FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Zhenyu Yang
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Mathematics)
GRADE / DEGREE

Department of Mathematics and Statistics
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Natural Parameter Values for Generalized Gene Adjacency
TITRE DE LA THÈSE / TITLE OF THESIS

David Sankoff
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

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

Richard Blue

Michael Newman

Patricia Evans (University of New Brunswick)

Yiqiang Zhao

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

NATURAL PARAMETER VALUES FOR GENERALIZED GENE ADJACENCY

Zhenyu Yang

Thesis Submitted to the Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
for the degree of
Doctor of Philosophy in Mathematics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

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

Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-65553-5
Our file Notre référence
ISBN: 978-0-494-65553-5

Abstract

As genomes of related species diverge through rearrangement mutations, groups of genes once tightly clustered on a chromosome will tend to disperse to remote locations on this chromosome or even onto other chromosomes. Even if most rearrangements are local, e.g., small inversions or transpositions, after a long enough period of time their chromosomal locations may reflect little or none of their original proximity. Given the gene orders in two modern genomes, then, it may be difficult to decide if some set of genes are close enough in both genomes to infer some ancestral proximity or some functional relationship. There are a number of formal criteria for gene clustering in two or more organisms, giving rise to cluster detection algorithms and statistical tests for the significance of clusters. These methods all depend on one or more arbitrary parameters as well as n, the number of genes in common in the two genomes. The various parameters control, in different ways, the proximity of the genes on the chromosome in order to be considered a cluster. Change the parameters and the number of clusters may change, as may the content of each cluster. We explore a two-parameter class of gene proximity criteria, and find natural values for these parameters. One has to do with the parameter value where the expected information contained in two genomes about each other is maximized. The other has to do with parameter values beyond which all genes are clustered. We analyse these using combinatorial and probabilistic arguments as well as simulations.

Acknowledgements

I would like to express my warmest thanks to my Ph.D. superviser, Dr. David Sankoff for leading me to enter this new amazing research area, comparative genomics. I very much appreciate for his guidance, encouragement and support over the years of my Ph.D study. Moreover, I would like to extend my appreciation to all the members of the Sankoff lab: Wei Xu, Ximing Xu, Chunfang Zheng, Robert Warren, Qian Zhu, Zaky Adam, Shengang Li, Adriana Muñoz and Ghada Badr for their collaboration, friendship and helpful discussions. Working with them make my life more colourful and exciting. Also, I would like to thank to my wife Zuojing Li and my family from the bottom of my heart for their wholehearted love, trust, support, encouragement and patience which are invaluable to me.

Dedication

To my dear wife Zuojing Li for her endless love and devoted support in my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Biological Background

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, **Comparative Genomics**, based on the macrostructure of entire genomes, rather than on the traditional comparison of a single gene or protein sequence in different organisms, to study the relationship of genome structure and function across different biological species.

One of the fundamental tasks in the comparative genomics is the identification of homologous genes in related genomes, pairs of genes, one in each genome, that are descended from a single gene in the ancestral genome, either through speciation or duplication. This is a prerequisite to many tasks in the comparative genomics.

Homologous genes are very useful tools in biology and bioinformatics. First, they help us to transfer knowledge of one genome to make inferences about another. Although increasing numbers of genome sequences are becoming available, most experimental studies are still carried out on a small set of model organisms. By determining how genes and genomic regions of poorly-studied organisms correspond to those of well-studied organisms, knowledge about one species can improve understanding of others. In particular, although humans are among the most well-studied organisms, many types of experimentation cannot be carried out on humans. Thus, transfer of knowledge from model organisms is essential for understanding human biological processes, and developing new disease treatments.

Moreover, homologous genes can help elucidate protein function and regulation. In bacteria, functionally related genes tend to be spatially clustered on the chromosome. Comparisons of gene order can identify sets of genes whose spatial arrangement is conserved, and that are more likely to be functionally related. Unlike sequence or structural homology methods, which primarily provide insight on the biochemical function of a protein, spatial clustering offers evidence of associations between proteins, such as physical interactions, or participation in the same pathway. These types of associations help identify the physiological or cellular role of a protein, complementing information derived from sequence comparisons. In bacteria, conserved gene order and content have been used for prediction of operons [4, 11, 25, 36, 38], horizontal transfers [19], and more generally to investigate the relationship between spatial organization and functional selection [17, 18, 22, 30, 32, 33].

Finally, we can construct a phylogeny tree of these genomes to represent the evolution history of species and estimate ancestor genomes based on the model species by comparing homologous genes in different genomes. 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 disrupt gene order and syntenic[1] structure [27].

However, as genomes of related species diverge through rearrangement mutations, large and small, groups of genes once tightly clustered on a chromosome will tend to disperse to remote locations on this chromosome or even on to other chromosomes. After a long enough period of time their chromosomal locations may reflect little or none of the original proximity. Given the gene orders in two modern genomes, then, it may be difficult to decide if some set of genes are close enough in both genomes to infer some ancestral proximity or some functional relationship.

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 genomes being compared. In practice it is useful to relax this stringent definition to some extent to detect the evolutionary signal in regions, that are "almost conserved", to avoid unstable estimates of the number of segments if these may be as small as one or two genes [7], and to diminish the effect of experimental error and other noise. The experimental errors can be attributed to gross mistakes in chromosomal assignment of genes, quantitative errors in gene positions as well as the errors occurring when integrating gene locations from different sources [20, 28]. So a less strict concept, *gene clusters*, is introduced, which are pairs of regions with similar, but not identical, gene content and gene order.

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

1.2 The Steps of Cluster Identification

To identify clusters in closely related genomes, map-based approaches are often used in which clusters are detected based on the locations of genomic *markers*, rather than direct comparison of the primary sequence. A marker-based approach to the identification of homologous segments typically involves the following four steps.

1.2.1 Gene and Chromosome

Genomic comparison using map-based approaches requires a set of markers, sequences with unique locations in the genome, as input data. When the markers are biological genes and the input data are genomic sequences, the marker identification problem reduces to the problem of gene finding. Many other types of markers can also be used. e.g., [23, 24, 29].

