrix} 1-2\beta_X & \gamma_X & 0 \\ 2\beta_X & 1-\beta_X/2-\gamma_X & 2\gamma_X \\ 0 & \beta_X/2 & 1-2\gamma_X \end{pmatrix}. \quad (2.1)$$

The subscript  $X$  represents the repeat family, i.e. either Alu or L1. Vector  $A^{X}_{\cdot,k}(t)$  ( $k > 0$ ) represents the frequencies of flanking region pairs in the  $k$ th age group with different configurations of the repeats from  $X$  family at evolution time  $t$ .

( $A^{X}_{1,k}(t)$ : (-/-);  $A^{X}_{2,k}(t)$ : (+/-);  $A^{X}_{3,k}(t)$ : (+/+).  $\sum_{i=1\sim 3} A^{X}_{i,k} = 1$ .)  $A^{X}_{\cdot,0}(t)$  represents

the configurations of repeat  $X$  in the flanking regions of the new duplications at evolution time  $t$ .

The changes in the distribution of repeat configurations over evolutionary time follows a Markov process, which can be written as:

$$A_{,1:k}^X(t + \Delta t) = (1 - \alpha) \cdot M_X \cdot A_{,1:k}^X(t) + \alpha \cdot M_X \cdot A_{,0:k-1}^X(t).$$

$A_{,0}^X(t)$  is a constant vector over all  $t$ . In the “null” model,  $A_{,0}^X(t) = R_X$ .

### *Unknown Parameters*

$h_1 \equiv$  The portion of duplications caused by repeat recombination;

$h_0 = 1 - h_1 \equiv$  The portion of duplications not caused by repeat recombination;

$f_1^X \equiv$  The fraction of the recombination mechanism mediated by repeats  $X$ ;

$f_0^X = 1 - f_1^X \equiv$  The fraction of the recombination mechanism mediated by repeats other than  $X$ .

### *Unknown Parameter Estimation*

Assuming that the mechanisms and their relative contributions to the segmental duplication events are well conserved over a long period of time in the

mammalian genomes, we have a constant  $\bar{A}_{\cdot,0}^X$ , i.e.  $A_{\cdot,0}^X(t) = A_{\cdot,0}^X(s) = \bar{A}_{\cdot,0}^X$ , for any  $t > 0$  and  $s > 0$ .  $\bar{A}_{\cdot,0}^X$  depends on the two unknown parameters,  $h_1$  and  $f_1^X$ , in the model:

$$\bar{A}_{\cdot,0}^X = h_0 \cdot R_X + h_1 \cdot \begin{pmatrix} f_0^X \\ 0 \\ f_1^X \end{pmatrix} \quad (2.2)$$

We wish to estimate  $h_1$  and  $f_1^X$  from the observed distribution of  $X$  repeat configurations  $A_{\cdot,l;k}^X$  in the duplication flanking regions. The process our model describes can be written as the following equations:

$$A_{\cdot,l;k}^X(t + \Delta t) = (1 - \alpha) \cdot M_X \cdot A_{\cdot,l;k}^X(t) + \alpha \cdot M_X \cdot A_{\cdot,0;k-1}^X(t) \quad (2.3)$$

i.e. for any  $i \geq 1$ ,

$$A_{\cdot,i}^X(t + \Delta t) = (1 - \alpha) \cdot M_X \cdot A_{\cdot,i}^X(t) + \alpha \cdot M_X \cdot A_{\cdot,i-1}^X(t) \quad (2.4)$$

After a sufficiently long period of time, the process will reach a stationary distribution:

$$A_{\cdot,l;k}^X(t + \Delta t) = A_{\cdot,l;k}^X(t) = \bar{A}_{\cdot,l;k}^X \quad (2.5)$$

i.e. for any  $i \geq 1$ ,

$$A_{\cdot,i}^X(t + \Delta t) = A_{\cdot,i}^X(t) = \bar{A}_{\cdot,i}^X \quad (2.6)$$

Combining Equations 2.4 and 2.6, we get

$$\bar{A}_{\cdot,i}^X = (1 - \alpha) \cdot M_X \cdot \bar{A}_{\cdot,i}^X + \alpha \cdot M_X \cdot \bar{A}_{\cdot,i-1}^X$$

After rearrangement, we get

$$\bar{A}_{\cdot,i}^X = \alpha \cdot (I - (1 - \alpha) \cdot M_X)^{-1} \cdot M_X \cdot \bar{A}_{\cdot,i-1}^X$$

Therefore, given a constant  $\bar{A}_{\cdot,0}^X$ , the stationary distribution can be written as: For any  $i \geq 1$ ,

$$\bar{A}_{\cdot,i}^X = \alpha^i \cdot (I - (1 - \alpha) \cdot M_X)^{-i} \cdot (M_X)^i \cdot \bar{A}_{\cdot,0}^X \quad (6)$$

When  $\alpha = 1$ ,

$$\bar{A}_{\cdot,i}^X = (M_X)^i \cdot \bar{A}_{\cdot,0}^X.$$

Assuming that the currently observed repeat configuration distribution in the duplication flanking regions is a good approximation of the stationary distribution:  $A_{\cdot,l;k}^X \approx \bar{A}_{\cdot,l;k}^X$ , we may compute the optimal  $h_1$  and  $f_1^X$  that result in a  $\bar{A}_{\cdot,0}^X$  which can best explain the observed data  $A_{\cdot,l;k}^X$ . To find the optimal values, we performed an exhaustive search in the parameter space  $[0,1] \times [0,1]$  with precision of 0.01. Given a parameter set of  $(h_1, f_1^X)$ , the corresponding  $\bar{A}_{\cdot,0}^X(h_1, f_1^X)$ , and the stationary distribution  $\bar{A}_{\cdot,l;k}^X(h_1, f_1^X)$  are computed. Out of the  $101 \times 101$  sets, the optimal  $(h_1, f_1^X)$  is chosen by minimizing the square difference between the stationary distribution and the observed distribution  $\|\bar{A}_{\cdot,l;k}^X(h_1, f_1^X) - A_{\cdot,l;k}^X\|_2$ .



### *Estimation Results*

Cross-validation is performed to test the performance of the model and the confidence intervals of the estimated parameters (see *Model Evaluation*). The data from 5'- and 3'- flanking region pairs are analyzed separately, and the results are combined. The estimated unknown parameters ( $h_1$  and  $f_1^X$ ) in each dataset and their confidence intervals are listed in terms of mean  $\pm$  standard deviation.

### **Model Evaluation**

We use a cross-validation method to test the performance of the model and the confidence intervals of the estimated parameters. The complete dataset is randomly partitioned into two equal-sized groups: An in-sample set to estimate the parameters and an out-of-sample set to cross-validate and measure significance of the estimated parameters. In the result section, we report the mean values of the parameters estimated in 50 independent trials, as well as their standard deviations. The goodness-of-fit was tested in the out-of-sample data using the parameters estimated from the in-sample data. We use the maximum  $\chi^2$  error between the predicted trajectories and the experimental data as goodness-of-fit statistics:

$$X^2 = \text{MAX}_{j=1:3} \left( \sum_{i=1:k} \left( \frac{(O_{j,i} - E_{j,i})^2}{E_{j,i}} \right) \right).$$

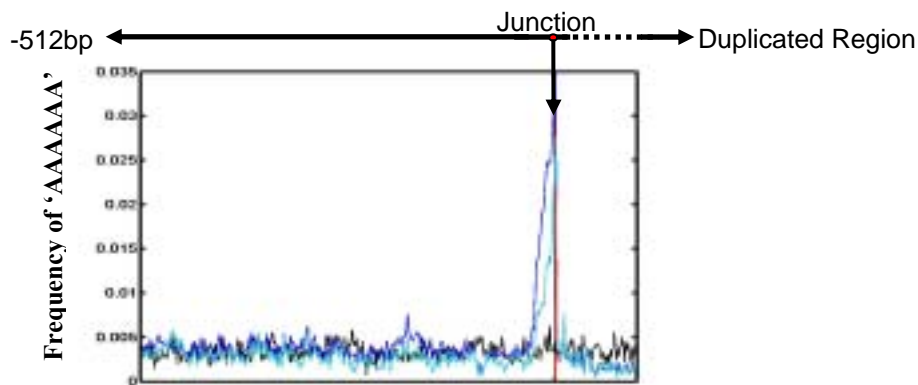
By  $O_{j,i}$  and  $E_{j,i}$  we denote, respectively, the observed and expected data trajectories. Index  $i$  indicates different duplication age groups. Index  $j$  indicates different repeat configurations [1: (-/-) ; 2: (+/-) ; 3: (+/+)]. The  $P$  value is computed as  $Pr(\chi_5^2 > X^2)$  for our model, and  $Pr(\chi_7^2 > X^2)$  for the null model. The degree of freedom is computed as:  $8(\text{age group number}) - 2(\text{free parameter number}) - 1 = 5$  for our model, and  $8(\text{age group number}) - 0(\text{free parameter number}) - 1 = 7$  for the null model.

### **Stability and Flexibility Computation**

The helix stability of DNA duplex is estimated by the average strand dissociation Gibbs free energy ( $\Delta G$ ) in overlapping 50bp windows, computed by the nearest neighbor model experimentally verified by Breslauer [26]. The DNA flexibility is estimated by the average twist angle in overlapping 50bp windows computed by the method in [148].

## 2.3 Mer Analysis on the Duplication Flanking Regions

In order to identify the possible sequence “signatures” left by the duplication mechanism, we analyzed the duplication flanking regions by computing the mer-frequency distributions in these sequences. All the flanking sequences were first aligned according to their position to the breakpoints. For each mer we create a profile of its copy number at each base-pair position in the aligned sequences. For example, for the mer ‘AAAAAA’, we computed how many times it appears at the first base-pair position in all the sequences examined, then how many times it appears at the second base-pair position, and so on (Figure 1). Assuming that the “signatures” contain characteristic sequences, and the distance between a certain “signature” and the breakpoints is conserved, we expect to identify the characteristic sequence by recognizing words that are overrepresented at specific positions.



**Figure 2.** The average frequencies of the 6-mer ‘AAAAAA’ in the flanking

sequences at different physical locations relative to the duplication breakpoint. All the flanking sequences are aligned in 5'-to-3' direction according to their breakpoint positions. The average frequency of 'AAAAAA' in the flanking sequences (blue) significantly increases in a small region immediately next to the breakpoints compared to the control set (black). The peak of the 'AAAAAA' frequency is only partially due to the enrichment of Alu sequences around the breakpoint. After all the flanking sequences containing Alus are removed from the analysis, the peak in the 'AAAAAA' frequency (cyan) continues to be highly significant.

Mers of length 5 and 6 are examined in this manner. The most striking observation was the strong enrichment of A(T)-rich words around the breakpoints (

Figure 2). There are other mers overrepresented at different positions as well. Interestingly, the majority of those overrepresented mers were found to align with each other and the aligned sequence turned out to be the Alu consensus sequence (Figure 3). Such a result reflects the intimate association between Alus and the recently duplicated regions—a point we will revisit in the following result section. It also attests to the power of the mer frequency analysis approach. When all the Alu-containing sequences were removed from the analysis, most mers were found not to be significant any longer, confirming that they were



**duplication flanking sequences are alignable and form a contiguous sequence (Query), which is highly similar to the second half of the Alu consensus sequence (Sbjct).**

Finally, we verify if we could discern any bias in the distribution of the sequences with those “enriched words,” particularly with respect to their chromosomal locations (e.g., intrachromosomal vs. interchromosomal duplication, pericentromeric/subtelomeric vs. euchromatic regions), but failed to find any with high enough statistical significance, suggesting that the mechanisms for interspersed segmental duplications are most likely the same for different regions in the genome.

## **2.4 Repeat Analysis on the Duplication Flanking Regions**

Inspired by the observation of Alu consensus sequence from the mer analysis, we further analyzed the repeat composition in the duplication flanking regions in more detail. Two assemblies of human genome (hg15 and hg16) were examined, as well as the two rodent genomes (mm3 and rn3) (see **Methods**). Consistent with the previous report on the human segmental duplications [13], we detected a significant over-representation of the repeats from the younger Alu subfamilies (AluY and AluS) in the flanking regions compared to random regions in

both human genome assemblies (**Figure 4**), but no significant over-representation of LINEs was detected. In the mouse and rat genomes, although no over-representation of the SINEs (B1, B2, ID and B4) is found (**Figure 4**), we detected a significant over-representation of the repeats from the younger LINE1 (L1) subfamilies in the flanking sequences compared to random regions (**Figure 4**). Therefore, both the human genome and the rodent genomes are enriched with the most recently active family of interspersed repeats in the duplication flanking regions (Alu in the human genome [183], and L1 in the rodent genomes [57][182]). The generality of the observation suggests that the recombination mediated by high-homology repeats may be a ubiquitous mechanism driving segmental duplications in all the mammalian genomes.

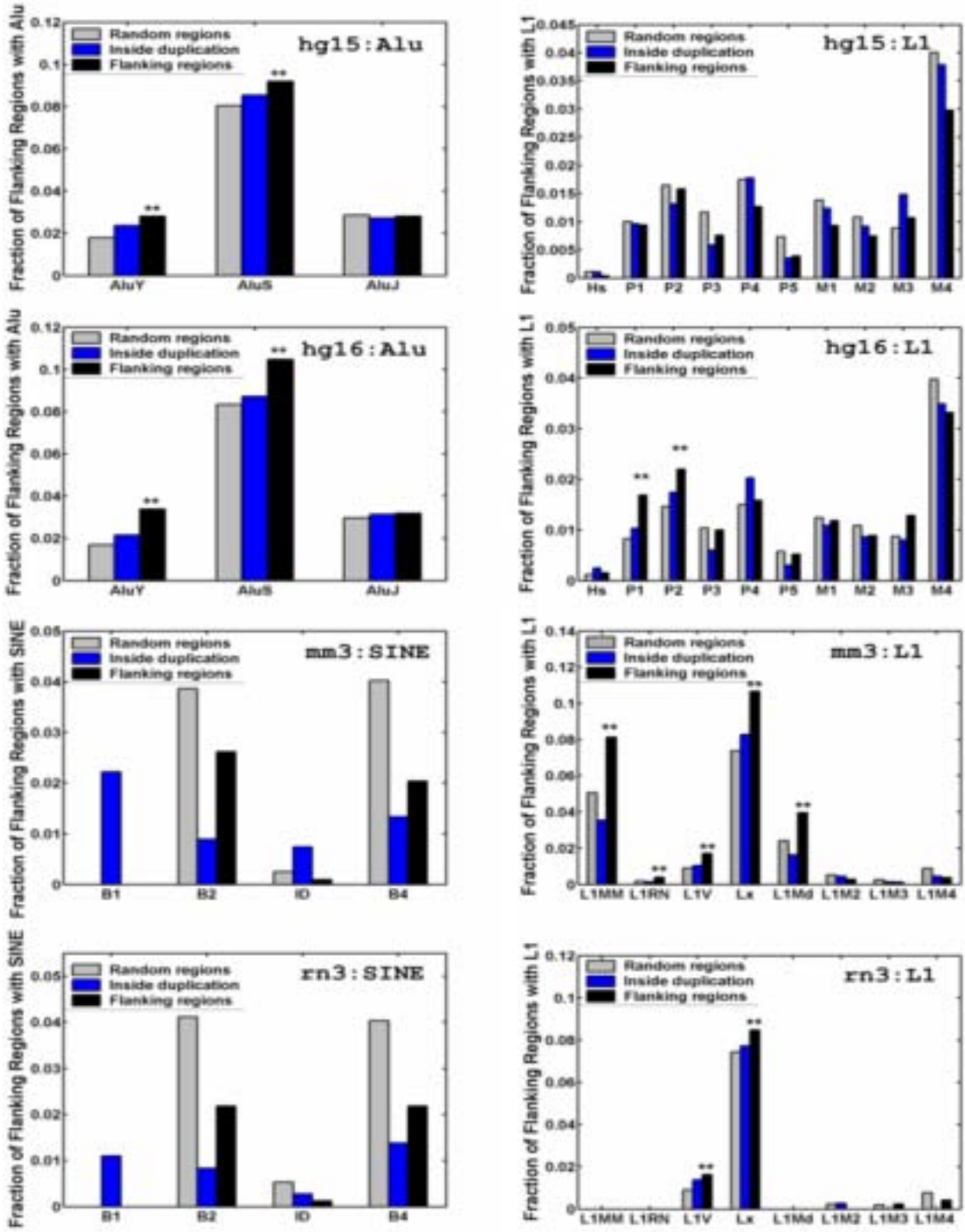
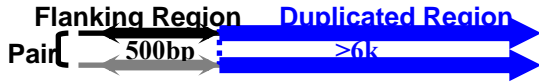


Figure 4. The appearance frequencies of various subfamilies of repeats detected in



the duplication flanking regions in the human (hg16 and hg15), mouse (mm3) and rat (rn3) genomes. The relationship between the flanking regions and the duplicated regions is shown in a pair of segmental duplications on top of the figure. In this dissertation the length of the flanking regions is 500bp, and the duplicated regions are longer than 6kb. The fractions of the flanking sequences containing different subfamily repeats are compared to the two control sets: sequences randomly selected from the whole genome, and sequences randomly selected from inside the duplication regions. The names of the different subfamilies of L1, Alu in the human genome, and SINEs in the rodent genomes are listed on the X-axis, roughly ordered according to their age (from younger to older). Two sample t-tests are used to test the statistical significance of the repeat overrepresentation in the flanking regions compared to the two controls respectively. \*\* The frequency in the flanking regions is significantly higher than both of the controls with  $p < 0.05$ . The statistics are based on the following sample sizes: hg16: random regions: 20918; inside the duplication regions: 13321; flanking sequences: 9788. hg15: random regions: 18864; inside the duplication regions: 9562; flanking sequences: 7652. mm3: random regions: 15824; inside the duplication regions: 6766; flanking sequences: 3288. rn3: random regions: 6274; inside the duplication regions: 3631; flanking sequences: 1652.

However, to test the above hypothesis, one needs to consider the highly active history of the over-represented repeats in the duplication flanking regions, and the reliability of the genome assembly and duplication mapping data.

Therefore, we conducted a detailed analysis on the hypothesis through a mathematical model that incorporates the evolutionary dynamic of the active repeats and minimizes the effect of assembly or mapping errors.

## **2.5 Error Analysis on the Current Mapping of Segmental Duplications in Human Genome**

Before analyzing the data in greater details, we first examined the degree of errors in the current mapping of the recent segmental duplications in the human genome that could be caused by the genome assembly, repeat annotation or duplication mapping procedure. In fact, as stated in [34], which reported the segmental duplication mapping in hg16, the accuracy of the duplication boundaries is within 500bp, and there are many mapped duplications involving assembly errors present in the earlier assembly version (hg15).

To assess the level of mapping and annotation errors, we conducted a detailed analysis on the “gap” and “shift” in the duplication flanking regions with matching repeats. Please see **Figure 5**, **Figure 6**, **Figure 7**. If the duplication mapping is accurate, the duplications caused by repeat recombination should contain no gaps or shifts; while the gaps and shifts in the duplications

caused by other mechanisms should have a distribution similar to those in the randomly paired genomic sequences. Furthermore, for duplications not caused by repeat recombination, the gap sequences should share no similarity and there should be no correlation between the shift sizes and the duplication ages. Therefore, homologous gaps, or association between duplication age and shift sizes are indications of map inaccuracies at the boundaries of the duplications caused by repeat recombination.

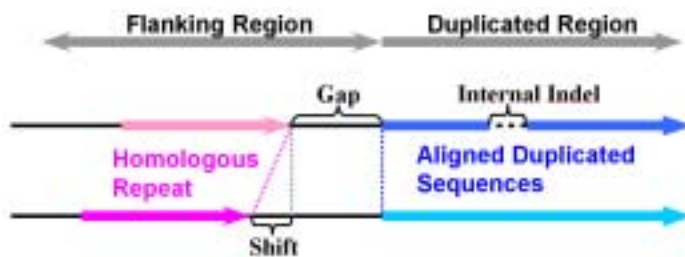


Figure 5. The figure represents schematically the definition of *gap*, *shift* and *internal indel*. *Gap* is the distance from the matching homologous repeats in the flanking regions to the duplication boundary. Gaps are expected when the duplication boundaries or repeat boundaries are not annotated precisely, or when there are concentrated accumulation of mutations in a small fragment within the duplicated region. *Shift* represents the difference in the positions of the matching homologous repeats in the flanking regions. Some of the large *shift* sizes may be caused by the random pairing of Alu repeats that are not related to duplications in the flanking regions. Others could be due to the insertions and deletions accumulated after the initial duplication event. *Internal indels* are the gaps in the alignment of the duplicated segments, which were caused by insertions and deletions after the duplication event.

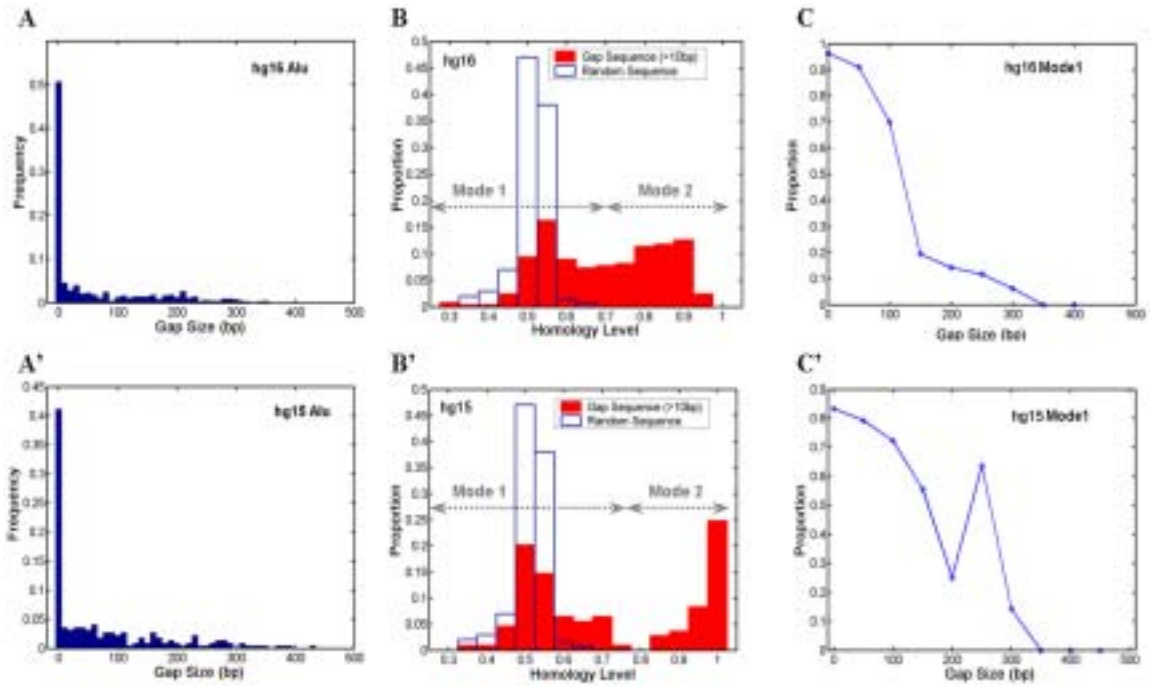


Figure 6. Analyses on the *gaps* between the matching homologous repeats in the flanking regions and the mapped duplication boundaries. We have measured the *gap* sizes in all the duplication pairs with matching Alu repeats (+/+) in their flanking regions. The distribution of the gap sizes is shown as a histogram (A & A'). Majority of the gaps have very small sizes (about 50% are smaller than 10bp). To characterize the larger gaps (>10bp), the gap sequences are aligned using dynamic programming (with score values: match=1; mismatch=-0.5; gap=-0.5). The homology levels (proportion of matched positions in the alignment) of the gap sequences estimated from the alignment results are displayed in B & B' (closed bars), in comparison with the estimated homology levels of random genomic sequences of similar sizes (open bars). The homology levels of the gap

sequences form a bimodal distribution (Mode1 and Mode 2). Mode 1 is similar to the random sequence results, and Mode 2 is significantly larger (close to 1). The presence of Mode 2 could be explained by the imprecision in the mapped duplication boundary positions. Mode 1 may have been caused by the random pairing of Alu repeats that are not related to duplications in the flanking regions, or by erosions in the duplicated sequences due to mutation accumulation after the initial duplication events. C & C' show how much (in proportion) of all the gaps within a particular size range has a relatively low homology level (those from Mode 1). There are very few long gaps that have low homology level, suggesting that most of the long gaps are due to inaccuracies in the boundary mapping. On the contrary, the proportion of the shorter gaps from Mode 1 is very large. Since it is more likely to get a by chance smaller fragment with low homology level under a fixed mutation rate by chance, the above observation is consistent with the presence of mis-mapped small fragments that are slightly more mutated but should be part of the duplicated regions. It is interesting to note that compared to hg15 dataset, in the hg16 dataset, the gap size distribution is more skewed towards zero; the bimodality in the distribution of the homology levels from larger gaps are less pronounced; and the proportion of larger gaps with low homology levels is smaller. These observations may suggest the improvement of duplication mapping in the later assembly version (hg16).

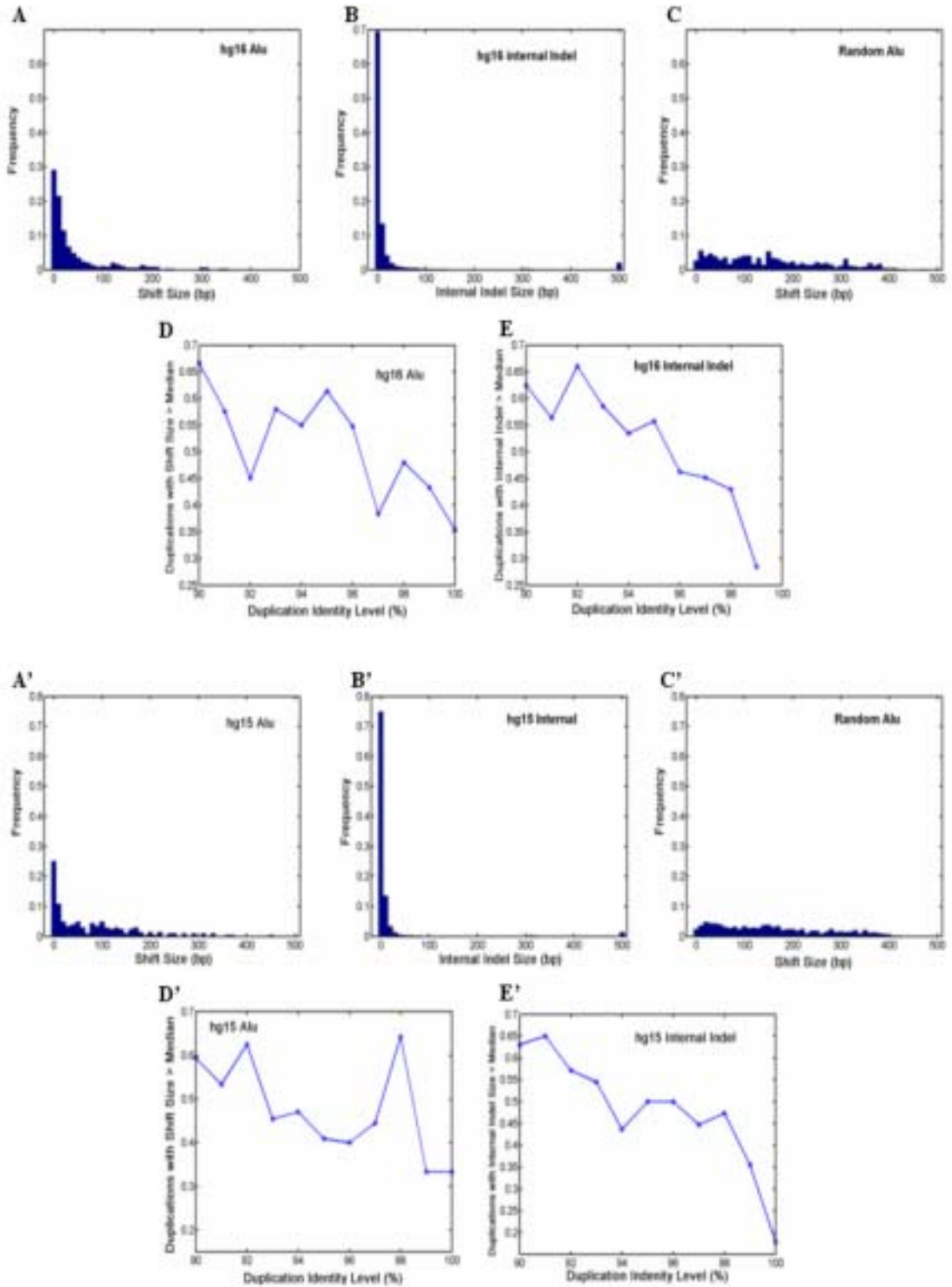


Figure 7. Analyses on the *shifts* between the positions of the matching

homologous repeats in the duplication flanking regions. The distribution of the shift sizes between the matching Alu repeats in the duplication flanking regions is displayed in (A & A'). In most cases, the shift sizes are small (about 70% is smaller than 40bp in hg16). The shifts in the positions of the matching repeats could have been caused by insertions and deletions after the initial duplication events, or the random pairing of Alu repeats that are not related to the duplication process in the flanking regions. To examine the expected shift sizes caused by insertion and deletion events after duplication, we analyzed the internal indel sizes in the duplication regions aligned by LAGAN [28] (B & B'). The shift sizes expected from the random pairing of Alus are computed from randomly paired genomic sequences of the same size (500bp) and are shown in C & C'. The distribution of shift size (A & A') has an intermediate shape between internal indel distribution (B & B') and random pairing distribution (C & C'): Both the shift size distribution and the internal indel distribution are skewed towards zero. However, the distribution of shift sizes is flatter, possibly due to the presence of randomly paired Alu repeats in the flanking regions (C & C'). Since we have included the random pairing case in our model (see Appendix), we expect that such cases will be removed as genomic background, and will not enter into our estimation of its contribution to the duplication process. To test our assumption that a considerable amount of the shifts are caused by insertions and deletions after duplication, we examined correlation between duplication age and shift size. As expected, shown in the age distribution of the duplications containing large internal indels

**(>median) (E & E'), older duplications (with lower duplication identity levels) are more likely to get larger insertion and deletions inside their duplicated regions. Similarly, older duplications also tend to have larger shift sizes (>median) in their flanking regions more often than younger duplications (D & D'). (The data in D & D' is more noisy due to a much smaller sample size than the internal indel dataset.) Therefore, the evolution of the shift in the flanking region is consistent with the evolution of insertion and deletion events inside the duplicated regions. Similar to our observation in Figure 6, in hg16 dataset, the shift size distribution is more similar to the internal indel distribution, and the correlation between duplication age and shift size is stronger than in hg15 dataset, suggesting a better duplication mapping because of the improvement on the assembly accuracy.**

We can summarize our observation by following conclusions:

- a) Gaps are rarely larger than 10 or 20 bps, and larger gaps often have high homologies. This is consistent with the fact that in some cases the boundary determined by mapping is incorrect and the true boundary is within the flanking region.
- b) The distribution of shift size in duplication flanking regions has an intermediate shape between the size distribution of indels inside the duplicated regions (internal indels) and the shift size distribution in randomly paired genomic regions: Both the shift size distribution in the



flanking regions and the internal indel size distribution are concentrated at zero and decay rapidly. The distribution of shift sizes in flanking regions is flatter, possibly due to the presence of randomly paired Alu repeats in the flanking regions. However, similar to the indel sizes inside the duplication regions, the lengths of the shifts in flanking regions are positively correlated with the age of the duplication event, suggesting that a considerable amount of the shifts are caused by insertions and deletions after duplication, and possibly lead to incorrectly mapped duplication boundaries.

