uOttawa

L'Université canadienne
Canada's university

**FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES**

uOttawa

L'Université canadienne
Canada's university

**FACULTY OF GRADUATE AND POSDOCTORAL STUDIES**

Ximing Xu

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Mathematics)

GRADE / DEGREE

Department of Mathematics and Statistics

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

The Statistical Analysis of Generalized Adjacency and GA-Clusters.

TITRE DE LA THÈSE / TITLE OF THESIS

Dr. David Sankoff

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Dr. Raluca Balan

Dr. Sanjoy Sinha

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

# THE STATISTICAL ANALYSIS OF GENERALIZED ADJACENCY AND GA-CLUSTERS

Ximing Xu

Thesis Submitted to the Faculty of Graduate and Postdoctoral Studies

In partial fulfillment of the requirements

for the degree of

Master of Science in Mathematics[1]

Department of Mathematics and Statistics

Faculty of Science

University of Ottawa



[1]The M.Sc. Program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

*Your file Votre référence*
*ISBN: 978-0-494-46505-9*
*Our file Notre référence*
*ISBN: 978-0-494-46505-9*

# Abstract

In this thesis I study a parametrized definition of gene clusters that permits control over the trade-off between increasing gene content versus conserving gene order within a cluster. This is based on the notion of generalized adjacency, which is the property shared by any two genes no farther apart, in the linear order of a chromosome, than a fixed threshold parameter $\theta$. We discuss the the statistical properties of generalized adjacency (**GA** ) and derive the limiting probability distribution of the number of **GA** for random genomes.We also propose a test for gene clusters satisfying the generalized adjacency criterion under the null hypothesis that the genes are ordered randomly along the genomes.

# Acknowledgements

I would like to express my warmest thanks to my supervisor, David Sankoff, for his constant encouragement, support and suggestions, which are invaluable to me. It has been a most rewarding research experience to work under his guidance. I also wish to thank Zhenyu Yang and Wei Xu for their helpful discussions through writing this thesis.

# Dedication

This thesis is dedicated to my uncle and aunt who supported me with free food and living place, as well as continual encouragement and lots of helpful suggestions for my graduate study.

# Contents

# List of Figures

# Chapter 1

# Introduction

The increasing availability of comprehensive linkage maps and complete genomic sequences from many prokaryotes, eukaryote organelles and more recently eukaryote nuclei has led to the burgeoning of a new area of **Comparative Genomics** based on the macrostructure of entire genomes rather than on the traditional comparison of a single homologous gene [1] or a protein sequence in different organisms. Comparing chromosomal gene order in two or more related species is an important approach to reconstruct chromosomal rearrangements in different evolutionary lineages and make phylogenetic inference; in addition, a genome self-comparison approach has been used to identify the evidence of large scale or whole genome duplication [12, 13, 15]. Other applications of this area include detecting operons, horizontal transfer, and functional selection in bacteria [5].

One of the fundamental tasks in the above studies is the identification of homologous chromosomal segments, pairs of chromosomal regions, one in each genome, that are descended from a single contiguous region in the ancestral genome, either through speciation or duplication.

[1] Homologs: Genes or features that share common ancestry, usually detected by structures of similarity.

## 1.1 Biological Background

Following speciation, offspring genomes initially have identical gene content and order. Similarly, a whole genome duplication yields a new genome with two identical copies of the ancestral genome. In both cases the two genome copies will invariably diverge over time. Genomes, containing the entire genetic complement of an organism, will evolve as the genes in them evolve through the processes of nucleotide substitution, insertion and deletion. Gene duplication, gene loss and horizontal transfer will also alter the gene complement, the set of genes appearing in the genome. In addition, larger scale *genome rearrangements* including translocation, transposition and inversion (**Figure** 1.1) disrupt gene order and syntenic[2] structure [14]. The evolutionary events result in closely related genomes sharing a few large conserved chromosomal segments while more distantly related genomes have many short segments.

According to the most stringent definition, conserved chromosomal segments are defined as any maximal contiguous chromosomal regions with the same gene content, order, and even orientation (the transcription direction associated with each gene) in two or more compared genomes. However, in practice it is useful to relax the stringent definition to some extent to detect evolutionary signal in conserved similar regions, to avoid unstable estimates of the number of segments if these may be as small as one or two genes [3], and to diminish the effect of experimental error and other noises. The experimental errors can be attributed to gross mistakes in chromosomal assignment of genes, quantitative errors in map positions as well as the errors occurring when integrating mapping results from different sources [8, 16]. A less strict concept is *gene clusters*, pairs of regions with similar, but not identical gene content and gene order.

[2]Two genes located on the same chromosome in a genome are said to be syntenic in that genome.

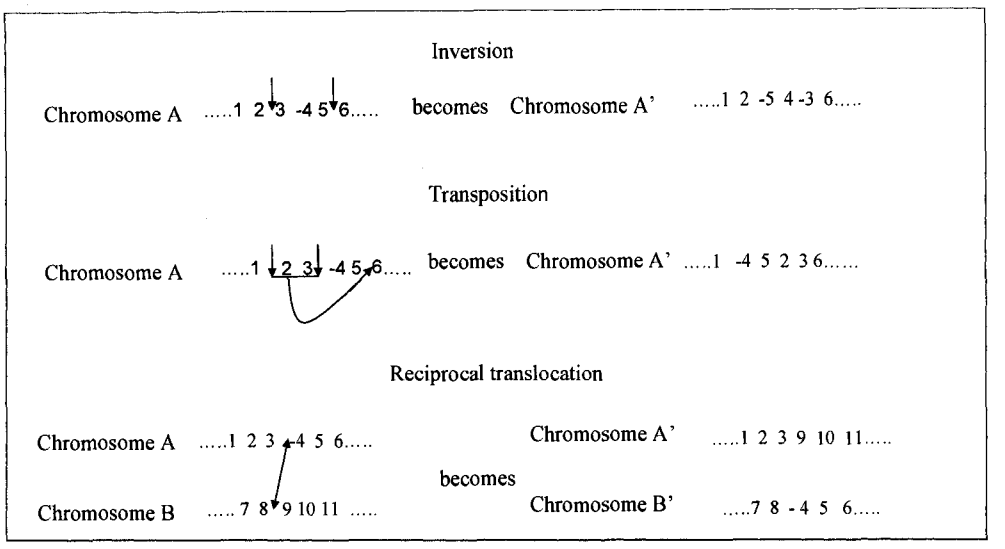



Figure 1.1: Schematic view of three most common genome rearrangement processes. Integers represent genes of interest and the sign indicates gene orientation. Black arrows represent breakpoints introduced into original genome. Inversion reverses both the order and the signs of genes between two breakpoints. Transposition removes a segment defined by two breakpoints and insert it at another breakpoint. Reciprocal translocation exchanges end segments of two chromosomes.

## 1.2 Cluster Identification

Gene clusters are usually detected based on the locations of *genomic markers*, rather than direct comparison of the primary sequence. Here, the *genomic markers* could be genes, motifs, anchors or, indeed, the results of any decomposition of the chromosome into disjoint ordered fragments identifiable in the two or more genomes being compared. Genes are commonly used as markers since their sequences tend to be conserved over long periods of evolutionary time, they usually exist in large numbers in a genome, and they are the unit of interest in many genomic studies [8].

In this thesis I assume that genes are used as markers and each gene has been

determined and its homologs in other genomes have also been identified. Then a genome is considered as a set of chromosomes, each represented as a sequence of genes. However, the discussion in this thesis is general enough to be applied to other types of markers as well. Gene orientation, which is not considered in most existing cluster definitions, is also disregarded in this thesis.

### 1.2.1 Cluster Definitions

Since in most cases evolutionary histories are unknown, it is difficult to characterize what such genomic regions suggesting common ancestry look like. Consequently, a number of definitions of *gene clusters* have been proposed, based on intuitive notions, characteristics of genomic data sets or the goals of studies [3, 4]. A number of algorithms have been developed to search for gene clusters following these definitions.

Criteria for identifying gene clusters, in two or more genomes entail a trade-off between increased content versus stricter order: if we require *genes* of the cluster to be ordered identically within different genomes, so that we can have great confidence that these are genuine, evolutionarily conserved or functionally determined configurations. However, only relatively small clusters are likely to satisfy this restrictive condition, and the corresponding analysis will miss large common genomic regions that only suffer small, perhaps insignificant, disruptions of common order. On the other hand, by allowing unrestricted scrambling of genes within the clusters (e.g., $r$-windows [3], max-gap [5] or "gene teams" [2]), we may be able to detect larger, more loosely structured groupings, but at least in the first analysis, must forgo accounting for local genome rearrangement, missing an important aspect of evolutionary history, and we relinquish the possibility of pinpointing extensive local conservation of order within the group.

Now I will introduce four commonly applied definitions of gene cluster: *conserved segment, common interval, r-window cluster* and *max-gap cluster* (also referred as "gene teams" [2]).