In this thesis, we assume that a genome consists of some linear unbroken chromosomes, the orientation of each gene is ignored and the distance between two genes is calculated using the number of genes between them. For instance, a genome $S = (g_1, g_2, \cdots, g_i)(g_{i+1}, \cdots, g_j)(g_{j+1}, \cdots, g_n)$, where g_k $(k = 1, 2, \cdots, n)$ is a gene of the genome S and braces represent the ends of chromosomes. In this example, the genome S has three chromosomes and n genes. This model assumes that genes do not overlap, and disregards the physical distance between genes. So, we do not need to deal with the variation in gene density that can lead to gene-rich and gene-poor regions of chromosomes when we compare two or more genomes.

1.2.2 Homology Detection

The most general definition of homology is that it designates a relationship of common descent between any entities, without further specification of the evolutionary scenario. Accordingly, entities related by homology, in particular, genes, are called homologs. In other words, two genes are homologous if they arose from a single gene in an ancestral genome. Orthologs and paralogs are two fundamentally different types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication. Two genes in different species are orthologous if they come from a single gene in the Most Recent Common Ancestor(MRCA) of the two species, and paralogous if they arose through a duplication event that preceded the divergence of the species[12, 13].

Figure 1.1 shows a hypothetical phylogenetic tree of a gene family. By gene duplication, the gene in MRCA becomes two gene copies, α and β. After speciation events occur, we can find α and β in species frog, chick and mouse. Then one gene copy from the gene in MRCA in one species is orthologous to the same gene copy in different species and paralogous to the different gene copy in different species. For instance, α gene in mouse is orthologous to α gene in chick and is paralogous to β in mouse, chick and frog. All six genes are homologous, as they arose from a single ancestral gene. All of them form a gene family.

1.2.3 Cluster Detection

Declarative and constructive gene cluster definitions

We can define clusters either by specifying precise characteristics that allow one to identify a cluster, or by giving a procedure for constructing clusters, without mentioning cluster properties. The former definition is called *declarative definition* and the latter one, *constructive definition.* Although a constructive definition makes it clear how to find clusters, it does not provide what the resulting clusters will look like. Unless some properties can be abstracted from the constructive idea, it is dif-

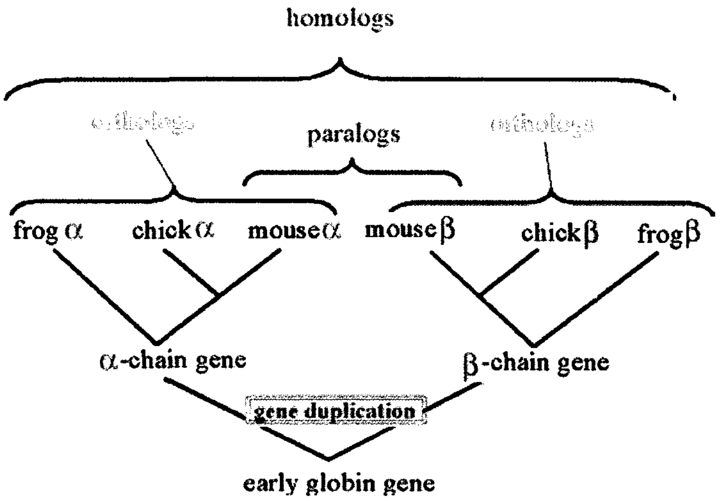


Figure 1.1: The relation of homologs, orthologs and paralogs (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html)

ficult to do analysis on the resulting cluster. A declarative definition, on the other hand, facilitates the analysis; however, it requires an additional search procedure to find clusters that satisfy the formal definition. No matter what cluster definition is used, it is necessary to verify that the constructive and declarative definitions are equivalent. Recently there has been a movement to formalize cluster definitions, and to develop precisely formulated search algorithms, so that correctness and efficiency of these algorithms can both be analyzed.

Cluster definitions

There are a number of formal definitions and criteria for gene cluster detection in two or more organisms, giving rise to cluster detection algorithms and statistical tests for the significance of clusters, e.g., [7, 16, 37]. Here we introduce five commonly applied criteria for gene clusters: *conserved segment, common interval, r-window cluster, max-gap cluster* (also referred as "gene teams" [1]) and *generalized adjacency gene* *cluster (GA cluster)* .

- The most conservative definition is *conserved segment*, a set of genes with the same gene content, order, and even orientation (the transcription direction associated with each gene) in two or more compared genomes [3, 20, 21, 31]. However, this stringent definition will exclude many regions that did indeed descend from a single ancestral region but have undergone a series of small rearrangements, insertions or deletions.

- A *common interval* defines a set of genes occurring contiguously in each of the genomes compared, ignoring gene order, but without allowing gene insertions and deletions. A number of researchers have developed search algorithms to efficiently find common intervals in genomic data [5, 14, 34]. However, this definition is still generally too strict, since gene duplication and loss are common when comparing distantly related genomes, and a single gene insertion or deletion in one genome may destroy a common interval.

- An *r-window cluster* is defined as a pair of windows in two genomes, each containing r genes, in which at least k genes are shared [7]. 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 gene cluster with 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 interval gene clusters. A max-gap cluster is *maximal* if it is not contained within any larger max-gap cluster.

- A *generalized adjacency gene cluster (GA cluster)* is a component of a graph, called *generalized adjacency graph.* Genes are represented as vertices in this graph and edges are added between vertices if the number of genes between the two corresponding genes are less than a given parameter θ in both genomes. Clearly, if $\theta = 1$, GA clusters reduce to conserved segments.

r-window clusters, *max-gap clusters* and *GA clusters* will be discussed further in Chapter 2.

The following is an **example** to illustrate the four different cluster criteria.

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

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

where the integers represent homologous gene pairs and the stars indicate genes with no homolog (or a remote 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}.

5. GA clusters

   (a) when θ=3: {1, 2, 3, 4, 5, 6, 7, 8};

   (b) when θ=2: {3, 4, 5} and {6, 7, 8};

The criteria r-window cluster, max-gap cluster and GA cluster all depend on one or two arbitrary parameters as well as n, the number of genes in common in the two genomes. The various parameters control, in different ways, the proximity of the genes on the chromosome in order to be considered a cluster.