- c) The number of large gaps and shifts reduces as assembly and mapping has improved (see **Figure 6**, **Figure 7**), further suggesting that they are possibly caused by the inaccuracies in the mapping and assembly.

The presence of these errors points to need for caution in the design and interpretation of the analyses on the duplication flanking regions. An analysis on sequences strictly at the mapped duplication boundaries underestimates or even worse, diminishes the signals left by the repeat recombination.

## **2.6 Markov Model of the Duplication Process**

The repeats that caused duplications by recombination should reside on the same side of the duplicated segment, in same orientation, and share enough

homologous sequences. Therefore, intuitively, we could directly estimate the contribution of repeat recombination to duplication by measuring the excessive level of such repeat configurations in the flanking regions of the newly duplicated segments before any erosion on the sequence occurs through mutation events. However, the newly duplicated segments are almost identical, therefore are most prone to genome assembly errors, making the estimations unreliable. On the other hand, if we use the “older” duplications, which are less prone to assembly errors; we could potentially over- or under-estimate the contribution of the repeats. —For instance, the actively amplifying transposable repeats can be inserted into the flanking regions after duplication, and form a configuration that falsely suggests a recombination event, resulting in overestimation of the hypothesis. Conversely, the repeats in the flanking regions can also lose its initial configuration after the recombination incident due to point mutations and deletions after duplication, consequently leading to underestimation of the hypothesis. Furthermore, the errors present in the current duplication mappings, as shown in the previous section, may also affect the observed pattern of repeat distribution.

To resolve the above dilemma, we incorporated the evolutionary dynamics of the repeats and the duplicated segments in our model. Over time, all the repeats in

the flanking regions, regardless of whether they have caused the duplication by recombination or not, are subject to changes in their configurations. Assuming that the mechanisms of segmental duplication and their relative contribution have been well conserved over time, the current repeat configuration in the flanking regions of duplications of different ages may be viewed as sampled from its stationary distribution. If the evolutionary rates of the repeats and the duplicated segments are known, the relative contribution of repeat recombination to segmental duplications can be estimated from the stationary distribution.

To minimize the effect of the errors in the current duplication mapping, in our model we use a flanking region size of about 500bp and count a repeat as present when its length is larger than 100bp. Such a less stringent criterion would allow the presence of small errors in the duplication mapping, thus avoiding underestimation. On the other hand, by taking into account the genomic background in our model, we “cancel” out the noises introduced by our less stringent criteria and avoid overestimation of counting randomly paired repeats not involved in recombination events.

Transitions within the Same Age Group			Transitions into the Next Age Group		
case (1a):		$(1-\alpha) \cdot (1-2\beta)$	case (1b):		$\alpha \cdot (1-2\beta)$
case (2a):		$(1-\alpha) \cdot 2\beta$	case (2b):		$\alpha \cdot 2\beta$
case (3a):		$(1-\alpha) \cdot \gamma$	case (3b):		$\alpha \cdot \gamma$
case (4a):		$(1-\alpha) \cdot (1-\beta/2-\gamma)$	case (4b):		$\alpha \cdot (1-\beta/2-\gamma)$
case (5a):		$(1-\alpha) \cdot \beta/2$	case (5b):		$\alpha \cdot \beta/2$
case (6a):		$(1-\alpha) \cdot 2\gamma$	case (6b):		$\alpha \cdot 2\gamma$
case (7a):		$(1-\alpha) \cdot (1-2\gamma)$	case (7b):		$\alpha \cdot (1-2\gamma)$

Table 2. The table lists all possible transitions between different states of the duplication flanking regions in a short evolution period  $\Delta t$ . The state of a flanking region pair is defined by both the configuration of the repeats (short arrows) in the flanking regions and the age group ( $k$ ) of the duplicated segments (long arrows), and is schematically displayed in the table. The left column lists all the possible transitions within the same age group ( $k$ ), and the corresponding transition probabilities. The right column lists all the possible transitions into the next (older) age group ( $k$  to  $k+1$ ), and the corresponding transition probabilities. The transition probabilities are expressed by the evolution rates of the repeats and duplicated segments:  $\alpha$ : the rate of duplicated segments evolving into an older age group in  $\Delta t$ ;  $\beta$ : the insertion rate of the repeat in the flanking regions by mechanisms such as retro-transposition in  $\Delta t$ ;  $\gamma$ : the decay rate of the repeats in the flanking

**regions due to mutations in  $\Delta t$ . (See Methods for details.)**

To explain the model, we begin by introducing some notations. In our model, each pair of the duplication flanking regions is assigned to a state specified by the configuration of the interspersed repeats in the flanking regions and the age of the duplication event. There are three possible repeat configurations in a pair of flanking regions (defined in **Figure 4**): The flanking regions may share a common repeat when they both contain a repeat from the same family in the same orientation and with sufficient length of homology (+/+) (see **Methods**); or one of them has a repeat, the other has no repeat or a repeat of different direction (+/-); or neither of them contains repeats (-/-). The ages of the duplication events are estimated by the sequence divergence level between the duplicated segments, and are grouped into bins with divergence interval  $\varepsilon$ . A flanking region pair is assigned to the age group  $k$ , if the corresponding duplicated segments have a divergence level of  $d$ , where  $k \cdot \varepsilon - \frac{1}{2}\varepsilon \leq d < k \cdot \varepsilon + \frac{1}{2}\varepsilon$ . The divergence interval is chosen to be  $\varepsilon=1\%$  based on sample size needed in each age group to draw statistical conclusions without being overly affected by corrupting noise (see **Appendix A** for details). This partition results in eight age groups, after omitting the duplications with extremely low divergence levels ( $d < 0.5\%$ ) because of their proneness to assembly errors. In the following text, we use the vector  $A^X_{\cdot,k}(t)$  ( $k > 0$ ) to represent the frequencies of flanking region pairs in the  $k$ th age

group with different configurations of the repeats from  $X$  family at evolution time  $t$ . ( $A^X_{1;k}(t)$ : (-/-);  $A^X_{2;k}(t)$ : (+/-);  $A^X_{3;k}(t)$ : (+/+).  $\sum_{i=1-3} A^X_{i;k} = 1$ .)  $A^X_{\cdot;0}(t)$  represents the configurations of repeat  $X$  in the flanking regions of the new duplications at evolution time  $t$ . Let  $h_I = 1 - h_0$  represent the fraction of the duplications caused by the repeat recombination mechanism, and among those let  $f^X_I = 1 - f^X_0$  represent the fraction mediated by repeat family  $X$ . (The product  $h_I f^X_I$  represents the relative contribution of the repeat family  $X$  to the duplications through the recombination-like mechanism.)  $A^X_{\cdot;0}(t)$  can be expressed using  $h_I$ ,  $f^X_I$ , and  $X$  repeat distribution in randomly paired sequences from the genome ( $R_X$ ) (for details, see Methods). Our model tests the following hypotheses: Null hypothesis ( $H_0$ ): recombination between repeats from family  $X$  does not contribute to segmental duplications, *i.e.*  $h_I f^X_I = 0$ ; alternative hypothesis ( $H_1$ ): recombination between repeats does contribute to segmental duplications, *i.e.*  $h_I f^X_I > 0$ .

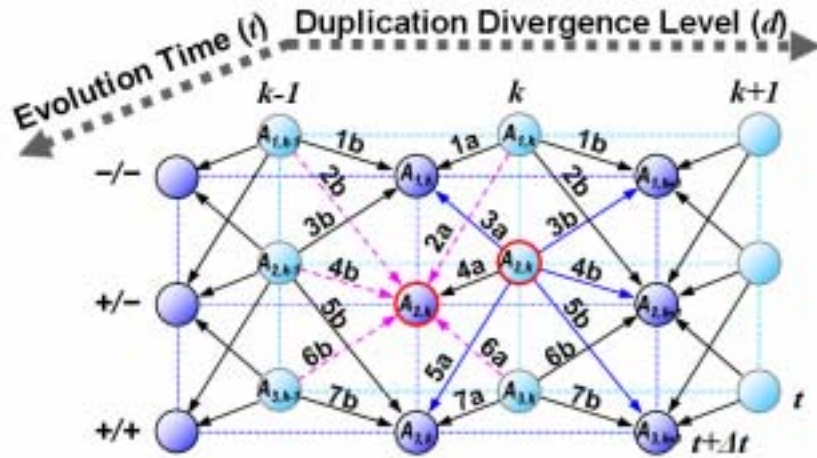


Figure 8. A schematic display of our mathematical model formulating the changes in the distribution of flanking region pairs over different states as a Markov process over evolution time. At a particular evolution time,  $t$ , the flanking region pairs are distributed over different states (circles), defined by the configuration of the repeats in the flanking region ( $--/--$ ,  $+/-$ , or  $+/+$ ) and the age group of the duplicated segments ( $k$ ). During evolution, in each time interval  $\Delta t$ , the flanking region pairs may change its state through many possible transitions (arrows). The change in the distribution of the flanking region pairs in a particular state at time  $t+\Delta t$  from time  $t$  depends on how much has entered into this state from other states, and how much has exited out of this state and into other states in interval  $\Delta t$  since time  $t$ . The in-flow and out-flow are the sum of the corresponding transition probabilities (1a~7a, 1b~7b), whose details can be found in Table 2. Take  $A_{2,k}$  (bold circled) for example, at evolution time  $t$  the flanking region pairs in state  $A_{2,k}$  can change into other states (grey arrows) in time interval  $\Delta t$ . At the same time, the flanking region pairs in other states can change into state  $A_{2,k}$  (dashed

arrows). The difference between  $A_{2,k}(t)$  and  $A_{2,k}(t+\Delta t)$  can be calculated by taking the difference between the sum of the outflows (grey arrows) and inflows (dashed arrows). Given enough evolution time, the process will reach the stationary state, in which the distribution over different states does not change with time any more, because each state has identical inflow and outflow. In the  $A_{2,k}$  example above, the sum of the grey arrows is equal to the sum of the dashed arrows in the stationary state.

The model describes the dynamically changing state distribution of the flanking regions as a Markov process over evolutionary time under the effect of accumulating mutations and repeat amplifications. Table 2 lists in details all the possible transitions between states in a small time interval ( $\Delta t$ ), as well as the corresponding transition probabilities expressed in the evolutionary rates of the repeats and duplicated segments. A schematic representation of the model, integrating the details in a small example, is displayed in **Figure 8**.

The model rests on two assumptions: First, the evolutionary dynamic rates and the mechanisms of segmental duplication as well as their relative contribution have been well conserved over a long period of evolutionary time. Second, the state distribution evolution in the flanking regions has reached its stationary state; *i.e.* despite the uninterrupted dynamic changes in the state of each individual



flanking region pairs, the distribution over different states among all the flanking region pairs stays unchanged. Formally, there exists a sufficiently large  $T$ , such that for any time  $t$  or  $s$  with  $t, s \geq T$ ,  $A^X_{\cdot,k}(t) = A^X_{\cdot,k}(s)$ , ( $k \geq 0$ ). For a detailed example of stationary states, see **Figure 8**. Under those assumptions, we can evaluate the two free parameters of the model ( $h_l$  and  $f^X_l$ ) based on the observed data, if the evolutionary rates are known (see Methods for details).

We apply the model to the duplication flanking regions in the human genome on the distribution of their states specified by repeats from Alu ( $X=Alu$ ) and L1 ( $X=L1$ ) families respectively, whose evolutionary rates have been well-characterized [105] (see Table 2). Two different datasets (hg15 and hg16) [13][34] are used. The free parameters in the model and their corresponding standard deviations are determined by cross-validation (see Methods). For both datasets (**Figure 9**), the model with the estimated parameters fits exceedingly well with the state distribution of the flanking regions specified by Alu repeats ( $p > 1 - 10^{-4}$  in the goodness-of-fit test, see Methods), while the null model (with  $h_l \cdot f^X_l = 0$ , see Methods) cannot explain the observed Alu distribution adequately ( $p = 0.04$ ). As expected, the null model can explain the L1 distribution in the flanking regions quite well ( $p = 0.86$ ), although the model with the estimated parameters can do slightly better ( $p > 1 - 10^{-4}$ ). See Table 3 for a list of the relative

contributions of Alu and L1 by recombination to the recent segmental duplications in human genome as estimated by the model.

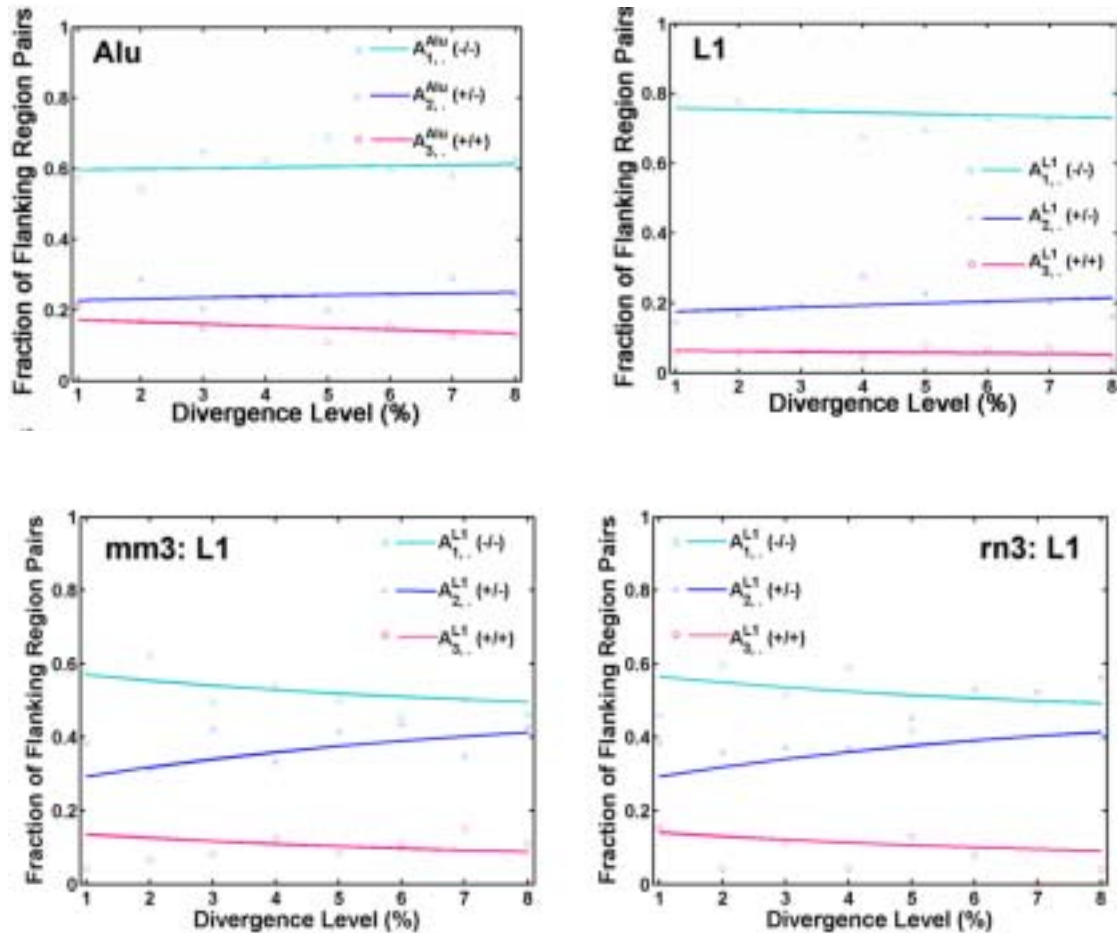


Figure 9. The fitting of the model to the distribution of Alu and L1 repeats in the duplication flanking regions in the human genome (hg16 shown here) and Mouse, Rat genomes. The fractions of flanking region pairs with different repeat distribution patterns are computed in each group of different sequence divergence levels ( $d$ ). We estimated the parameters and fitted our model to the distribution of Alu and L1 in the flanking region sequence pairs respectively. The various

symbols represent the real data, and the smooth lines are the theoretical trajectories of the model for the optimal choices of the parameters  $h_l$  and  $f_l$ . The total number of flanking regions pairs is hg16: 4894; mm3:1644; rn3: 826.

To further measure the significance of the contribution to the duplication process by the recombination in these two repeat families, we compared the estimated contribution ( $h_l f_l^X$ ) from the original dataset (*flanking*) to three control datasets: The permuted dataset (*permute*) is created by randomly switching the partners in the flanking region pairs while preserving the total repeat frequencies. The *outside* and *inside* datasets are obtained from positions farther outside or inside of the duplicated regions respectively. The results are listed in **Table 3**. As anticipated by the model, the estimated contributions in the *permute* and *outside* data, where random distribution is expected, are very close to zero; whereas in the *inside* data, where no random distribution is expected, the estimations are very close to one (**Table 3**). The contribution of Alu recombination to the duplication ( $h_l f_l^{Alu}$ ) estimated from *flanking* data is about 12%, which is significantly higher than the estimation from the *permute* and *outside* datasets. However, the contribution of L1 recombination estimated from the *flanking* is much lower, and do not differ significantly from either the *permute* or *outside* dataset. The estimated unknown parameters ( $h_l$  and  $f_l$ ) in each dataset

and their confidence intervals are listed below. All data are shown in terms of mean±standard deviation. Alu family repeats: hg15:  $h_I = 0.278 \pm 0.042$ ;  $f_I^{Alu} = 0.445 \pm 0.082$ ; hg16:  $h_I = 0.315 \pm 0.032$ ;  $f_I^{Alu} = 0.416 \pm 0.042$ ; L1 family repeats: hg15:  $h_I = 0.355 \pm 0.041$ ;  $f_I^{L1} = 0.084 \pm 0.026$ ; hg16:  $h_I = 0.330 \pm 0.048$ ;  $f_I^{L1} = 0.214 \pm 0.039$ .

<b>Dataset</b>	<b>Flanking</b>	<b>Permute</b>	<b>Inside</b>	<b>Outside</b>
Alu(hg15)	12.1±1.4%	0.5±2.3%	91.8±2.2%	3.8±0.8%
Alu(hg16)	12.9±1.0%	0.2±1.3%	92.5±1.3%	3.7±0.7%
L1(hg15)	3.1±1.0%	0.4±1.5%	92.1±1.8%	2.7±1.0%
L1(hg16)	6.9±1.1%	0.8±2.0%	92.7±1.4%	2.7±1.0%

**Table 3. The contribution of repeat recombination, estimated by the model from the datasets in different regions. *Flanking*: the original dataset from the duplication flanking regions. *Permute*: the permuted dataset from the flanking regions. *Inside*: the dataset from regions inside the duplication. *Outside*: the dataset from regions outside the duplication far (>3000bp) away from the breakpoint. All the data are shown in mean±standard deviation.**

The hg15 dataset and hg16 dataset were independently mapped by different research groups using different strategies [13][34], and it has been shown that the earlier map (hg15) contains more artifacts caused by assembly errors than the later one [13]. In spite of such differences, the model still gives consistent results between the two assemblies. It is also reassuring to find

that for both repeat families, the model estimated that the fraction of the duplications caused by recombination-like mechanism ( $h_1$ ) is about 30%, even though their contributions to the duplication mechanisms are quite different. The consistency in the parameter values suggests the robustness of our model against errors in assembly, mapping and annotation. This robustness is mostly due to the parsimony of the model, and the way in which the model accounts for a reasonable amount of errors and efficiently removes the corrupting noise.

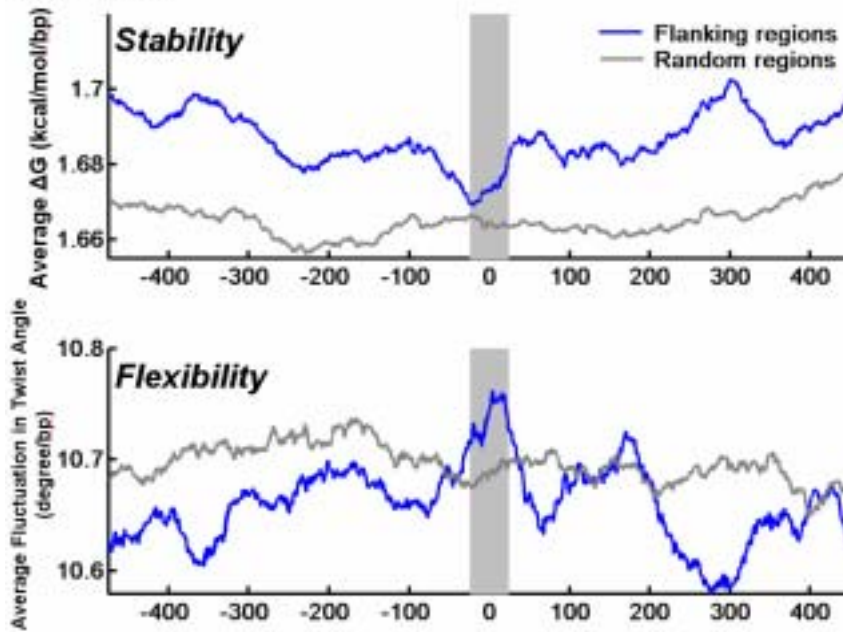
For the mouse and rat genomes, a good estimation of the evolutionary dynamic parameters of the interspersed repeats is still lacking. Furthermore, the available duplication mappings in the rodent genomes are likely to be less accurate due to the unfinished status of the genome assemblies [57][182]. Those factors prevented us from applying the model accurately to the rodent datasets as we did for the Alu and L1 repeats in the human genome. However, if one approximates the mutation rates in the rodent genomes by doubling the corresponding rates in the human genome and the rodent L1 insertion rate by tripling the human L1 insertion rate, then it is possible to reach a fairly good fitting for the L1 distribution in both the mouse and the rat datasets (**Figure 9**). The contribution of the L1 repeats to the recent segmental duplications through recombination-like mechanism is then estimated at about 10% in the rodent genomes.

In conclusion, in all the mammalian genomes examined, our model estimates that about 10~12% of the recent segmental duplications were caused by the recombination between the most active interspersed repeat elements in the genome (Alu in human, and L1 in rodents). The results from the model further suggest that the segmental duplications are likely to be caused by multiple mechanisms, and a large fraction (~70%) of the duplications are caused by some unknown mechanism independent of the interspersed repeat distributions, which is consistent with the conclusions of [192].

## **2.7 Physical Instability in the Duplication Flanking Regions**

Apart from the repeats, in our mer-based statistical analysis we also discovered an enrichment of DNA sequences that are physically unstable around the duplication boundaries in our mer analysis. The physical properties of the DNA duplex plays an important role as the initial step in many molecular processes, as shown in transcription [18], replication [135], and the large genome rearrangement events that originated from the chromosomal “fragile” sites [115][118]. Therefore, it is possible that similar properties can initiate or facilitate the segmental duplication process in the mammalian genomes.

## Human



## Mouse

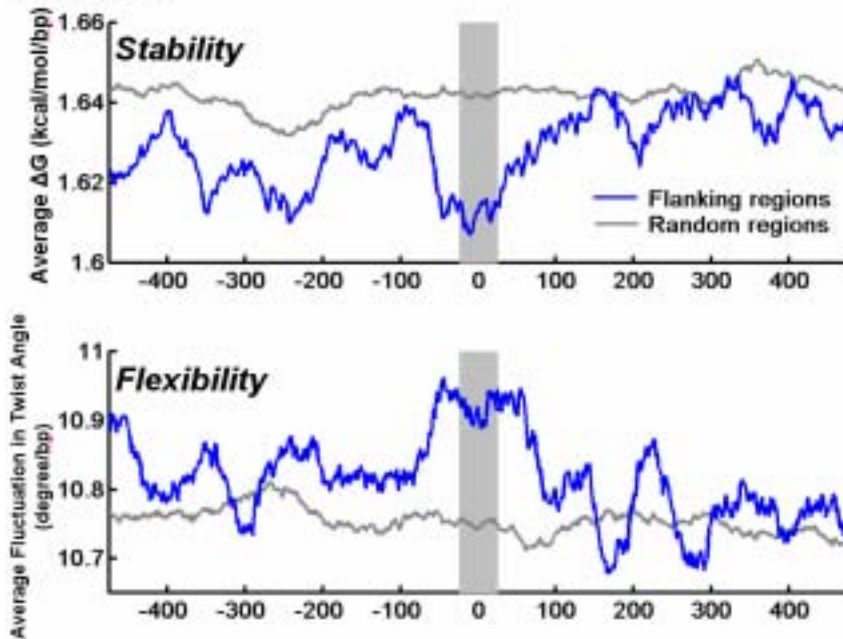


Figure 10. The helix stability and the DNA flexibility in the repeat-less flanking sequences in the human (hg16 shown here) and mouse genomes. The average

helix stabilities in the flanking regions around the duplication junction (darker line) and the repeat-less random genomic regions (lighter line) are estimated by the average  $\Delta G$  in overlapping 50bp windows. The average DNA flexibility in the flanking regions around the duplication junction (darker line) and the repeat-less random genomic regions (lighter line) are estimated by the average fluctuation in the helix twist angle in overlapping 50bp windows. The shaded regions indicate the duplication junction where there is a slight decrease in the helix stability and a slight increase in the DNA flexibility. The mapped duplication boundary is at 0bp; the negative bp positions are coordinates outside the duplicated region; and the positive bp positions are coordinates inside the duplicated region.

To explore possible repeat-independent explanations and to avoid the bias introduced by the AT-rich regions in Alus and L1s, we analyzed the flanking sequences that do not contain any repeats for their helix stability [26] and DNA flexibility [148] (See **Methods** for details). These two features are suggested to be the specific characteristics of the “fragile” sites in the genome, where genetic rearrangements frequently occur [115][118]. In both the mouse and human datasets, there is a slight decrease of the average helix stability and increase of the average DNA flexibility at the duplication junction compared to the other regions either inside or outside the duplicated segments (see **Figure 10**). To test the significance of these observations, we counted the number of duplication



junctions (-250bp to +250bp flanking the boundary) that contain sequence sites with both exceptionally low helix stability and exceptionally high flexibility. The criteria for recognizing such a site is that the average  $\Delta G$  in its centering 50bp window is smaller than 1.3kcal/mol/bp (the bottom 0.5% of random genomic sequences) and the average fluctuation in twist angle is larger than  $14^\circ$  per bp (the top 0.5% of random genomic sequences) [115][118]. We found enrichment of such potential “fragile” sites in the repeat-less duplication flanking regions from all of the three mammalian genomes compared to the randomly selected genomic regions (**Table 4**). The enrichment of these characteristic sites is statistically significant in all the datasets, except in the rat genome where it is just on the verge of being significant. Interestingly, the significance level increases with the degree of finishing of the genome assemblies, suggesting that the lack of significance in the rat genome could be explained by its most primitive status of the current assembly. It is worth noting that because of the presence of errors in the current duplication mappings, the proportions of the duplications with “fragile” site like properties may be underestimated here.

The methods we used to calculate helix stability [26] assumes that the free energy of a duplex results from the sum of its nearest-neighbor interactions with some pre-assigned initiation free energy for different base pairs, and

represents the free energy at a particular condition (25°C in a pH7 1M NaCl buffer). The method and data used to compute DNA flexibility were obtained by computational method minimizing the free energy without the consideration of the backbone [148]. We chose those methods to be consistent with the previous work. However, more updated methods and measurements [147][190] can be used to repeat the analysis. With improved mapping and measurement of flexibility and stability, a more exact picture will emerge.

<b>Genome</b>	<b>Flanking</b>	<b>Random</b>	<b>Fold</b>	<b>P-value</b>
Human(hg16)	4.82% (2052)	1.99% (2964)	2.42	$<10^{-7}$
Human(hg15)	3.81% (2863)	2.41% (5280)	1.58	$<10^{-5}$
Mouse(mm3)	3.68% (815)	2.51% (2632)	1.47	$<0.05$
Rat(rn3)	4.21% (570)	2.76% (1123)	1.53	0.07

**Table 4. The enrichment of the “fragile” sites in the repeat-less duplication flanking sequences in different mammalian genomes. The table lists the fractions of the flanking and random regions containing “fragile” sites, the total number of sequences examined (in parenthesis), and the folds of enrichment (Fold). The significance of the enrichment (*p*-value) is computed using two-sample test for binomial proportions.**

The over-representation of sequences with physical features similar to the “fragile” sites in the duplication flanking regions suggests that segmental duplications may share a mechanism linked to genetic instability. While this is the first evidence for the hypothesis that some repeat-independent

mechanism is involved in the recent mammalian segmental duplications, the hypothesis needs to be explored further.

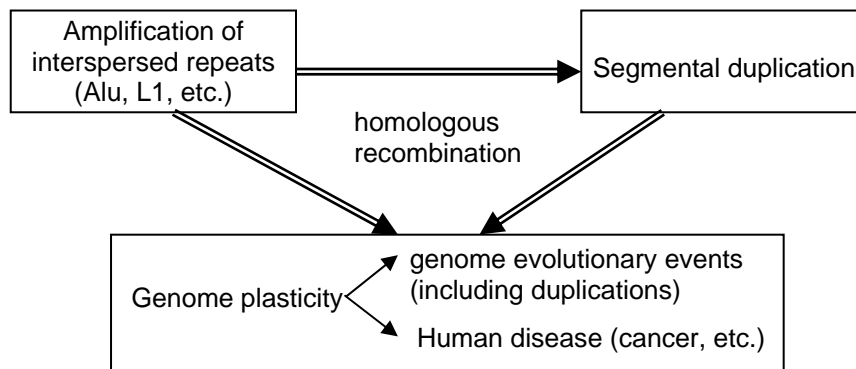
## 2.8 Summary

From previous studies [34] as well as our detailed analysis on *gaps* and *shifts* in the duplication flanking regions (see Figure 6, Figure 7), we conclude that the current map of segmental duplications is still tainted with errors from assembly, mapping and annotation. In the presence of these errors, an analysis on sequences strictly at the mapped duplication boundaries will underestimate or even diminish the signals left by the repeat recombination. Using a flanking region size that allows some *gaps* and *shifts* helps us to minimize the effect of these errors on our analysis. In addition, by incorporating our knowledge of the related evolutionary processes in the dynamic model, it was possible to decrease the effect of random noise. Therefore, in spite of the nature of the data, our method was found to be quite robust. Of course, the accuracy of the results will increase with the finishing stages of the genome assembly and the improvement on the mapping and annotation schemes.