- The most conservative definition is **conserved segment** as defined in Section 1.1. However, the stringent definition will exclude many regions that did indeed descend from a single ancestral region but have undergone a series of small rearrangements.

- A **common interval** is defined to be a set of genes occurring contiguously in each of the genomes compared, ignoring gene order, but without allowing gene insertions and deletions.

- An $r$-**window cluster** is defined as a pair of windows, each containing $r$ genes, in which at least $k$ genes are shared. This definition allows rearrangements as well as a limited number of insertions and deletions. If $k$=$r$, an $r$-window cluster reduces to a common interval of size $k$. How to best choose the values of $r$ and $k$ is a problem in practice.

- A **max-gap cluster** is a set of marked genes where the number of intervening genes between adjacent marked genes in each genome compared is not larger than a given gap parameter, $g$. This definition also ignores gene order and allows insertions and deletions, but does not constrain the maximum length of the cluster. When $g$=0, max-gap clusters reduces to common intervals. A max-gap cluster is maximal if it is not contained within any larger max-gap cluster.

Now I provide an **example** to illustrate the four different cluster definitions.

Given two genomes: $G_1 = 1\ \ 2\ \ *\ \ 3\ \ 4\ \ 5\ \ *\ \ *\ \ 6\ \ 7\ \ *\ \ 8$

and $G_2 = 2\ \ *\ \ 6\ \ 7\ \ *\ \ 8\ \ 1\ \ *\ \ 5\ \ 3\ \ 4$

where the integers denote homologous gene pairs and the stars indicate genes with no homolog in the other genome, then we can find

1. Conserved segments: {3, 4} and {6, 7}.
2. Common intervals: {3, 4, 5} and {6, 7}.
3. $r$-window clusters
   (a) when $r$=5, $k$=3: {1, 3, 4}, {3, 4, 5} and {6, 7, 8};
   (b) when $r$=6, $k$=4: {1, 3, 4, 5}.
4. Maximal max-gap clusters
   (a) when $g$=2: {1, 2, 3, 4, 5, 6, 7, 8};
   (b) when $g$=1: {6, 7, 8} and {3, 4, 5}.

In **Chapter** 3, I will introduce a parametrized definition of gene clusters, **GA-cluster**, that allows us to control the emphasis placed on conserved order within a cluster [21] and hence to systematically explore the details of the content/order trade-off. I have presented some of the methods used in this chapter elsewhere [20].

### 1.2.2 Statistical Tests for Gene Clusters

Once gene clusters have been identified, it is imperative to determine the significance of gene clusters by rejecting the hypothesis that such a cluster could have occurred by chance. A number of tests have already been proposed based on randomization, combinatorial analysis, as well as formal statistical frameworks.

The significance of gene clusters does not only depend on how "closely" the genes in the cluster distribute in the genomes, which is what most existing tests focus on, but also on the searching strategies, *i.e.* how the cluster was found. Durand and Sankoff [3] have modeled three common search strategies.

- **Reference region**. Given a set of prespecified genes, which may be found contiguously located in a *reference* genome or share a particular functional property, our goal is to search along the whole genome and find a cluster, which contains all or part of these genes.

- **Window sampling**. Given a pair of windows of genes, a cluster may be identified by determining whether they share a significant number of homologs. These windows are selected may because they share a pair of previously known homologs of particular interest.

- **Whole genome comparison**. Given two genomes, our goal is to identify all gene clusters locating in close proximity in both genomes. Individual gene clusters are usually found in this scenario.

The searching strategies have been systematically discussed in the statistical tests for $r$-window clusters [3] and max-gap clusters [5].

A lot of other factors also contribute to cluster significance, such as gene order and orientation, however, most statistical tests for gene clustering ignore them since the interactions between these factors are extremely complex, and how to combine them in one test is still an open problem ([5], [9], [3], [17]). Recently, some statistical tests have began to take multiple (more than two) genomes and gene families into account ([10], [3]).

In **Chapter 2** I will discuss statistical tests for *r-window* clusters and *max-gap* clusters in more detail. In **Chapter 4**, under the null hypothesis that the genes are

ordered completely randomly on the genomes, I propose tentative statistical tests for GA-clusters through **whole genome comparison** based on a simplified model of genomes. I have presented some of the methods used in Chapter 4 elsewhere [20].

# Chapter 2

# Statistical Tests for *r-window* and *max-gap* Clusters

*r-window* and *max-gap* cluster definitions have been widely used in empirical studies, however, very few formal statistical models have been developed to test the significance of clusters, while most current approaches estimate the distributions of test statistics based randomization. Recently, Durand and Sankoff [3] constructed statistical tests for *r-window* clusters under different scenarios; Hoberman, *et al.* [5] present analytical statistical models for clusters satisfying max-gap criterion. Discussions in this chapter are mainly based on the two papers.

We will not consider the genomes with circular chromosomes. A genome is modeled as a set of $n$ genes, ordered by their positions in the genome: $G=\{1, 2, \ldots, n\}$, ignoring physical distances between genes. We assume genes do not overlap and the distance between two genes is simply the number of genes between them.

## 2.1 Statistical Tests for *max-gap* Clusters

In the rest of this thesis we use the term *max-gap* clusters as shorthand for *maximal max-gap* clusters. Hoberman, *et al.* [5] develop statistical tests for max-gap clusters found in two different searching strategies, **reference region** and **whole genome comparison**. Due to the nature of max-gap definition, a max-gap cluster cannot be identified through the local searching strategy, **window sampling** [9].

### 2.1.1 Reference Region

In this scenario, $m$ genes are prespecified (or marked), each of which has exact one homolog in the genome of $n$ genes, our goal is to test the significance of a cluster containing all or part of the $m$ specified genes, identified through searching the whole genome. The null hypothesis is that the $m$ genes are randomly distributed in the genome.

1. **Complete cluster case**. The test statistic is the maximum gap between the marked genes. Under the null hypothesis, the probability of observing that $m$ marked genes form a max-gap cluster with maximum gap less than or equal to $g$ (equation(2) in [5]) is:

$$P(n,m,g) = \frac{1}{\binom{n}{m}} \begin{cases} (n - w_{mg} + 1 + \frac{w_{mg}-m}{2})(g+1)^{m-1}, & \text{if } w_{mg} \leq n+1 \\ d_0(m,g,n) & \text{otherwise.} \end{cases} \tag{2.1.1}$$

   where $w_{mg} = m + (m-1)g$ and $d_0(m,g,n) = \sum_{i=0}^{(n-m)/(g+1)} (-1)^i \binom{m-1}{i} \binom{n-i(g+1)}{m}$

2. **Incomplete cluster case**. The maximum gap value $g$ is fixed in advance. Unlike complete cluster case, the $m$ marked genes can form different clusters in the same genome, and the size of the largest cluster is used as test statistic.

Under the null hypothesis, the probability that this test statistic is not less than the observed value $h$ (equation(4) in [5]) is:

$$Q(n, m, h, g) = 1 - \frac{\eta[n, m, g+1, 0]}{\binom{n}{m}} \tag{2.1.2}$$

where $\eta[n, m, j, c]$ can be calculated as the following recursion formula:

$$\eta[n, m, j, c] = \begin{cases} 0, & \text{if } c = h \text{ or } n < m \\ 1, & \text{else if } m = 0 \\ \eta[n-1, m, j+1, c] + \eta[n-1, m-1, 0, c+1], & \text{else if } j \leq g \\ \eta[n-1, m, j+1, c] + \eta[n-1, m-1, 0, 1], & \text{otherwise.} \end{cases}$$

Based on Equation 2.1.1, the probability of observing a complete cluster as a function of $m$ for $n = 1,000$ is shown in Figure 2.1 (FIG.2. in [5]). The probability of observing a complete cluster is an increasing function of $g$, however, it does not increase monotonically with $m$, as expected. When $m$=$n$, we can always observe a complete cluster for any value of $g$. The similar trend is also found for the probability of observing an incomplete cluster as a function of $m$ given $h = \frac{m}{2}$. So, for max-gap cluster definition without constraints on the length or order, larger clusters do not always imply greater significance, which disagrees with "a widespread belief that cluster significance grows with the number of homologs in the cluster" [5].