1.2.4 Cluster Significance Test

In the previous section, we introduced the basic steps for identifying gene clusters, from determining the position of genes in the genome to designing cluster definitions and criteria. After these steps, we can create the algorithm to find gene clusters. However, it is not possible to estimate a gene clustering algorithm's accuracy, sensitivity or specificity, because we do not know what the true evolutionary relationships are. Over time, processes of genome mutation and rearrangement cause the relationships among formerly adjacent genes to become more and more similar to the statistical background. Thus, to evaluate putative gene clusters, it is imperative to test and reject the hypothesis that the observed similarities could have occurred by chance. So cluster significance statistical tests are necessary to evaluate the accurate identification of ancient segmental homologies.

Moreover, statistical models also enable the principled selection of search parameters. Cluster definitions and criteria are based on one or more parameters which are fixed and defined by users. For example, the gap parameter g in max-gap cluster and the parameters r and k which represent the window size and the smallest number of gene shared in the r-window cluster definition. All of these parameters are user-specified. If parameters selected are too strict, many significant clusters will not be detected. On the other hand, very liberal parameter values may lead to biologically meaningless clusters being detected. A statistical model can be used to determine the range of parameter values within which a cluster will still be significant.

1.3 Cluster Properties and Searching Strategies

1.3.1 Cluster Properties

As we have mentioned, there are many different gene clustering criteria. It is difficult for us to compare the different criteria without enumerating desirable cluster characteristics. Furthermore, the significance of gene clusters we obtain using the clustering criteria also depends on the cluster properties. Hoberman and Durand [15] listed some important cluster properties. In this section, I will present some of the most commonly used ones.

- **Size:** Almost all approaches to gene clustering evaluation consider gene cluster size, *i.e.* the number of marked genes contained within the cluster. Here the "marked" gene is one in some predetermined subset of genes.

- **Length:** The length of a gene cluster is the total number of marked and unmarked genes contained within it.

- **Density:** The density of a gene cluster is the cluster size divided by the cluster length. The majority of existing methods attempt to find regions that are densely populated with marked genes.

- **Order:** Some gene clustering criteria such as *conserved segments* consider the order of gene in the cluster, *i.e.* the genes of a gene cluster in one genome must be in identical or opposite order in the other genome. Other criteria, however, release this constrain.

- **Orientation:** Conserved spatial organization in bacterial genomes often points to functional associations between genes. In particular, clusters of genes in close proximity, with the same orientation, often indication operons.

1.3.2 Search Strategies of Gene Clustering

Researchers are sometimes interested in the genes in a particular region of a genome and search one or more other genomes for similar regions. Or, gene clusters may be found in "fishing expeditions" for clusters in the whole genome comparisons. So the significance of a gene cluster depends not only on the characteristics of the cluster, but also on the way it was found. The larger the search space, the less significant the cluster. However, most statistical tests do not consider the search space size, and most people represent cluster results without providing the details of the search procedure during which it was detected. Durand and Sankoff [7] characterized the following three most common search strategies:

1. **Reference set**: Given a set of genes of interest, the goal is to identify subsets of these genes that are located in close proximity in the genome. For example, the genes of interest share a particular functional or regulatory property. Using this gene set as reference set, we look for clusters of reference set genes in a genome. In this case, the search space is the entire genome.

2. **Window sampling**: Given two chromosomal regions, the goal is to determine whether the regions share a significant number of homologs, in order to obtain

evidence that they descended from a single region in an ancestral genome. In many cases, these windows are selected because they contain a pair of known homologs of particular interest. This search scenario may be used, for example, to determine whether a particular set of paralogs were duplicated through a large scale event, or to assess whether the gene order around a pair of orthologs has been conserved. In window sampling, the search space is confined to the two regions of interest.

3. **Whole genome comparison**: Given two genomes, the goal is to identify all clusters of genes that appear in both genomes. In this case, appropriate tests must be used to avoid exaggerating cluster significance due to a much larger search space.

1.4 Thesis Overview

As I mentioned, there are a number of formal criteria for gene clustering in two or more organisms, giving rise to cluster detection algorithms and statistical tests for the significance of clusters, e.g., [7, 16, 37]. These methods all depend, however, on one or two arbitrary parameters as well as n, the number of genes in common in the two genomes. The various parameters control, in different ways, the proximity of the genes on the chromosome in order to be considered a cluster. Change the parameters and the number of clusters may change, as may the content of each cluster.

In this thesis, we define a two-parameter class of gene proximity criteria, where two genes are said to be one-way (i, j)-adjacent if they are separated by $i - 1$ genes on a chromosome in one of the genomes and $j - 1$ genes in the other or either-way (i, j)-adjacent if they are separated by $i - 1$ genes on a chromosome in either one of the genomes and $j - 1$ genes in the other. And also we define a (θ, ψ)-adjacency cluster in terms of a graph. These definitions are inspired by previous work [37, 39] on (in our present terminology) (θ, θ) clusters.

As with other clustering criteria, the quantities θ and ψ will at first seem to be arbitrary parameters in our definition of a cluster. However, research on statistical properties of the generalized adjacency gene clustering criteria will enable us to remove some of this arbitrariness, by finding "natural values" for θ and ψ as a function of n, the total number of genes in the genomes.

In Chapter 2, I represent in some detail techniques of two previous gene clustering models, the r-window model and the max-gap model and some results of research under these models. Then I introduce the generalized adjacency model. I also represent some related work by my colleagues [37, 39].

In Chapter 3, we develop in more detail the generalized adjacency model defined by Zhu *et al.* [39] and introduce a class of two-parameter generalized adjacency gene clustering models. We pay particular attention to $(1, j)$ adjacencies and $(1, \theta)$ clusters since these may be of interest in genome similarity studies in that they depend upon how much genes that are strictly adjacent in either of the genomes being compared are separated in the other genome. We then move to the more generalized criteria, (i, j) adjacencies and (θ, ψ) clusters. We start by defining a wide class of similarities (or equivalently, distances) between two genomes in terms of a class of weights on the (i, j)-adjacencies, namely any system of fixed-sum non-negative weights ω non-increasing in i and j, representing decreasing weight with increasing separation of the genes on the chromosome. In any pair of genomes, in order to maximize the sum of the weights, we prove a theorem showing that the solution reduces to a uniform weight on gene separations up to certain values of θ and ψ, and zero weight on larger separations. By using simulations to investigate the expected value of the optimal θ and ψ under a uniform measure on the space of genomes, it is found that the parameters are slowly-growing functions of n, the length of the genomes and this behaviour is closely modeled by a theorem in the theory of record times of series of random i.i.d. variables.