The human genome has significantly more interspersed segmental duplications than the rodent genomes [57][182][183]. It was suggested that

the difference is due to the recent burst of primate Alu retro-transposition activity [10]. However, the preliminary estimations from our model suggest that the relative contribution from the most active repeats through the recombination-like mechanism remains more or less constant in the human and rodent genomes. Therefore, the answer to why the genomes have different amount of segmental duplications is to be sought elsewhere (for example, the difference in the tolerance for large duplications, the difference in effective population sizes, or the finishing stage of the genome assembly [153]).

Although recombination between repeats cannot explain all the segmental duplications, it did contribute to the process significantly, as we have estimated using our model. If we view the amplification of repeats also as a duplication process, then duplication, as a dynamic process, has demonstrated its avalanche effect on the genome evolutionary dynamics by creating a positive feedback mechanism through the recent segmental duplications in the human genome (**Figure 11**). The amplifying interspersed repeats in the human genome can drive evolution both directly by homologous recombination between themselves, and indirectly by causing other duplications which can lead to further recombinations.



**Figure 11. The schematic representation of the propagating effects of duplication during recent human genome evolution. The recent amplification (duplication) of interspersed repeats (Alu, L1, etc.) populated a significant portion of the human genome with segmented regions of high levels of similarity. These interspersed homologous repeat sequences provided potential sites for recombinations between different parts of the genome, thus conferring a basis for genome plasticity and causing various genome evolutionary events, including large segmental duplications. The segments duplicated in this manner could in turn mediate yet more recombinations between their homologous copies, leading to further evolutionary events, including duplications in multiple scales. Hence, duplication creates a positive feedback dynamic, and makes genomic evolution a self-driven process.**

Segmental duplications have been shown to be associated with both genome rearrangement events during species evolution [6][11] and the copy number

fluctuations [107][134][104][160] and other rearrangements [101] in genomic sequences during cancer development. Therefore, some of the mechanisms used by segmental duplications, such as recombination mediated by interspersed repeats [97][164] may be shared by other genomic rearrangement events. Suggested by the “fragile” sites we found in the duplication flanking sequences and their association with the breakpoints of the syntenic blocks [6][11], perhaps, another common mechanism could be correlated to the specific physical properties in the DNA sequences. In fact, it has been suggested that segmental duplications in yeast are caused by breakage-induced-replications induced by replication fork stalling at the AT-rich replication termination sites [164]. These are topics of future research that may rely on mathematical models akin to the ones proposed here.

## **Chapter 3**

### **Statistical Structure of the Genomes**

#### **3.1 Introduction and Related Work**

Genome evolution is a dynamic process driven mainly by the changes in the genomic sequences. It can further lead to changes in the cellular information at higher levels (transcriptome, proteome, interactome, etc.). Various historical evolutionary events leave their “signatures” in the present sequences, which can be deciphered by statistical analyses on a family of genomes that are currently available. Although different organisms can experience completely different selection pressure, the dominating evolutionary processes may leave common signatures in the genome structures of all the organisms.

A survey of the literature reveals many interesting statistical analyses of various kinds on genomic and proteomic data. Among the large collection of results, the most interesting ones are those pointing to a universality seen through statistical characteristics, shared by data from all organisms, from different cellular processes, as well as at various scales.

**Long-range correlation (LRC) in genomic sequences:** Correlations are the result of interactions between different constituents of a system. When a sequence has LRC, it suggests that the interactions extend within the entire system. Since last decade, LRC between single nucleotides have been found to be persistent (positively correlated) and pervasive in the genomic sequences of various species from different kingdoms, and in different regions of the genomic sequences [128][30][126][195].

**Nonrandom distribution of various elements on genomes:** The physical locations of the various DNA elements on the genome are not randomly distributed. For example, fractality (fractional dimension) has been detected in the juxtaposition of coding and non-coding regions in higher eukaryotic genomes [3]. The physical locations of Alu's in human genome also show a highly nonrandom pattern [161]. (This deviation from randomness may be explained by the distribution of the target sequences for insertion).

**Linguistic features in genomic and proteomic sequences:** It was shown that the mer (oligonucleotides of a particular length) frequency distributions in DNA sequences, both coding and non-coding, and amino acid sequences have a statistical feature similar to Zipf's law, as observed in natural languages



[111][56]. Such a distribution, when compared with a random distribution, has an over-representation of words of both high and low frequencies.

### **Scale-free property of cellular pathways and protein interaction networks:**

The large-scale gene networks and protein interaction networks in some model organisms are presently available, e.g. metabolic networks in *E. coli* [84], protein interaction networks in yeast [151] and *H. pylori* [139]. The topology of those networks was found to be characteristic of a group of graphs known as scale-free networks [83][113]. Scale-free networks are characterized by their “hubby” structures associated with a power-law distribution of their connectivities, and can be created by an evolution process following a “rich gets richer” rule.

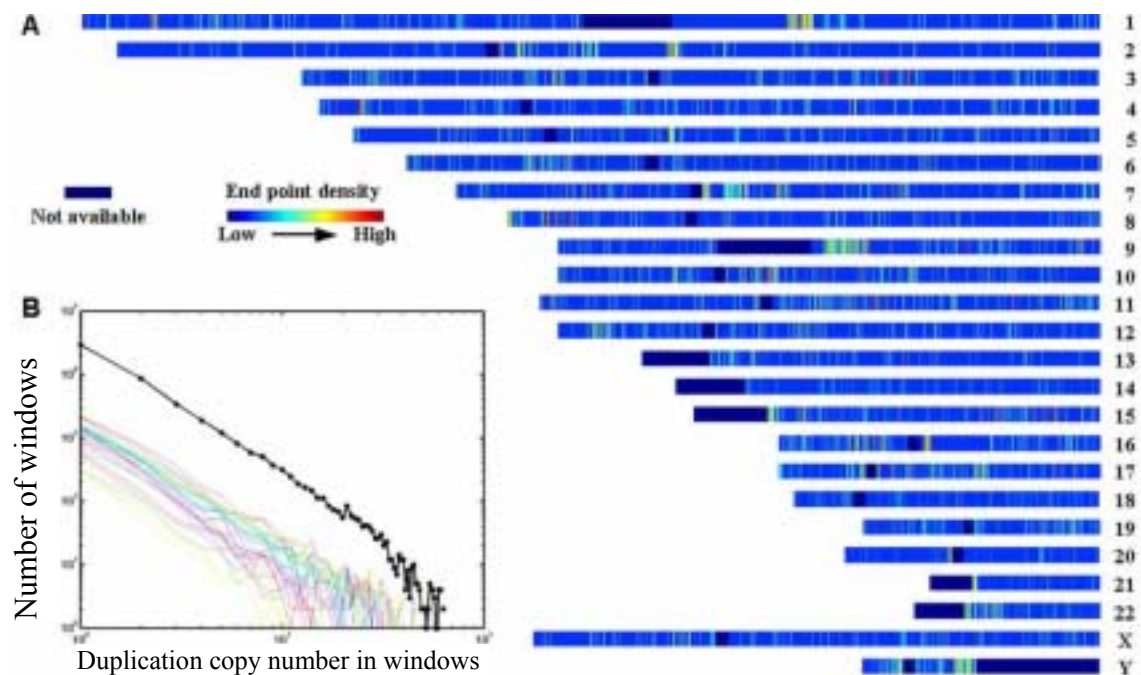
At first glance, the statistical features discussed above may seem to be unrelated at one glance. In fact, there is a generic and deep connection among them – all of them can be created by an evolutionary process with positive-feedback mechanism, such as the duplication process. *The duplication process we refer to here as well as in the following chapters has the general definition of a process that leads to the duplication of a sequence segment. Therefore, it includes more specific processes such as tandem duplication, segmental duplication, and even transposition.* Although some models based on duplication have been

proposed to explain the observed statistical features [186][79][20][174][19], no systematic analysis on different scales across a wide variety of species has been performed. To fully appreciate the significance of duplication process at many levels of biology, we have systematically surveyed the distributions of various genomic components at different scales (from mers, peptides, to protein families), and in different genomes (from bacteria, archaea, to eukaryota). To study if there are alternative mechanisms other than duplication that can cause the observed shape in the distributions, we also examined the effect of repeats and selection on the mer copy number distribution. But neither repeats nor selection can fully explain the observed pattern in mer distribution, suggesting that duplication is the main mechanism that has caused these statistical features in the genome. A better understanding of the genome statistical structure can also provide new insights into the study of comparative genomics, which can potentially lead to better phylogenomic methods.

### **3.2 Analysis on the Distribution of the Segmental Duplications in Human Genome**

Recently, intensive large segmental duplications (both intra- and inter-chromosomal) have been reported in the assembled human genome, and the potential large duplicated regions (>500bp, >95% identity) have been

mapped out in pairs under the standard sequence homology criteria [13][34][12][35][170]. In the previous chapter we have studied their potential mechanisms by statistical analysis and modeling on their duplication flanking sequences. To further study the dynamics of the segmental duplication process, we examined how the duplication frequency is distributed along the human genome. The average duplication copy number in each non-overlapping window of a fixed size is computed along the genome. The histogram of the duplication copy number in each window over a chromosome or over the whole genome gives a power-law distribution (**Figure 12**), implying that the genomic regions previously duplicated tend to be duplicated more often.



**Figure 12.** The distribution of the potential duplication 'hot-spots' on the

human genome. **A. The distribution of the duplicated segment end-points on the chromosomes (over windows size of 1Kb). The 'hot-spot' density is color coded (see the color bar). The dark areas represent chromosomal regions where no reference sequences are available. There is a tendency for areas with high densities to cluster together on the chromosomes. B. The distribution of the duplication copy numbers on a log-log scale. The X-axis shows the duplication copy number in non-overlapping windows of size 1Kb, starting from 1. The Y-axis indicates the number of non-overlapping 1Kb windows containing a given copy number of duplications. It is clear that the  $\text{Log}_{10}(\text{duplication copy number})$  and  $\text{Log}_{10}(\text{number of windows})$  form a linear relationship, both in the whole-genome range (thick black points) and in individual chromosome range (multiple thin colored lines).**

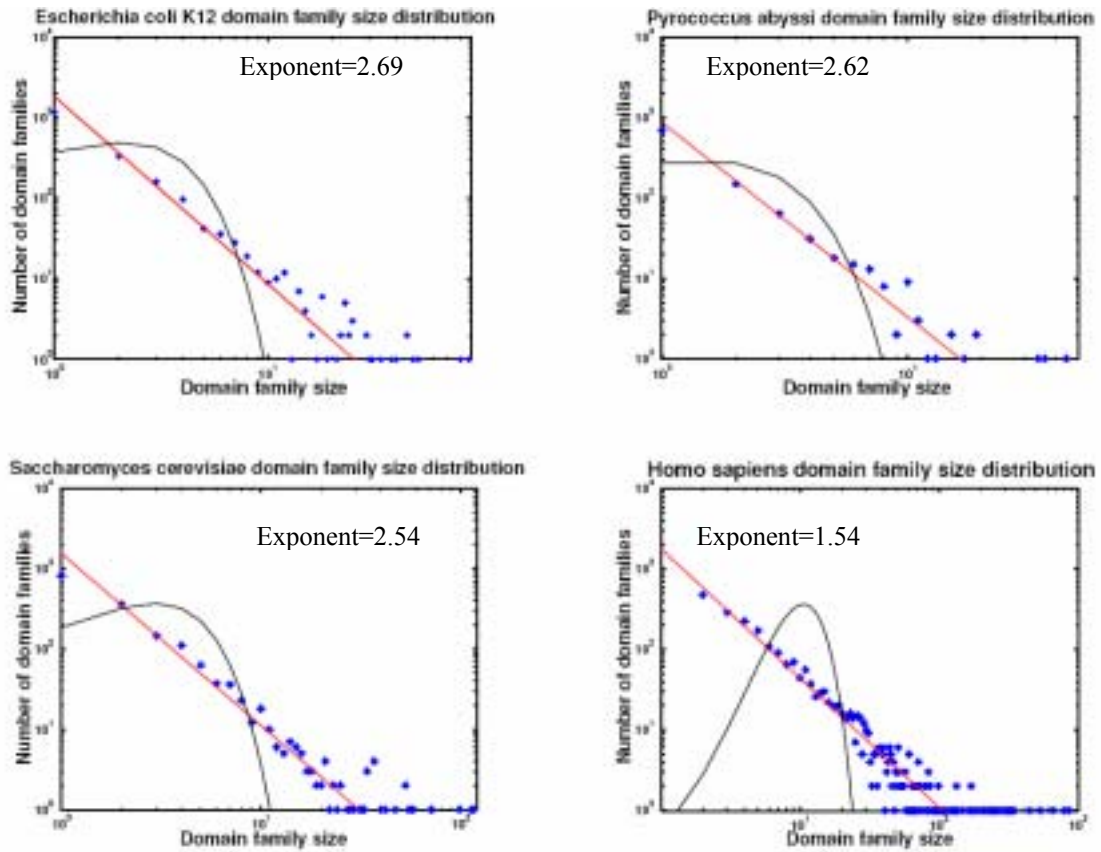
To further verify this interpretation, we performed a correlation test (detrended fluctuation analysis) [129] on the number of duplication breakpoints in the non-overlapping windows along the genome in relation to the distances between them. The result of the test showed a positive correlation between neighboring windows, suggesting that over the evolutionary history, consecutive segmental duplications occur favorably near or on some previously duplicated segments, and are absent elsewhere. Such observation is consistent with the positive feedback dynamics of the duplication process that helps to shape the statistical structure of the genome.

The power law distribution of the duplication frequency is reminiscent of the connectivity distribution of the scale-free network. Along with its definition [84], it was suggested that the scale-free network structures can be created by a process in which the probability of adding a new edge to a node in the graph is proportional to the present number of edges (connectivity) of the node. The similarity of the copy number distribution of segmental duplication to the connectivity distribution of a scale-free network suggests that the probability of duplicating a segment in the genome is roughly proportional to the current copy number of the segment.

### **3.3 Analysis on the Distribution of the Protein Domain Family Sizes**

To test if such characteristic dynamics also has long-term impact on the proteome level, we analyzed the distribution of the protein domain family sizes in various genomes. The protein domain families in different organisms are extracted from the protein family database InterPro [5]. The classification of domain families is based on sequence signature and homology. A family of protein domains found by this method can be viewed as a cluster of amino acid sequences from a proteome that share enough similarity with each other and

have maintained their critical sequences. When the histograms of the sizes of those protein domain families from various organisms were plotted on a log-log scale, a linear relationship was observed in all cases, including in *E. coli* K12, *P. abyssi*, *S. cerevisiae*, *H. sapiens* shown in **Figure 13**. Therefore, the domain family size distribution, or more generally, the copy numbers of homologous amino acid sequences, seems to follow a power-law distribution. It is consistent with the observation in the copy number distribution of the segmental duplications, which suggests that the copy number-dependent duplication process may have contributed to the distribution of the protein domain family sizes.



**Figure 13. Protein domain family size distribution in some genomes examined. The plots show the protein domain family size distribution in the corresponding proteomes (dots). The protein domain family data is extracted from InterPro database. The plots are on a log-log scale. The almost linear shape (straight line) on such a plot indicates a power law relationship between a domain family size and the number of domain families of that size. The exponents of the fitted power law are listed. Therefore, the protein domain family size distributions are also characterized by an over-representation of large size families when compared with uniformly random distributions (black curves).**

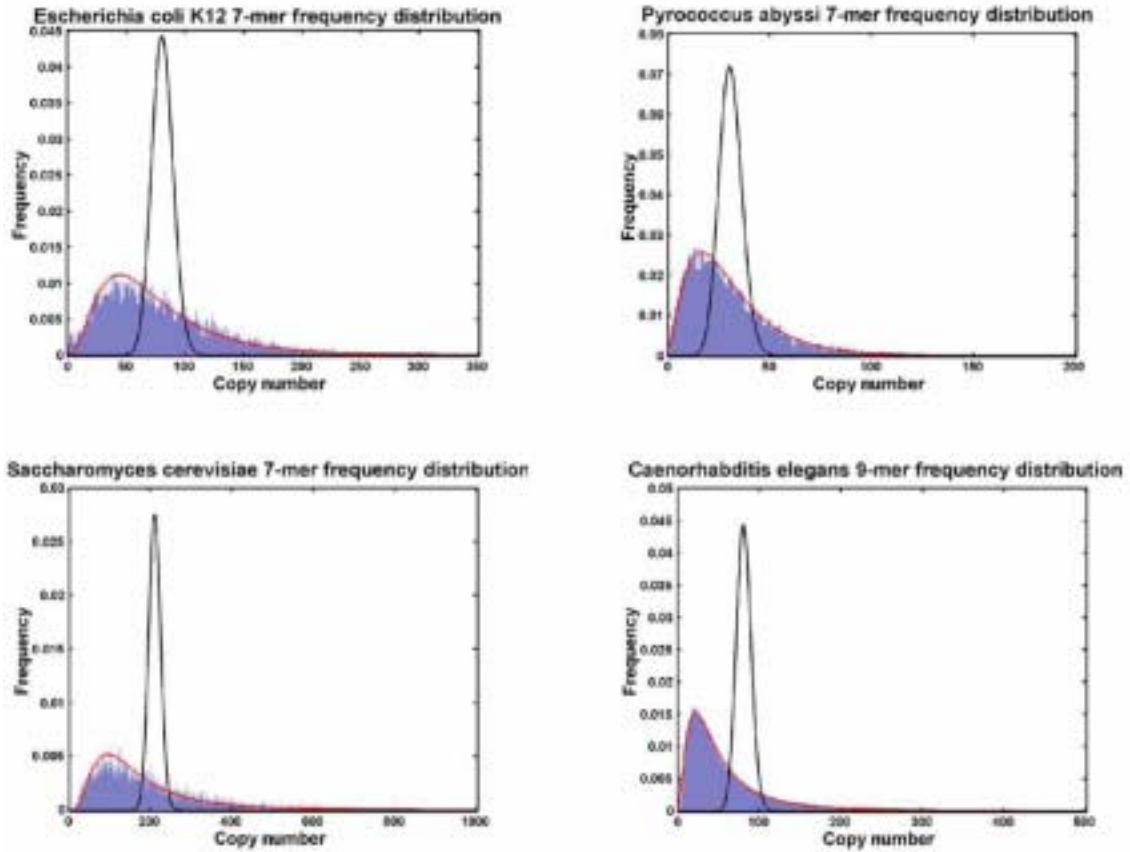
However, although the protein domain family size distributions from different organisms all follow a power-law distribution, the actual exponent in the power-law is different among different species Figure 13. Such quantitative variation may reflect the difference in the rate and scale of the duplication process, as well as other evolutionary processes, such as deletion and substitution. We will use a parsimonious genome evolution model to estimate such differences in the later chapters.

### **3.4 Analysis on the Distribution of the Mer Frequencies**

To study the statistical features of smaller sequence elements, we performed a large-scale non-overlapping mer-frequency distribution analysis in different whole genomic sequences. The experiment was conducted on all the reasonable mer-sizes, covering almost all the presently available whole genomic sequences and including various organisms from all the kingdoms. To avoid the complication of inversions, we treated two inversely complimentary mers as one species. (For example, 5'-ATCG-3' and 5'-CGAT-3' are counted as one mer species, i.e., their frequencies are combined.) Therefore, for mer size  $m$ , there are  $4^m/2$  species of  $m$ -mers. From our results, it is clear that the mer-frequency distributions from all the genomic data examined deviate from the random distribution (see **Figure 14** for some of the results). Furthermore, they are all

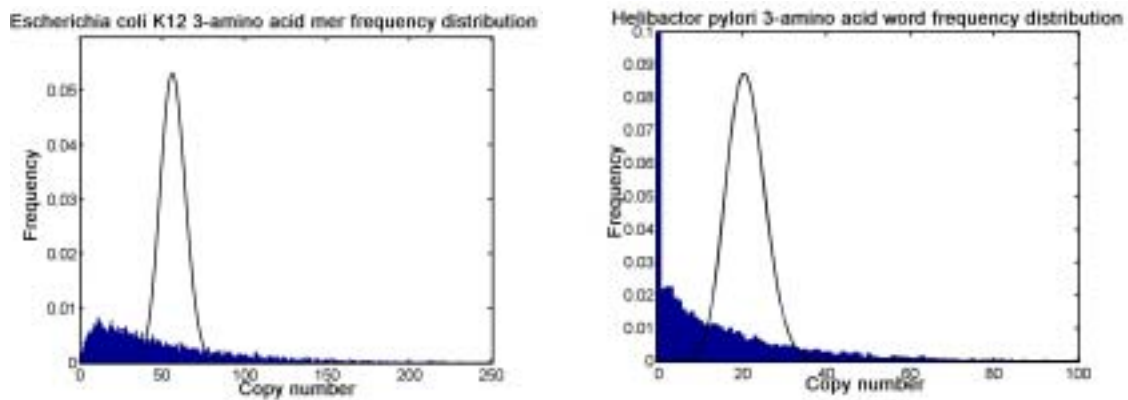


characterized by the same type of deviation — over-representation of high frequency mers.



**Figure 14. Mer-frequency distribution in some genomes examined. The plots shows the non-overlapping mer frequency distribution (bars) in the genomes of a eubacteria (*E. coli K12*), an archea (*P. abyssi*) and two eukaryote (*S. cerevisiae* and *C. elegans*). When compared with the expected distribution from a random sequence of the same length (black line), the distributions from real sequences consistently show an over-representation of high-copy mers.**

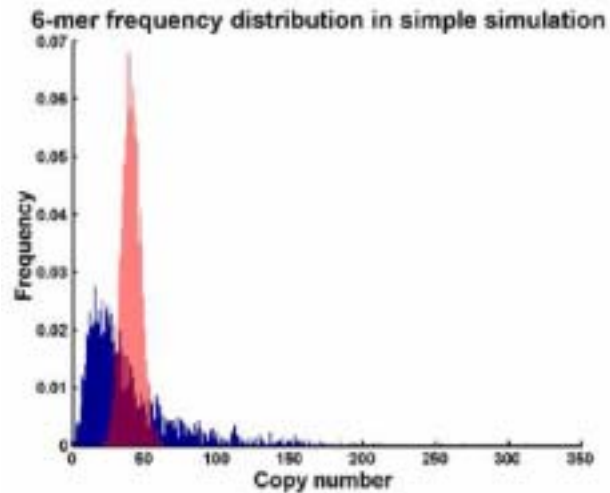
We have also examined the mer-frequency distribution in just the coding sequences, and the distribution of amino-acid mer-frequency in the corresponding proteome sequences. (For a length of  $m$ , there are  $20^m$  species of different amino-acid mers.) Both results share the same type of deviation from the random distribution that is observed in the whole genomic sequences (Figure 15).



**Figure 15. Amino acid mer-frequency distribution in *E. coli* K12 and *H. pylori* (bars). When compared with the expected distribution from a random sequence of the same length (black line), the distributions from real sequences consistently show an over-representation of high-copy mers.**

Despite the observation that the mer copy number distributions do not always follow a power-law distribution (although one can show that the tail distribution is similar to power-law), one finds that the types of the pattern observed in the

mer distribution is consistent with those observed in the protein domain family size distribution in that the high frequency elements are overrepresented. Although on a different scale, the mer distribution may also be caused by the positive feedback dynamics suggested by the segmental duplication distribution in the human genome, in which the duplication probability is proportional to the current copy number of the element. To examine this hypothesis, a simple simulation of “evolution by duplication” was performed, where a short random sequence (1000bp) was allowed to evolve to a final length of 500Kb by duplicating fragments randomly chosen from itself. The deviation in the mer-frequency distribution of the final sequence from a random sequence closely resembles the pattern seen in real genomes (see **Figure 16**). Therefore, the statistics of mers in genomes and short amino-acid words in proteomes could be simply due to the duplication processes during genome evolution. We will introduce a formal model for genome evolution in next chapter to explain the mer copy number distribution more carefully.

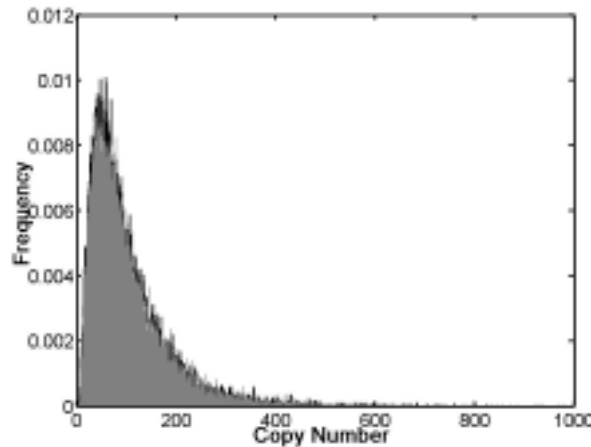


**Figure 16.** The 6-mer frequency distribution of the resulting sequence of a simple “evolution by duplication” simulation. The initial condition is a random sequence of length 1000bp. The sequence is evolved through multiple iterations until it reaches a length of 500Kb. In each iteration, a fragment of length uniformly randomly distributed from 1 to 100bp is randomly chosen from the sequence, duplicated, and re-inserted randomly into the sequence. The dark bars in the plot show the 6-mer frequency distribution of the final sequence from the simulation. The light bars show the 6-mer frequency distribution of a random sequence of the same length.

### **3.5 The Effect of Repeats and Selection on the Mer Distributions**

One may argue that the overrepresentation of high frequency elements are due to selection constrains or excessive repetitive elements. However, since the

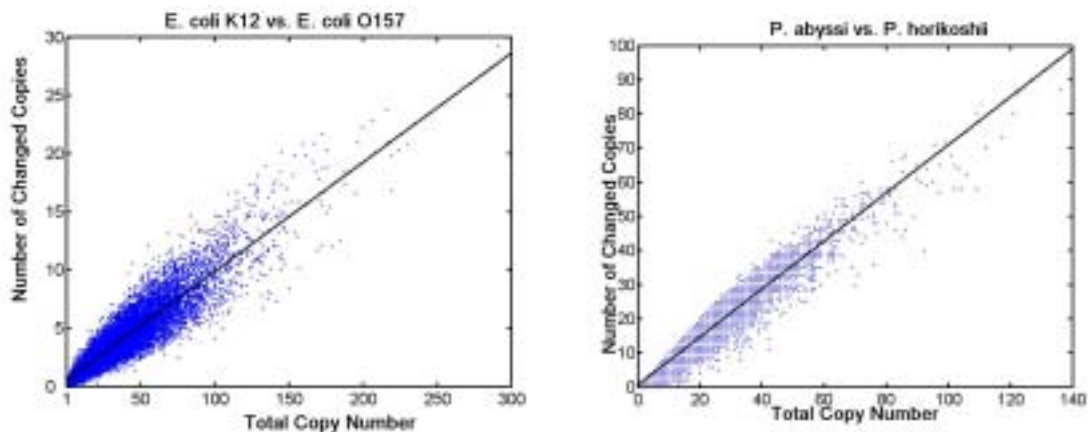
prokaryotes, such as *E. coli* and *P. abyssi*, do not contain many repetitive elements, but still show the same qualitative pattern in their composition distribution, it is unlikely that the repeats are solely responsible for the “fat-tail” in the mer frequency distribution. To examine the hypothesis in eukaryotes, we compared the mer copy number distribution in *S. cerevisiae* before and after all the repeats in the genomes have been masked and found no significant changes in the shape of the distribution (**Figure 17**).



**Figure 17. The 7-mer copy number distribution in *S. cerevisiae* before (grey bars) and after (black line) the removal of repetitive elements from the analysis.**

To examine the hypothesis that the deviation in the mer copy number distribution from the random distribution is caused by differential selection constraints on mers with different copy numbers, we examined the rate of mer copy number changes in evolution by comparing closely related species. We have

aligned the genomic sequence of *E. coli* K12 to another strain *E. coli* O157 using MUMmer [40]. The total copy number of each 7-mer in the strain K12 as well as the number of copies that have experienced mutations in O157 compared to K12 is recorded. If the selection constrain is different for mers with different copy numbers, we expect to see a difference in the mutational frequencies of mers with different copy numbers. However, as shown in **Figure 18**, the number of changed copies and the total copy number of each mer is linearly correlated, indicating a constant mutational frequency for mers of different copy numbers. Such observation indicates that most of the mutational changes are homogenously distributed among different mers and likely to be neutral; therefore, the number of changed copies for each mer is roughly proportional to its total copy number. A similar picture is also seen in *P. abyssi* when it is compared to a slightly distant relative *P. horikoshii*. Therefore, it is unlikely that the skewed distribution of mer copy numbers is mainly caused by natural selection. However, one cannot rule out the possibility of natural selection effect on the composition of the genome. Unfortunately, there is no method currently available to measure such effects reliably.



**Figure 18.** The relation between the total copy number of each 7-mer in *E. coli* K12 and *P. abyssii* and the number of copies that have experienced changes when compared to the genomes of their close relatives *E. coli* O157 and *P. horikoshii*, respectively. Each point represents one particular 7-mer. The number of changed copies is roughly linear (fitted by the black line) to the total copy number, indicating a homogeneous natural selection pressure on mers with different copy numbers.

### 3.6 Summary

We have examined the copy number distribution of various genome components on very different scales and in many different genomes. Consistent with previous studies, all genomes examine are characterized by the overrepresentation of the high frequency elements. The distributions of larger scale components (segmental duplications and protein domain families) follow a power-law

distribution. The power law property, one of the definitive features of the scale-free networks, can be created by a simple duplication process called “preferential attachment” that favors high copy number elements. The distribution of mers, although do not follow power-laws, can also be explained by an extremely simple simulation process that includes duplication process.

By examining the mer distribution before and after the removal of the repeat elements in the genome, we have shown that the presence of the repetitive elements cannot fully explain the statistical characterization of the mer distribution in the genomes. In addition, a linear relationship between the copy number of a mer in a genome and the number of copies mutated in a closely related genome implies that there is no obvious difference in the selection pressure for mers with different copy numbers. Therefore, it is unlikely that either repeats or selection can fully explain the observed distribution pattern. A duplication process with the positive feedback “rich gets richer” dynamics may be the main force carving out the common qualitative features in the statistical structures of various genomes. In the next chapter, we will examine the quantitative difference in the distributions of these components in different genomes.