### 2.1.2 Whole Genome Comparison

In this scenario, we only consider pair-wise comparison. We are given two genomes, $G_1$ and $G_2$, each containing $n$ genes sharing $m$ genes. The null hypothesis is that the $m$ genes are randomly distributed in $G_1$ as well as in $G_2$. Like reference region scenario, we also assume that each of the $m$ genes in $G_1$ has exact one homolog in $G_2$, and

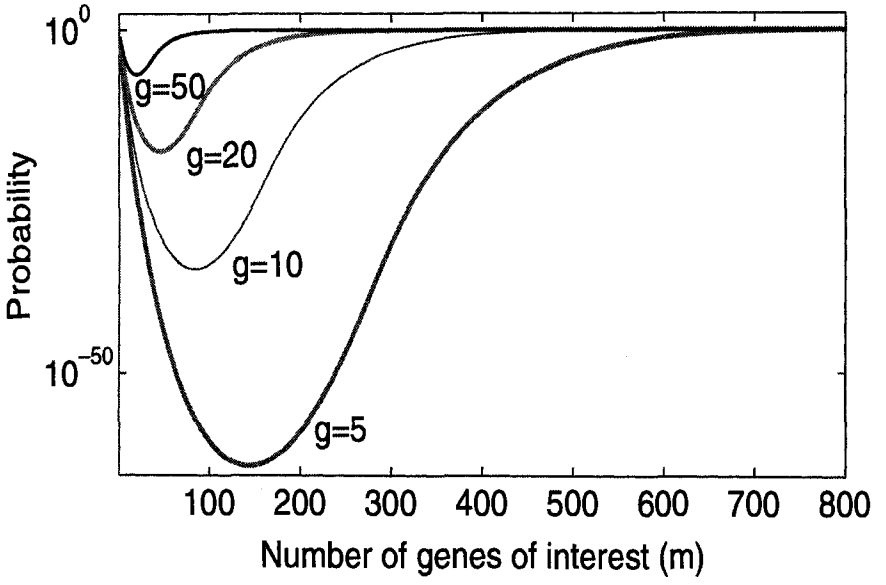



Figure 2.1: (FIG.2. in [5])Probability of a complete max-gap cluster of m marked genes in a genome of size n=1,000 as a function of m, for g={5, 10, 20, 50}.

vice versa. Through whole genome comparison between $G_1$ and $G_2$, the probability of observing a complete max-gap cluster of $m$ common genes is $[P(n, m, g)]^2$, where $P(n, m, g)$ is defined in Equation 2.1.1. If $G_1$ and $G_2$ are closely related and share a high percentage of genes, this quantity can approach 1 (*e.g.* $m = n$). Consequently, in whole genome comparison scenario, rather than calculate the probability of observing a cluster of size at least $h$, Hoberman, *et al.* [5] tried to determine the probability of observing a cluster of exactly size $h$, and present the upper and lower bounds for this quantity.

## 2.2 Statistical Tests for $r$-*window* Clusters

Durand and Sankoff [5] present a comprehensive analysis on statistical tests suitable for $r$-window clusters. We begin with a simple model under reference region scenario.

Given a genome $G$ of $n$ genes, and $m$ prespecified genes each of which has exact one homolog in $G$, our goal is to test the significance of a $r$-window cluster containing all of the $m$ genes, identified through searching the whole genome. The null hypothesis is that the $m$ genes are randomly distributed in a genome with $n$ genes. The probability of observing that the $m$ genes span at most $r$ slots in $G$ (equation(2) in [3]) is given as follows. If this quantity is less than a given level $\alpha$, then we reject the null hypothesis, and think the observed $r$-window cluster is significant.

$$q(n,m,r) = \frac{(n-r)\binom{r-1}{m-1} + \binom{r}{m}}{\binom{n}{m}}$$

### 2.2.1 Gene Family Model

Since virtually all genomes contain gene families, sets of genes with similar sequence and function, that arose through duplication of genetic material [5], and identifying true homologs is still much debated, a more realistic and complicated model, the **Gene Family Model** has also been studied.

We assume that homology relationships have already been determined and the genes in a genome can be partitioned into non-intersecting gene families. Every gene in gene family $f_j$ is homologous to all the other members in the same family. Genes in different families cannot be homologous. Define the gene family size, $\phi_{ij}$, as the number of genes in gene family $f_j$ in genome $G_i$. Let $\mathcal{F} = \{f_j\}$ be the set of all gene families in genomes under consideration and $n_f = |\mathcal{F}|$ the number of total gene families. Let $n_i$ be the number of genes in genome $G_i$.

Based on the *gene family model*, Durand and Sankoff [5] have constructed statistical tests against null hypotheses of random gene order, taking incomplete clusters, multiple genome comparison and self-genome comparison into account. However, the treatment presented is mainly of theoretical interest and the expressions for calcu-

lating the *p-values* of their test statistics are not computationally tractable. They also give the formulae for calculating the expected number of clusters of a given type under different cases, which can be used as informal tests.

### 2.2.2 Window Sampling

Unlike max-gap clusters, $r$-window clusters can be obtained by window sampling.

1. **Two genomes without gene families.** Given two genomes $G_1$ and $G_2$ each containing the same set of $n$ genes, and a pair of windows of length $r$, $W_1$ and $W_2$, drawn from $G_1$ and $G_2$ respectively. Under the null hypothesis that genes are randomly distributed in $G_1$ and $G_2$, the probability that $W_1$ and $W_2$ share at least $m$ genes (equation (22) in [3]) is:

$$P(n, r, m) = \sum_{i=m}^{r} \frac{\binom{r}{i}\binom{n-r}{r-i}}{\binom{n}{r}}$$

2. **Two genomes with gene families.** Under gene family model, a **$r$-window cluster** is redefined as a pair of windows, each containing $r$ genes, in which at least $k$ gene families are shared. We are given two genomes $G_1$ and $G_2$ containing $n_1$ and $n_2$ genes respectively and having the same set of gene families $\mathcal{F}$. Two windows of length $r$, $W_1$ and $W_2$ are selected from $G_1$ and $G_2$ respectively. Durand and Sankoff [5] give an expression for calculating the probability that $W_1$ and $W_2$ share at least $m$ genes (equation(23) in [3]), but it is computationally intractable. Based on further assumption of fixed gene family size, *i.e.*, all $\phi_{ij}$ (as defined in Section 2.2.1) take the same value, $\phi$, Raghupathy and Durand [9] provided a computationally tractable expression for this probability using generation function approach:

$$Q(m) = \sum_{k=m}^{r} [\binom{n_f}{k} p_1(k) \sum_{l=m}^{k} \binom{k}{l} p_2(l)]$$

$$\text{where, } p_1(k) = \binom{n_1}{r}^{-1} (-1)^k \sum_{i=\frac{r}{\phi}}^{l} [(-1)^i \binom{k}{i} \binom{i\phi}{r}]$$

$$p_2(l) = \binom{n_2}{r}^{-1} \sum_{z=max(0,r-k\phi)}^{r-l} (-1)^l \sum_{i=\lceil \frac{r-z}{\phi} \rceil}^{l} [(-1)^i \binom{l}{i} \binom{i\phi}{r-z}] \binom{n_2 - k\phi}{z}$$

# Chapter 3

# Statistical Properties of Generalized Adjacency

Zhu *et al.* [21] presented a new parametrized definition of gene clusters that allows us to control the emphasis placed on conserved order within a cluster and hence to systematically explore the details of the content/order trade-off. The basis for this is the notion of ***generalized adjacency***, which is the property shared by any two genes no farther apart, in the linear order of a chromosome, than a fixed threshold. Then a ***generalized adjacency cluster*** in two or more genomes is just a maximal set of genes, where in each genome these genes form a connected chain of generalized adjacencies. Increasing the size of the threshold relaxes the degree of common ordering required, within a cluster, in different genomes.

Nevertheless, for any fixed threshold, evolutionary rearrangements continue to disrupt the orders of genes on chromosome and will create, alter or destroy *generalized adjacency clusters.* Since even pairs of randomly constructed genomes may have some *generalized adjacency clusters* in common, the question arises of whether the number or size of these common clusters is significantly larger than the random case. To

answer such questions, in this chapter we study the statistical properties of *generalized adjacency*; in Chapter 4 we will propose tentative tests for *generalized adjacency clusters* identified through *whole genome comparison* under the null hypothesis that the genes are ordered completely randomly on the genomes.

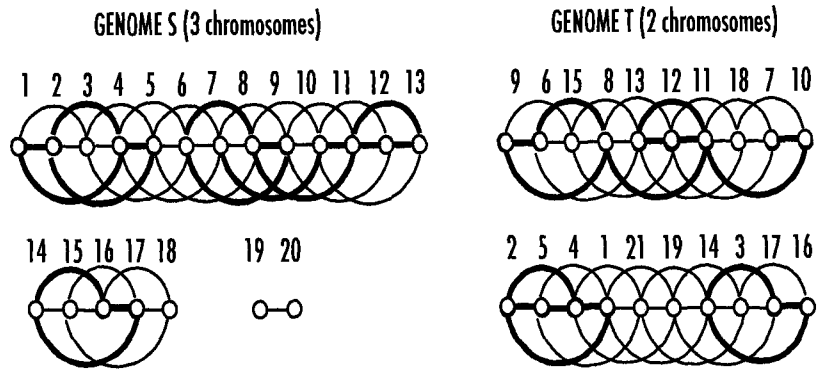
## 3.1 Definitions

An undirected graph is represented as $G = (V, E)$, where $V$ is a set of **vertices** or nodes of this graph, and $E$ denotes the set of unordered pairs of vertices, called **edges**.