In Chapter 4, we study the second set of "natural" parameter values determined by a kind of percolation behaviour of the (θ, ψ) generalized adjacency gene clusters. Beyond certain values of these parameters, it is no longer surprising, revealing or significant to find large groups of genes clustering, because all clusters rapidly coalesce together even in pairs of random genomes. So tests of significance are no longer meaningful.

First, we calculate the expected number of (i, j)-adjacent gene pairs, prove that this number converges to a Poisson variable, and then find an expression for the expected number of (i, j) adjacencies with $i \leq \theta$ and $j \leq \psi$. Moreover, we extend this to the expected number of three and four genes which are in the same generalized adjacent gene cluster. At last, we discuss the percolation threshold which serves as an upper bound on the meaningful choices θ and ψ if we are willing to disregard the criteria of Chapter 3, and prefer to search for clusters more widely dispersed on chromosomes. Most analytical results on percolation pertain to completely random (Erdős-Rényi) graphs. The graphs associated with (θ, ψ) generalized adjacency gene clusters manifest delayed percolation, so the use of Erdős-Rényi percolation values would be a "safe" but conservative way of avoiding dangerously high values of the parameters. We show how to translate known results on Erdős-Rényi percolation back to generalized adjacency clusters. we also introduce random bandwidth-limited graphs and use simulations to compare the delays of generalized adjacency and bandwidth-limited percolation with respect to Erdős-Rényi percolation in order to understand what structural properties of generalized adjacency are responsible for the delay.

In Chapter 5, we lift the constraint that the genomes must have the same gene content and look for some properties of generalized adjacency gene clusters which can be used in more real genome comparison cases.

In Chapter 6, we extend the two parameter generalized gene clustering criterion from two-genome comparison to multiple genome comparison and discuss the interpretation of generalized adjacency of two genes in multiple genomes.

Chapter 7 represents the conclusion of my thesis and some open questions.

Chapter 2

Gene Clustering Models and Their Statistical Tests

Various clustering definitions and algorithms have been widely used in empirical studies, however, very few formal statistical models have been developed to test the significance of clusters. Most approaches have simply estimated the distributions of test statistics based on randomization. In contrast, Durand and Sankoff [7] undertook a mathematical approach constructing statistical tests for *r-window* clusters under different scenarios; Hoberman *et al.* [16] represented analytical statistical models for clusters satisfying max-gap criterion. Zhu *et al.* [39] represented a new parameterized definition of gene clusters, *generalized adjacency gene cluster (GA cluster)* and Xu *et al.* [37] initiated research into the statistical properties of this model. In this chapter, I will discuss these three models. Genomes are 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 genomes are linear, genes do not overlap and the distance between two genes is simply the number of genes between them.

2.1 r-Window Gene Clustering Model

In 2003, Durand and Sankoff [7] introduced a new definition of a gene cluster, r-window gene cluster, and presented a comprehensive analysis on statistical tests suitable for this cluster definition. They begin with a simple model under a reference region scenario.

For a given genome G with n genes and m genes prespecified, they assume every prespecified gene has exactly one occurrence in G. The 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 chosen by chance, *i.e.* the probability of choosing any m genes in n genes is the same. Then, the probability of observing that the m genes span at most r slots in G (equation(2) in [7]) is

$$q(n,m,r) = \frac{(n-r)\binom{r-1}{m-1} + \binom{r}{m}}{\binom{n}{m}} \tag{2.1.1}$$

So if this quantity is less than a given significance level α, then the null hypothesis is rejected, and the observed r-window cluster is significant, which suggests or confirms that genes in the cluster share some evolutional or functional properties.

2.1.1 Gene Family Model

In real genomes, where genes may occur in two or more almost identical copies, Durand and Sankoff [7] also propose a more realistic and complicated model, the *Gene Family Model.*

They assume that homology relationships have already been determined and the genes in a genome can be partitioned into non-intersecting gene families, sets of genes with similar sequences. Every gene in a gene family is homologous to all the other members in the same family while genes in different families cannot be homologous.

The size of a gene family, represented by ϕ_{ij}, is defined as the number of genes in the gene family, say f_j, in the genome G_i. Let $\mathcal{F} = \{f_j\}$ be the set of all gene families in genomes under consideration and $n_f = |\mathcal{F}|$ which is the total number of gene families. Denote by n_i be the number of genes in the genome G_i.

Based on the gene family model, Durand and Sankoff [7] construct statistical tests against null hypotheses of random gene order, taking incomplete clusters, multiple genome comparison and genome self-comparison into account. However, the treatment presented is mainly of theoretical interest and the expressions for calculating 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 an informal test.

2.1.2 Window Sampling

r-window clusters can be obtained by window sampling. Here I represent only the window sampling model for two genomes.

1. **Without gene families**

   For two genomes G_1 and G_2 containing the same set of n genes, given a pair of windows W_1 and W_2 of length r, 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 is :

$$P(n,r,m) = \sum_{i=m}^{r} \frac{\binom{r}{i}\binom{n-r}{r-i}}{\binom{n}{r}} \tag{2.1.2}$$

   which is equation (22) in [7].

2. **With gene families**

   Under the gene family model, an r-window cluster is redefined as a pair of windows, each containing r genes, in which at least k gene families are shared. Durand and Sankoff [27] studied 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 with length r, W_1 and W_2, are selected from G_1 and G_2 respectively. They obtained an expression of the probability that W_1 and W_2 share at least m genes, but it is computationally intractable. Using a generating function method, Raghupathy and Durand [26] provided a computationally tractable expression of this probability by constraining all ϕ_{ij} (as defined in Section 2.1.1) to take on the same value, ϕ. The generating function is

$$Q(m) = \sum_{k=m}^{r} [\binom{n_f}{k} p_1(k) \sum_{l=m}^{k} \binom{k}{l} p_2(l)] \tag{2.1.3}$$

where

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

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

2.2 Max-Gap Gene Clustering Model

In the rest of this thesis I use the term *max-gap* clusters as shorthand for *maximal* max-gap clusters. Hoberman, *et al.* [16] developed 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, max-gap clusters cannot be identified by **window sampling** [26].