## Chapter 4

# A Polya's Urn Model for Genome Evolution

### 4.1 Introduction and Related Work

In the past few years, with the increasing availability of whole-genome sequencing data, detailed statistical analyses of the sequenced genomes have been carried out. It is now apparent that genomes are neither random nor deliberately and accurately sculpted. The seemingly random non-coding regions have nonrandom compositions and long-range correlations, whereas the more conserved coding regions are subject to constant mutations and tolerant of enough polymorphisms. Although there is a huge diversity on the sequence level in the genomes of different organisms, as we have shown in the last chapter, some statistical characteristics of genome composition and structure are found to be generic. The results from genomic data analysis at different scales, at different levels, and in different organisms repeatedly show the same pattern (over-representation of high-frequency elements), which are observed as the “fat-tails” in the histograms of mer frequencies, gene family sizes, and duplication copy numbers [111][56][193]. These statistical features are further reflected in higher-level cellular processes, such as protein-protein interaction networks,

metabolic networks, and genetic pathways [84][151][139][83].

These observations consistently suggest a generic evolutionary dynamic involving positive feedback as postulated by S. Ohno's theory of “evolution by duplication”. Although some models based on duplication have been proposed to explain the observed statistical features, all of them are applied to a specific level of cellular process: [186] on microbial gene family size distribution; [79] on 6-mer frequency distribution; [138] on protein family size distribution; [20] on expression network topology; and [127] on protein interaction network topology. Since the protein interaction networks, as well as other higher-level cellular processes, are encoded in genomic sequences, the evolution of their topology is rooted in the genomic sequence changes. Therefore, we believe that a more general model of “evolution by duplication” at genomic level can explain the common pattern observed at various scales and different cellular information levels, and enable the quantification of the importance of duplication and/or deletion process relative to other evolutionary processes.

In this chapter, we will first introduce a simple version of our genome evolution model for the copy number distribution of non-overlapping mers in various genomes. Different from the previously proposed “minimal”

models [186][79], our model also includes deletion as one of the three essential processes in genome evolution besides substitution and duplication. However, the number of free parameters does not increase. The parsimony of the model, especially the necessity of including deletion, is attested by its superior performance in fitting the copy number distributions of mers of different sizes. A more refined model applicable to the distributions of overlapping mers will be examined and applied to estimate the evolutionary distance in the next chapter.

## **4.2 Polya's Urn Model for Genome Evolution**

Our genome evolution model can be viewed as an extension of Polya's urn model [86], and considers genome evolution as a non-stationary Markov process. A genomic sequence evolves by three elementary processes under lineage-specific rates. Similar to most of the previous models [186][79], in our model, we assume that most sequence evolution events are neutral and the mutational rates are mostly homogeneously distributed along the genome. Such assumptions can be made if one believes that most of the mutational events are neutral or near neutral given the effective population size [38]. Under these assumptions, the probability of a particular genome component getting involved in any evolutionary process is proportional to its copy number in the genome. This is consistent with our observation that for a particular mer the number of copies changed during

evolution is linearly related to its total copy number in the genome, and is further supported by the clustered distribution pattern of the recent segmental duplications in the human genome. Therefore, the genome components with high copy numbers are more likely to be duplicated, deleted or substituted, just as the properties controlled by the rules of game in Polya's Urn. The copy number distribution of the different components is decided by the relative rate and scale of the three processes.

Models with such a simplified assumption may not faithfully reflect the biological reality especially without considering the effect of natural selection. But so far no good models for natural selection are available. Given the complexity of the genome evolution problem, and the evidences that most of the time the changes in the genomic sequences are neutral, such simplifications, which lead to a mathematically traceable model, are necessary to make the first step towards the understanding of the quantitative behavior in the genome evolution process. As the model gets more complicated by incorporating more context-dependent processes, these assumptions can be gradually relieved.

### **4.3 A Simple Model for Non-overlapping Mers: The Parsimony of the Genome Evolution Model**

Recently, several research groups have proposed genome evolution models to explain the statistical features in the distributions of various genome components. All of them are motivated by Ohno's theory [124], incorporating two basic processes: duplications and point mutations. DeLisi *et al.* [186] described a simple model to explain the gene family size distributions in various microbes. Very recently, Lee *et al.* [79] proposed another minimal model that was able to fit the 6-mer distributions in several bacterial genomes. Both models assume that the currently observed genomes have evolved neutrally from some small primordial genome with randomly distributed components under homogeneous evolutionary rates.

Our model [193] for genome evolution, although also motivated by Ohno's theory, incorporates not only point mutations and duplications, but also deletions. By applying different models to the statistical distributions of the non-overlapping mers in various genomes, we found that deletions play a role no less critical than mutations or duplications. The effect of deletion process cannot simply be replaced by a reduction in duplication rate and/or an increase in point mutation rate. Deletion occupies a unique role in evolution as we found

that the omission of the deletion process leads to an inadequate model. These conclusions from model analyses are consistent with biological experimental results [63][132][131], which show that deletions happen as often as duplications, and their contribution in shaping the genome composition is significant. Our model, which considers all three processes, is able to fit the distributions of mers of different sizes from a wide range of scale. It also applies equally well to eukaryotic genomes. However, models with only substitution and duplication processes but no deletions, such as Lee's model [79], can only explain the distribution of genome components on a specific scale, and usually is not applicable to all genomes.

The effect of genome evolution on the distribution of the non-overlapping mers in the genome can be considered in a graph evolution setting. The genome under evolution can be represented by a directed Eulerian graph<sup>1</sup>. Each pair of inverse-complementary mer species of a particular length is represented by a node<sup>2</sup>. Whenever two non-overlapping mers are immediately adjacent to each other in

---

<sup>1</sup> Eulerian graph is a group of graphs in which each node has even number of edges (for directed Eulerian graph, for each node the number of incoming edges is equal to the number of outgoing edges). In these graphs, one can find Eulerian in which each edge is visited exactly once.

<sup>2</sup>To avoid the complication of inversions, we treated two inversely complimentary mers as one species. (For example, 5'-ATCG-3' and 5'-CGAT-3' are counted as one mer species, i.e., their frequencies are combined.) Therefore, for mer size  $m$ , there are  $\frac{4^m}{2}$  species of  $m$ -mers.

the genome, they are connected by an additional directed edge. Without loss of generality, the edges are always directed from the 5' end to the 3' end. In a graph created in this manner, the number of directed edges from node  $i$  to node  $j$  ( $l_{i,j}$ ) indicates how many times the  $i^{\text{th}}$  mer is immediately adjacent to the 5' end of the  $j^{\text{th}}$  mer in the genome. Due to the Eulerian property of the graph, each node has identical in- and out-degrees. We use  $k_i$  to represent both the out-degree ( $l_i^{\text{out}}$ ) and the in-degree ( $l_i^{\text{in}}$ ) of the node  $i$ , which are equal to the copy number of the corresponding mer in the genome. For mers of size  $m$ , and a genome of length  $G$ , the graph will have a total of  $N = \frac{4^m}{2}$  nodes and  $E = \frac{G}{m} = \sum_{i=1}^N l_i$  edges. Each possible Eulerian path in the nontrivial (non-singleton) connected component encodes a genome with the same mer composition. However, the genomes represented by the same graph do not necessarily have the same arrangement of mers.

The evolution of a genome is modeled as a non-stationary discrete Markov process on the graph. The model assumes that all the presently existing genomes originated from a very small primordial genome with uniformly randomly distributed mers. Thus, the initial graph is a random graph with a small average degree. In each time step, one of the three possible processes occurs:

*duplication* of a chosen mer (with probability  $\mu_r$ ), *deletion* of a chosen mer (with probability  $\mu_d$ ), or *substitution* of a chosen mer by another mer (with probability  $\mu_s$ ) (Figure 19). Therefore,  $\mu_r + \mu_d + \mu_s = 1$ .

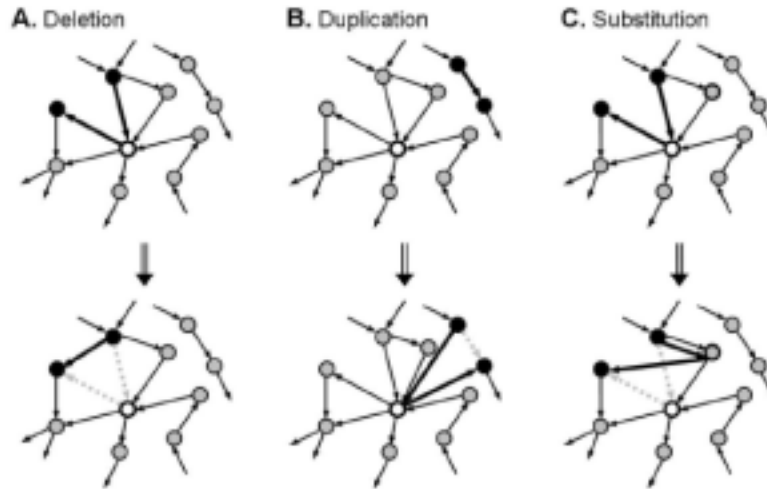


Figure 19. The three processes occurring during graph evolution: *deletion*, *duplication*, and *substitution*. In each process, the target node (clear circle) is chosen with preference for nodes with larger degrees: If the  $i$ -th node has degree  $l_i$ , the probability of it being chosen is proportional to  $\frac{l_i}{\sum_{i=1}^N l_i}$ . In deletion (A), a pair of edges of the target node (thick black arrows), one incoming and one outgoing, is randomly chosen and deleted, and a new edge (thick black arrow) is added between the ascendant and descendant nodes (black filled circles). In duplication (B), new edges are added between the target node and the ascendant/descendant nodes (black filled circles) of an edge (thick black arrow) randomly chosen to be deleted. In substitution (C), a randomly chosen pair of edges of the target node (thick black arrows), one incoming and one outgoing, is rewired to the



**randomly chosen substitute node (gray filled circle with thick boundary). Note that all the processes during graph evolution preserve the equality of the in-degree and out-degree of each node.**

To avoid extinction, we let  $\mu_r > \mu_d$ . During graph evolution, let  $l_i(t)$  and  $E(t)$  indicate the copy number of  $i^{\text{th}}$  mer and the total number of mers in the evolving genome at  $t^{\text{th}}$  iteration. If we assume that the target mers for any process is chosen uniformly randomly from the genome, then the probability of  $i^{\text{th}}$  mer species being chosen for a process in the next iteration is proportional to its frequency in the genome in the current iteration ( $\propto \frac{l_i(t)}{E(t)}$ ). Such a strategy implements a “rich gets richer” dynamic rule. If a mer undergoing substitution is modeled as changing into any other mer with equal probability after substitution<sup>3</sup>, then with this simplifying assumption, we can write down the difference equation describing the expected probability distribution for the copy number of the  $i^{\text{th}}$  mer:

$$P(l_i(t) = n) = P(l_i(t-1) = n-1)P(l_i(t) = n | l_i(t-1) = n-1)$$

---

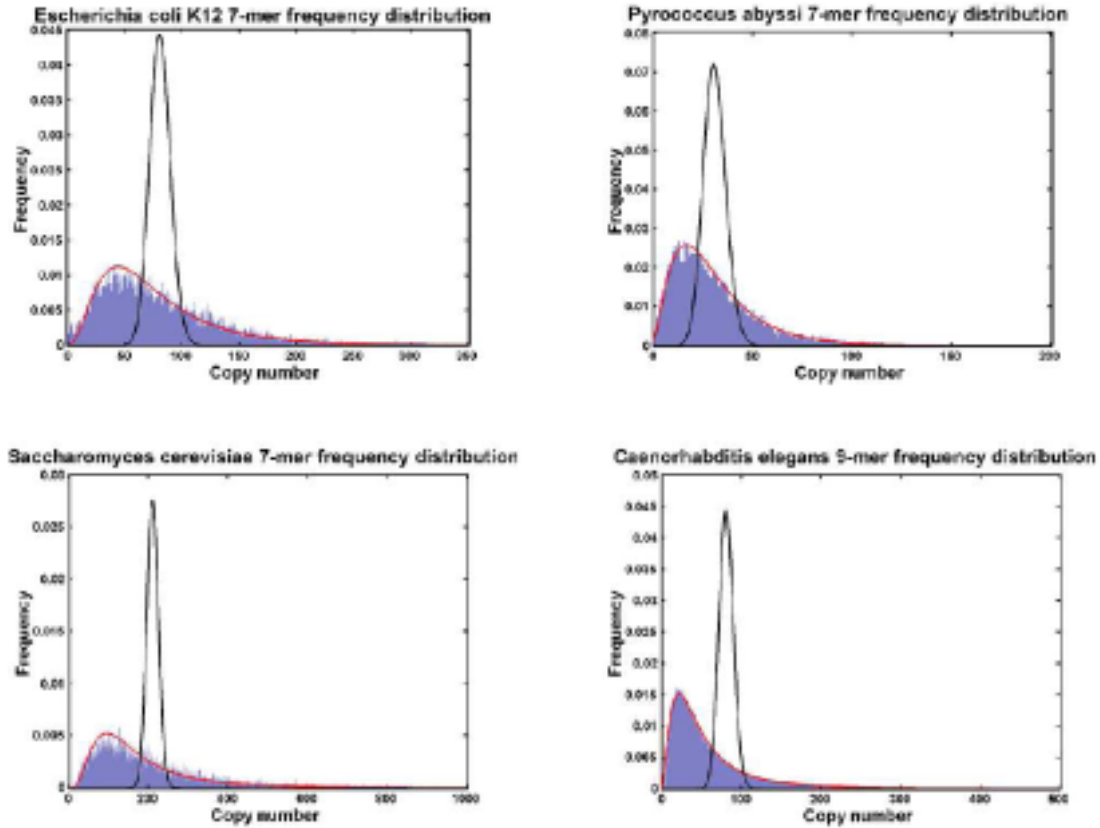
<sup>3</sup>We approximate the probability of a specific mer being chosen to substitute another mer during substitution as  $\frac{1}{N-1}$ . This approximation follows if we assume that the frequency of the nearest neighbors (with 1bp mismatch) of the  $i^{\text{th}}$  mer is  $\frac{3m}{N-1}$  in the rest of the genome excluding the  $i^{\text{th}}$  mer. Since only when the mutation occurs at the mismatched position and results in a particular base pair out of three possible substitution outcomes, will it change into the  $i^{\text{th}}$  mer, the probability of such an event can be approximated by  $\frac{3m}{N-1} \cdot \frac{1}{m} \cdot \frac{1}{3} = \frac{1}{N-1}$ .

$$\begin{aligned}
& +P(l_i(t-1) = n)P(l_i(t) = n | l_i(t-1) = n) \\
& +P(l_i(t-1) = n+1)P(l_i(t) = n | l_i(t-1) = n+1) \\
= & P(l_i(t-1) = n-1) \left( \mu_r \frac{n-1}{E(t-1)} + \left(1 - \frac{n-1}{E(t-1)}\right) \frac{\mu_s}{N-1} \right) \\
& +P(l_i(t-1) = n) \left( 1 - \frac{n}{E(t-1)} - \left(1 - \frac{n}{E(t-1)}\right) \frac{\mu_s}{N-1} \right) \\
& +P(l_i(t-1) = n+1) \left( \mu_d \frac{n+1}{E(t-1)} + \mu_s \frac{n+1}{E(t-1)} \right) \\
= & P(l_i(t-1) = n-1) \left( \left(\mu_r - \frac{\mu_s}{N-1}\right) \frac{n-1}{E(t-1)} + \frac{\mu_s}{N-1} \right) \\
& +P(l_i(t-1) = n) \left( 1 - \left(1 - \frac{\mu_s}{N-1}\right) \frac{n}{E(t-1)} - \frac{\mu_s}{N-1} \right) \\
& +P(l_i(t-1) = n+1) \left( \mu_d \frac{n+1}{E(t-1)} + \mu_s \frac{n+1}{E(t-1)} \right). \tag{4.1}
\end{aligned}$$

When the mer size is sufficiently large, each mer species only accounts for a very small fraction of the genome; we assume that the copy number of each mer species evolves independently, and the genome size  $E(t)$  grows deterministically with time in a linear fashion with a rate of  $(\mu_r - \mu_d)$ . These approximations have been validated by Monte Carlo simulations. Therefore, the above equation can be viewed as an expression of the copy number distribution of all possible mers in a genome.

We fit our model to the mer frequency distribution in real genomes by

numerical simulations. The initial condition is set as a random sequence of length 1kb. The iteration proceeds until the graph size reaches the corresponding size of the real genome under study. The model has only two free parameters, but it is able to fit the distributions of mers over a wide range of scales (Figure 20).



**Figure 20. Mer-frequency distribution in some genomes examined. The plots shows the non-overlapping mer frequency distribution (grey bars) in the genomes of a eubacteria (*E. coli K12*), an archaea (*P. abyssi*) and two eukaryota (*S. cerevisiae* and *C. elegans*), and the distribution from random sequence of the same length (black line). Our simulation results (grey line) from the simple graph model closely fit the “real” mer frequency distribution. Given that we only have**

two free parameters ( $\mu_s$  = single point mutation probability and  $\mu_r / \mu_d$  = ratio of probabilities of duplication over deletion, see below) in the model, the data-fitting is extremely convincing.

The comparison between our full model and the model without deletion process reveals that deletion process is as essential as point mutations and duplications. When deletion is omitted, the model can still fit the 6-mer frequency distribution quite well — a result consistent with Lee, *et al.* [186]. However, this model can no longer fit the frequency distribution of mers of other sizes (Figure 21).

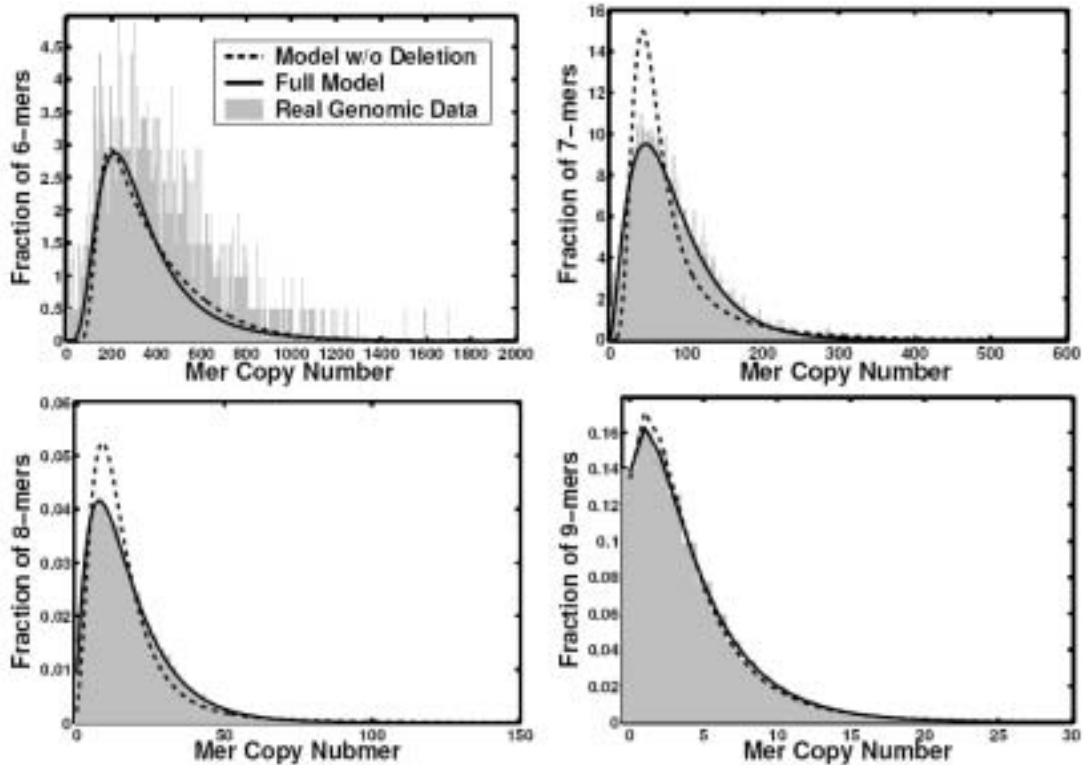


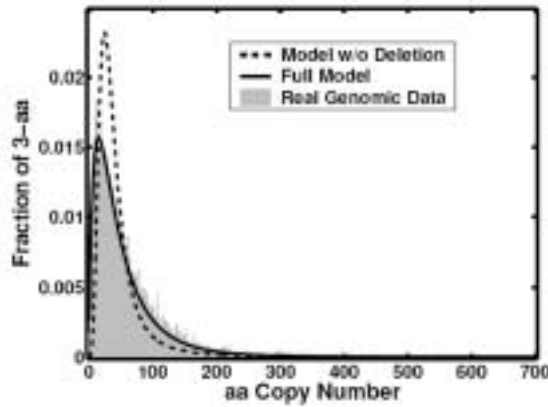
Figure 21. Our model (black solid lines) is fitted to the distributions of different mer sizes in *Escherichia coli* K12 genome (gray bars). The results are

**compared between the full model (black solid lines) and the model without deletion (black dotted lines). Both models fit quite well to 6-mer distribution. However, for other mer sizes, the full model, which includes deletion, evidently does much better than the other, which only incorporates point mutation and duplication.**

Our model fits not only the distributions of nucleotide mers in genomic sequences but also the distributions of amino acid mers in protein sequences<sup>4</sup>. Those results on the amino acid level again proves the essential role of deletion in the model (Figure 22). Therefore, although deletion can be neglected when modeling the distributions of large functional units, such as gene families [186], it has a significant effect in modeling the statistical features at a smaller scale. The diminished role of deletions on gene family level may be due to the strong selection pressure against deletions of large sizes. But in a scalable and more generalizable model, deletion remains irreplaceable.

---

<sup>4</sup>When the model is applied to aa frequency distributions, each node in the graph represents a different peptide of length  $w$ , and there are  $20^w$  nodes in the graph. The initial condition is a random peptide sequence of 300 single amino acids



**Figure 22.** The model is also successfully applied to the distribution of 3-amino acid mers computed from all *Escherichia coli K12* proteins (gray bars). The comparison of the results between the full model (black solid lines) and the model without deletion (black dotted lines) confirms once again the importance of deletion process in the model.

It is worth noting that the model parameter  $\mu_s$  (point mutation probability) is significantly lower when fitted to amino acid mer distributions than to the distributions of mers of corresponding sizes (three times the amino acid mer size). For example, the  $\mu_s$  estimated for *E. coli* K12 nucleotide 9-mers is 0.2810, while that for the amino acid 3-mers is 0.0192; and the  $\mu_s$  estimated for *H. pylori* nucleotide 6-mers is 0.007, while for the amino acid 2-mers is 0.0024. Such differences can be explained by the purifying selection in coding regions and the degeneracy of amino acid codons. The successful application of the model on amino acid mer frequency distributions imply an expected, yet

important phenomenon — the evolution processes and their resulting statistical structures on the genomic level are well-reflected on protein level. Naturally, we expect our model to generalize, in order to explain the statistical features in higher-level genomic or cellular processes, such as protein-protein interaction networks, signaling pathways, etc.

In our empirical studies, the model is applied to various mer lengths and to genomes from organisms of various domains: eubacteria, archaea, unicellular and multicellular eukaryota. The fitted values of the two free parameters ( $\mu_s$  and  $\frac{\mu_r}{\mu_d}$ ) in some of the studied genomes are listed in Table 5.

<b>Eubacteria</b>	<b>6-mer</b>		<b>7-mer</b>		<b>8-mer</b>	
<b>Genomes</b>	$\mu_s$	$\mu_r / \mu_d$	$\mu_s$	$\mu_r / \mu_d$	$\mu_s$	$\mu_r / \mu_d$
<i>M. genitalium</i>	0.0117	1.03	0.0477	1.14	0.1725	1.67
<i>M. pneumoniae</i>	0.0410	1.10	0.1241	1.61	0.3106	2.81
<i>H. influenzae</i>	0.0136	1.03	0.0366	1.10	0.1546	1.58
<i>S. subtilis</i>	0.0095	1.02	0.0320	1.09	0.1406	1.50
<i>E. coli</i> K12	0.0101	1.01	0.0210	1.02	0.0708	1.05
<b>Archaea</b>	<b>6-mer</b>		<b>7-mer</b>		<b>8-mer</b>	
<b>Genomes</b>	$\mu_s$	$\mu_r / \mu_d$	$\mu_s$	$\mu_r / \mu_d$	$\mu_s$	$\mu_r / \mu_d$
<i>P. abyssi</i>	0.0155	1.02	0.0674	1.09	0.2028	1.51
<i>P. furiosus</i>	0.0111	1.02	0.0303	1.03	0.1132	1.21
<i>S. solfataricus</i>	0.0072	1.02	0.0338	1.05	0.0670	1.15
<i>S. tokodaii</i>	0.0059	1.02	0.0190	1.05	0.0637	1.22
<b>Eukaryotic</b>	<b>8-mer</b>		<b>9-mer</b>		<b>10-mer</b>	
<b>Genomes</b>	$\mu_s$	$\mu_r / \mu_d$	$\mu_s$	$\mu_r / \mu_d$	$\mu_s$	$\mu_r / \mu_d$
<i>S. cerevisiae</i>	0.0568	1.18	0.1944	2.00	0.3704	2.97
<i>C. elegans</i>	0.0112	1.06	0.0350	1.30	0.1307	2.97
<i>A. thaliana</i>	0.0051	1.02	0.0131	1.05	0.0519	1.24
<i>D. melanogaster</i>	0.0114	1.04	0.0393	1.20	0.1728	2.78

**Table 5. Graph model parameters ( $\mu_s, \mu_r / \mu_d$ ) fitted to the mer-frequency distribution data (6 to 8-mer for prokaryotic genomes and 8 to 10-mer for eukaryotic genomes) from the whole genome analysis. Different mer lengths are shown for prokaryotes and eukaryotes because of the large difference in their genome sizes.**

The fitted parameter values in the table show some interesting properties. First, the point mutation probabilities ( $\mu_s$ ) increase monotonically with the mer length ( $m$ ) in each genome. This may reflect the scaling effect introduced by fixing the size of duplications and deletions in the model as the size of one mer ( $m$ ). However, in the related biological processes, while one point mutation always changes one mer to another, the size of a duplication or deletion event may be larger than the mer size in the model, leading to changes in the copy numbers of more than one mer. For a duplication or deletion event of a certain size, when the mer size increases, the number of mers affected by the event decreases. Therefore, the relative probability of point mutation of longer mers tends to be bigger than those of shorter ones. Second, the model fits various distributions nicely when  $\mu_r / \mu_d$  is set to be larger than 1, and the values of  $\mu_r / \mu_d$  grow with the mer lengths in each genome. These results validate our assumption ( $\mu_r > \mu_d$ ), but also suggest that the probability of duplication decays more slowly than the probability of deletion when the length of the duplicated/deleted



fragment increases. Therefore, duplication events of large sizes are more likely to occur than deletion events of comparable sizes. Third, not so infrequently, the relative point mutation rate  $\mu_s$  as well as the ratio  $\mu_r / \mu_d$  tends to be anti-correlated with the genome size (in **Table 5**, genomes in different domains are listed according to their genome sizes in an ascending manner). These observations are to be expected if the sizes of the fragments in both duplication and deletion events are bigger in larger genomes.

The fitted model parameters to a genome imply the relative frequencies of point mutation, duplication and deletion events over the evolution history of the genome. For example, *A. thaliana* genome has been reported to have gone through several rounds of large-scale duplication events, accompanied by massive gene loss relatively recently [62][173]. On the contrary, no recent large-scale duplications or deletions are detected in *S. cerevisiae*<sup>5</sup>, *C. elegans*, or *D. melanogaster*. Consistent with those genome studies, in *A. thaliana* the relative point mutation rate  $\mu_s$  and the ratio between duplication and deletion  $\mu_r / \mu_d$  are much lower compared to other eukaryotic genomes, indicating duplication and deletion events of higher rate and larger scale.

---

<sup>5</sup>Although *S. cerevisiae* also went through a large-scale duplication, the event is far more ancient, and the duplicated segments have significantly diverged.

The genomes in the Table 5 are separated into prokaryotics and eukaryotics. The prokaryotic genomes are further divided into eubacteria (upper half) and archaea (lower half). It is interesting to notice that although the parameter values vary among the organisms from the same domain, the most dramatic variations are observed between prokaryotic and eukaryotic genomes. More specifically, in eukaryotic genomes, the relative point mutation rates ( $\mu_s$ ) are much smaller, as well as the ratios between duplication and deletion probability ( $\mu_r / \mu_d$ ) for a certain mer length. We hypothesize that this might be due to the differences in the efficiency of various evolution machineries acting at the molecular level between prokaryotes and eukaryotes, or between haploid and diploid genomes, such as DNA repair efficiency, recombination rate, and tolerance of deletions or insertions.

The model for genome evolution we presented here does not compensate for the effect of natural selection. Evolution is modeled as a purely stochastic process, which assumes that all the genome-altering events are neutral. In spite of the important role that natural selection plays on evolution of genomes, our approximate model is still capable of explaining distributions of various scales and in different organisms without implicitly modeling selection force. This may imply that most of the events during genome evolution are actually neutral,

which is consistent with our analysis results described in the previous chapter where we observed homogeneous selection pressure onmers with different copy numbers. A more interesting implication is that natural selection acts not only on individual gene level, it may also act by tuning the relative frequencies of the basic stochastic processes (deletion, duplication and point mutation) in evolution. In that case, it is likely that the variation in the model parameter values across different organisms further reflects the differences in the organisms' interaction with their environment.

As important as the effect of natural selection on genome evolution, the model also suggests that evolution is not a delicately tuned process. It is rather full of chances at each step. Similar to the dynamics in Polya's Urn model [86], a small change in some earlier steps will possibly lead to dramatic difference in the later ones. Each genome can be viewed as the current position of a random walk during such a non-stationary Markov process. When two species diverge from the common ancestor, it can be viewed as a branching process. Therefore, the structures of each genome, such as mer frequency, can be seen as a signature more or less unique to the organism in a high-dimensional genome composition space. We think this is the biological basis underlying the successful application of mer frequency spectrum in several biological sequence classification

studies [102]. In the next chapter, we describe a phylogeny method that estimates the evolutionary distance by measuring the differences of such stochastic processes in different sequences.

## **4.4 Summary**

Motivated by our observations on the statistical structure of the genomes from last chapter, we propose a simple genome evolution model to explain the common patterns found in the distribution of various sequence elements on different scales. The model contains three basic processes: substitution, duplication and deletion. Compared to the previously proposed “minimum” genome evolution model [186][79], our model has an additional deletion process. The necessity of the deletion process in the parsimonious model is proved by the capability of our model in fitting the distributions of mers of different sizes, and the limitation of a model without the deletion process to fit only 6-mer distributions.