Let $V_X$ be the set of genes (treated as vertices) in the genome $X$. These genes are partitioned among a number of total orders called **chromosomes**. For genes $g$ and $h$ in $V_X$ on the same chromosome in $X$, let the pair $gh \in E_X$ if the number of genes intervening between $g$ and $h$ in $X$ is less than $\theta$, where $\theta \geq 1$ is a fixed **neighbourhood parameter**.

Given two genomes $X$ and $Y$, and the graphs $G_X = (V_X, E_X)$ and $G_Y = (V_Y, E_Y)$ with a non-null set of vertices (genes) in common $V_{XY} = V_X \cap V_Y$. We say a subset of $C \subseteq V_{XY}$ is a ***generalized adjacency cluster*(GA-cluster** for short) if it consists of the vertices of a maximal connected subgraph of $G_{XY} = (V_{XY}, E_X \cap E_Y)$. If $gh \in E_{XY}$, where $g \in V_{XY}$, $h \in V_{XY}$ and $E_{XY} = E_X \cap E_Y$, then we call there is a **generalized adjacency**(**GA** for short) in genomes $X$ and $Y$.

Our definition of generalized adjacency clusters is illustrated in Figure 3.1. This definition of clusters decomposes the genes in the two genomes into identical sets of disjoint generalized adjacency clusters of size greater or equal to 2, and possibly different sets of singletons belonging to no cluster, either because they are in $V_{XY}$, but not in $E_X \cap E_Y$, or because they are in $V_X \cup V_Y \setminus V_{XY}$. For simplicity, we do not




**Generalized Adjacency Clusters:**

θ= 2 : {2,4,5}, {6,8}, {11, 12 13}, {16, 17}
θ= 3 : {1,2,4,5}, {6,7,8,9,10, 11, 12 13}, {14, 16, 17}
θ= 4 : {1,2,3,4,5}, {6,7,8,9,10, 11, 12 13}, {14, 16, 17}

Figure 3.1: Graphs constructed from two genomes using parameter $\theta = 3$. Thick edges determine generalized adjacency clusters. Clusters listed for $\theta = 2$ and $\theta = 4$ as well.

attempt to deal with duplicate genes in this paper, *i.e.* we do not take gene families into account, and we also assume $V_X = V_Y = V_{XY}$. In practice, depending on the relative emphasis to be placed on order rearrangement versus gene insertion/deletion, we can delete all genes in $V_X \cup V_Y \setminus V_{XY}$ *before* calculating $E_X$ and $E_Y$, so as to exclude the effect of the markers unique to $X$ or unique to $Y$.

When $\theta = 1$, a GA-cluster reduces to a *conserved segment* conserving exactly the same gene content and order (or reversed order) in both genomes. When $\theta = \infty$, the definition returns simply all the synteny sets, namely the sets of markers in common between two chromosomes, one in each genome.

## 3.2 The number of GAs in two random genomes

In our model, each genome can be represented as a permutation of the first $n$ positive integers. We denote by $I$ the *reference genome* $1, 2, \ldots, n$ and by $R$ the *random genome* sampled from all $n!$ possible genomes, each with probability of $\frac{1}{n!}$.

Given two random genomes, we can always relabel genes to convert either of the genome to *reference genome*, without changing any property (*e.g.* number, location and size) of generalized adjacencies and GA-clusters except the "names" of genes, since two genes adjacent in two genomes are still adjacent in both genomes after conversion while two genes without sharing common adjacency can not be adjacent in both genomes after conversion either. It is also easy to show that the distribution of the number of GAs between two random genomes is the same as the distribution when one genome is fixed as $I$ while the other genome is randomly picked from the set of $n!$ possible genomes.

Let $n_2 = |E_I \cap E_R|$ denote the number of common edges, i.e. the number of **GAs**. For a random genome $R = r_1, r_2, \ldots, r_n$, if $r_h = i$, we define the *position* of $i$ in $R$ to be $g_i = h$. Then

$$|E_I \cap E_R| = |\{1 \leq i < j \leq n \mid j - i \leq \theta, |g_i - g_j| \leq \theta\}|. \tag{3.2.1}$$

Next we will study the probability distribution of $n_2$.

### 3.2.1 For large $\theta$

A potential problem with generalized adjacency clustering, which it shares with other methods such as max-gap, is that beyond certain values of $\theta$, instead of large clusters being statistically significant, the absence of such clusters becomes significant. We examine these cases first, before analyzing the more useful, smaller values of $\theta$.

1. $\theta \geq n-1$. In this case $n_2 = |E_I| = |E_R| = \binom{n}{2}$, so that $P[n_2 = \binom{n}{2}] \equiv 1$. This is because all pairs $g_i$ and $g_j$ , $i \neq j$, satisfy $|i-j| \leq \theta$ and $|g_i - g_j| \leq \theta$.

2. $\theta = n-2$

   (a) If two ends of the genome are occupied by 1 and $n$, which occurs with the probability of $\frac{2}{n(n-1)}$, $n_2 = \binom{n}{2} - 1$.

   (b) If $|(g_1, g_n) \cap \{1, n\}| = 1$, which occurs with the probability of $\frac{4(n-2)}{n(n-1)}$, $n_2 = \binom{n}{2} - 2$.

   (c) If $|(g_1, g_n) \cap \{1, n\}| = 0$, which occurs with the probability of $\frac{(n-2)(n-3)}{n(n-1)}$, $n_2 = \binom{n}{2} - 2$

   Thus, $P[n_2 = \binom{n}{2} - 1] = \frac{2}{n(n-1)}$ and $P[n_2 = \binom{n}{2} - 2] = \frac{(n-2)(n+1)}{n(n-1)}$

3. $\theta = n-k$, where $k$ is a positive integer and smaller than $\frac{n}{2}$. In this case,

$$|E_I| = |E_R| = k(n-k) + \frac{(n-k)(n-k-1)}{2}.$$

   Now, $|E_I \cap E_R| \geq |E_I| - \frac{k(k-1)}{2}$, because the number of the pairs $(g_i, g_j)$, $i \neq j$ satisfying both $|i-j| \leq \theta$ and $|g_i - g_j| > \theta$ can not be greater than $\frac{k(k-1)}{2}$. Then

$$n_2 \geq \binom{n}{2} - 2\binom{k}{2}$$

So, for small $k$, the value of $n_2$ can be very large, and the information on gene order is weak.

### 3.2.2 For small $\theta$

$\theta = 1$. The definition of generalized adjacency reduces to the ordinary notion of adjacency. In this case $n_2$ has been used to measure the similarity of two genomes. The asymptotical distribution of $n_2$ was proved to be Poisson distribution with mean

value 2 by Wolfowitz, 1944 [18]. The exact expression for its probability distribution has already been given by Kaplansky [6] and Robbins [11]. Recently this problem is restudied by Xu and Sankoff [19], using generating function approach, and generalized to more realistic cases taking multi-chrosome, circular chromosome and gene orientation into account.

In our model the probability that $n_2$=k is of the following form (equation(2) in [6])

$$P(n_2 = k) = \frac{2^k e^{-2}}{k!}\left[1 - \frac{k^2 - 3k}{2n} + \frac{k^4 - 8k^3 + 9k^2 + 22k - 16}{8n(n-1)}\right] + O(n^{-3})$$

Hence, it is easy to get

$$\lim_{n\to\infty} P(n_2 = k) = \frac{2^k e^{-2}}{k!},$$

the Poisson distribution with mean value 2.

We now present our main analytical results. We first examine the expected value $\mathbf{E}(n_2)$ of the number of adjacencies common to $I$ and $R$.

**Proposition 3.2.1** *For $\theta \geq 1$,*

$$\mathbf{E}(n_2) = 2\theta^2 - \frac{4n\theta^3 - \theta^2(1+\theta)^2}{2n(n-1)},$$

*so that for a given $\theta$*

$$\lim_{n\to\infty} \mathbf{E}(n_2) = 2\theta^2$$

**Proof:** Counting the total number of edges in $E_I$, we have

$$|E_I| = (n-\theta)\theta + \sum_{i=1}^{\theta-1} i = n\theta - \binom{\theta+1}{2}$$

Each of these edges has the same probability

$$p = \frac{2(n-2)!}{n!}\sum_{i=1}^{\theta}(n-i)$$

of occurring in $E_R$. Thus

$$\begin{aligned}\mathbf{E}(n_2) &= \sum_{i=1}^{|E_I|}(p\cdot 1+(1-p)\cdot 0)\\ &= |E_I|p\\ &= 2\theta^2-\frac{4n\theta^3-\theta^2(1+\theta)^2}{2n(n-1)}.\end{aligned}$$

.

### 3.2.3 The limiting probability distribution of $n_2$

We can say more about the limiting behaviour of $n_2$. We begin by introducing some theorems on convergence in distribution.