2.2.1 Reference Region

In this scenario, m genes are prespecified (or marked), each of which has exact one homolog in the genome of n genes, the motivation is to test the significance of a cluster containing all or part of the m specified genes, identified through searching the whole genome. In this test, the null hypothesis is that the m genes are distributed in the genome randomly. Hoberman, *et al.* [16] represented two test statistics for the difference cases in the reference region scenario: the *complete cluster* case and the *incomplete cluster* case.

In the complete cluster case, they calculated what the probability of observing that m marked genes form a max-gap cluster with maximum gap less than or equal to g under the null hypothesis mentioned above. The test statistic is (equation(2) in [16]):

$$P(n,m,g) = \frac{1}{\binom{n}{m}} \begin{cases} (n - w_{mg} + 1)(g+1)^{m-1} + (\frac{w_{mg}-m}{2})(g+1)^{m-1}, & w_{mg} \leq n+1 \\ \\ d_0(m,g,n), & \text{otherwise.} \end{cases} \tag{2.2.1}$$

where $n - w_{mg} + 1 = m + (m-1)g$ is the number of ways of placing the first marked gene, $(g+1)^{m-1}$ is the number of ways of placing the remaining marked genes (or equivalently, the number of ways of choosing $m-1$ gaps each between 0 and g), and the last term is the number of ways of constructing a max-gap cluster within the last $w-1$ genes in the genome. Furthermore,

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

Unlike the complete cluster case, in the incomplete cluster case with fixed maximum gap value g , the m marked genes can form different clusters in the same genome, and the size of the largest cluster is used as test statistic. Hoberman, *et al.* used dynamic programming to count those permutations which do not contain a cluster of size h or larger and subtract to obtain the probability of observing at least one incomplete cluster.

i.e.

$$Q(n, m, h, g) = 1 - \frac{\eta[n, m, g+1, 0]}{\binom{n}{m}} \tag{2.2.2}$$

The algorithm is defined by the following recursion relation (equation(4) in [16]):

$$\eta[n, m, j, c] = \begin{cases} 0 & \text{if } c = h \text{ or } n < m \\ 1 & \text{else if } m = 0 \\ \eta[n-1, m, j+1, c] + \eta[n-1, m-1, 0, c+1] & \text{else if } j \leq g \\ \eta[n-1, m, j+1, c] + \eta[n-1, m-1, 0, 1] & \text{otherwise} \end{cases} \tag{2.2.3}$$

Based on Equation 2.2.1, the probability of observing a complete cluster as a function of m for $n = 1,000$ is shown in Figure 2.1 (Figure 2. in [16]). The probability of observing a complete cluster is an increasing function of g, however, it does not increase monotonically with m, as expected. For example, when m=n, a complete cluster for any value of g can always be observed. 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 the max-gap cluster definition without constraints on the length or order, larger clusters do not always imply greater significance, which contradicts "a widespread belief that cluster significance grows with the number of homologs in the cluster" [16].

2.2.2 Whole Genome Comparison

In this scenario, Hoberman, *et al.* only consider pair-wise comparison. Given two genomes, G_1 and G_2, each containing n genes only m of which are in common, the

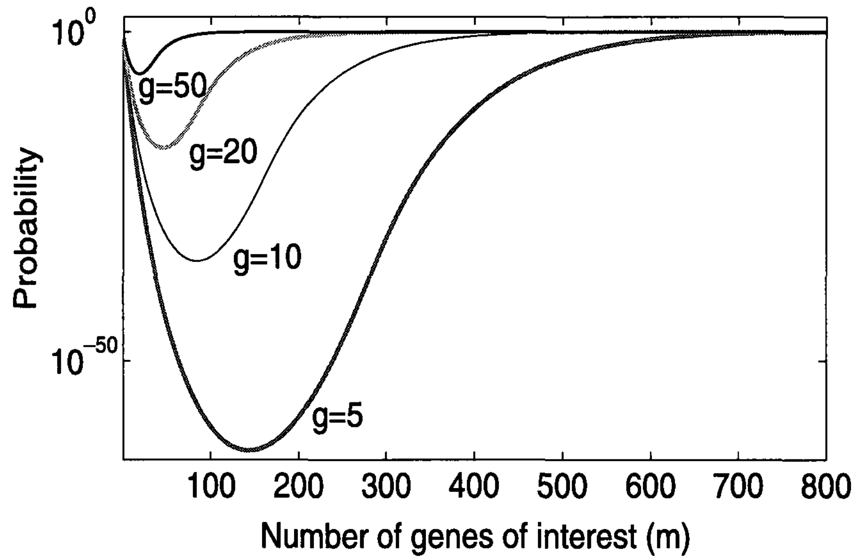


Figure 2.1: (Figure 2. in [16])Probability of a complete max-gap cluster of m marked genes in a genome of size $n = 1,000$ as a function of m, for $g = \{5, 10, 20, 50\}$.

null hypothesis is that the m genes are randomly distributed in G_1 as well as in G_2. Like the reference region scenario, they also assume that each of the m genes in G_1 has exactly one homolog in G_2, and vice versa. Of course, $n - m$ genes in each genome have no homologs in the other. 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 2.2.1. 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 the whole genome comparison scenario, rather than calculate the probability of observing a cluster of size at least h, Hoberman, *et al.* [16] tried to determine the probability of observing a cluster of exactly size h, and represent the upper and lower bounds for this quantity.

2.3 Generalized Adjacency Gene Clustering Model

Zhu *et al.* [39] represented a new parameterized 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*[1], 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 gene cluster (GA cluster)*[2] in two or more genomes is just a maximal set of genes, where in each genome these genes form a connected component 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 gene clusters. Since even pairs of randomly constructed genomes may have some generalized adjacency gene clusters in common, the question arises of whether the number or size of these common clusters is significantly larger than the random case. Xu *et al.* [37] studied the statistical properties of generalized adjacency to answer such questions.