We have shown that our simple model is able to fit the statistical distributions on different scales and on different sequences (nucleotides and amino acids) from wide range of organisms, including eubacteria, archaea and eukaryota. The general and unifying nature of our model suggests a universal minimal

set of mechanisms (deletion, duplication and substitution) that are driving genome evolution. Ultimately, these basic schemes can be viewed as the results of selection not just on genomes, but also on the evolutionary processes and their modulations. These processes persist possibly because their combination balances the plasticity against the robustness of not just the genome sequences, but also the cellular and inter-cellular structures. These features of genomic processes hold the answer to how genomes can be both stable, and yet paradoxically mutable and adaptive.

However, in this simple model studied here, all the duplication and deletion events are implicitly assumed to have the size of a fixed mer length. In the next chapter, we introduce a more refined model that allows duplication and deletion events of different sizes, and use it to study the evolutionary distance between two sequences.

## **Chapter 5**

# **A Novel Alignment-Independent Method for Estimating Phylogenomic Distances**

### **5.1 Introduction and Related Work**

Most of the conventional phylogeny methods can be roughly categorized into two types: those that rely on detailed sequence alignment results, and those that rely on gene order. Those methods are relatively accurate once the correct alignment results or gene orders are given, and can provide high-resolution details on the evolutionary events in the sequences. However, both types have their limitations. For example, both types require a common set of orthologous sequences without paralogs. This requirement is difficult to satisfy especially for more divergent genomes, not to mention that it introduces bias in the sampling procedure. In the alignment-dependent methods, the sequence alignment procedure implies the assumption that there is contiguity in the conservation between the homologous sequences. However, this assumption can be violated if there are any rearrangement events in the sequences, which occurs in divergent genomes and frequently in pathogen genomes. Moreover, most of the alignment-dependent methods use distance measures based on mismatches in the

alignment results (Kimura, Jukes-Cantor, HKY, Gamma, *etc.*) (for a general introduction see [122]). Fast evolving regions and insertion/deletion events that are not alignable are usually discarded. Since the divergence in sequences happens on different time and size scales, such approach may lead to bias in the results, even if we ignore the variability in the alignment results depending on which algorithm is used (reviewed in [27]).

The phylogeny methods based on gene order compute the genome evolutionary distance as the number of genome rearrangement events [145][119][167][61]. Those methods do not rely on the details of the sequence alignment results, but the order of the orthologous genes in the genomes. However, gene order methods are computationally expensive, and only applicable to small datasets with a limited number of genes (such as mitochondria or chlorophyll genomes). Since those methods usually require a common set of orthologous genes, they are affected by common evolutionary events such as gene duplication, deletion, fusion and rapid divergence. Furthermore, they are highly sensitive to the errors in the genome assembly or gene annotation, which are common in the genomes sequenced by shot-gun techniques.

To overcome the problems in the traditional methods, a new class of

alignment-independent methods has been developed in recent years to complement existing approaches. According to the scale of the components studied by those methods, two directions have emerged: those that work on gene scale and those that work on mer (oligonucleotides or oligopeptides) scale. The gene-scale methods [65][78] mostly take the maximum parsimonious approach based on the presence and absence of a particular gene or gene family. These methods are applicable to very divergent genomes. However, they are prone to the errors or biases introduced by homologous alignments in ortholog recognition and genome annotation processes, and are further complicated by the presence of paralogs. There are a few methods currently available that are based on the mer frequency statistics in the sequences [137][162]. Although they are relatively robust and often produce consistent results with the commonly accepted phylogenetic relations, the distance measurements used (Singular Value Decomposition (SVD) [162] and normalized vector angle [137]) are in normalized forms, and do not have explicit biological interpretations. Therefore, how well those distance functions will scale to genomes of very different sizes is still unknown.

We propose a new method as complementary to the traditional approach, and provide an accompanying theoretical explanation for mer-based analyses.



The method provides information about genome evolution at a very coarse level, and cannot reveal high-resolution details of individual evolutionary events. However, like other alignment-independent methods, it is barely affected by the assembly or annotation error, and free of the possible bias introduced by the alignment algorithm. The method is derived naturally from our genome evolution model described in the previous chapter and treats the genome evolution process as a non-stationary Markov process. Through a step-by-step development, the method offers an explanation on why and how the mer-based methods work under a parsimonious genome evolution model, thus providing the missing theoretical support for the class of alignment-independent methods. Furthermore, unlike the previous methods, the distance measured by our method has explicit biological meaning: the total number of substitution, duplication and deletion events that occurred since the divergence of the two genomes. Therefore, it is expected to scale properly for genomes of different sizes. We have tested our method on *in silico* evolved sequences, as well as some real biological sequences. Our method can recover the “true” evolutionary distances quite faithfully when the divergence level between the sequences is lower than 35%, a threshold similar to the alignment-based methods.

Furthermore, since our method only relies on the mer statistics of the

sequences, it not only is independent of the sequence alignment, but also does not require a completed sequence. The mer-based method can be applied to unassembled sequencing data, such as the unassembled reads from shotgun sequencing with a suitable coverage. If the mer statistics can be measured directly in array-based experiments, we can even apply our method to estimate the phylogenomic distance in a sequencing-independent manner. Most of the current array-based phylogeny methods measure the similarity level of the unknown species/strains to the known species/strains by the presence or absence of the “signature” sequences from the genomes of the known species/strains [179][181][47]. Such methods assume that the unknown genome is closely related to the known genomes, or we have the completed genomes of enough species/strains to estimate the phylogenetic position of the unknown organism. In contrast to these methods, our method supports a comparative array-based approach that does not require any prior knowledge of the two genomes under consideration. The evolutionary distance between the two genomes can be measured by their mer-statistics, which can be read off from the array signals. These features suggest many applications for our method, such as fast sequencing-independent phylogenomic mapping of a large population, reconstruction of the cellular phylogenomic relationships in solid tumor tissues,

or estimation of the relation between evolution and geometry in metagenomic materials.

## **5.2 A Refined Model for Genome Evolution**

In the previous chapter we studied a simple genome evolution model for non-overlapping mers that treats each  $m$ -mer as an inseparable unit of genome structure. Through its simplicity and graphic representability, this model allowed us to test its parsimony, and suggests the necessity of all three evolutionary processes in the model, especially the deletion process. However, genomic sequences are made of contiguous base pairs. Therefore, we have further refined our model for genome evolution, in which evolutionary events occur at the level of individual base pairs. Genome evolution in this model is a continuous-time discrete-state non-stationary Markov process.

Similar to the simple model described in the previous section, in the refined model, the three elementary evolutionary processes (substitution, duplication and deletion) occur under some lineage-specific constant rate. Therefore, although the model is additive or cumulative over time, we do not assume a constant rate among different lineages. But in this model, the sizes of the duplicated or deleted segments follow certain distributions, i.e. duplications or deletions of

different sizes (in base pairs) occur with different probabilities (**Figure 23**). According to Li's and Graur's work [67][125], the sizes of the indel (duplication and deletion) events most likely follow a power-law distribution, so that the occurrence probability of duplications or deletions of a particular size is inversely proportional to a polynomial (in a simplest form, a power) of the event size. Expressed in mathematical notations:

$t$ : Evolution time (of an arbitrary unit);

$G(t)$ : Size of the genome at time  $t$  (bp for nucleotides, and amino acid for proteins);

$p_s$ : rate of substitution per bp;

$p_r$ : rate of starting a duplication per bp;

$p_d$ : rate of starting a deletion per bp.

$X$ : Maximal size of deletion or duplication allowed;

$x$ : Size of a deletion or duplication event;

$f_r(x)$ : Probability of a duplication size of  $x$ , where  $\sum_{x=1: X} (f_r(x)) = 1$ ;

$f_d(x)$ : Probability of a deletion size of  $x$ , where  $\sum_{x=1: X} (f_d(x)) = 1$ ;

$f_r(x)$  and  $f_d(x)$  are power-law distribution functions:

$$f_r(x) = c_r x^{-b_r} ; f_d(x) = c_d x^{-b_d} ;$$

The probabilities of getting a duplication or deletion event of size  $x$

at a particular base position are  $p_r f_r(x)$  and  $p_d f_d(x)$ , respectively.

$L_r$  : Average duplication size;  $L_r = \sum_{x=1: X} (f_r(x)x)$ ;

$L_d$  : Average deletion size;  $L_d = \sum_{x=1: X} (f_d(x)x)$ .

$g$  : Genome growth rate ( $g = p_r L_r - p_d L_d$ );

and we assume that the genome size changes deterministically as:

$$G(t) = G(0)e^{gt}.$$

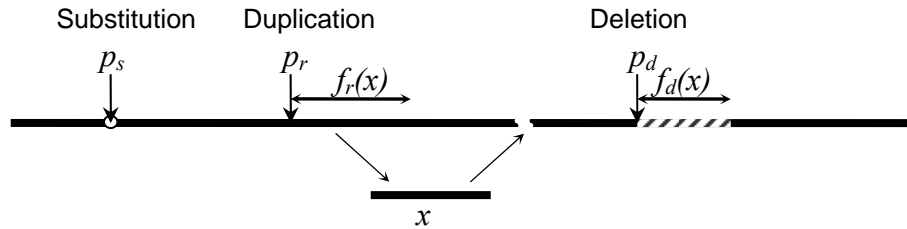


Figure 23. The refined Polya's Urn model for genome evolution.

## 5.2 Rationale

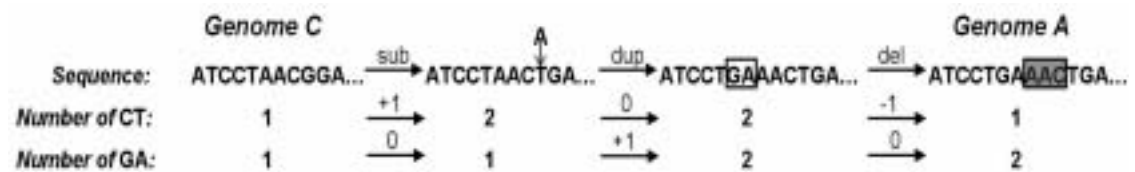
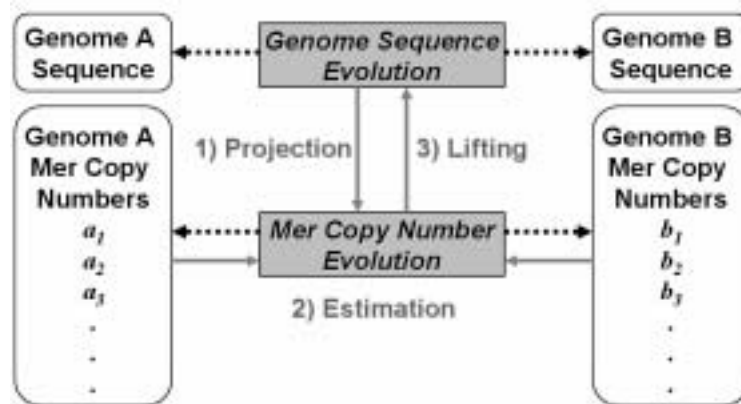


Figure 24. The evolutionary events on the genomic sequence level can be projected onto each individual mer as the changes in their copy numbers. A simple example of genome evolution from an ancestral genome C to genome A is demonstrated

in the figure. Three sequential evolutionary events occurred on the sequence: one substitution (sub: A→T), one duplication (dup: GA), and one deletion (del: AAC). The copy numbers of the two 2-mers: CT and GA, change accordingly.

The method is designed based on a simple observation: The evolution in the genomic sequence can be projected onto each individual mer (Projection), whose copy number changes as a stochastic evolutionary process governed by the dynamics of the evolutionary events on the sequence level (see **Figure 24** for an example). Therefore, the sequence evolution and the mer copy number evolution are tightly correlated. If we can write down such correlation and the form of the mer copy number evolutionary process, we can first estimate the dynamics of mer copy number evolutionary process (Estimation), and then estimate the corresponding events which happened on the sequence level using the mer copy number information (Lifting) (**Figure 25**).



**Figure 25.** The development of our method can be summarized in three

**steps: 1) Projection: Based on the genome evolution model, the evolution on the genomic sequence level can be projected onto the mer space as the evolution in their copy numbers. The relation between the parameters in mer copy number evolution and genome evolution are established using mathematical induction. 2) Estimation: Given the form of the mer copy number evolution process and the mer copy numbers in the two genomes, we can estimate the parameters of the process using the maximum likelihood method (ML). 3) Lifting: Once the parameters for mer copy number evolution are known, we can compute the evolutionary distance on the sequence level based on the relation between the parameters in the two processes established in Projection step.**

These three steps and a reasonable genome evolution model are necessary for the development of a phylogenetic method based on mer-statistics because of the following two reasons: First, due to the complicated dynamics of the evolutionary process, simple distance metrics, such as Euclidean distance cannot provide an appropriate distance measurement from the mer copy numbers. Second, even if some of the simple distance functions can be empirically generated, it is difficult to assess their generality because of the lack of biological interpretation on the nature of the function.

### 5.3 Methods

In our method, we use as a prior the parsimonious genome evolution model we proposed previously. Although the model is highly simplified compared to the real genomic evolutionary events, it possesses sufficient and necessary power to explain the distribution of mer copy numbers in various genomes. In the model we assume neutral evolution and homogeneous mutation rate throughout the genome. Therefore, the probability of a particular mer getting involved in any evolutionary process is proportional to its copy number in the genome.

Based on this model, we will relate the mer copy number evolution process to the genome evolution process, and show how to estimate the number of evolutionary events from mer copy number statistics from the sequences.

#### **Projection of Genome Evolution onto Mer Copy Number Evolution**

Based on our model of genome evolution, we deduce the form of the mer copy number evolutionary process. Some of the notations are listed below:

$m$  : Mer size (bp for nucleotides, and amino acid for proteins);

$N$  : Number of different mers ( $N = 4^m$  for nucleotids, and  $N = 20^m$  for amino acids);



$l(t)$  : Copy number of a specific mer at time  $t$ .

The mer copy number evolutionary process is described as a non-stationary Markov process expressed in the parameters governing the genome evolution model. The parameters can be estimated from the mer copy number distributions in the genomes under comparison using the Maximum Likelihood (ML) approach.

### *Sequences with Uniformly Random Distribution*

We first deal with the simplest case in which the components (mers and sub-mers) are randomly distributed in the sequence with uniform base composition. In these random sequences, the expected frequency of a mer of length  $m$  is  $\frac{1}{4^m}$  for nucleotides and  $\frac{1}{20^m}$  for amino acids. Thus, the expected frequency of a mer only depends on its length, and does not depend on the frequencies of other mers or the subsequences inside the mer. We also assume that all the point mutation patterns have equal rates.

#### 1. Nucleotides

Since we ignore the inversion events, we treat an oligonucleotide and its reverse complement as one mer by adding their copy numbers together. During genome evolution, the duplicated DNA segments are inserted randomly

into the genome in the forward or reverse direction with equal probabilities. Since when the mer size is sufficiently large (and the expected mer frequency is thus low), it is highly unlikely to get two copies of the same mer in one short sequence, we assume the copy number of a specific mer can change at most by one in a short evolutionary time interval ( $\Delta t$ ). The conditional probabilities of copy number changes can be written as below:

$$\begin{aligned}
& P(l(t + \Delta t) = l(t) + 1 | l(t), G(t)) \\
&= l(t)^2 \frac{1}{G(t)} \left( -\sum_{x=m}^X p_r f_r(x) (x-m+1)(m-1) + \sum_{x=1}^X p_d f_d(x) (x-m+1) \sum_{i=1}^{m-1} \frac{1}{4^i} \right) \Delta t + \\
& l(t) \left( \sum_{x=m}^X p_r f_r(x) \left( (x-m+1) - 2 \sum_{i=1}^{m-1} \frac{1}{4^i} \right) - \sum_{x=1}^{m-1} p_r f_r(x) 2 \sum_{i=1}^x \frac{1}{4^i} \right) \Delta t - \\
& l(t) \left( \sum_{x=1}^X p_d f_d(x) \left( 2 \frac{(x+m-1)(m-1)}{N} + \sum_{i=1}^{m-1} \frac{1}{4^i} \right) \right) \Delta t + \\
& \frac{G(t)}{N} \left( \sum_{x=m}^X p_r f_r(x) 4(m-1) + \sum_{x=1}^{m-1} p_r f_r(x) 2(x+m-1) + \sum_{x=1}^X p_d f_d(x) 2(m-1) + 2p_s \right) \Delta t
\end{aligned}$$

$$\begin{aligned}
& P(l(t + \Delta t) = l(t) - 1 | l(t), G(t)) \\
&= l(t)^2 \frac{1}{G(t)} \left( -\sum_{x=m}^X p_r f_r(x) \left( (x-m+1)(m-1) + 2(m-1)(m-\frac{3}{2}) \right) \right) \Delta t + \\
& l(t) \left( \sum_{x=m}^X p_r f_r(x) \left( (m-1) - 2 \sum_{i=1}^{m-1} \frac{1}{4^i} \right) + \sum_{x=1}^{m-1} p_r f_r(x) \left( m-1 - 2 \sum_{i=1}^x \frac{1}{4^i} \right) \right) \Delta t + \\
& l(t) \left( \sum_{x=m}^X p_d f_d(x) \left( (x-m+1) \left( 1 - \frac{2(m-1)}{4^m} \right) + 2(m-1) \sum_{i=1}^{m-1} \frac{1}{4^i} \right) \right) \Delta t + \\
& l(t) \left( \sum_{x=1}^{m-1} p_d f_d(x) \left( x+m-1 - \sum_{i=1}^x \frac{1}{4^i} \right) + mp_s \right) \Delta t
\end{aligned}$$

$$\begin{aligned}
& P(l(t + \Delta t) = l(t) | l(t), G(t)) \\
&= 1 - P(l(t + \Delta t) = l(t) + 1 | l(t), G(t)) - P(l(t + \Delta t) = l(t) - 1 | l(t), G(t)) \tag{5.1}
\end{aligned}$$

After omitting the lower order terms  $\mathcal{O}\left(\frac{l(t)}{G(t)}\right)$ , the evolution of mer copy number

can be written as:

$$\begin{aligned}
P(l(t + \Delta t) = l(t) + 1 | l(t), G(t)) &= (p_1(m)l(t) + q_1(m)G(t)) \Delta t \\
P(l(t + \Delta t) = l(t) - 1 | l(t), G(t)) &= p_{-1}(m)l(t) \Delta t \\
P(l(t + \Delta t) = l(t) | l(t), G(t)) &= 1 - ((p_1(m) + p_{-1}(m))l(t) - q_1(m)G(t)) \Delta t
\end{aligned}$$

The parameters in mer copy number evolution,  $p_1(m)$ ,  $q_1(m)$  and  $p_{-1}(m)$ ,

are functions of mer size  $m$ , and can be expressed using the parameters in the genome evolution model:

$$\begin{aligned}
 p_1(m) &= -\frac{2}{3}p_r - \frac{1}{3}p_d + \sum_{x=m}^X p_r f_r(x)(x-m+1) \\
 q_1(m) &= \frac{1}{N} \left( 2(m-1)p_r + 2(m-1)p_d + 2p_s + 2 \sum_{x=1}^{m-1} p_r f_r(x)x + 2(m-1) \sum_{x=m}^X p_r f_r(x) \right) \\
 p_{-1}(m) &= \left(m - \frac{5}{3}\right)p_r + \left(m - \frac{5}{3}\right)p_d + mp_s + L_d
 \end{aligned} \tag{5.2}$$

## 2. Amino Acids

There are two major differences between the nucleotide model and the amino acid model. First, while the nucleotide model uses a four-letter alphabet, the amino acid model uses twenty. Second, while duplicated segments can be inserted into the nucleotide sequences in both directions, for amino acids there is only one direction since the insertion in the opposite direction can change the coded amino acids. These differences lead to small changes in the representation of the mer copy number in amino acid evolution. However, the general form is not changed. For amino acid:

$$\begin{aligned}
p_1(m) &= -\frac{2}{19}p_r - \frac{1}{19}p_d + \sum_{x=m}^X p_r f_r(x)(x-m+1) \\
q_1(m) &= \frac{1}{N} \left( (m-1)p_r + (m-1)p_d + p_s + \sum_{x=1}^{m-1} p_r f_r(x)x + (m-1) \sum_{x=m}^X p_r f_r(x) \right) \\
p_{-1}(m) &= \left(m - \frac{21}{19}\right)p_r + \left(m - \frac{21}{19}\right)p_d + mp_s + L_d
\end{aligned} \tag{5.3}$$

Therefore, for sequences with uniformly random distribution, the parameter  $p_{-1}(m)$  in mer copy number evolution is linearly related to the mer size  $m$ , and the total mutational rate  $p_s + p_r + p_d$ .

#### *Sequences with Correlation*

However, for real biological sequences, the behavior of  $p_{-1}(m)$  deviates from the linear relationship with mer size  $m$  (exhibits a quadratic-like behavior, see Figure 26), indicating that the random model cannot be naively applied to real biological sequences where correlation exists. Such deviation is probably caused by the over-simplified assumption in our random model. Namely, the frequency of the mers is only dependent on their length, whereas in real sequences it depends on the frequencies of its submers and subsequences (see Figure 27). Notice that if we use a real genomic sequence as our initial common ancestral sequence in the *in silico* simulation experiments, it counts for part of the deviation in behavior.

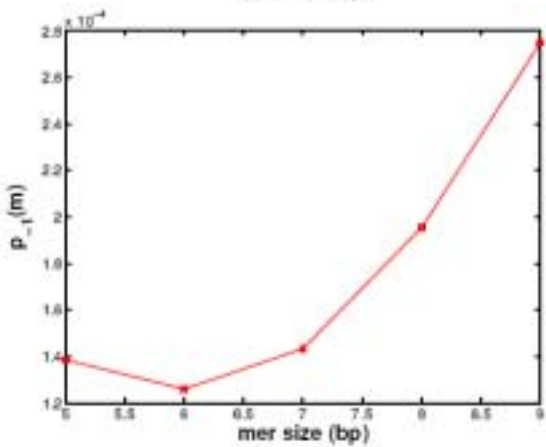
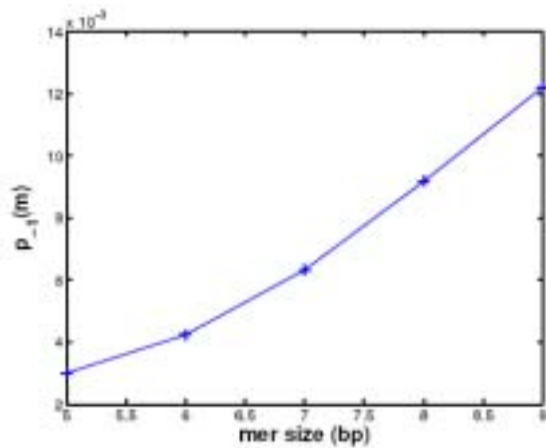
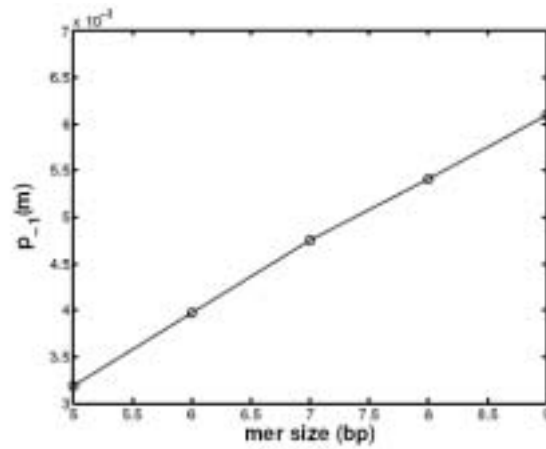


Figure 26. The relation between estimated  $p_{-1}(m)$  and the mer size  $m$  according to the random sequence model. The upper-most panel shows the result from an *in silico* simulation using purely random sequences as common ancestral genomes.

The middle panel shows the result from an *in silico* simulation using the genomic sequence of a bacterial phage (Phi6) as the common ancestral genome. The lower-most panel shows the result of comparing two phage genomes (Phi6 and Phi8).

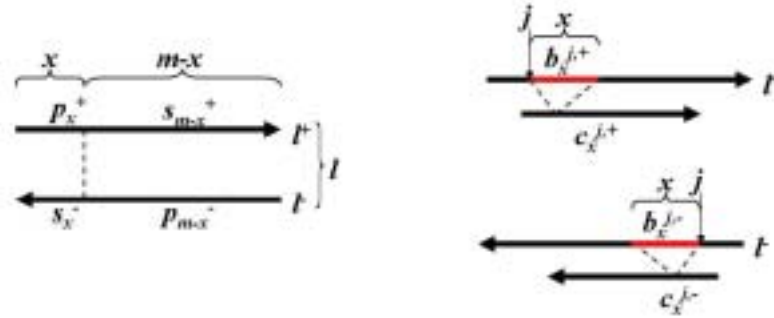


Figure 27. Notations for the copy numbers of the submers and subsequences in a length- $m$  mer. + and – superscripts indicate the direction of the mer as forward or reverse complements, respectively.  $p_x$ : copy number of the length- $x$  prefix submer;  $s_x$ : copy number of the length- $x$  suffix submer;  $b_x^j$ : copy number of the length- $x$  substring starting at position  $j$ ;  $c_x^j$ : copy number of the combined substrings excluding the length- $x$  substring starting at position  $j$ .

Therefore, we further refined our model for the nucleotides to incorporate the correlation in the sequences. In this model, the expected copy number of a mer would depend on the copy numbers of its sub-mers and sub-sequences, and the conditional probabilities of the copy number changes can be described as:

$$\begin{aligned}
& P(l(t+\Delta t) = l(t) + 1 | l(t), G(t)) \\
&= \sum_{x=m}^X \left( p_r f_r(x) (x-m+1) l(t) \left(1 - \frac{(m-1)l(t)}{G(t)}\right) \right) \Delta t + \\
& \sum_{x=m}^X \left( p_r f_r(x) \sum_{i=1}^{m-1} (p_i^+ s_i^+ + p_i^- s_i^- + p_i^+ p_{m-i}^- + s_i^+ s_{m-i}^- - \frac{1}{2} l(t) (p_i^+ + s_i^+ + p_{m-i}^- + s_{m-i}^-)) \frac{1}{G(t)} \right) \Delta t + \\
& \sum_{x=1}^{m-1} \left( p_r f_r(x) \frac{1}{2} \sum_{j=2}^{m-x} (c_x^{j,+}(t) + c_x^{j,-}(t)) (b_x^{j,+}(t) + b_x^{j,-}(t)) \frac{1}{G(t)} \right) \Delta t + \\
& \sum_{x=1}^{m-1} \left( p_r f_r(x) \sum_{i=1}^x (p_i^+ s_i^+ + p_i^- s_i^- + p_i^+ p_{m-i}^- + s_i^+ s_{m-i}^- - \frac{1}{2} l(t) (p_i^+ + s_i^+ + p_{m-i}^- + s_{m-i}^-)) \frac{1}{G(t)} \right) \Delta t + \\
& \sum_{x=1}^X \left( p_d f_d(x) \sum_{i=1}^{m-1} (p_i^+ s_i^+ + p_i^- s_i^-) \left(1 - \frac{l(t)(x+m-1)}{G(t)}\right) \frac{1}{G(t)} \right) \Delta t + \\
& p_s l(t) \Delta t
\end{aligned}$$

$$\begin{aligned}
& P(l(t+\Delta t) = l(t) - 1 | l(t), G(t)) \\
&= \sum_{x=m}^X \left( p_r f_r(x) l(t) \sum_{i=1}^{m-1} \left(1 - \frac{(x-m+1)l(t)}{G(t)} - \frac{1}{2} \frac{p_i^+ + s_i^+ + p_{m-i}^- + s_{m-i}^-}{G(t)}\right) \right) \Delta t + \\
& \sum_{x=1}^{m-1} \left( p_r f_r(x) l(t) (m-1 - \sum_{i=1}^x \frac{1}{2} (p_i^+ + s_i^+ + p_{m-i}^- + s_{m-i}^-)) \frac{1}{G(t)} \right) \Delta t + \\
& \sum_{x=m}^X \left( p_d f_d(x) l(t) (x-m+1) \left(1 - \sum_{i=1}^{m-1} \left(\frac{p_i^+ s_i^+ + p_i^- s_i^-}{G(t)^2}\right)\right) \right) \Delta t + \\
& \sum_{x=m}^X \left( p_d f_d(x) \left(2(m-1)l(t) - l^+(t) \sum_{i=1}^{m-1} \left(\frac{p_i^+ + s_i^+}{G(t)}\right) - l^-(t) \sum_{i=1}^{m-1} \left(\frac{p_i^- + s_i^-}{G(t)}\right)\right) \right) \Delta t + \\
& \sum_{x=1}^{m-1} (p_d f_d(x) l(t) (m-x-1)) \Delta t + \\
& \sum_{x=1}^{m-1} \left( p_d f_d(x) \left(2xl(t) - l^+(t) \sum_{i=1}^x \left(\frac{p_i^+ + s_i^+}{G(t)}\right) - l^-(t) \sum_{i=1}^x \left(\frac{p_i^- + s_i^-}{G(t)}\right)\right) \right) \Delta t + \\
& p_s l(t) m \Delta t
\end{aligned} \tag{5.4}$$

After a series of simplifications (see Appendix B), we get



$$\begin{aligned}
& P(l(t + \Delta t) = l(t) + 1 | l(t), G(t)) \\
&= \left( \sum_{x=m}^X p_r f_r(x) x \right) \left( l(t) \left( 1 - \frac{(m-1)l(t)}{G(t)} \right) \right) \Delta t + \\
& \left( \sum_{x=m}^X p_r f_r(x) \right) \left( -(m-1)l(t) \left( 1 - \frac{(m-1)l(t)}{G(t)} \right) + \frac{2\sigma(t) - l(t)\delta(t)}{G(t)} \right) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_r f_r(x) x \right) \left( \frac{1}{m-1} \frac{\sigma(t)}{G(t)} \right) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_r f_r(x) \right) \left( \frac{\sigma(t)}{G(t)} - l(t) \frac{\delta(t)}{G(t)} \right) \Delta t + \\
& \left( \sum_{x=1}^X p_d f_d(x) x \right) \left( -l(t) \frac{\sigma(t)}{G(t)^2} \right) \Delta t + \\
& \left( \sum_{x=1}^X p_d f_d(x) \right) \left( \left( 1 - \frac{(m-1)l(t)}{G(t)} \right) \frac{\sigma(t)}{G(t)} \right) \Delta t + \\
& p_s l(t) \Delta t
\end{aligned}$$

$$\begin{aligned}
& P(l(t + \Delta t) = l(t) - 1 | l(t), G(t)) \\
&= \left( \sum_{x=m}^X p_r f_r(x) x \right) \left( -(m-1) \frac{l(t)^2}{G(t)} \right) \Delta t + \\
& \left( \sum_{x=m}^X p_r f_r(x) \right) \left( l(t) \left( (m-1) \left( 1 + \frac{(m+1)l(t)}{G(t)} \right) - \frac{\delta(t)}{G(t)} \right) \right) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_r f_r(x) x \right) \left( -\frac{l(t)}{m-1} \frac{\delta(t)}{G(t)} \right) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_r f_r(x) \right) (l(t)(m-1)) \Delta t + \\
& \left( \sum_{x=m}^X p_d f_d(x) x \right) \left( l(t) \left( 1 - \frac{\sigma(t)}{G(t)^2} \right) \right) \Delta t + \\
& \left( \sum_{x=m}^X p_d f_d(x) \right) \left( l(t) \left( -(m-1) \left( 1 - \frac{\sigma(t)}{G(t)^2} \right) + 2(m-1) - \frac{\delta(t)}{G(t)} \right) \right) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_d f_d(x) x \right) l(t) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_d f_d(x) \right) \left( l(t) \left( m-1 - \frac{\delta(t)}{G(t)} \right) \right) \Delta t + \\
& p_s l(t) m \Delta t
\end{aligned} \tag{5.5}$$

where

$$\begin{aligned}
\sigma(t) &= \sum_{i=1}^{m-1} \left( p_i^+(t) s_i^+(t) + p_i^-(t) s_i^-(t) \right) \\
&= \sum_{i=1}^{m-1} \left( p_i^+(t) p_{m-i}^-(t) + s_i^+(t) s_{m-i}^-(t) \right) \\
&= \left( \frac{l(t)}{2} \right)^{\frac{m-2}{m-1}} \left( (p_1^+(t))^{\frac{m}{m-1}} \left( \frac{m}{2} - 1 + \frac{m}{2} \frac{p_1^+(t)}{p_1^-(t)} \right) + (p_1^-(t))^{\frac{m}{m-1}} \left( \frac{m}{2} - 1 + \frac{m}{2} \frac{p_1^-(t)}{p_1^+(t)} \right) \right); \\
\delta(t) &= \sum_{i=1}^{m-1} \left( p_i^+(t) + s_i^+(t) \right) \\
&= \sum_{i=1}^{m-1} \left( p_i^-(t) + s_i^-(t) \right) \\
&= p_1^+(t) \frac{1 - \frac{l(t)}{2p_1^+(t)}}{1 - \left( \frac{l(t)}{2p_1^+(t)} \right)^{\frac{1}{m-1}}} + p_1^-(t) \frac{1 - \frac{l(t)}{2p_1^-(t)}}{1 - \left( \frac{l(t)}{2p_1^-(t)} \right)^{\frac{1}{m-1}}};
\end{aligned} \tag{5.6}$$

### Mer Copy Number Evolution Parameter Estimation

The parameters in the mer copy number evolutionary process are estimated using Maximum Likelihood method (ML). From the above equations, we can see that the evolution of a mer copy number is a non-stationary Markov process. The Markov transitional matrix for mers of size  $m$  is a tridiagonal matrix defined by

$$P_{i+1,i}(\mathbf{m}, t) = P(l(t + \Delta t) = i + 1 | l(t) = i, \mathbf{G}(t));$$

$$P_{i-1,i}(\mathbf{m}, t) = P(l(t + \Delta t) = i - 1 | l(t) = i, \mathbf{G}(t));$$

The values in the matrix change over time with the genome size  $G(t)$ . The formula of  $P_{i+1,i}(\mathbf{m}, t)$  and  $P_{i-1,i}(\mathbf{m}, t)$  are listed in (5.1) for non-correlated sequences, and in (5.5) for sequences with correlations, respectively. Let

$\alpha_k(t)$  denote the probability of a mer evolving into  $k$  copy numbers at time  $t$ .