**Theorem 3.2.2** (Theorem 30.1 in [1]). *let $\mu$ be a probability measure on the line having finite moments $\alpha_k = \int_{-\infty}^{\infty} x^k \mu(dx)$ of all orders. If the power series $\sum_k \alpha_k r^k/k!$ has a positive radius of convergence, then $\mu$ is the only probability measure with the moments $\alpha_1$, $\alpha_2$,....*

A probability measure is called *determined by its moments* if it satisfies the conclusion of Theorem 3.2.2.

**Theorem 3.2.3** (Theorem 30.2 in [1]). *Suppose that the distribution of $X$ is determined by its moments, that the $\mathbf{X}_n$ have moments of all orders, and that $\lim_n \mathbf{E}[\mathbf{X}_n^r] = \mathbf{E}[\mathbf{X}^r]$ for r=1,2,.... Then the distribution of $\mathbf{X}_n$ converges to the distribution of $X$.*

From Theorem 3.2.2 and Theorem 3.2.3, it is easy to get

**Corollary 3.2.4** (Theorem 2 in [19]). *For probability distributions of $X_n$, if their $k^{th}$ factorial moment, $\mathbf{E}[\boldsymbol{X}_{(k)}] = \int_{-\infty}^{\infty} x(x-1)\cdots(x-(k-1))\mu(dx)$, converges to $\lambda^k$, then their probability distributions converge to **Poisson** distribution with mean $\lambda$.*

**Theorem 3.2.5** *For $\theta \geq 1$, $n_2$ converges in distribution to a Poisson distribution with mean value $2\theta^2$.*

**Proof:** Here we only prove the limiting probability distribution of $n_2$ is $Poisson(8)$ for $\theta = 2$. Using same approach we could get the required results for general cases. Define random variables

$$y_i = \begin{cases} 1, & \text{if there exists a cluster of size 2 starting with gene } i \text{ in the random genome } R, \\ 0, & \text{otherwise, } i = 1, 2, \ldots, n. \end{cases}$$

The probability that gene $i$ will be the initial element of a cluster of size greater than two is $O(\frac{1}{n^2})$, and hence that the probability of the occurrence of a cluster of size greater than two anywhere in the genome is $O(\frac{1}{n})$. So the limiting distribution of $n_2$ is the same as that of

$$y = \sum_{i=1}^{n} y_i,$$

provided either exists.

Now we consider the $k^{th}$ factorial moment of $y$,

$$E\left[y\left(y-1\right)\ldots\left(y-k+1\right)\right] = \sum_{i_1, i_2, \ldots, i_k \in V} E\left[y_{i_1} y_{i_2} \ldots y_{i_k}\right], \tag{3.2.2}$$

where V={1,2,3,4,...,n}, and no two elements are equal for all $k$ tuples, $i_1$, $i_2$,..., $i_k$. Equation 3.2.2 holds because its both sides represent the expectation of the number of ways to choose $k$ non-zero elements from $\{y_1, y_2, \ldots, y_n\}$. By Corollary 3.2.4, to prove the limiting probability distribution of $y$ is $Poisson(8)$, we only need to show that its $k^{th}$ factorial moment converges to $8^k$.

$E\left[y_{i_1} y_{i_2} \ldots y_{i_k}\right]$ is just the probability that $y_{i_1} = 1, y_{i_2} = 1, \ldots, y_{i_k} = 1$, simultaneously, since all $y_i$ can only take two values, 1 and 0, and

$$\begin{aligned} E\left[y_{i_1} y_{i_2} \ldots y_{i_k}\right] &= P(y_{i_1} = 1, y_{i_2} = 1, \ldots, y_{i_k} = 1) \cdot 1 + 0 \\ &= P(y_{i_1} = 1, y_{i_2} = 1, \ldots, y_{i_k} = 1) \end{aligned}$$

This probability is either zero (*e.g.* when $i_1 = 1$ is the initial element of the cluster $1, 2$ and $i_2 = 2$ is the initial element of the cluster $2, 1$, or when some of them form a cluster with size larger than 2) or is of the form of $\left(\frac{8}{n}\right)^k + O\left(\frac{1}{n^{k+1}}\right)$ (except the $k$-tuples which contain at least one of the four genes, $1, 2, n-1$ and $n$), where the number 8 comes from the fact that gene $i$ ($i = 3, 4, 5, \ldots, n-2$) can be the initial element of a cluster of size 2 in the following 8 forms: $(i, i-2)$, $(i, i-1)$, $(i, i+1)$, $(i, i+2)$, $(i, \star, i-2)$, $(i, \star, i-1)$, $(i, \star, i+1)$ and $(i, \star, i+2)$, where $\star$ denotes a gene.

Moreover, the ratio of the number of $k$-tuples $i_1, i_2, \ldots, i_k$ for which the probability is zero or which contain at least one of the four genes $1, 2, n-1$ and $n$, to the number of $k$-tuples for which the probability is the form of $\left(\frac{8}{n}\right)^k k! + O\left(\frac{1}{n^{k+1}}\right)$ is $O(\frac{1}{n})$. So,

$$\begin{aligned}
\lim_{n\to\infty} E\left[y\left(y-1\right)\ldots\left(y-k+1\right)\right] &= \lim_{n\to\infty} \sum_{i_1, i_2, \ldots, i_k \in V} E\left[y_{i_1} y_{i_2} \ldots y_{i_k}\right] \\
&= \lim_{n\to\infty} \binom{n}{k} k! \left(\left(\frac{8}{n}\right)^k + O(\frac{1}{n^{k+1}})\right) \\
&= 8^k.
\end{aligned}$$

## 3.3 Experiments

We generated $10,000$ random permutations for $n = 100$ and calculated $n_2$ for various values of $\theta$. In Figure 3.3, we compare the simulated distribution of $n_2$ (with means indistinguishable from $2\theta^2 - \frac{4n\theta^3 - \theta^2(1+\theta)^2}{2n(n-1)}$ in each case) to the Possion distribution with mean value $2\theta^2$, for $\theta = 2, 5$ and 10. For fixed $n$, the difference is larger as $\theta$ increases, though as $n$ increases the Poisson is the limiting distribution. We also did simulation for other values of $n$ and $\theta$, and some results are shown in Figure 3.3. We can find that for $\theta = 5$, the difference becomes smaller as $n$ increases.

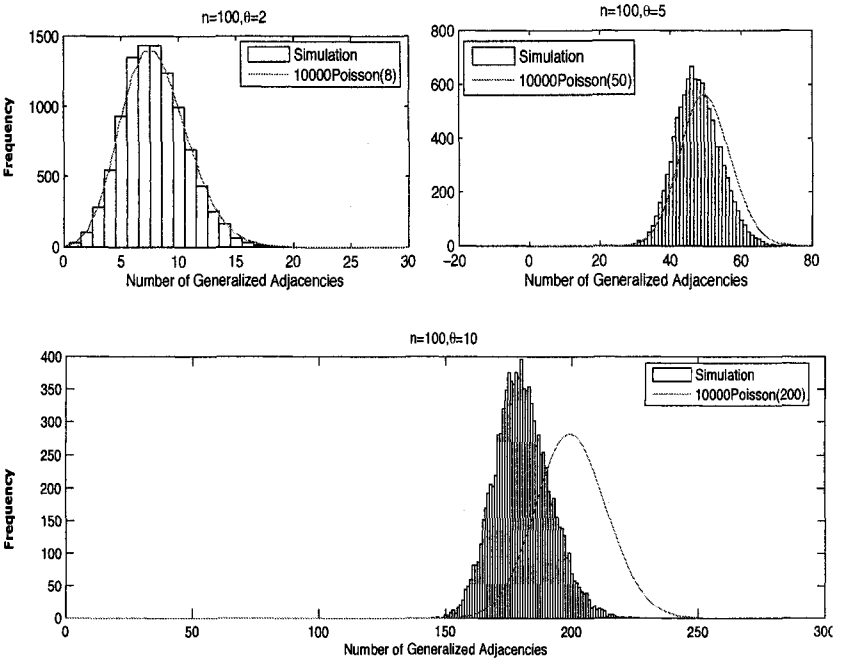



Figure 3.2: Empirical distribution of $n_2$ compared to the related Poisson distribution for $\theta$=2, 5 and 10.