2.3.1 GA Cluster Definition

Definition 2.3.1. *[37] Let V_X to be the set of markers in the genome X. These markers are partitioned among a number of total orders called* ***chromosomes****. For markers g and h in V_X on the same chromosome in X, let $gh \in E_X$ if the number of genes intervening between g and h in X is less than θ, where $\theta \geq 1$ is a fixed*

[1] all the gene pairs which are (i, j)-*adjacency* in my definition of *two-parameter generalized adjacency*, where $i \leq k$ and $j \leq k$, are generalized adjacencies with parameter k in [39]

[2] It is called (θ, θ)-*adjacency cluster* in my definition of *two-parameter generalized adjacency.*

neighbourhood parameter.

Consider the graphs $G_S = (V_S, E_S)$ and $G_T = (V_T, E_T)$ with a non-null set of vertices in common $V = V_S \cap V_T$. We say a subset of $C \subseteq V$ is a ***generalized adjacency gene cluster*** *if it consists of the vertices of a maximal connected subgraph of $G_{ST} = (V, E_S \cap E_T)$.*

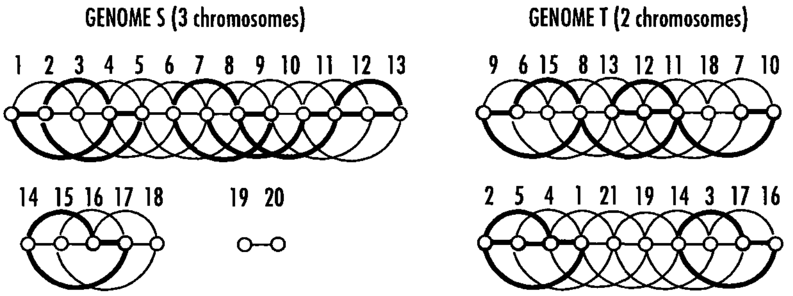
This definition of clusters, illustrated in Figure 2.2, decomposes the genes in the two genomes into two identical sets of disjoint generalized adjacency gene clusters of size greater or equal to 2, and possibly different sets of singletons belonging to no cluster, either because they are in V_{XY}, but not in $E_X \cap E_Y$, or because they are in $V_X \cup V_Y \setminus V_{XY}$. For the simplicity of the properties study, Xu and Sankoff [37] did not attempt to deal with duplicate genes in their paper, *i.e.* they did not take gene families into account, and also assumed $V_X = V_Y = V_{XY}$. In practice, depending on the relative emphasis to be placed on order rearrangement versus gene insertion/deletion, they 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.

2.3.2 Some Statistical Properties of GA Clusters

Each genome can be represented as a permutation of the first n positive integers. Then represent 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 $\frac{1}{n!}$.

Let $n_2 = |E_I \cap E_R|$ represent the number of common edges, i.e. the number



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

of generalized adjacencies. For a random genome $R = r_1, r_2, \ldots, r_n$, if $r_h = i$, they define the *position* of i in R to be $g_i = h$. Then

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

Proposition 2.3.2 (Proposition in [37]). *For $\theta \geq 1$,*

$$\mathbf{E}(n_2) = 2\theta^2 - \frac{4n\theta^3 - \theta^2(1+\theta)^2}{2n(n-1)}, \tag{2.3.1}$$

so that for a given θ

$$\lim_{n \to \infty} \mathbf{E}(n_2) = 2\theta^2 \tag{2.3.2}$$

Proposition 2.3.3 (Proposition in [37]). *For $\theta \geq 1$, n_2 converges in distribution to a Poisson distribution with parameter $2\theta^2$.*

2.3.3 Square-root Law for Parameter Selection

Based on 10,000 pairs of random genomes of size 100, varying parameter θ, the frequency of different maximal cluster sizes is shown in Figure 2.3. Figure 2.4 is the result of the same calculation when the genome size n is 1,000. 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 $n = 1,000$.

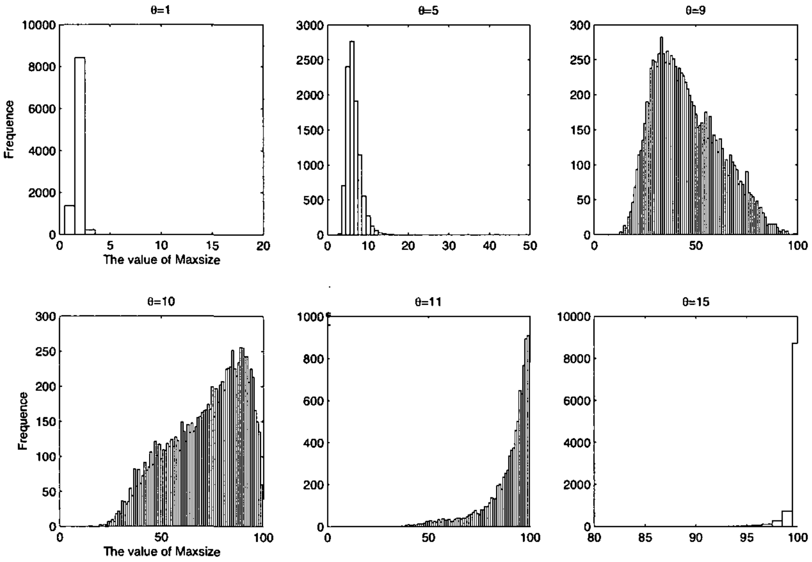


Figure 2.3: (Figure 4 in [37])Histograms for $k_{\max}$ when $n = 100$.

On the basis of 10,000 pairs of random genomes, we determined the change-point θ^* (a value of θ, 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 2.1), and in Figure 2.5 plotted these points against $\sqrt{n}$. This suggests that the change-point is approximately $\theta^* = \sqrt{n}$.