Then given a small evolutionary time interval  $\Delta t$ :

$$\begin{pmatrix} \alpha_0(t + \Delta t) \\ \alpha_1(t + \Delta t) \\ \alpha_2(t + \Delta t) \\ \vdots \\ \vdots \end{pmatrix} = M(t) \begin{pmatrix} \alpha_0(t) \\ \alpha_1(t) \\ \alpha_2(t) \\ \vdots \\ \vdots \end{pmatrix}$$

$$M(t) = \begin{pmatrix} 1 - P_{1,0}(m, t) & P_{0,1}(m, t) & 0 & \dots \\ P_{1,0}(m, t) & 1 - P_{0,1}(m, t) - P_{2,1}(m, t) & P_{1,2}(m, t) & \dots \\ 0 & P_{2,1}(m, t) & 1 - P_{1,2}(m, t) - P_{3,2}(m, t) & \dots \\ 0 & 0 & P_{3,2}(m, t) & \dots \\ \vdots & \vdots & \vdots & \dots \end{pmatrix}$$

Over an evolutionary time  $T$ :

$$\begin{pmatrix} \alpha_0(T) \\ \alpha_1(T) \\ \alpha_2(T) \\ \vdots \\ \vdots \end{pmatrix} = M_T \begin{pmatrix} \alpha_0(0) \\ \alpha_1(0) \\ \alpha_2(0) \\ \vdots \\ \vdots \end{pmatrix}, M_T = \prod_{i=0}^{T/\Delta t} M(i\Delta t). \quad (5.7)$$

The optimal parameters  $\bar{\theta}$  for a pair of genomes (A and B) are those that maximize the likelihood of observing the mer copy numbers in one genome (B) assuming that they have evolved from the corresponding mer copy numbers in the other genome (A) through the nonstationary Markov evolutionary process described above.

We denote the copy number of the 1st, 2nd, ..ith, ..Nth mer in genome A as

$\vec{a} = \{a_1, a_2, \dots, a_i, \dots, a_N\}$ , and those in genome B as  $\vec{b} = \{b_1, b_2, \dots, b_i, \dots, b_N\}$ , respectively.

Then, the above statement can be written as:

$$\vec{\theta} = \arg \max_{\vec{\theta}} \left( P(\vec{b} | \vec{a}, Model) \right).$$

If we assume that the evolution of different mers are independent of each other, then:

$$\begin{aligned} \vec{\theta} &= \arg \max_{\vec{\theta}} \left( \prod_{i=1}^N P(b_i | a_i, Model) \right) \\ &= \arg \max_{\vec{\theta}} \left( \sum_{i=1}^N \log P(b_i | a_i, Model) \right) \end{aligned} \quad (5.8)$$

where  $P(b_i | a_i, Model) = \alpha_{b_i}(T)$ ; given  $M_T$  and  $\alpha(0)$ , in which  $\alpha_{a_i}(0) = 1$  and  $\alpha_k(0) = 0$  and for all  $k \neq a_i$ .

During the parameter estimation procedure, we artificially choose a genome growth rate  $g$ . Given a  $g$  value, the evolution time between two genomes A and B is:

$$T = \frac{|\log(G_A) - \log(G_B)|}{|g|}. \quad (5.9)$$

$G_A$  and  $G_B$  are the lengths of the sequences A and B, respectively. The artifact introduced by an arbitrary  $g$  will be cancelled out during the evolutionary distance computation step (discussed later). The log likelihood for a particular set of parameters is computed as described above, and the optimal

parameter set is chosen by searching for maximum log likelihood using simplex method.

When no correlation in the sequences is considered,

$$\vec{\theta} \equiv \{p_1, p_{-1}, q_1\};$$

and when correlation is considered,

$$\vec{\theta} \equiv \{p_r, p_d, p_s, B_d\}.$$

### Estimation of Genome Evolutionary Distance

Notice that for both random nucleotide sequences and amino acid sequences, the parameter  $p_{-1}(m)$  is linearly related to the mer size  $m$ , and can be written as:

$$p_{-1}(m) = (p_r + p_d + p_s)m + C$$

$C$  is a constant:  $L_d - \frac{5}{3}(p_r + p_d)$  for nucleotides, and  $L_d - \frac{21}{19}(p_r + p_d)$  for amino acids. If we estimate the value of  $p_{-1}(m)$  for different mer sizes  $m$ , we can get the sum of the mutation rates,  $(p_r + p_d + p_s)$ , by linear regression between  $p_{-1}$  and  $m$ . For correlated sequences, we can get the sum  $(p_r + p_d + p_s)$  directly from parameter estimation<sup>6</sup>. The evolutionary distance between the two genomes can

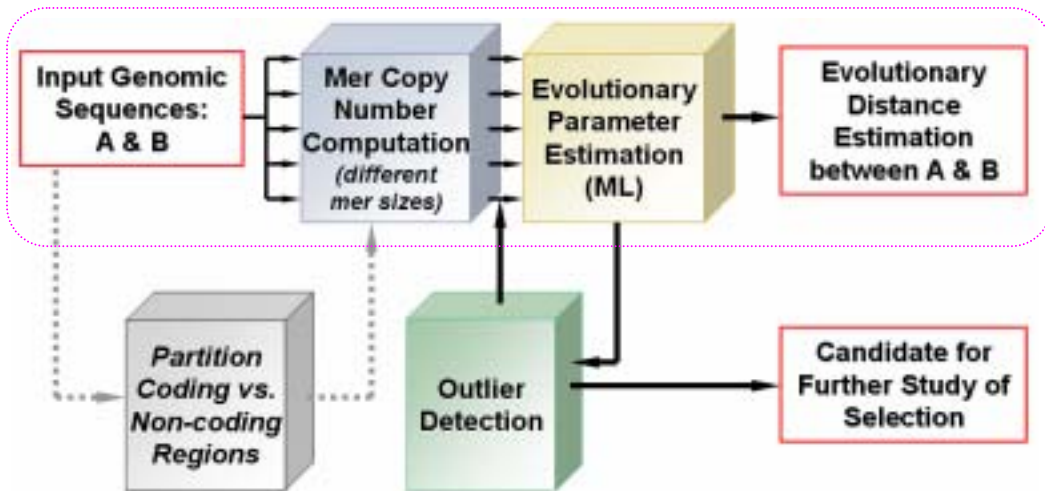
---

<sup>6</sup> So far, in both methods (with or without consideration of sequence correlation), we can only accurately estimate the total mutation rate  $(p_s + p_r + p_d)$ . The individual rates cannot be separately estimated. This is possibly because some of the terms in (5.5) are much smaller than the other ones, effectively decreasing the degree of freedom in the system.

then be estimated by the total number of evolutionary events computed as follows:

$$\begin{aligned}
 D &= \int_{t=0}^T (p_r + p_d + p_s)G(t)dt \\
 &= \int_{t=0}^T (p_r + p_d + p_s)G(A)e^{gt} dt \\
 &= \frac{(p_r + p_d + p_s)}{|g|} |G(B) - G(A)|
 \end{aligned}
 \tag{5.10}$$

### Method Summary



**Figure 28.** The design of our method. The procedures are described in the text. The procedures within the frame have already been implemented.

The procedures for estimating the evolutionary distance between two genomes are represented schematically in **Figure 28**: Given two input genomic sequences, we first compute the (overlapping) copy numbers of all the mers of a

particular size. The parameters for the mer copy number evolutionary process are estimated by the ML method given the mer copy numbers in the two sequences and the model of the mer copy number evolution. Since the neutral evolution and homogeneous mutation rate assumptions can be violated in real biological sequences, we will perform an outlier detection step after the parameter estimation. If most of the mers evolve neutrally and have similar mutation rates, the exceptional mers will contribute very low likelihood values given the parameters estimated for neutral homogeneous evolution. The parameters are reevaluated after these outliers are omitted. These two steps can be performed iteratively until it meets a certain convergence criterion. The parameters are recorded in accordance to the mer size. The procedures described above are repeated for different mer sizes. Finally, the evolutionary distance between the two genomes is computed. The detected outliers may be subjected to more detailed analysis to determine the reasons leading to its abnormality, for example, natural selection, active transposition, *etc.* The analysis can be applied to the coding sequences, but can also be applied to the non-coding sequences if two genomes are closely related.



## 5.4 Method Verification

### Simulation Data

#### *Pair-wise distance*

To verify and evaluate our method, we estimate the total number of substitution, duplication and deletion events that separate two genomes evolved from a common ancestral genome *in silico*. We first create an artificial genome (C). Using C as an ancestral genome, we evolve it through two lineages (A and B) under a specific set of genome evolutionary rates based on our genome evolution model (Figure 29). During the *in silico* evolution, we record the total number of substitution, duplication and deletion events that occurred since the divergence of the two lineages. This is the true evolutionary distance between the two genomes A and B. We compare the true distance to the distance estimated using our method based on the mer copy number statistics in genome A and B.

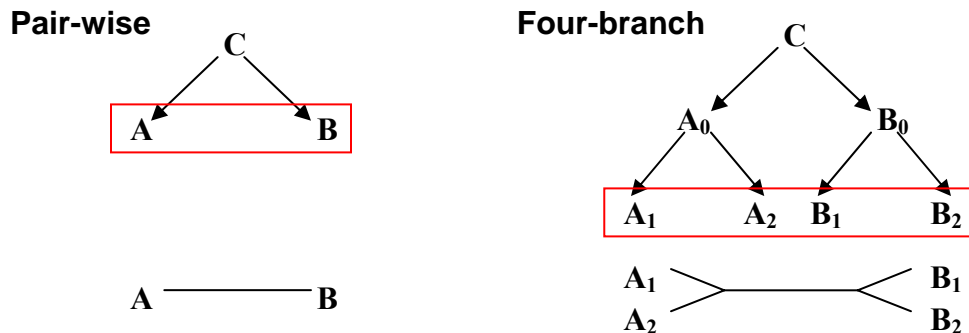


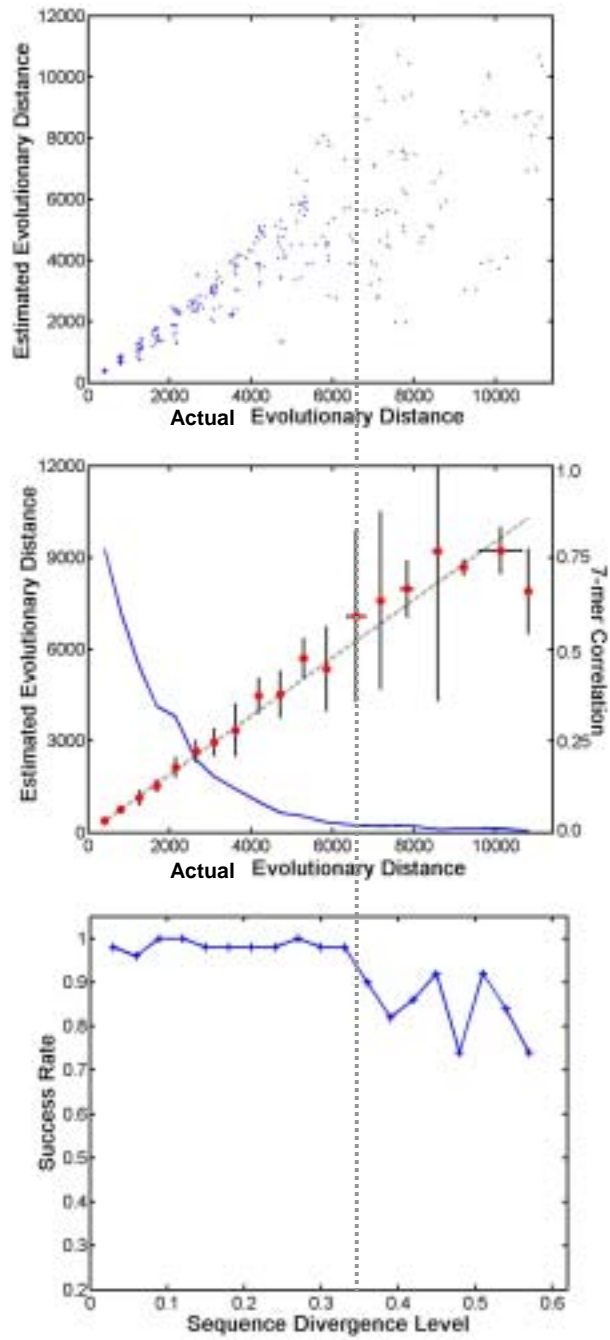
Figure 29. The scheme for *in silico* evolution simulation. The artificial

**sequences are evolved from a common ancestral genome (C) following the above tree structure, and the number of evolutionary events occurred on each branch is recorded. The evolutionary distance between the artificial sequences on the leaves of the tree (framed) is estimated using our method based on their mer statistics. For the quartet scenario, an unrooted tree is generated using the Neighbor Joining Method based on the pair-wise distances.**

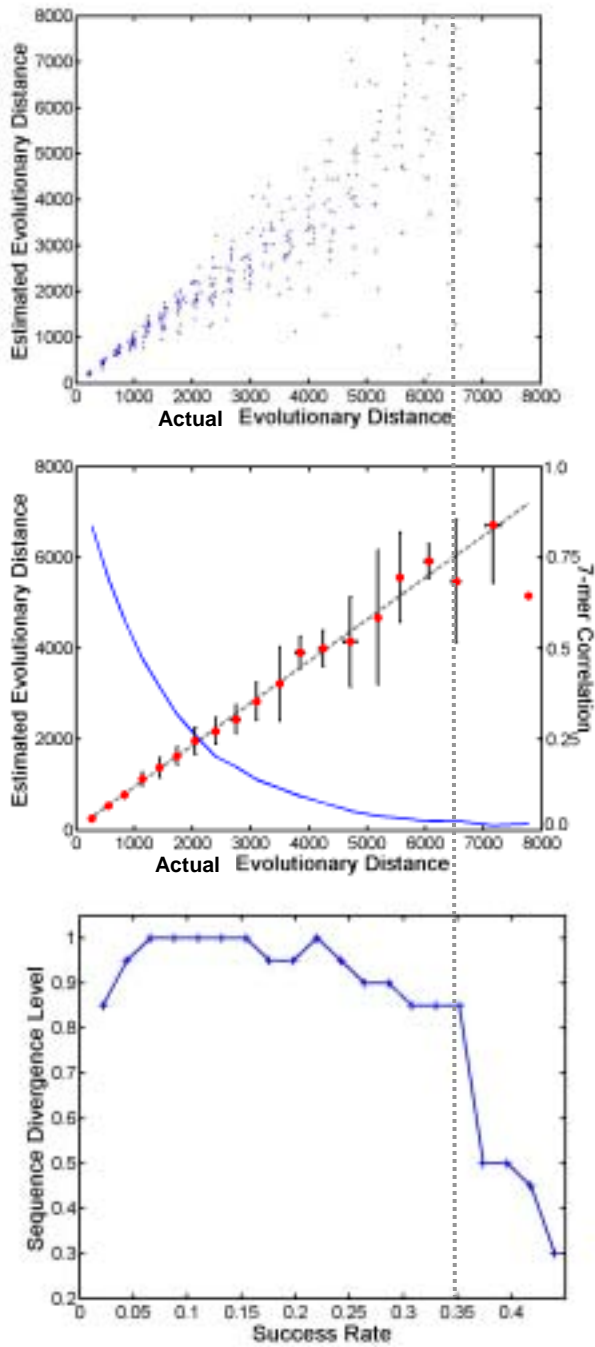
We first test our method without consideration of the correlation in the sequences on artificial sequences with random distributions. Two scenarios are tested: homogeneous mutation rates where the mutation rates for genome A and B are the same, and lineage-specific mutation rates where the mutation rates for genome A and B are different after their divergence from genome C. As seen in **Figure 30**, in both scenarios, the estimation given by our method is quite accurate when there is still a detectable level of correlation in the mer copy numbers between the two genomes, which corresponds to roughly 35% sequence divergence level given the set of mutation rates we used. The test also shows that our method is robust against lineage specific events, and does not require homogeneous mutation rates among the lineages. The 35% divergence level in non-coding regions represents quite closely related evolutionary relationship. However, our method can also be applied to coding and protein sequences, in

which 35% divergence level will imply species that are quite remotely related.

**A**



**B**

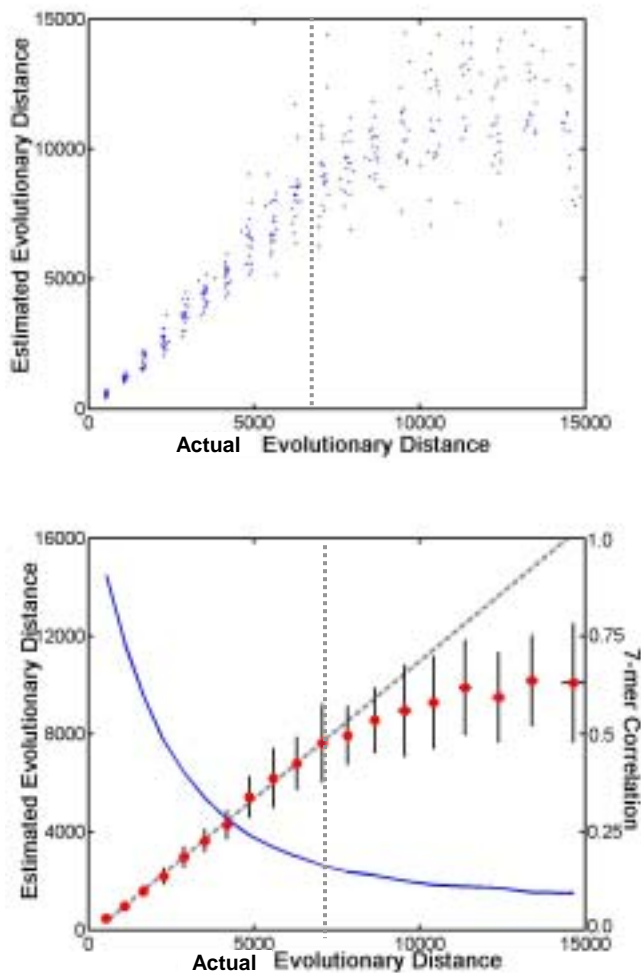


**Figure 30. Evaluation of the method without consideration of sequence correlation using *in silico* simulated data. We tested our method under two scenarios: homogeneous lineage mutation rate (A) and lineage specific mutation rate (B).**

The genome evolution parameters in A are:  $p_s=1 \times 10^{-6}$ ,  $p_r=5 \times 10^{-7}$ ,  $p_d=5 \times 10^{-7}$  for both lineages; the genome evolution parameters in B are:  $p_s^A=5 \times 10^{-6}$ ,  $p_r^A=5 \times 10^{-7}$ ,  $p_d^A=5 \times 10^{-7}$ ;  $p_s^B=1 \times 10^{-6}$ ,  $p_r^B=5 \times 10^{-7}$ ,  $p_d^B=5 \times 10^{-7}$ . 20 independent simulations are performed for each different expected sequence divergence level. The top panel shows the actual evolutionary distance and the estimated distance of each individual experiment measured by the total number of substitution, duplication and deletion events. In the middle panels, each circle represents the average of the successful trials out of the 20 simulations, and the horizontal and vertical bars represent the corresponding standard deviations in the actual and estimated distances respectively. A trial is successful if there is a good linear regression between the estimated  $p_{-1}(m)$  values and mer sizes ( $m$ ) (p-value<0.01). The success rate (number of successful trials/20) is plotted in the lower panels. As shown in the figure, our method can estimate the evolutionary distance quite accurately (fitted with the grey line of slope 1) when the sequence divergence level is below 35% (dotted line). But when the divergence level goes above 35%, the success rate drops dramatically as does the estimation accuracy, thus suggesting a possible limit of applicability of our method. This limit coincides with the disappearance of the correlation in the mer copy numbers from the two genomes, as shown in the upper panels (curve).

We also tested our method with consideration of the correlation in the sequences on artificial sequences with random distributions (Figure 31). The accuracy

levels of the two different methods are similar, although instead of failing to give an estimate, the correlation method starts to underestimate the distance (“long branch attraction”) when the divergence level is higher than 35%. However, the correlation method behave superior to the un-correlated method on real genomic sequences (see the next section).



**Figure 31.** Evaluation of the method after taking into consideration the sequence correlation using *in silico* simulated data. 50 independent simulations are

performed for each different expected sequence divergence level. The left panel shows the actual evolutionary distance and the estimated distance of each individual experiment. The right panels show the comparison between the actual evolutionary distance and the estimated distance, measured by the total number of substitution, duplication and deletion events. Each circle represents the average of the 50 simulations, and the horizontal and vertical bars represent the corresponding standard deviations in the actual and estimated distances respectively. The correlation method starts to underestimate the evolutionary distance (long branch attraction) when the divergence level between the two sequences goes above 35% (dotted line). The genome evolution parameters are:  $p_s=1 \times 10^{-6}$ ,  $p_r=5 \times 10^{-7}$ ,  $p_d=5 \times 10^{-7}$  for both lineages.

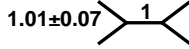
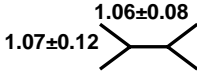
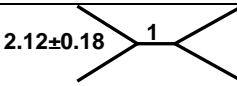
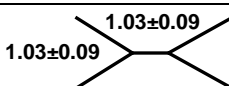
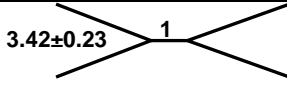
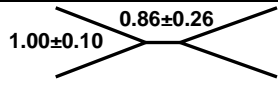
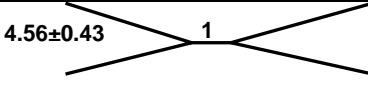
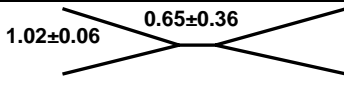
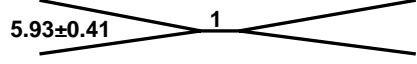
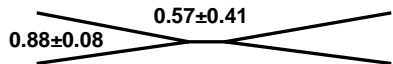
### *Quartet*

Ultimately, we would like to use our method to infer the phylogenomic relationship among different genomes in an alignment- as well as sequencing-independent manner. Before applying our method to real genomic sequence data, we first evaluated the ability of our method to recover the correct topology of the phylogeny tree using *in silico* evolution. Artificial sequences that form a four-branch unrooted tree are generated in our *in silico* evolution experiments. A common ancestral sequence (C) is evolved into two intermediate ancestral lineages ( $A_0$ ,  $B_0$ ), and the final sequences ( $A_1$ ,  $A_2$ ,  $B_1$ ,  $B_2$ ) used for

evolutionary distance estimation are formed by further evolving each intermediate ancestral sequence into two lineages respectively (see **Figure 29**). The number of evolutionary events occurred during artificial evolution on each branch is recorded. The pair-wise evolutionary distance between the final evolved sequences are estimated using our method, and a phylogeny tree is constructed using the Neighbor Joining method.

To estimate the sensitivity of our method in resolving short internal branches in the phylogeny trees, we generated quartet datasets with different length ratios between the internal and external branches. In our experiments, the method recovers the correct topology as well as the branch lengths quite faithfully when the internal branch length is at least 25% of the external branches (Table 6), suggesting that our method is quite robust.



True tree	Different topology	Inferred tree
	0 (15)	
	0 (15)	
	0 (15)	
	0 (15)	
	6 (15)	

**Table 6.** The sensitivity of our method evaluated using the *in silico* evolved quartet datasets. The true topology of the four-branch tree is shown in the left column with the relative length of the external branch to the internal branch indicated as mean±std (standard deviation). The evolutionary distances between each pair of external sequences are estimated using our method, and phylogenies are constructed. The number of times that the inferred tree has a different topology from the true tree out of the total number of experiments (in parenthesis) is recorded in the second column. When the correct topology is inferred, the relative lengths of the branches in the inferred tree compared to the branch lengths in the true trees are computed and listed in the third column, in the form of mean±std. All sequences are evolved under the parameters:  $p_s=1 \times 10^{-6}$ ,  $p_r=5 \times 10^{-7}$ ,  $p_d=5 \times 10^{-7}$  for all lineages.

## Real Genomic Data

### *Pair-wise distance*

The statistical structure of the real genomic sequences is much more complex than the artificially generated sequences, since they also contain correlations, base biases, region-specific rates, *etc.* To test if our method is applicable to real biological sequences, we chose a few pairs of orthologous sequences without rearrangements for which global alignment is applicable, and compared the estimation given by our methods (with or without the consideration of correlation in the sequence) to the results from the sequence alignment method (LAGAN) [28]. Since the sequences are chosen to be closely related, we may assume that the incidents of convergent or reverse evolution events are rare, and the number of mismatches and indels in the sequence alignment results can approximate the actual evolutionary events that occurred after the divergence of the two sequences. **Table 7** lists the estimations using the sequence alignment method and our alignment-independent methods with or without consideration of the sequence correlation. The numbers of evolutionary events computed from our correlation method are quite similar to those from the alignments, suggesting that our correlation method is also valid for real biological sequences. Within our expectation, because it disregards the correlated structure in the real

biological sequences, the method without consideration of the correlation in the sequences is ill-behaved when applied to the real sequences, constantly underestimating the distance or even failing to estimate in some cases.

Genomes (divergence)	Number of Events		
	Alignment	Un-correlated	Correlated
<i>D. melanogaster</i> vs. <i>D. pseudoobscura</i> orthologous sequence (35%)	3962	Fail	4543
Mouse vs. Rat orthologous sequence (11%)	471	457	491
Human vs. Baboon orthologous sequence (6.2%)	490	380	513
Human vs. Chimp orthologous sequence (1.6%)	127	101	142

**Table 7. Comparison of our methods with the alignment results for closely related sequences without rearrangements.**

Also observed is that the estimations from our correlation method are always slightly higher than the event counts from the alignment results, especially for the sequence pairs with slightly higher divergence level such as *D. melanogaster* vs. *D. pseudoobscura*. This may suggest that the event counts from the alignments can be viewed as a weighted parsimonious counting, and underestimate the real event number because of the occurrence of convergent or reverse evolution.

### *Quartet*

We also tested our method on real genomic sequences to determine its efficiency on phylogeny tree inference. As benchmark, we chose several orthologous regions from the multiple mammalian (human, macaque, mouse and rat) sequence alignments generated by MULTI-LAGAN [28]. The divergence level between the four species in these aligned regions is roughly as follows: Human vs. Macaque: ~5%; Human or Macaque vs. Mouse or Rat: ~35%; Mouse vs. Rat: ~12%. Unrooted four-branch trees are generated based on the number of event counts from sequence alignments or the evolutionary distances estimated by our method. The topology of the trees and their branch lengths are compared. In all three cases, the method faithfully recovers the topology (Table 8), however, the internal branches is slightly underestimated in some cases, mostly because the method starts to underestimate the distance when the divergence level in the sequence pairs is larger than 35%.

Examples	Tree built from alignment event counts	Tree built from our method
1		
2		
3		

**Table 8.** The phylogeny trees inferred from the orthologous sequences of four mammalian species. In the left column, the trees are inferred by the pair-wise distance between sequence pairs measured by the total event counts in the sequence alignments. The trees in the right column are inferred by the pair-wise evolutionary distances estimated by our method. The trees are constructed by Neighbor Joining method.