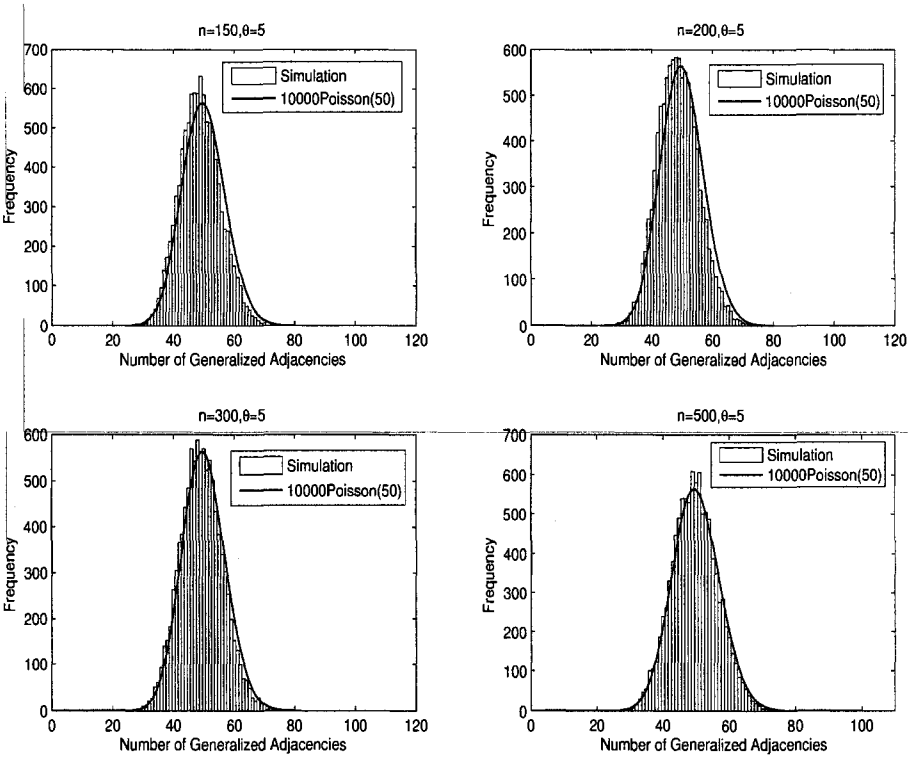



Figure 3.3: Empirical distribution of $n_2$ compared to the related Poisson distribution for $\theta$= 5, $n$= 150, 200, 300 and 500

# Chapter 4

# Testing GA-Clusters

In this Chapter we use the same model in Section 3.2. Given an observed GA-cluster of size $k$ through comparison between two genomes, we may want to know whether the expected number of GA-clusters of size at least $k$ or the probability that at least one GA-cluster can be observed, under the null hypothesis of random gene order, is significantly small.

## 4.1 Clusters of larger size

We use $n_k$ to denote the number of connected components[1] of size $k$ in $E_I \cap E_R$, with no disjointness requirement or restriction against the component being contained in a larger cluster. If a connected component of size $k$ is not contained within any larger connected component, then it is a GA-cluster. In this chapter, we try to find some analytical results for the expectation of $n_k$, rather than the expected number of GA-clusters of size $k$.

We have already studied the distribution of $n_2$ in Chapter 2. We now consider

[1]In graph theory, two vertices $g$ and $h$ are called collected if there is an edge between them, and a connected component is a maximal connected subgraph of $G$.

the expectation of $n_3$. Extending the counting approach we used in the proof of Proposition 3.2.1, we can list all the connected components of size 3 in genome $I$ and calculate the probability it is also in $R$. Adding all the probabilities together, we find

$$\mathbf{E}(n_3) = \frac{\theta^2}{n}(5\theta^2 - 2\theta - 1) + O(\frac{1}{n^2})$$

Similarly, with additional effort, we find that

$$\mathbf{E}(n_4) = \frac{\theta^2}{n^2}(\frac{124}{9}\theta^4 - \frac{95}{6}\theta^3 - \frac{8}{9}\theta^2 + \frac{29}{6}\theta + \frac{1}{9}) + O(\frac{1}{n^3})$$

but the number of different kinds of components of size 5 precludes extending our method, based on listing all possibilities, to $n_5$ and beyond.

Despite the fact that we have only partial results for $n_k$, we can still use standard statistical methods to test for the relatedness of two genomes or the significance of a GA-cluster, especially if $\mathbf{E}(n_4)$ is small.

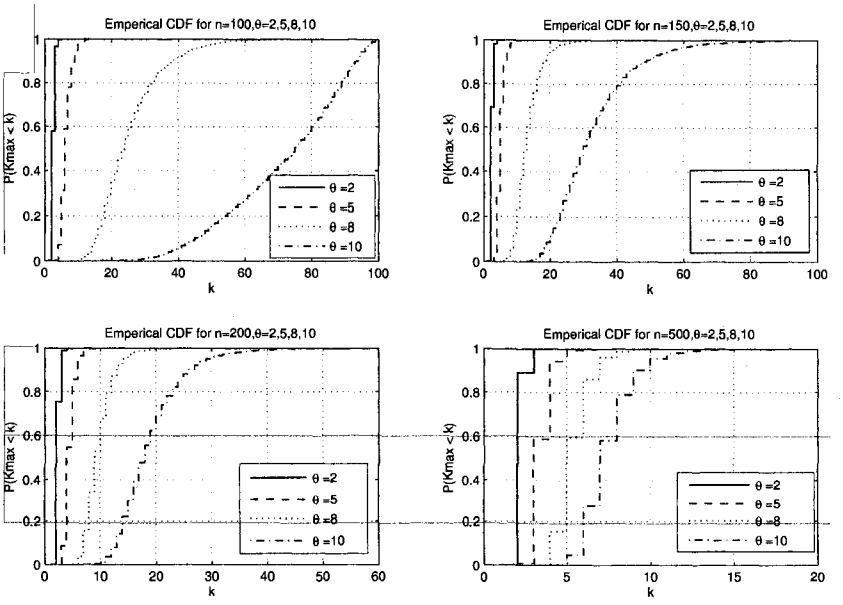



Figure 4.1: Empirical cumulative distribution functions for $k_{\max}$ as a function of $k$, for fixed $n$ and various values of $\theta$

## 4.2 The maximum size of GA-cluster

The ideal and intuitive statistic to use to test the relatedness of genomes or to detect clusters would be the size of the largest cluster $k_{\max}$, and the probability of observing at least one cluster of size at least $k$ is $k = \mathrm{P}(k_{\max} > k)$, which is not suitable for max-gap clusters, as stated in Chapter 2.

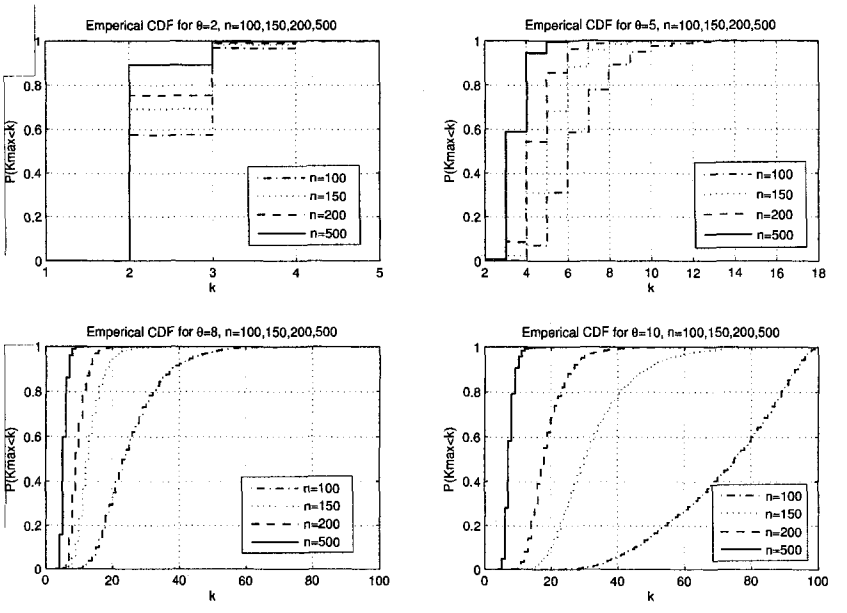



Figure 4.2: Empirical cumulative distribution functions for $k_{\max}$ as a function of $k$, for fixed $\theta$ and various values of $n$

While analytical techniques have not produced useful information about the distribution of $k_{\max}$, it is a straightforward matter to simulate random genomes and estimate this distribution empirically. Figure 4.1 shows the cumulative distribution functions for $k_{\max}$ as a function of $\theta$ for a number of different fixed $n$. This kind of result can be directly used for testing.

Figure 4.1 shows that $\mathrm{P}(k_{\max} < k)$ is an decreasing function of $\theta$ for fixed $n$ and $k$ while it is an increasing function of $n$ given $\theta$ and $k$, shown in Figure 4.2. Of

particular interest is the dramatic change in the structure of the function between $\theta = 8$ and $\theta = 10$, when $n = 100$. Suddenly the mass of the distribution shifts from values around 20 to values around 75. We will investigate this phenomenon in more detail in next section.

## 4.3 Square-root law for parameter selection

As exemplified in Figures 4.3 and 4.4, each based on 10,000 pairs of random genomes, it is remarkable how quickly the distribution changes between $\theta = 9$ and $\theta = 10$ for $n = 100$, and between $\theta = 31$ and $\theta = 33$ for $n = 1000$.