By calculating how much of the probability mass falls to the right of $0.5n$, for each value of θ, Figure 2.6 is obtained, showing that the change behaviour, in proportion

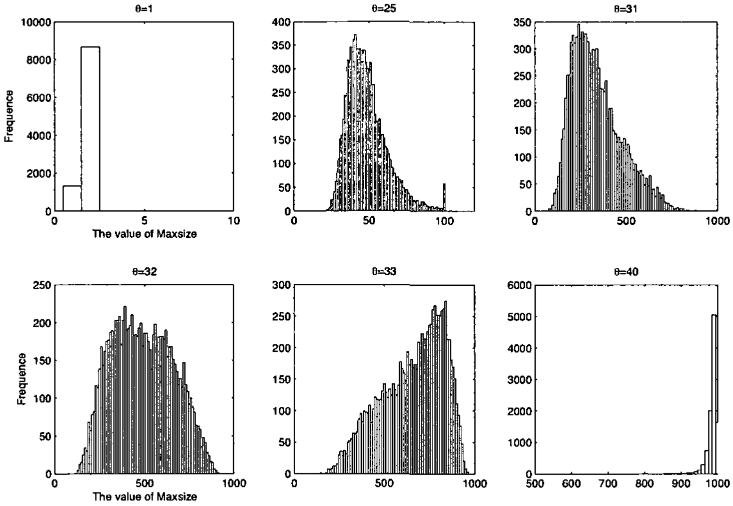


Figure 2.4: (Figure 5 in [37])Histograms for $k_{\max}$ when $n = 1000$

Table 2.1: Change-point as a function of n

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

to $\sqrt{n}$, tends to a sharp "cut-off" at or near $\theta = \sqrt{n}$.

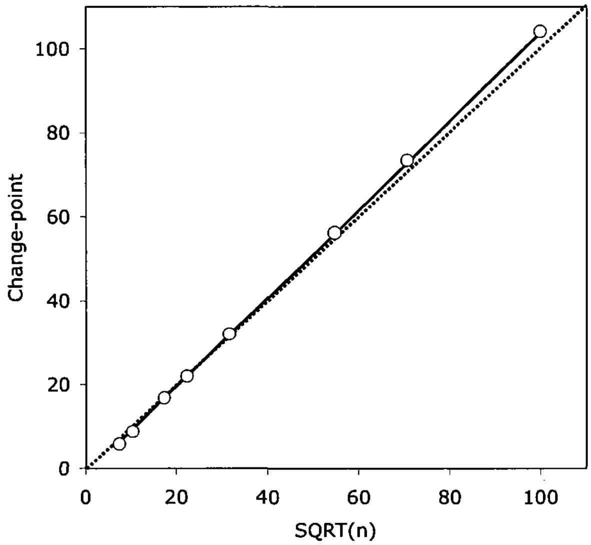


Figure 2.5: (Figure 6 in [37])Change-point for $k_{\max}$ as a function of $\sqrt{n}$. Dotted diagonal represents $\sqrt{n}$

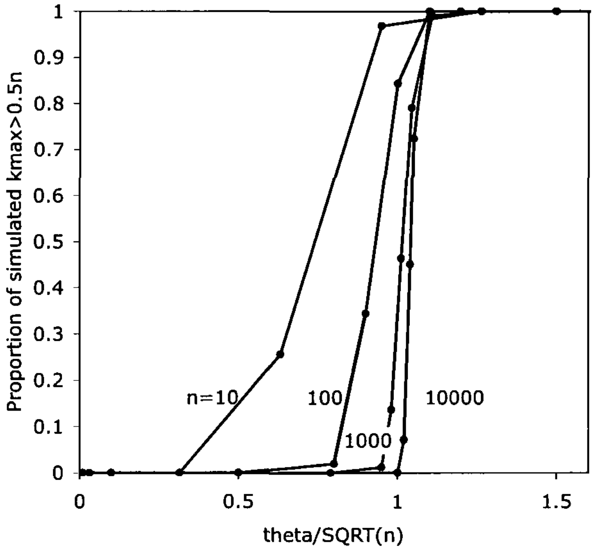


Figure 2.6: (Figure 7 in [37])Histograms for $k_{\max}$ when $n = 1,000$

Chapter 3

The Two Parameter Generalized Adjacency Model

In the previous chapter, I presented the one-parameter GA cluster criterion introduced by Zhu *et al.* [39] and also some statistical results on this criterion found by Xu and Sankoff [37]. However, some biologists are more interested in figuring out how far apart two genes in one genome tend to be if they are strictly adjacent in another one. Furthermore, using one parameter to compare multiple chromosomal genomes may lose cluster sensitivity when we do pairwise chromosomal comparison because of the large difference between chromosomes. Thus we need to extend the one-parameter GA cluster criterion to solve these kinds of problems.

In this chapter and the following chapters, we extend the GA cluster definition to a more general *two-parameter generalized gene cluster*, and study some properties under this new definition.

3.1 Basic Definitions

Definition 3.1.1. *Let S be a genome with n distinct genes, i.e., a permutation of the integers from $\{1, \ldots, n\}$. Two genes (integers) x and y are* ***i-adjacent*** *in S, written $|xy|_S = i$ or $x \overset{i}{\sim} y$ in S, if there are $i-1$ genes between them in S. E.g., 1 and 4 are 2-adjacent in the genome (2 1 3 4). We also write $x \sim y$ if $x \overset{1}{\sim} y$, i.e., if x and y are adjacent in the usual sense.*

Definition 3.1.2. *We say two common genes x and y of genomes S and T are* ***one-way (i, j)-adjacent*** *if they are i-adjacent in the genome S and j-adjacent in the genome T. While genes x and y are* ***either-way (i, j)-adjacent*** *if they are i-adjacent in either one of the genomes and j-adjacent in the other.*

Definition 3.1.3. *Let ω_{ij} be the (i, j)-adjacency* ***weight*** *on two genes that are one-way (i, j)-adjacent, i.e., i-adjacent in one genome and j-adjacent in the other, such that*

1. $0 \leq \omega_{ij} = \omega_{ji},\ i, j \in \{1, 2, \ldots, n-1\}$

2. $\sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \omega_{ij} = 1$

3. $\omega_{ij} \geq \omega_{kl}$ if

   (a) $\max(i, j) < \max(k, l)$

   (b) $\max(i, j) = \max(k, l)$ *and*
   $\min(i, j) < \min(k, l)$

Definition 3.1.4. *The* ***distance*** *between two genomes S and T is then*

$$d(S, T) \quad = \quad 2(n-1) - \sum_{i=1}^{n-1} \left(n_{ii}\omega_{ii} + \sum_{j=1}^{n-1} n_{ij}\omega_{ij} \right). \tag{3.1.1}$$

where ω_{ij} is the weight of two genes x and y are one-way (i,j)-adjacent in genomes S and T; n_{ij} is the total number of one-way (i,j)-adjacent gene pairs, (x,y) that are i-adjacent in S and j-adjacent in T.