## 5.5 Study on the Modulation of Substitution and Indel Events

Many investigations have been conducted to examine the variation in the rates of different evolutionary events along a genome [157][74][188][140]. In these studies, the rates are found to be highly correlated among different

mutational processes, including substitution, deletion and duplication. It is now well accepted that instead of a universal molecular clock, the substitution rate, as well as duplication and deletion rates varies along different evolutionary lineages [132]. The differences in these rates contribute to generate different statistic composition and structures in various genomes. For example, a higher deletion rate has been suggested to be responsible for the smaller size of some genomes [131][132], and a higher transposon amplification rate is one of the factors that caused human genome expansion [105].

Some studies have been done to examine the correlation between the duplication and deletion rates of the functional units (genes or transposons) in the genome and the substitution rate. Studies on microbial genomes suggest that the free living bacteria with large effective population sizes have a relative higher rate in genome rearrangement, including gene duplication and deletion events, relative to substitutions [166]. A recent study on alpha-proteobacteria [24] could not find any correlation among the rates of gene duplication or deletion and substitution. Although the topology of the phylogeny trees built by maximum parsimony (MP) based on indel events are usually consistent with the trees built based on substitutions [91], to test the consistency on the branch lengths, one would need an efficient ML method to estimate the branch length based on indel events.

However, currently such a method is lacking. Therefore, in large, whether and how the rate of the indel (deletion and duplication) events on various scales changes with regard to the substitution rate along the evolutionary history of the genome is still a mostly unexplored problem.

Based on our knowledge from molecular biology, the rates of the different evolutionary processes may change in many different ways. For example, some DNA surveillance machinery [100] can target both mismatches (caused by substitutions) and “bubble” structures (caused by indel events). If the relative efficiency of the DNA surveillance machinery in eliminating substitutions and indels is well conserved during evolution, as well as the relative effect of the cellular or environmental factors on the mutational events, the changes in the substitution and indel rates should co-vary along the evolutionary time in different lineages. The co-variation in the mutational rates can also be caused by changes in the effective population sizes. On the other hand, some of the DNA surveillance machineries and external factors may be more likely to affect a specific type of mutation [76]. If these more specific factors are differentially regulated in different lineages, then the relative rate of the substitution and indel events can change. However, because of the constraints on how much mutation a genome can tolerate without losing its fitness, the total rate of the various

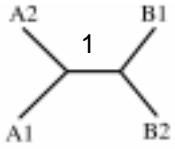
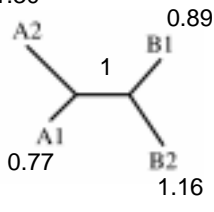
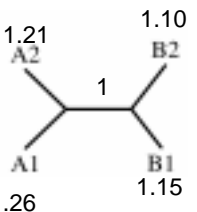
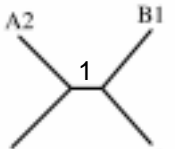
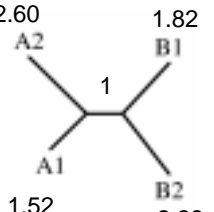
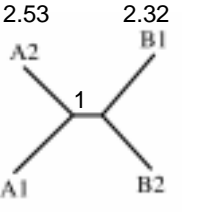
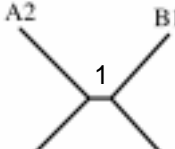
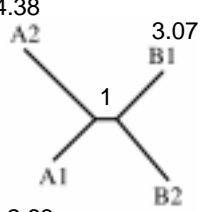
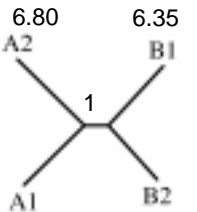
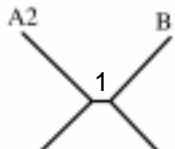
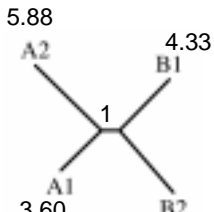
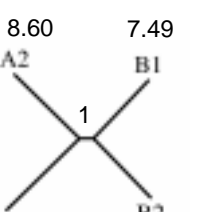
evolutionary events may be under strict control. In such a case, the substitution and indel rates may not change in a tightly correlated manner, but their sum may remain constrained.

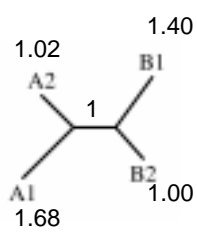
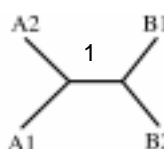
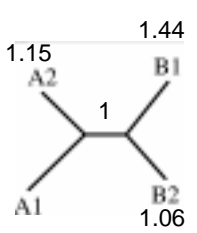
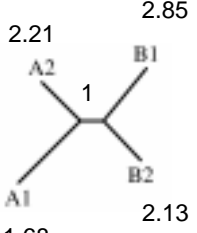
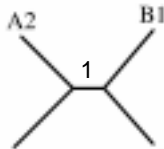
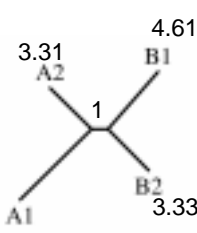
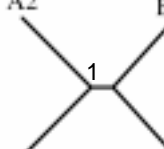
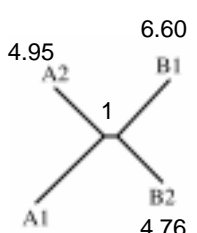
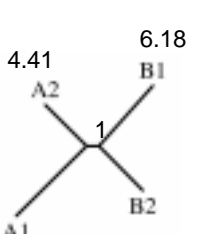
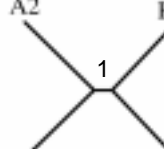
The ability of our method to detect evolutionary distance between two sequences as the sum of the substitution and indel events makes it possible to study the modulation of the substitution and indel rates in different sequences when the evolutionary distance estimation from our method is compared to that from the methods based on substitution only.

To demonstrate the application of our method in the study of the modulation of the substitution and indel events, we simulated different scenarios (**Table 9**) of substitution-indel modulations in a set of quartets. In the first scenario, the substitution rates in each lineage remain the same, while the indel rates are much larger in two of the lineages. The substitution-based method gives a tree with equal branch lengths, ignoring the accelerated indel rates in two of the lineages. However, our method generates a tree with longer branches for two of the lineages with higher indel rates, because the evolutionary distance estimated using our method also incorporates indel events. In the second scenario, the substitution and indel rates vary among different lineages, but



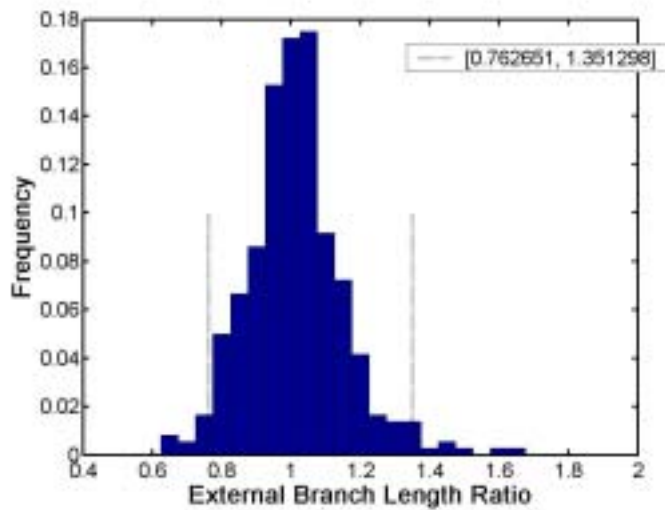
compensate each other in total mutational rate. In this case, the substitution-based method generates a tree with unequal branch lengths, while our method gives a tree with equal branch lengths. The two scenarios are simulated in the *in silico* evolution, and the trees generated using different methods are compared.

	True tree	Substitution-based method		Our method			
Scenario 1	 $1.24 \pm 0.14$	0		5		0	$A_1A_2$
		2		1		1	$A_1B_1$
		0		5		0	$A_1B_2$
		1		3		0	$A_2B_1$
		0		0		0	$A_2B_2$
		0		2		0	$B_1B_2$
		0		5		0	$A_1A_2$
	 $2.53 \pm 0.33$	0		5		0	$A_1A_2$
		0		1		1	$A_1B_1$
		0		5		0	$A_1B_2$
		0		4		1	$A_2B_1$
		0		0		0	$A_2B_2$
		0		2		0	$B_1B_2$
		0		5		0	$A_1A_2$
	 $4.05 \pm 0.45$	0		5		0	$A_1A_2$
		0		1		0	$A_1B_1$
		0		5		0	$A_1B_2$
		0		4		0	$A_2B_1$
		0		0		0	$A_2B_2$
		0		1		0	$B_1B_2$
		0		5		0	$A_1A_2$
	 $5.47 \pm 0.43$	0		5		0	$A_1A_2$
		0		1		0	$A_1B_1$
		0		5		0	$A_1B_2$
		0		4		0	$A_2B_1$
		0		1		0	$A_2B_2$
		0		1		1	$B_1B_2$
		0		5		0	$A_1A_2$

	True tree		Substitution-based method		Our method			
Scenario 2		5		0		2	A <sub>1</sub> A <sub>2</sub>	
		0		0		0	A <sub>1</sub> B <sub>1</sub>	
		5		0		3	A <sub>1</sub> B <sub>2</sub>	
		4		0		2	A <sub>2</sub> B <sub>1</sub>	
		0		0		3	A <sub>2</sub> B <sub>2</sub>	
		3		0		3	B <sub>1</sub> B <sub>2</sub>	
		2.21				5		0
	0	0	2		A <sub>1</sub> B <sub>1</sub>			
	5	0	3		A <sub>1</sub> B <sub>2</sub>			
	2	0	1		A <sub>2</sub> B <sub>1</sub>			
	0	0	0		A <sub>2</sub> B <sub>2</sub>			
	2	0	0		B <sub>1</sub> B <sub>2</sub>			
	3.31		5			0		
	0		0	0		A <sub>1</sub> B <sub>1</sub>		
	5		0	3		A <sub>1</sub> B <sub>2</sub>		
	4		0	2		A <sub>2</sub> B <sub>1</sub>		
	0		0	0		A <sub>2</sub> B <sub>2</sub>		
	3		0	1		B <sub>1</sub> B <sub>2</sub>		
	4.41			4			0	
	0	0		0	A <sub>1</sub> B <sub>1</sub>			
	4	0		0	A <sub>1</sub> B <sub>2</sub>			
	4	0		1	A <sub>2</sub> B <sub>1</sub>			
	0	0		0	A <sub>2</sub> B <sub>2</sub>			
	2	0		0	B <sub>1</sub> B <sub>2</sub>			

**Table 9. The two different scenarios of differential modulation of the substitution and indel events. The trees inferred based on the substitution in sequence alignments using maximum likelihood (ML) method and the tree inferred using the evolutionary distances estimated from our method are compared to the true tree side by side. Each row represents the average of five independent quartet simulations. The average lengths of the external branches relative to the internal branches are marked. The internal branch corresponds to approximately 5% sequence divergence level. Neighbor Joining method is used to construct the trees from the evolutionary distance estimations. The number of trees (out of five trials) with significant differences in a particular pair of external branches is listed. Two external branch lengths are significantly different if their length ratio falls in the upper or lower 2.5% of the ratio distribution empirically computed from simulated quartets with expected equal external branches.**

Some of the results are listed in **Table 9**. To measure the significance of the length difference among the external branches, we simulated quartet data with expected equal length external branches (see Table 6), and computed the distribution of the ratios between the external branches from the four-branch unrooted tree (Figure 32). The lengths of two external branches are significantly different if their ratio falls into the upper or lower 2.5% of the distribution. The numbers of trees in each experiment that have significant difference in a particular pair of external branches are listed in Table 9.



**Figure 32. The empirical distribution of the ratio among external branches computed from the simulated quartet datasets where equal external branch lengths are expected. The dotted lines show the upper and lower 2.5% of the distribution, whose explicit values are listed on the upper right part of the plot.**

Our method, although tends to underestimated the relative length of the internal branch when the divergence level between the sequences increases, is able to reconstruct the relative lengths of the external branches quite faithfully to the true tree, and effectively incorporates the influence of lineage-specific variations in the relative substitution and indel rates. However, the trees constructed using substitution-based method cannot reflect the effect of indel events. The comparison between the trees in their branch lengths gives an estimate on

the relative rate of indel events with regard to substitutions. For example, in the second row of scenario 1, according to the true tree, the quartet should have equal external branch lengths (no significant branch length difference is detected). Consistent with the true tree, in most of the cases there is no significant differences between the external branch lengths in the trees inferred using our method. However, in the trees inferred by substitution-based method, some of the branch pairs are found to be significantly different. For instance,  $A_2$  branch is not significantly different from the  $A_1$  branch based on our method, but is consistently longer in the substitution-based trees. Such an observation suggests that the substitution rate in  $A_2$  lineage is significantly higher than that in the  $A_1$  lineage; however, since there is no difference in their total mutation rates, the indel rate in  $A_1$  lineage is much higher than that in the  $A_2$  lineage. Therefore, our method allows one to investigate the difference in the modulation of the substitution and indel processes in difference sequences.

## **5.6 Summary**

From the parsimonious genome evolution model, we have derived an alignment and even sequencing-independent method for evolutionary distance estimation based on mer statistics. The method treats genome evolution as a non-

stationary Markov process, and the evolutionary distance is measured by the total number of substitution, duplication and deletion events since the divergence of the two sequences, and estimated by ML given the mer statistics of the two sequences. During the development of the method, we first introduced a method suitable for sequences with random distributions. To incorporate the correlation between the copy number of a mer and its sub-mer and sub-sequences, we developed a more complicated but more realistic method applicable to real genomic data with complex structures.

Compared with the other alignment-independent methods that are currently available [65][78][137][162], our method offers an explanation of why genome composition-based method works. In addition, its estimated evolutionary distance has explicit biological meanings. Furthermore, since the method allows growth or reduction of the genome size, it automatically takes into account the mer copy number distribution shifts between genomes of different sizes. Whereas most of the other phylogenetic methods do not consider the changes in the sequence sizes, it is not clear how well they can scale when the two sequences under comparison have very different lengths.

We have evaluated our method using both simulation data and real genomic

sequence data to test its ability in recovering both the pair-wise evolutionary distance and the topology of the underlying quartet phylogenetic tree. The method can faithfully recover the true distance and phylogeny tree topology when the divergence level between the sequences is under 35%.

One of the major differences between our method and most of the commonly used phylogenetic methods is that our evolutionary distance measurement also incorporates the number of indel events besides substitution, while most of the other methods are based purely on the substitution changes. Therefore, when combined with the substitution-based method, one can use our method to study the modulation of the indel events relative to substitutions in different sequences. We have demonstrated some examples from *in silico* evolution. When more computationally efficient implementation of the method is available, we can apply it to whole-genomic sequences.

Finally, although so far our method has been described based on mer statistics, it can be easily generalized to the statistics of other genomic components, such as protein domains, gene families, etc., by changing the unit size in the method from a small mer to a larger functional unit while keeping the same evolutionary model.



# Chapter 6

## Summary

### 6.1 Summary

In this thesis, we have examined the duplication process in genome evolution from different angles using various Markov models, including the mechanisms of the segmental duplications in the mammalian genomes; the analysis and modeling of the duplication effect on the statistical structures of various genomes; and the measurement of evolutionary distance incorporating duplication events.

#### **Mechanisms of the Recent Segmental Duplications in Mammalian Genomes**

To uncover the mechanisms for the recent segmental duplications in the mammalian genomes, we analyzed the duplication flanking sequences with various methods in great details, and modeled the evolution of the flanking sequences after duplication as a Markov process that incorporates the death and birth of the sequence elements involved in the duplication process [194]. From our analyses, we found that about 12% of the recent segmental duplications in the mammalian genomes were caused by recombination mechanism between

homologous interspersed repeats that are most recently active, demonstrating the dynamic interaction between duplicated sequences of different scales. We also found that a part of the segmental duplications which cannot be explained by the repeat recombination mechanism are correlated with physical instabilities around their duplication breakpoints. Similar physical properties have been found at the fragile sites that usually lead to genetic instabilities [115][118]. On the other hand, segmental duplications have also been found to be involved in large scale genome rearrangements [11][6][51][85]. The association between such “fragile” sites and the duplicated segments suggests a complex interaction between duplication and other genome evolutionary events.

Similar processes have also been found in developing cancer cells [106][107][2][134], which may be caused by similar mechanisms. Currently, with techniques such as array-CGH, the copy number fluctuations in the cancer genome, i.e. duplication and deletion events, can be detected. When enough resolution is reached on the breakpoint locations of these events, similar sequence analyses to the ones we used here can be applied to suggest their mechanisms and dynamics.

## **Analysis and Modeling of the Duplication Effect on the Statistical Structure of the Genomes**

The analyses on the statistical structure of a wide range of genomes showed confound effect of the duplication process on various scales [193]. The distribution of the genome components of different scale and in various genomes are all featured by the over-representation of the high-frequency elements. An examination on the effect of selection on the mer copy number changes found no correlation between the selection pressure and the copy number of the mers in the genomes. This suggests that most of the evolutionary events affecting mer copy numbers are neutral and the mutational rates are mostly homogeneous along the genome. Therefore, the observed statistical structures are likely to be explained by the intrinsic dynamics of the various evolutionary processes instead of the effect of natural selection.

To explain these observations, we developed a parsimonious genome evolution model on the sequence level that has three elementary processes: substitution, duplication and deletion [193]. The parsimony of the model, especially the necessity of the deletion process has been demonstrated by the difference between the full model and the model without the deletion process on their ability in explaining the distribution of mers of different sizes. The model was

applied to various genomes. The fitted parameters reflect the average relative rate of the three processes over the evolutionary history of the genomes.

### **Alignment-Independent Phylogeny Method**

Based on our parsimonious genome evolution model, we further developed an alignment-independent method for measuring evolutionary distances, estimated as the total number of substitution, duplication and deletion events. Comparing to other currently available methods, the independence of our mer statistics based method from alignments and even fully assembled sequences makes it applicable to sequences with intensive rearrangements, unassembled sequence fragments, or even sequencing-independent array-based data. These features will have important applications to large scale phylogeny mapping efforts, such as different individuals in a mixed population or cells from different geometric compartments of a solid tumor.

We have evaluated our method for its resolution in both pair-wise and quartet distance inference using both simulated and real genomic sequences. The method recovered the evolutionary distances quite faithfully when the divergence level between the sequences is less than 35%, but start to show long-branch attraction effect for higher divergence levels. It can consistently resolve the correct

topology of the quartets when the internal branch length is  $>25\%$  of the external branch lengths. Furthermore, we have demonstrated, using simulated data, that by comparing the results from a substitution-based method, one can use our method to study the differences in the modulation of the substitution and indel events in various lineages. Note that the indel events detected by our method include different scales, and are not limited to transposon insertion and deletions. Therefore, it provides an unprecedented approach to study how the relative frequencies of different mechanisms in the genome evolution process are regulated.

In summary, duplication is a multi-mechanism evolutionary process and has a positive-feedback dynamics. Its relative rate and scale compared to other evolutionary events has a profound effect on the statistical structure of the genomes on many different levels. Because of such complexity, one needs to consider the dynamics and interactions of the duplication process when studying its mechanisms and effects. But with proper models, one can examine these problems both qualitatively and quantitatively.

## **6.2 Future Work**

### **Further Examination of the Segmental Duplication Dynamics**

Both the clustered distribution of the segmental duplications [193] and the “duplication in duplication” mosaic structures in some of the recently duplicated segments [8] suggest that a positive feedback dynamics in the duplication process. Although such a dynamics can be explained by a Polya’s Urn type of model, the actual mechanism is still not completely understood. Apart from the recombination mechanism between the amplified homologous repeats that can lead to duplications of a larger scale, one of the plausible hypotheses is that the presence of older duplications can induce more new duplications by providing highly homologous and long recombination seeds. To test such a hypothesis, one needs to annotate duplications of a wide range of different ages. Although this has been proved to be a very challenging task, it can provide deep insights and possible explanations on how the positive feedback dynamics of duplications are formed, as well as how they affect other rearrangement events. Therefore, we are currently developing a new Bayesian-based comparative genomics approach aiming to map out the less recent segmental duplications in the mammalian genomes. A Markov model similar to the one we used on the more recent duplications can be used to dissect the complicated interaction between

duplications of different ages and between duplications and other genome rearrangement events, such as synteny group formations. Furthermore, when duplications of low enough homology level can be detected, one can also start examining the effect of duplication on speciation events.

### **Incorporating Heterogeneous Mutation Rate in the Alignment-Independent Phylogeny Method**

The alignment-independent method we proposed in this thesis assumes neutral evolution and homogeneous mutation rates along the genome. However, in real genomic sequences, the mutational rates change from region to region [157][74][188][140]. To account for such phenomenon, we plan to extend the current method to be hyper-parametric. Similar to what has been done to model the region-specific substitution rate, instead of using constants, we can treat the observed mutation rates on each mer as drawn from a underlying distribution, such as a Gamma distribution [122]. The parameters of the distribution can be estimated from a sufficiently large dataset.

## **Bayesian Framework for the Alignment-Independent Phylogeny Method**

The current parameter estimation step in our phylogeny method takes a maximum likelihood approach because of the lack of the prior knowledge on the general distribution of the various mutation rates. However, as the estimations on the mutational rates in different lineages start to accumulate, they can be stored in a database. A Bayesian approach then can be taken, assuming that the already estimated mutational rates provide an approximation to the true global distribution, and can be taken as a reasonable prior.

## **Generalization and Potential Application of the Alignment-Independent Phylogeny Method**

The mer-based alignment-independent method we developed can be easily generalized to model the copy number fluctuations in other genomic components, such as protein domains or gene families. Such generalization will provide a unifying explanation on how the class of composition-based phylogeny methods work, and may lead to a hierarchical method that can compare and combine the phylogenetic inferences computed from different levels of genomic structure.

Finally, with our method, one can design array-based experiments to



reconstruct the phylogenetic relationships in a mix population. For example, by whole genome hybridization, the copy number of different mers in a particular genome drawn from a population can be estimated from the signals on the chip, from which we can apply our method to estimate the evolutionary distance between any individuals in a sequencing-independent way. Similarly, the method can be applied to the tumor cells separated from different geometric areas of the tumor tissue to find whether there is a relation between the micro-evolutionary process in the developing tumor cells and their relative geometric positions in the tumor tissue.

## Appendices

### Appendix A. Choice of Duplication Divergence Interval Size

From empirical studies, we found that the repeat configuration (+/+) has the lowest frequencies (about 10%) in each age group. Assuming that the number of (+/+) configurations in each age group follows a Binomial distribution  $\sim S(N, p)$  with a true probability  $p$  close to 0.1, we choose a minimal age group size of 100. This choice makes the probability of getting an estimated frequency  $\hat{p}$  that is  $\hat{p} < 0.5p$  or  $\hat{p} > 1.5p$  smaller than 0.1. The probabilities are computed as

$$\sum_{i=0}^{\lfloor Np(1-\delta) \rfloor} \binom{N}{i} p^i (1-p)^{N-i} + \sum_{i=\lceil Np(1+\delta) \rceil}^N \binom{N}{i} p^i (1-p)^{N-i} .$$

The error can also be

bounded from above by a Chernoff Bound [33].

N	$p = 0.1, \delta = 0.5$	$p = 0.1, \delta = 0.25$	$p = 0.5, \delta = 0.5$	$p = 0.5, \delta = 0.25$
10	0.419	0.419	$6.54 \times 10^{-2}$	$6.54 \times 10^{-2}$
50	0.169	0.156	$1.98 \times 10^{-4}$	$1.64 \times 10^{-2}$
100	$9.71 \times 10^{-2}$	0.126	$3.72 \times 10^{-7}$	$3.31 \times 10^{-3}$
500	$2.06 \times 10^{-4}$	$2.51 \times 10^{-2}$	$5.96 \times 10^{-13}$	$5.65 \times 10^{-9}$
1000	$2.83 \times 10^{-7}$	$4.53 \times 10^{-3}$	$3.69 \times 10^{-13}$	$3.70 \times 10^{-13}$

**Table 10. The probability of getting an MLE (assuming iid Bernoulli random variable) estimation  $\hat{p}$  outside a specific error bound  $\delta$  of the true probability  $p$ , i.e.  $\hat{p} < p(1-\delta)$  or  $\hat{p} > p(1+\delta)$ , in a sample of size  $N$ .**

## Appendix B. Empirical Simplification of Mer Copy Number Evolution

For most genomes, we observe that the frequency of a nucleotide mer is similar to its reverse complementary. Therefore, we assume  $l^+(t) \approx l^-(t) \approx \frac{1}{2}l(t)$ . Similarly,  $p_i^+(t) \approx s_{m-i}^-(t)$  and  $s_{m-i}^+(t) \approx p_i^-(t)$ . From empirical observation, we also found that the correlation between the copy numbers of the mers and their sub-mers and sub-sequences can be modeled by expressing the sub-mer/sequence copy numbers as an geometric function decided by the mer copy number and the frequency of the end nucleotide in the mer ( $p_1(t)$  and  $s_{m-1}(t)$ ):

$$p_i^+(t) = l^+(t)(r_p^+(t))^{m-i}; p_i^-(t) = l^-(t)(r_p^-(t))^{m-i};$$

$$s_i^+(t) = l^+(t)(r_s^+(t))^i; s_i^-(t) = l^-(t)(r_s^-(t))^i;$$

where

$$r_p^+(t) = \left(\frac{p_1^+(t)}{l^+(t)}\right)^{\frac{1}{m-1}}; r_p^-(t) = \left(\frac{p_1^-(t)}{l^-(t)}\right)^{\frac{1}{m-1}};$$

$$r_s^+(t) = \left(\frac{s_{m-1}^+(t)}{l^+(t)}\right)^{\frac{1}{m-1}}; r_s^-(t) = \left(\frac{s_{m-1}^-(t)}{l^-(t)}\right)^{\frac{1}{m-1}}.$$

When the approximations are applied,

$$r_s^-(t) \approx r_p^+(t) \approx \left(\frac{2p_1^+(t)}{l(t)}\right)^{\frac{1}{m-1}}; r_s^+(t) \approx r_p^-(t) \approx \left(\frac{2p_1^-(t)}{l(t)}\right)^{\frac{1}{m-1}}.$$

Approximate the base frequency at time  $t$  by

$$p_1^+(t) = p_1^+(0) + (p_1^+(T) - p_1^+(0)) \frac{t}{T};$$

$$p_1^-(t) = p_1^-(0) + (p_1^-(T) - p_1^-(0)) \frac{t}{T};$$

and approximate

$$\begin{aligned} \sum_{i=1}^x (p_i^+(t)s_i^+(t) + p_i^-(t)s_i^-(t)) &= \sum_{i=1}^x (p_i^+(t)p_{m-i}^-(t) + s_i^+(t)s_{m-i}^-(t)) \\ &\approx \frac{x}{m-1} \sigma(t); \\ \sum_{i=1}^x (p_i^+(t) + s_i^+(t)) &= \sum_{i=1}^x (p_i^-(t) + s_i^-(t)) \\ &\approx p_1^+(t) \frac{1 - \frac{l(t)}{2p_1^+(t)}}{1 - \left(\frac{l(t)}{2p_1^+(t)}\right)^{\frac{1}{m-1}}} + p_1^-(t) \frac{1 - \frac{l(t)}{2p_1^-(t)}}{1 - \left(\frac{l(t)}{2p_1^-(t)}\right)^{\frac{1}{m-1}}}; \end{aligned}$$

$$\sum_{j=2}^{m-x} (c_x^{j,+}(t) + c_x^{j,+}(t))(b_x^{j,+}(t) + b_x^{j,-}(t)) \approx \frac{m-x-1}{m-1} \sigma(t).$$

We have

$$\begin{aligned}
& P(l(t + \Delta t) = l(t) + 1 | l(t), G(t)) \\
&= \left( \sum_{x=m}^X p_r f_r(x) x \right) \left( l(t) \left( 1 - \frac{(m-1)l(t)}{G(t)} \right) \right) \Delta t + \\
& \left( \sum_{x=m}^X p_r f_r(x) \right) \left( -(m-1)l(t) \left( 1 - \frac{(m-1)l(t)}{G(t)} \right) + \frac{2\sigma(t) - l(t)\delta(t)}{G(t)} \right) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_r f_r(x) x \right) \left( \frac{1}{m-1} \frac{\sigma(t)}{G(t)} \right) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_r f_r(x) \right) \left( \frac{\sigma(t)}{G(t)} - l(t) \frac{\delta(t)}{G(t)} \right) \Delta t + \\
& \left( \sum_{x=1}^X p_d f_d(x) x \right) \left( -l(t) \frac{\sigma(t)}{G(t)^2} \right) \Delta t + \\
& \left( \sum_{x=1}^X p_d f_d(x) \right) \left( \left( 1 - \frac{(m-1)l(t)}{G(t)} \right) \frac{\sigma(t)}{G(t)} \right) \Delta t + \\
& p_s l(t) \Delta t
\end{aligned}$$

$$\begin{aligned}
& P(l(t + \Delta t) = l(t) - 1 | l(t), G(t)) \\
&= \left( \sum_{x=m}^X p_r f_r(x) x \right) \left( -(m-1) \frac{l(t)^2}{G(t)} \right) \Delta t + \\
& \left( \sum_{x=m}^X p_r f_r(x) \right) \left( l(t) \left( (m-1) \left( 1 + \frac{(m+1)l(t)}{G(t)} \right) - \frac{\delta(t)}{G(t)} \right) \right) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_r f_r(x) x \right) \left( -\frac{l(t)}{m-1} \frac{\delta(t)}{G(t)} \right) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_r f_r(x) \right) (l(t)(m-1)) \Delta t + \\
& \left( \sum_{x=m}^X p_d f_d(x) x \right) \left( l(t) \left( 1 - \frac{\sigma(t)}{G(t)^2} \right) \right) \Delta t + \\
& \left( \sum_{x=m}^X p_d f_d(x) \right) \left( l(t) \left( -(m-1) \left( 1 - \frac{\sigma(t)}{G(t)^2} \right) + 2(m-1) - \frac{\delta(t)}{G(t)} \right) \right) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_d f_d(x) x \right) l(t) \Delta t + \\
& \left( \sum_{x=1}^{m-1} p_d f_d(x) \right) \left( l(t) \left( m-1 - \frac{\delta(t)}{G(t)} \right) \right) \Delta t + \\
& p_s l(t) m \Delta t
\end{aligned}$$

## Bibliography

- [1] Achaz, G., Netter, P, Coissac, E. (2001) Study of intrachromosomal duplications among the eukaryote genomes. *Mol. Biol. Evol.* **18**, 2280-2288.
- [2] Albertson D.G., Collins C., McCormick F., Gray J.W. (2003a). Chromosome aberrations in solid tumors. *Nat Genet.* **34**:369.
- [3] Almirantis, Y., Provata, A., (2001). An evolutionary model for the origin of non-randomness, long-range order and fractality in the genome, *Bioessays*, **23**:647.
- [4] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol.* **215**, 403-10.
- [5] Apweiler, R. *et al.* (2000) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research* **29**, 37-40.
- [6] Armengol, L., Pujana, M.A., Cheung, J., Scherer, S.W. & Estivill, X. (2003) Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Human Molecular Genetics* **12**, 2201-2208.
- [7] Baake, E., von Haeseler, A. (1999) Distance Measures in Terms of Substitution Processes. *Theo. Pop. Biol.* **55**, 166-175.