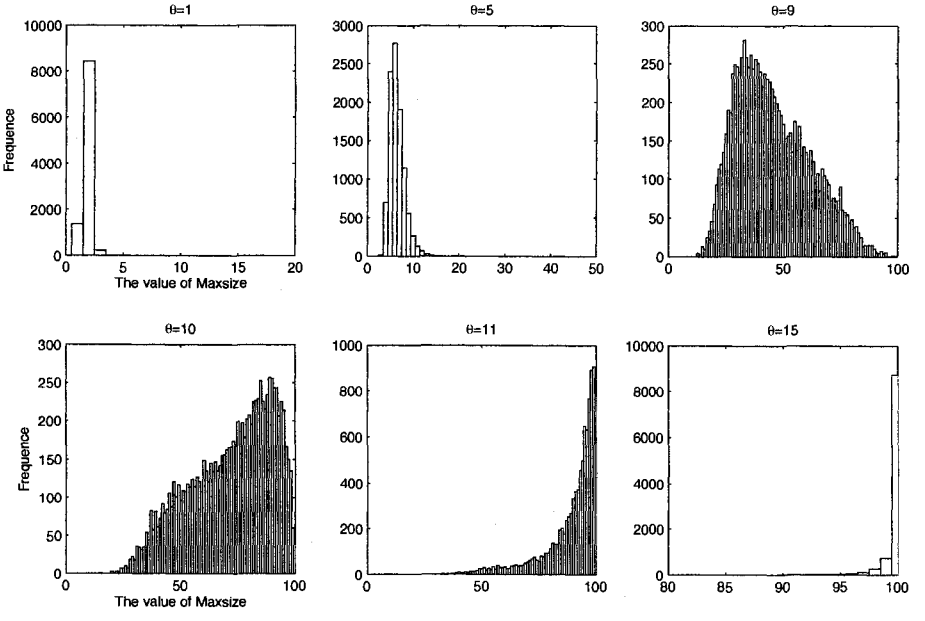



Figure 4.3: Histograms for $k_{\max}$ when n=100

On the basis of 10,000 pairs of random genomes, we determined the change-point $\theta^*$ (a value of $\theta$, after which the average of $k_{max}$ jumps from below $0.5n$ to above $0.5n$ immediately and dramatically) for a range of values of $n$(shown in Table 4.1 ), and in Figure 4.5 plotted these points against $\sqrt{n}$. This suggests that the change-point

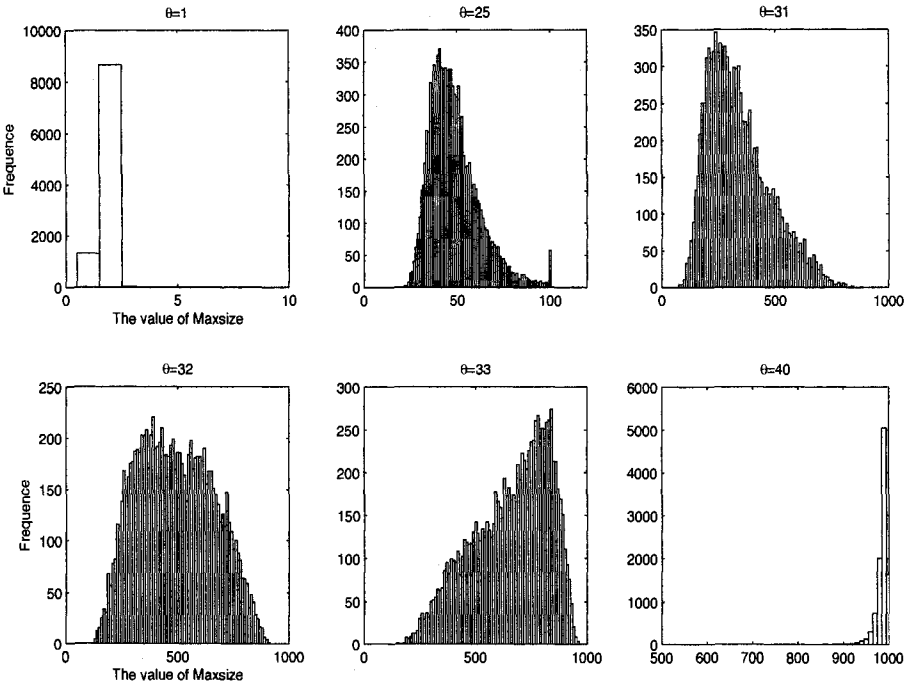



Figure 4.4: Histograms for $k_{\max}$ when n=1000

satisfies $\theta^* = \sqrt{n}$ or some similar relation. To characterize the abruptness of the

Table 4.1: Change-point for different $n$

| The value of $n$ | 50 | 100 | 300 | 500 | 1000 | 3000 | 5000 | 10000 |
|---|---|---|---|---|---|---|---|---|
| Change-point | 6 | 9 | 17 | 22 | 32 | 56 | 73 | 104 |

change around the change-point, we calculated how much of the probability mass falls to the right of $0.5n$, for each value of $\theta$. Figure 4.3 shows that the change behaviour, in proportion to $\sqrt{n}$, tends to a sharp "cut-off" at or near $\theta = \sqrt{n}$.

Testing significance of clusters is not only useful for making right decision but also helpful for selection of parameters for cluster definitions. In practice the value of $\theta$ we choose to characterize a GA-cluster should not be greater than $\sqrt{n}$, based on the knowledge of the cut-off behaviour.

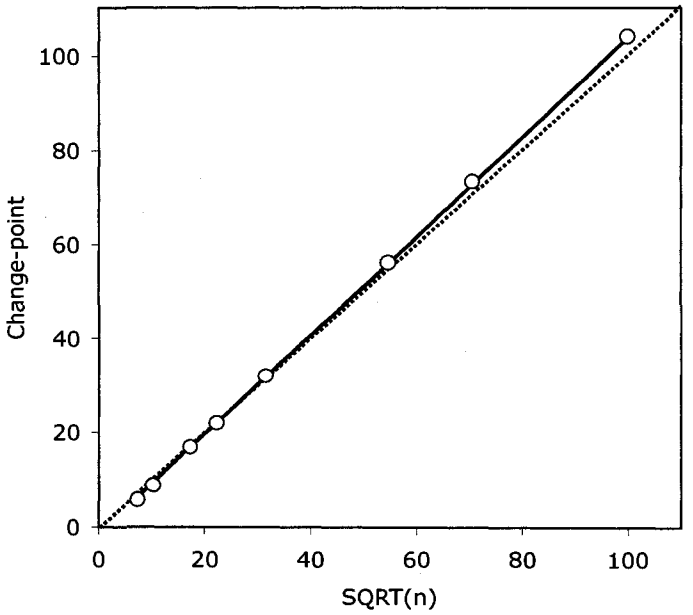



Figure 4.5: Change-point for $k_{\max}$ as a function of $\sqrt{n}$. Dotted diagonal represent exact $\sqrt{n}$.

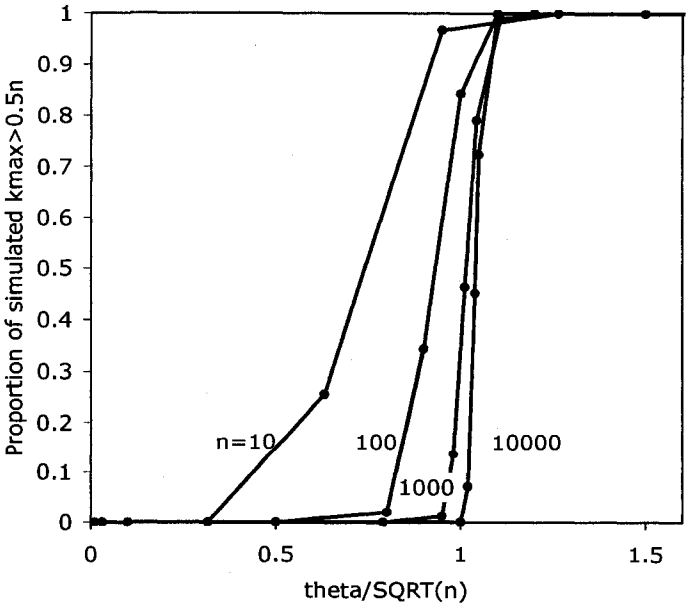



Figure 4.6: Cutoff for maximum size cluster.

# Chapter 5

# Discussions and Future Directions

I have studied a new definition of gene clusters, **GA-clusters**, that allows us to control the emphasis placed on conserved order within a cluster by adjusting the *neighbourhood parameter* $\theta$. It is easy to show that every GA-cluster with parameter $\theta$ satisfies the max-gap criterion with parameter $\theta - 1$, but not vice versa. A simple example is: when $\theta = 1$ ($\theta - 1 = 0$), a GA-cluster reduces to a *conserved segment* conserving exactly the same gene content and order (or reverse order) while a max-gap cluster reduces to a *common interval* which only keep the same gene content.

Besides gene order, most existing cluster definitions characterize clusters according to their size, length and density. Hoberman and Durand[4] also summarized additional "desirable" properties for cluster definitions, such as *nestedness* and *disjointness*. Next I will explore some properties of GA-cluster in more detail.

## 5.1 GA-Cluster Properties

- **Size** is the number of genes constituting a cluster. Like max-gap cluster definition, GA-cluster definition does not constrain the maximum size of the cluster, allowing it to be as large as the whole genome length.

- **Length** is the total number of genes a cluster spanning in a particular genome. GA-cluster definition does not constrain the length of the cluster either. In Figure 3.1, when $\theta = 2$, the size of the cluster $\{2, 4, 5\}$ is 3, and its length is 4 for Genome S and 3 for Genome T.

- **Density**: GA-cluster definition constrain "local density" that the gap between any two adjacent genes in a cluster must be less than the neighbourhood parameter, $\theta$. For a GA-cluster, the *global density*, defined as its size $k$ divided by its length, can be as small as $\frac{k}{k+(k-1)(\theta-1)} = \frac{k}{(k-1)\theta+1}$.

- **Nestedness**: A cluster of size $k$ is *nested* if it contains a "sub-cluster" of size $h$, for $h$=1, 2, ..., $k-1$. Max-gap cluster definition does not satisfy this property[4, 2] while GA-cluser definition does. Suppose a GA-cluster of size $k$ is determined by a set of $m$ different common edges $(a_1, b_1)$, $(a_2, b_2)$, ..., $(a_m, b_m)$, where "( )" denotes *edge*. By definition of GA-cluster, for any $i$=1, 2,..., $m$, $(a_i, b_i)$ must share at least one end-point with at least one of other edges. We can delete some edges (if required), to get a set of "new" common edges satisfying that at least one of them, denoted as $J$, share exact one endpoint with others , and they still produce the same GA-cluster of size $k$. Delete the endpoint unique to $J$, then we get a sub-cluster of size of $k-1$. Repeat this step, we can get sub-clusters of any size from 2 to $k-2$.

- **Disjointness**: Two clusters are disjoint if they do not share common genes. Obviously, all GA-clusters and max-gap clusters identified through whole genome comparison are disjoint, without taking gene families into account. $r$-window cluster definition dose not have this property. For example, in the **example** given in Section 1.2.1, when $r = 5$, $k = 3$, the $r$-window clusters $\{1, 3, 4\}$ and $\{3, 4, 5\}$ are not disjoint.

- **Isolation**: For any genome under whole genome comparison, the clusters are *isolated* if the maximum gap within clusters is always less than the minimum gap between any two adjacent clusters. This property is satisfied for max-gap clusters, but not for GA-clusters. For example, in Figure 3.1, when $\theta = 3$, in Genome S, the maximum gap within clusters is 1 (*e.g.* there is one gene intervening between gene 2 and gene 4), but there is no gene intervening between GA-clusters {1, 2, 4, 5} and {6, 7, 8, 9, 10, 11, 12, 13}.

- **Symmetry**: For some cluster definitions, the clusters identified through whole genome comparison can be different depending which genome is selected as a "reference" genome[4]. For $r$-window, max-gap and GA-cluster definitions, once two genomes are given, the set of clusters is uniquely determined.

## 5.2 Future Directions in testing gene clusters

The advent of the new biological area, *Comparative Genomics* has presented a lot of statistical problems such as gene clustering tests discussed here. How to design effective and computationally tractable statistical tests incorporating genomic information, such as gene content, order and orientation, is still an open problem. Even only taking gene order [17] or gene content [9, 3, 5], to construct statistical tests is very difficult. The *generating function* seems a powerful tool to deal with some enumeration problem to derive distributions of tentative test statistics [19, 9].

In this thesis, based on a very simplified genome model, we have begun the investigation of statistics related to GA-clusters. The behaviour of the number of clusters for a given $n$ and $\theta$ seems amenable to analytical investigation, as we have demonstrated with a number of new results. The distribution of $k_{\max}$, a tool for suggesting the biologically most interesting clusters, does not seem as accessible, but

is easily simulated. Knowledge of the cut-off behaviour serves to delimit the region for meaningful tests to $\theta$ suitably less than $\sqrt{n}$. In the future we should explore the distribution of $k_{\max}$ further; we also should construct statistical tests based on more realistic and complicated genome models, taking multi-chromosome, circular chromosome and gene duplication into account.

# Bibliography


[1] Billingsley, P. (1995), *Probability and Measure*, 3rd edition, John Wiley and Sons.

[2] Bergeron, A., Corteel, S. and Raffinot, M.(2002), The algorithmic of gene teams.In WABI 2002, Gusfield, D., Guigo, R., Eds., *Lecture Notes in Computer Science*, vol.2452, pp. 464-476.

[3] Durand, D. and Sankoff, D.(2003),Tests for gene clustering, *Journal of Computational Biology* , **10**, pp. 453-482.

[4] Hoberman, R. and Durand, A. (2005), The Incompatible Desiderata of Gene Cluster Properties, "Proceedings of the 3rd RECOMB Workshop on Comparative Genomics", Mclysaght and Huson, eds., *Lecture Notes in Bioinformatics*, pp. 73-87

[5] Hoberman, R.,Sankoff,D. and Durand, D. (2005), The Statistical Analysis of Spatially Clustered Genes under the Maximum Gap Criterion, *Journal of Computational Biology* , **12**, No.8, pp. 1081-1100.

[6] Kaplansky, I. (1945), The Asymptotic Distribution of Runs of consecutive elements, *The Annals of Mathematical Statistics*, **16**, No.2, pp. 200-203.

[7] Nadeau, J. and Taylor, B. A. (1984), Lengths of chromosomal segments conserved in comparative maps, *Mamm Genome*, **9**, No.6, pp. 491-495.

[8] Nadeau, J. and Sankoff, D. (1998), Counting on comparative maps, *Trends in genetics:TIG*, **14**, No.12, pp. 495-501.

[9] Narayanan Raghupathy and Durand, D. (2007), Individual Gene Cluster Statistics in Noisy Maps, In *Comparative Genomics, RECOMB International Workshop (2005)*, McLysaght,A., et al. Eds., *Lecture Notes in Computer Science*, vol. 3678, Springer-Verlag, pp. 106-120.

[10] Narayanan Raghupathy, Hoberman, R. and Durand, A. (2007), Two Plus Two Does not Equal Three: Statistical Tests for Multiple Genome Comparison, *Asiapacific Bioinformatics Conference (2007)*, pp.215-225.

[11] Robbins, D. P. (1980),The probability that neighbors remain neighbors after random rearrangements, *The Annals of Mathematical Statistics*, **87**, No.2, pp. 122-124.

[12] Sankoff, D. (1999), Comparative mapping and genome rearrangement, In *From Jay Lush to Genomics: Visions for Animal Breeding and Genetics*, Dekkers, J.C.M ., Lamont, S.J.and Rothschild, M.F., Eds., pp. 124-134.

[13] Sankoff,D. (2002), Short inversions and conserved gene clusters, *Bioinformatics* ,**18**, pp. 1305-1308.

[14] Sankoff,D. (2003), Rearrangements and chromosomal evolution, *Current opinion in genetics & development*, **13**, No.6, pp. 583-587.

[15] Sankoff,D. and El-Mabrouk, N. (2002), Genome Rearrangement, In *Current Topics in Computational Biology (2002)*, Jiang,T., Smith,T., Xu,Y. and Zhang, M., Eds., MIT Press, pp. 135-155.

[16] Sankoff,D., Ferretti, V. and Nadeau, J.H. (1997), Conserved segment identification, *Journal of Computational Biology* , **4**, pp. 559-565.

[17] Sankoff,D. and Haque, L. (2005), Power boost for cluster tests, In *Comparative Genomics, RECOMB International Workshop(2005)*, McLysaght,A.and Huson, D.H., Eds., *Lecture Notes in Computer Science*, vol. 3678, Springer-Verlag, pp. 121-130.

[18] Wolfowitz,J. (1944), Note on runs of consecutive elements, *Annals of Mathematical Statistics*, **15**, pp. 97-98.

[19] Xu,W., Alain,B. and Sankoff, D. (2008), Poisson adjacency distributions in genome comparison: multichromosomal,circular,signed and unsigned cases. To appear in *Bioinformatics* 24 (18).

[20] Xu, X. and Sankoff, D. (2008), Tests for Gene Clusters Satisfying the Generalized Adjacency Criterion, A. L. C. Bazzan, M. Craven, and N. F. Martins Eds., BSB 2008, LNBI 5167, Springer-Verlag, pp. 152-160.

[21] Zhu, Q., Adam, Z., Choi,V. and Sankoff,D. (2008), Generalized gene adjacencies graph bandwidth and clusters in yeast evolution, ISBRA 2008.