- [8] Babcock, M., Pavlicek, A., Spiteri, E., Kashork, C. D., Ioshikhes, I., Shaffer, L. G., Jurka, J. & Morrow, B. E. (2003) Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. *Genome Res.* **13**, 2519-2532.
- [9] Bafna, V., Pevzner, P. (1995) Sorting permutations by transpositions. In *Proc. 6<sup>th</sup> Annual ACM-SIAM Symp. On Disc. Alg.* Pp614-623. ACM Press.
- [10] Bailey, J. A., Liu, G. & Eichler, E. E. (2003) An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**, 823-834.
- [11] Bailey, J.A., Baertsch, R., Kent, W.J., Haussler, D. & Eichler, E.E. (2004) Hotspots of mammalian chromosomal evolution. *Genome Biology* **5**, R23.
- [12] Bailey, J.A., Church, D.M., Ventura, M., Rocchi, M. & Eichler, E.E. (2004) Analysis of segmental duplications and genome assembly in the mouse. *Genome Research* **14**, 789-801.
- [13] Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W. & Eichler, E.E. (2002) Recent segmental duplications in the human genome. *Science* **297**, 1003-1007.
- [14] Balmain, A., Gray, J., Ponder, B. (2003). The genetics and genomics of cancer. *Nat Genet.* **33 Suppl**:238.
- [15] Baptiste, E., Philippe, H. (2002) The Potential Value of Indels as Phylogenetic Markers: Position of Trichomonads as a Case Study.

- Mol. Biol. Evol.* **19**, 972-977.
- [16] Batzer, M.A. & Deininger, P.L. (2002) Alu repeats and human genomic diversity. *Nature Reviews* **3**, 370-379.
- [17] Baudot, A., Jacq, B., Brun, C. (2004) A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein-protein interaction network. *Genome Biol.* **5**, R76.
- [18] Benham, C.J. (1993) Sites of predicted stress-induced DNA duplex destabilization occur preferentially at regulatory loci. *Proc. Natl. Acad. Sci. USA* **90**, 2999-3003.
- [19] Berg, J., Lassig, M., Wagner, A. (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol Biol.* **4**, 51.
- [20] Bhan, A., Galas, D.J., Dewey, T.G. (2002) A duplication growth model of gene expression networks. *Bioinformatics* **18**, 1486-93.
- [21] Blanchette, M., Green, E.D., Miller, W., Haussler, D. (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.* **14**, 2412-23.
- [22] Blanchette, M., Schwikowski, B., Tompa, M. (2002). Algorithms for phylogenetic footprinting. *J. Comp. Biol.* **9**, 211-223.
- [23] Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K., Ovcharenko, I., Pachter, L., Rubin, E.M. (2003) Phylogenetic

- shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391-1394.
- [24] Boussau, B., Karlberg, E.O., Frank, A.C., Legault, B.A., Andersson, S.G.E. (2004) Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proc. Natl. Acad. Sci. USA* **101**, 9722-9727.
- [25] Bray N, Dubchak I, Pachter L. (2003) AVID: A global alignment program. *Genome Res.* **13**, 97-102.
- [26] Breslauer, K.J., Frank, R., Blocker, H. & Marky, L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* **83**, 3746-3750.
- [27] Brocchieri, L. (2001) Phylogenetic inferences from molecular sequences: review and critique. *Theoret. Pop. Biol.* **59**, 27-40.
- [28] Brudno M, Do, CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721-31.
- [29] Brudno, M., Chapman, M., Gottgens, B., Batzoglou, S., Morgenstern, B. (2003) Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics* **4**, 66.
- [30] Buldyrev, S.V., Goldberger, A.L., *et al.* (1993). Fractal landscapes and molecular evolution: modeling the myosin heavy chain gene family, *Biophysical Journal*, **65**:2673.

- [31] Castillo-David, C.I., Hartl, D.L. (2002) Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Mol. Biol. Evol.* **19**, 728-235.
- [32] Cavalcanti, A.R.O., Ferreira, R., Gu, Z., Li, W.H. (2003) Patterns of gene duplication in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. *J. Mol. Evol.* **56**, 28-37.
- [33] Chernoff, H. (1952) A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *Annals of Mathematical Statistics* **23**, 493-507.
- [34] Cheung, J., Estivill, X., Khaja, R., MacDonald, J.R., Lau, K., Tsui, L.C. & Scherer, S.W. (2003) Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biology*, **4**, R25.
- [35] Cheung, J., Wilson, M.D., Zhang, J., Khaja, R., MacDonald, J.R., Heng, H.H.Q., Koop, B.F., L.C. & Scherer, S.W. (2003) Recent segmental and gene duplications in the mouse genome. *Genome Biology*, **4**, R47.
- [36] Conant, G.C., Wagner, A. (2003) Asymmetric Sequence Divergence of Duplicate Genes. *Genome Res.* **13**, 2052-2058.
- [37] Coyne, J.A., Orr, A. (1997) Patterns of speciation in *Drosophila* revisited. *Evolution* **51**, 295.
- [38] Crow, J.F. "Introduction to Population Genetic Theory." Harper & Row, 1996.

- [39] Dehal, P., Predki, P., Olsen, A.S., Kobayashi, A., Folta, P., *et al.* (2001) Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science*. **293**, 104-11.
- [40] Delcher AL, Phillippy A, Carlton J, Salzberg SL. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478-83.
- [41] Denver, D.R., Morris, K., Lynch, M., Thomas, W.K. (2004) High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* **430**, 679-82.
- [42] Dermitzakis, E.T., Clark, A.G. (2001) Differential selection after duplication in mammalian developmental genes. *Mol. Biol. Evol.* **18**, 557-562.
- [43] Dermitzakis, E.T., Reymond, A., Scamuffa, N., Ucla, C., Kirkness, E., Rossier, C., Antonarakis, S.E. (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302**, 1033-1035.
- [44] Difilippantonio, M.J., Petersen, S., Chen, H.T., Johnson, R., Jasin, M., Kanaar, R., Ried, T., Nussenzweig, A. (2002) Evidence for Replicative Repair of DNA Double-Strand Breaks Leading to Oncogenic Translocation and Gene Amplification. *J. Exp. Med.* **196**, 469-480.
- [45] Dubchak, I., Brudno, M., Loots, G.G., Pachter, L., Mayor, C., Rubin, E.M., Frazer, K.A. (2000) Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10**, 1304-1306.
- [46] Dunham, M.J., Badrane, H., Ferea, T., Adams, J., Brown, P.O.,

- Rosenzweig, F., and Botstein, D. (2002). Characteristic Genome Rearrangements in Experimental Evolution of *Saccharomyces cerevisiae*. *PNAS*, **99**, 16144-16149.
- [47] Dunman, P.M., Mounts, W., McAleese, F., *et al.* (2004) Uses of *Staphylococcus aureus* GeneChips in genotyping and genetic composition analysis. *J. Clin. Micro.* **42**, 4275-4283.
- [48] Durand, D. (2003) Vertebrate evolution: doubling and shuffling with a full deck. *Trends Genet.* **19**, 2.
- [49] Durrett, R. "Probability: theory and examples." (1996) 2<sup>nd</sup> Edt. Duxbury Press.
- [50] Ejima, Y. & Yang, L. (2003) Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. *Human Mol. Gen.* **12**, 1321-1328.
- [51] Emanuel B.S., Shaikh T.H. (2001). Segmental duplications: an 'expanding' role in genomic instability and disease. *Nat Rev Genet.* **2**,791.
- [52] Felsenstein, J., Department of Genome Science, University of Washington.
- [53] Fitch, D.H., Bailey, W.J., Tagle, D.A., Goodman, M., Sieu, L., Slightom, J.L. (1991) Duplication of the gamma-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. *Proc Natl Acad Sci U S A.* **88**, 7396-400.
- [54] Force, A., Lynch, M., Pickett, K.B., Amores, A., Yan, Y., Postlethwait,

- J. (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics*, **151**:1531-1545.
- [55] Friedman, R, Hughes, A.L. (2003) The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol. Biol. Evol.* **20**, 154-163.
- [56] Ganapathiraju, M., Weisser, D., *et al.* (2004). Comparative n-gram analysis of whole-genome protein sequences, *Appl Bioinformatics*. **3**, 193-200.
- [57] Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493-521.
- [58] Gibson, T.J., Spring, J. (1998) Genetic Redundancy in Vertebrates: Polyploidy and Persistence of Genes Encoding Multidomain Proteins. *Trends Genet.* **14**, 46-49.
- [59] Glazko, G.V., Nei, M. (2003) Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* **20**, 424-34.
- [60] Goldman, N., Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725-736.
- [61] Gramm, J. & Niedermeier R. (2002) Breakpoint Medians and Breakpoint Phylogenies: A Fixed Parameter Approach. *Bioinformatics* **18**:S128-S139.

- [62] Grant, D., Cregan, P., Shoemaker, R.C. (2000) Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. *Proc. Nat. Acad. Sci.* **97**, 4168-4173.
- [63] Gregory, T.R. (2003) Insertion-Deletion Biases and the Evolution of Genome Size. *Gene* **324**, 15-34.
- [64] Gu, X, Wang, Y., Gu, J. (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature Genet.* **31**, 205-209.
- [65] Gu, X. & Zhang, H. (2004) Genome Phylogenetic Analysis Based on Extended Gene Contents. *Mol. Biol. Evol.* **21**:1401-1408.
- [66] Gu, X. and Li, W.H. (1998) Estimation of Evolutionary Distances Under Stationary and Nonstationary Models of Nucleotide Substitution. *Proc. Natl. Acad. Sci.* **95**, 5899-5905.
- [67] Gu, X., Li, W.H. (1995) The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *J. Mol. Evol.* **40**, 44-73.
- [68] Gu, X., Zhang, Z. Huang, W. (2005) Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc. Nat. Acad. Sci.* **102**, 707-712.
- [69] Gu, Z., Cavalcanti, A., Chen, F.C., Bouman, P. and Li, W.H. (2002). Extent of Gene Duplication in the Genomes of Drosophila, Nematode



- and Yeast. *Mol. Biol. Evol.* **19**:256-262.
- [70] Gu, Z., Rifkin, S.A., White, K.P., Li, W.H. (2004) Duplicate genes increase gene expression diversity within and between species. *Nat Genet.* **36**, 577-9.
- [71] Gu, Z., Steinmetz, L.M., Gu, X., Scharfe, C., David, R.W., Li, W.H. (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* **42**, 63-66.
- [72] Gumucio, D.L., Heilstedt-Williamson, H., Gray, T.A, Tarle, S.A., *et al.* (1992) Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol. Cell. Biol.* **12**, 4919-4929.
- [73] Hannenhalli, S., Pevzner, P.A. (1999) Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *J. ACM.* **46**, 1-27.
- [74] Hardison, R.C, Roskin, K.M., *et al.* (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13**, 13-26.
- [75] Hasegawa, M., Kishino, H., Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA *J. Mol. Evol.* **22**, 160-174.
- [76] Hoare, S. Zou, Y., Purohit, V., *et al.* (2000) Differential incision of bulky carcinogen-DNA adducts by the UvrABC nuclease: comparison of incision rates and the interactions of Uvr subunits with lesions of

- different structures. *Biochemistry* **39**, 12252-12261.
- [77] Hooper, S.D., Berg, O.G. (2003) On the nature of gene innovation: duplication patterns in microbial genomes. *Mol. Biol. Evol.* **20**, 945-954.
- [78] House, C. H. & Fitz-Gibbon, S. T. (2002) "Using Homolog Groups to Create a Whole-Genomic Tree of Free-Living Organisms: An Update." *J. Mol. Evol.* **54**:539:547.
- [79] Hsieh, L., Luo, L., Ji, F., Lee, H.C. (2003) Minimal model for genome evolution and growth. *Phys. Rev.* **90**, 018101.
- [80] Hughes, A.L, Green, J.A., Garbayo, J.M. Roberts, R.M. (2000) Adaptive Diversification within a Large Family of Recently Duplicated, Placentally Expressed Genes. *Proc. Natl. Acad. Sci. USA* **97**, 3319-3323.
- [81] Hughes, A.L. (1994) The Evolution of Functionally Novel Proteins After Gene Duplication. *Proc. Biol. Sci.* **256**, 119-124.
- [82] Jaillon, O., *et al.* (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**, 946-957.
- [83] Jeong, H., Mason, S.P., Barabasi, A.L., Oltvai, Z.N., (2001). Lethality and centrality in protein networks. *Nature*, **411**:41.
- [84] Jeong, H., Tombor, B., Albert, R., Oltvai, A.N., Barabasi, A.L., (2000). The large-scale organization of metabolic networks. *Nature*, **407**:651.
- [85] Ji, Y., Eichler, E.E., Schwartz, S. and Nicholls, R.D. (2000).

- Structure of Chromosomal Duplicons and Their Role in Mediating Human Genomic Disorders, *Genome Research*, **10**, 597-610.
- [86] Johnson, N.L. "Urn models and their Application", Wiley, 1977.
- [87] Jukes, T.H. and Cantor, C.R. (1969) pp21-132. Evolution of protein molecules. Academic, New York.
- [88] Kafri, R., Bar-Even, A., Pilpel, Y. (2005) Transcription control reprogramming in genetic backup circuits. *Nat Genet.* **37**, 295-9.
- [89] Kalofus, K.J., Jackson, A.R., Milosavljevic, A. (2004) Pash: efficient genome-scale sequence anchoring by positional hashing. *Genome Res.* **14**, 672-678.
- [90] Kapitonov, V.V. & Jurka, J. (1996) The age of Alu subfamilies. *J. Mol. Evol.* **42**, 59-65.
- [91] Kawakita, A., Sota, T., Ascher, J.S., Ito, M., Tanaka, H., Kato, M. (2003) Evolution and Phylogenetic Utility of Alignment Gaps Within Intron Sequences of Three Nuclear Genes in Bumble Bees. *Mol. Biol. Evol.* **20**, 87-92.
- [92] Kellis, M., Birren, B.W., Lander, E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**, 617-624.
- [93] Kent, W.J. (2002) BLAT-the BLAST-like alignment tool.

- Genome Res.* **12**, 656-664.
- [94] Kent, W.J., Zahler, A.M. (2000) Conservation, regulation, synteny and introns in a large-scale *C.briggsae-C.elegans* genomic alignment. *Genome Res.* **10**, 1115-1125.
- [95] Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120.
- [96] Kolodrubetz, D., Kruppa, M. Burgum, A. (2001) Gene Dosage Affects the Expression of the Duplicated NHP6 Genes of *Saccharomyces cerevisiae*. *Gene* **272**, 93-101.
- [97] Kolomietz, E., Meyn, M.S., Pandita, A. & Squire, J.A. (2002) The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. *Genes, Chromosomes & Cancer* **35**, 97-112.
- [98] Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I, Koonin, E.V. (2002) Selection in the evolution of gene duplications. *Genome Biol.* **3**, R0008.1-9.
- [99] Koszul, R., Caburet, S., Dujon, B., & Fischer, G. (2004) Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* **23**, 234-243.
- [100] Kunkel, T.A., Erie, D.A. (2004) DNA Mismatch Repair. *Annu. Rev. Biochem.*
- [101] Lengauer, C., Kinzler K.W., and Vogelstein B. (1998). Genetic

- Instabilities in Human Cancers. *Nature*, **396**:643-649.
- [102] Leslie, C.S., Eskin, E., Cohen, A., Weston, J., Noble, W.S. (2004) *Bioinformatics* **20**, 467-76.
- [103] Leveugle, M., Prat, K., Perrier, N., Birnbaum, D., Coulier, F. (2003). ParaDB: a tool for paralogy mapping in vertebrate genomes. *Nucleic Acids Res.* **31**:63-7.
- [104] Li, S., Zhang, L., Kern, W.F., Andrade, D., Forsberg, J.E., Bates, F.R. & Mulvihill, J.J. (2002) Identification of t(15;17) and a segmental duplication of chromosome 11q23 in a patient with acute myeloblastic leukemia M2. *Cancer Genet. Cytogenet.* **138**, 149-152.
- [105] Liu, G., Zhao, S., Bailey, J.A., Sahinalp, S.C., Alkan, C., Tuzun, E., Green, E.D. & Eichler, E.E. (2003) Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Research* **13**, 358-368.
- [106] Lucito R., Nakimura M., West J.A., Han Y., Chin K., Jensen K., McCombie R., Gray J.W., Wigler M. (1998). Genetic analysis using genomic representations. *Proc Natl Acad Sci U S A.* **95**:4487.
- [107] Lucito R., West J., Reiner A., Alexander J., Esposito D., Mishra B., Powers S., Norton L., Wigler M. (2000). Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Res.* **10**:1726.
- [108] Lynch, M. (2002) Gene Duplication and Evolution. *Science*

- 297**, 945-947.
- [109] Lynch, M., Conery, J.S., (2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, **290**, 1151.
- [110] Ma, B., Tromp, J., Li, M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18**, 440-445.
- [111] Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., *et al.* (1994). Linguistic features of noncoding DNA sequences. *Physical review letters*, **73**, 3169.
- [112] Margulies, E.H., Blanchette, M., Haussler, D., Green, E.D. (2003) Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507-2518.
- [113] Maslov, S., Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, **296**:910.
- [114] Maslov, S., Sneppen, K., Eriksen, K.A., Yan, K,K, (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol Biol.* **4**, 9.
- [115] Matsuyama, A., Shiraishi, T., Trapasso, F., Kuroki, T., Alder, H., Mori, M., Huebner, K. & Croce, C.M. (2003) *Proc. Natl. Acad. Sci. USA* **100**, 14988-14993.
- [116] McConnell, K.J., and Beveridge, D.L. 2001. Molecular Dynamics Simulation of B'-DNA: Sequence Effects on A-Tract-Induced Bending and

- Flexibility. *J. Mol. Biol.* **314**:23-40.
- [117] McLysaght, A., Hokamp, K., Wolfe, K.H. (2002) Extensive genomic duplication during early chordate evolution. *Nat Genet.***31**, 200-4.
- [118] Mishmar, D., Rahat, A., Scherer, S.W., Nyakatura, G., Hinzmann, B., Kohwi, Y., Mandel-Gutfroind, Y., Lee, J.R., Drescher, B., Sas, D.E., *et al.* (1998) Molecular characterization of a common fragile site (FRA7H) on human chromosome 7 by the cloning of a simian virus 40 integration site. *Proc. Natl. Acad. Sci. USA* **95**, 8141-8146.
- [119] Moret B. M. E., Wang, L-S., Warnow, T., Wyman, S. K. (2001) New Approaches for Reconstructing Phylogenies from Gene Order Data. *Bioinformatics* **17**:S165-S173.
- [120] Muse, S. (1996) Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* **13**, 105-114.
- [121] Nadeau J.H., Sankoff D. (1997). Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics.* **147**:1259.
- [122] Nei, M. & Kumar, S. "Molecular Evolution and Phylogenetics." Oxford University Press, 2000.
- [123] Nowak, M.A., Boerlijst, M.C., Cooke, J., Smith, J.M. (1997) Evolution of Genetic Redundancy. *Science* **388**, 167-171.
- [124] Ohno, S. "Evolution by gene duplication." (1970). Springer-

Verlag.

- [125] Ophir, R., Graur, D. (1997) Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene*. **205**, 192-202.
- [126] Ouyang, Z., Wang, C., She, Z.S. (2004) Scaling and hierarchical structures in DNA sequences. *Physic. Rev.* **93**, 078103.
- [127] Pastor-Satorras R., Smith E., Sole R.V. (2003). Evolving protein interaction networks through gene duplication. *J Theor Biol.* **222**:199.
- [128] Peng, C.K., Buldyrev, S.V., *et al.* (1992) Long-range correlations in nucleotide sequences. *Nature* **356**:168.
- [129] Peng, C.K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., Goldberger, A.L. (1994) Mosaic organization of DNA nucleotides. *Physical Review E.* **49**, 1685-1689.
- [130] Pennacchio, L.A., Rubin, E.M. (2001) Genomic strategies to identify mammalian regulatory sequences. *Nature Rev.* **2**, 100-109.
- [131] Petrov, D.A. (2002) Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.* 2002 **61**, 531-44.
- [132] Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L., Shaw, K.L. (2000) Evidence for DNA loss as a determinant of genome size. *Science* **287**, 1060-2.
- [133] Pevzner, P. & Tesler, G. (2003) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.*



13:37-45.

- [134] Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D. and Brown, P.O. (1999). Genome-wide Analysis of DNA Copy-Number Changes Using cDNA Microarrays. *Nature Genetics*, **23**:41-46.
- [135] Prashanth, A.K. & Benham, C. (2003) *Genome Informatics*, CSHL, pp9.
- [136] Prince, V.E., Pickett, F.B. (2002). Splitting Pairs: The Diverging Fates of Duplicated Genes. *Nature Rev. Gen.* **3**:827-837.
- [137] Qi, J., Wang, B., Hao, B-I. (2004) Whole Proteome Prokaryote Phylogeny Without Sequence Alignment: A K-String Composition Approach. *J. Mol. Evol.* **58**:1-11.
- [138] Qian J., Luscombe N.M., Gerstein M. (2001). Protein family and fold occurrence in genomes: power-law behavior and evolutionary model. *J Mol Biol.* **313**:673.
- [139] Rain, J.C., Selig, L., *et al.* (2001).The protein-protein interaction map of *Helicobacter pylori*. *Nature*, **409**:211.
- [140] Reich, D.E., Schaffner, S.F., Daly, M.J., McVean, G., Mullikin, J.C., Higgins, J.M., Richter, D.J., Lander, E.S., Altshuler, D. (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135-142.
- [141] Remm, M., Storm, C.E., Sonnhammer, E.L. (2001).

- Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol.* **314**:1041-52.
- [142] Riehle, M.M., Bennett, A.F. and Long, A.D. (2001). Genetic Architecture of Thermal Adaptation in *Escherichia coli*. *PNAS*, **98**:525-530.
- [143] ———Roussas, G.G. “A course in mathematical statistics.” (1997) 2<sup>nd</sup> Edt. Academic Press.
- [144] Samonte, R.V., and Eichler, E.E. (2002). Segmental Duplications and the Evolution of the Primate Genome. *Nature Reviews*, **3**, 65-72.
- [145] Sankoff, D., Blanchette, M. (1998) Multiple genome rearrangement and breakpoint phylogeny. *J. Comp. Biol.* **5**, 555-570.
- [146] Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., Cedergren, R. (1992) Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. U S A.* **89**, 6575-9.
- [147] SantaLucia, J. Jr, Allawi, H.T., Seneviratne, P.A.. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry.* **35**, 3555-62.
- [148] Sarai, A., Mazur, J., Nussinov, R., Jernigan, R.L. (1989) Sequence dependence of DNA conformational flexibility. *Biochemistry* **28**, 7842-7849.
- [149] Schwartz, S, Kent, W.J., Smit, A., Zhang, Z., Baersch, R., Hardison, R.C., Haussler, D., Miller, W. (2003) Human-Mouse alignments with

- BLASTZ. *Genome Res.* **13**, 103-107.
- [150] Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Rouck, J., Gibbs, R., Hardison, R., Miller, W. (2000) PipMaker-a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577-586.
- [151] Schwikowski, B., Uetz, P., Fields, S. (2000). A network of protein-protein interactions in yeast, *Nature Biotech*, **18**:1257.
- [152] Seoighe, C., Wolfe, K.H. (1999) Yeast genome evolution in the post-genome era. *Curr Opin Microbiol.***2**, 548-54.
- [153] She, X., Jiang, Z., Clark, R.A., Liu, G., Cheng, Z., Tuzun, E., Church, D.M., Sutton, G., Halpen, A.L., Eichler, E.E. (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927-930.
- [154] Sieber O.M., Heinimann K., Tomlinson I.P. (2003). Genomic instability--the engine of tumorigenesis? *Nat Rev Cancer.* **3**:701.
- [155] Simillion, C. Vandepoele, K., Montagu, M.C.E., Zabeau, M., Van de Peer, Y. (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc. Nat. Acad. Sci.* **99**, 13627-13632.
- [156] Smit, A.F.A., Toth, G., Riggs, A.D. & Jurka, J. (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* **246**, 401-417.
- [157] Smith, N.G.C, Webster, M.T., Ellegren, H. (2002) Deterministic

- mutation rate variation in the human genome. *Genome Res.* **12**, 1350-1356.
- [158] Smith, P.D., Moss, S.E. (1994) Z-DNA-Forming Sequences at a Putative Duplication Site in the Human Annexin VI-Encoding Gene. *Gene* **138**, 239-242.
- [159] Snel, B., Bork, P., Huynen, M.A. (2002) The identification of functional modules from the genomic association of genes. *Genome Res.* **12**, 17-25.
- [160] Squire, J.A., Pei, J., Marrano, P., Beheshti, B., Bayani, J., Lim, G., Moldovan, L. & Zielenska, M. (2003) High-resolution mapping of amplifications and deletions in pediatric osteosarcoma by use of CGH analysis of cDNA microarrays. *Genes Chrom. Cancer* **38**, 215-225.
- [161] Stenger, J.E., Lobachev, K.S., *et al.* (2001). Alu distribution in human genome. *Genome Research*, **11**:12-27.
- [162] Stuart, G. W., Moffett, K., Baker, S. (2002) Integrated Gene and Species Phylogenies from Unaligned Whole Genome Protein Sequences. *Bioinformatics* **18**:100-108.
- [163] Suter, B., Schnappauf, G. and Thoma, F. 2000. Poly(dA·dT) Sequences Exist as Rigid DNA Structures in Nucleosome-Free Yeast Promoters *in vivo*. *Nucleic Acid Research*, **28**:4083-4089.
- [164] Swensen, J., Hoffman, M., Skolnick, M. H. & Neuhausen, S. L. (1997) Identification of a 14 kb deletion involving the promoter region of BRCA1 in a breast cancer family. *Hum. Mol. Genet.* **6**, 1513-1517.

- [165] Swofford (1993) Illinois Natural History Survey, Champaign, IL
- [166] Tamas, I., Klasson, L, Canback, B., Naslund, A.K. *et al.* (2002) 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376-2379.
- [167] Tang, J. & Moret B. M. E. (2003) Scaling up Accurate Phylogenetic Reconstruction from Gene-Order Data. *Bioinformatics* **19**:i305-i312.
- [168] Tesler, G. (2002) GRIMM: genome rearrangements web server. *Bioinformatics* **18**, 492-3.
- [169] Ting, C.T., Tsaur, S.C., Sun, S., Browne, W.E., Chen, Y.C., Patel, N.H., Wu, C.I. (2004) Gene duplication and speciation in *Drosophila*: evidence from the Odysseus locus. *Proc. Natl. Acad. Sci.* **101**, 12232-5.
- [170] Tuzun, E., Bailey, J.A., & Eichler, E.E. (2004) Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Research* **14**, 493-506.
- [171] Van de Peer Y., Taylor J.S., Braasch I., Meyer A. (2001). The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *J Mol Evol.* **53**:436.
- [172] ———Varadhan, S.R.S. “Probability theory.” (2001) American Mathematical Society.
- [173] Vision, T.J., Brown, D.G., Tanksley, S.D. (2000) The origins of genomic duplications in *Arabidopsis*. *Science* **290**, 2114-2117.

- [174] Wagner, A. (1994) Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proc Natl Acad Sci U S A.* **91**, 4387-91.
- [175] Wagner, A. (1998) The Fate of Duplicated Genes: Loss or New Function? *BioEssays* **20**, 785-788.
- [176] Wagner, A. (2002) Asymmetric Functional Divergence of Duplicate Genes in Yeast. *Mol. Biol. Evol.* **19**, 1760-1768.
- [177] Walsh, B. (2003) Population-genetic models of the fates of duplicate genes. *Genetica* 118:279-94.
- [178] Walsh, J.B. (1995) How Often Do Duplicated Genes Evolve New Functions? *Genetics* **139**, 421-428.
- [179] Wang, D., Coscoy, L., Zylberberg, M., Avila, P.C., Boushey, H.A., Ganem, D., DeRisi, J.L. (2002) Microarray-based detection and genotyping of viral pathogens. *Proc. Natl. Acad. Sci. USA*, **99**, 15687-15692.
- [180] Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., Lawrence, C.E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat Genet.* **26**, 225-8.
- [181] Watanabe, T., Murata, Y. Oka, S., Iwahashi, H. (2004) A new approach to species determination for yeast strains: DNA microarray-based comparative genomic hybridization using a yeast DNA microarray with 6000 genes. *Yeast* **21**, 351-365.

- [182] Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562.
- [183] Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.
- [184] Wolfe, K.H., Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708-13.
- [185] Wu, C.-I. (2001) The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851.
- [186] Yanai I., Camacho C.J., and DeLisi C. (2000). Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. *Physical Review Letters* **85**, 2641.
- [187] Yang, S., Doolittle, R.F., Bourne, P.E. (2005) Phylogeny determined by protein domain content. *Proc. Natl. Acad. Sci. USA* **102**, 373-378.
- [188] Yang, S., Smit, A.F, Schwartz, S., Chiaromonte, F., Roskin, K.M., Haussler, D., Miller, W., Hardison, R.C. (2004) Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. *Genome Res.* **14**, 517-527.
- [189] Yang, Z. (1997) PAML: a program package for phylogenetic analysis

- by maximum likelihood. *Comput Appl Biosci.* **13**, 555-6.
- [190] Yeramian, E., Jones, L. (2003) GeneFizz: A web tool to compare genetic (coding/non-coding) and physical (helix/coil) segmentations of DNA sequences. Gene discovery and evolutionary perspectives. *Nucleic Acids Res.* **31**, 3843-9.
- [191] Zhang, J., Nei, M. (2000) Positive Selection in the Evolution of Mammalian interleukin-2 Genes. *Mol. Biol. Evol.* **17**, 1413-1416.
- [192] Zhang, L., Lu, H.H., Chung, W.Y., Yang, J., Li, W-H. (2005) Patterns of segmental duplication in the human genome. *Mol. Biol. Evol.* **22**, 135-141.
- [193] Zhou, Y., Mishra, B. (2004) Models of Genome Evolution. *Modeling in Molecular Biology*, 287-304, Natural Computing Series, Springer.
- [194] Zhou, Y., Mishra, B. (2005) *Proc. Natl. Acad. Sci. USA* **102**, 4151-4056.
- [195] Zhou, Y., Paxia, S., Rudra, A., Mishra, B. (2002). A Random Walk Down the Genome: a case study of DNA evolution in *Valis*. *Computer IEEE Press*, **35(7):73-79**.