Since $\omega_{ij} = \omega_{ji}$, let $n'_{ij} = n_{ij} + n_{ji}$ $(1 \leq j \leq i \leq n-1)$. Then n'_{ij} is the total number of either-way (i,j)-adjacent gene pairs in genomes S and T if $j < i$ and twice of the total number of either-way (i,j)-adjacent gene pairs in genomes S and T when $i = j$. Thus we can simplify the distance (3.1.1) to obtain the genome distance under the either-way adjacent gene pairs as

$$d(S,T) \quad = \quad 2(n-1) - \sum_{i=1}^{n-1}\sum_{j=1}^{i} n'_{ij}\omega_{ij}. \tag{3.1.2}$$

Definition 3.1.5. *Let E_S^θ be the set of all i-adjacencies in S, where $1 \leq i \leq \theta$. We define a subset of $C \subseteq V$ to be a* **one-way (θ,ψ)-generalized adjacency gene cluster**, *or* **one-way (θ,ψ)-adjacency cluster**, *if it consists of the vertices of a connected component of the* **generalized adjacency graph**, *$G_{ST}^{\theta\psi} = (V, E_S^\theta \cap E_T^\psi)$.*

Figure. 3.1 illustrates how genomes $S = 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9$ and $T = 2\ 1\ 5\ 7\ 8\ 3\ 6\ 4\ 9$ determine the one-way $(1,3)$ clusters $\{1,2\}$, $\{3,4\}$, $\{6,7,8\}$ and the one-way $(3,1)$ clusters $\{1,2\}$, $\{3,4,6\}$, $\{5,7,8\}$.

Generalizing definition 3.1.5, we obtain the definition of an either-way cluster as follows

Definition 3.1.6. *Let E_S^θ be the set of all i-adjacencies in S, where $1 \leq i \leq \theta$. We define a subset of $C \subseteq V$ to be an* **either-way (θ,ψ)-generalized adjacency gene cluster**, *or* **either-way (θ,ψ)-adjacency cluster**, *if it consists of the vertices of a connected component of the* **generalized adjacency graph**, *$G_{ST}^{\theta\psi} = (V, (E_S^\theta \cap E_T^\psi) \cup (E_S^\psi \cap E_T^\theta))$.*

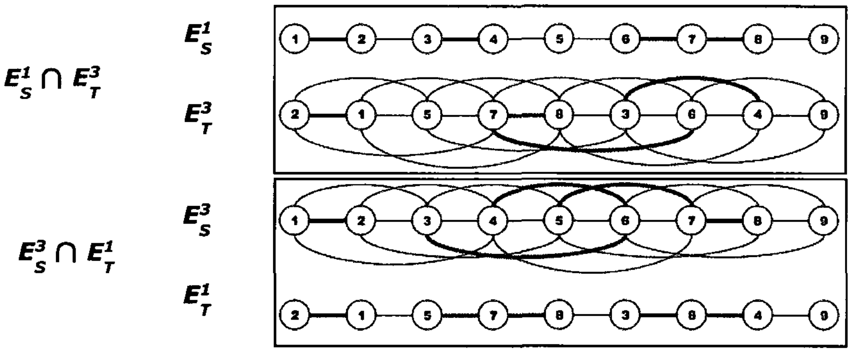


Figure 3.1: Determination of $(1, 3)$ clusters and $(3, 1)$ clusters.

Figure. 6.2 illustrates how genomes $S = 1\;2\;3\;4\;5\;6\;7\;8\;9$ and $T = 2\;1\;5\;7\;8\;3\;6\;4\;9$ determine the $(1, 3)$ clusters $\{1, 2\}$ and $\{3, 4, 5, 6, 7, 8\}$.

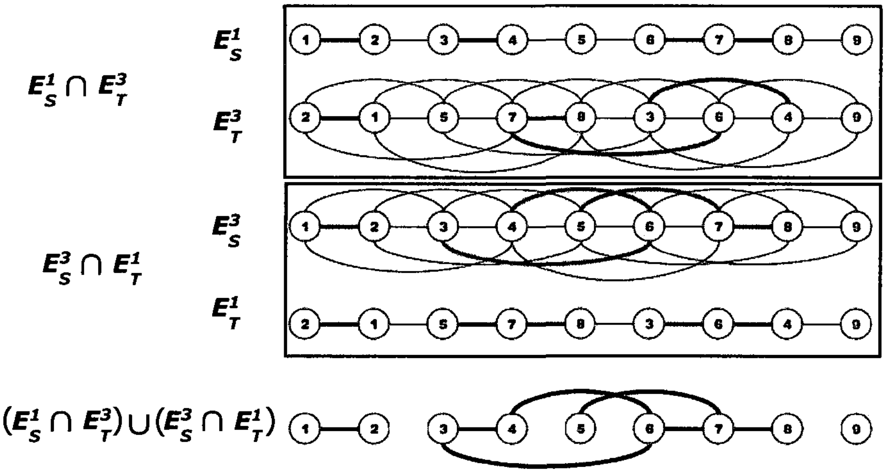


Figure 3.2: Determination of $(1, 3)$ clusters (or $(3, 1)$ clusters).

Based on the definitions above, we get the GA cluster in previous work [37, 39] to be a particular case of my definition, namely a (θ, θ)-adjacency cluster. Like other gene cluster definitions, the choice of parameter values is arbitrary. One of the main goals in this thesis is to remove some of this arbitrariness. In the following section, I will represent some statistical properties of adjacencies and adjacency clusters to reduce their arbitrariness, by finding "natural values" for θ and ψ as a function of n, the number of genes in the genomes. Here I only consider the whole genome comparison scenario since conclusions obtained in the whole genome comparison scenario also hold in the reference genes set scenario. I will begin with either-way $(1, j)$-adjacent gene pairs and either-way $(1, \theta)$-adjacency clusters. These may be of particular interest in genome similarity study, as mentioned at the beginning of this chapter. After that, I will extend it to the more general case.

3.2 A "Natural" Weight Function

3.2.1 Weight in $(1, j)$ Generalized Adjacency Model

In this model, we fix one parameter of either-way (i, j)-adjacency and either-way (θ, ψ)-adjacency cluster to 1. Then the definitions of weight and genome distance in Section 3.1 will be represented as follows:

Definition 3.2.1. *Let $(1, j)$-adjacency* ***weight*** *ω_i be any non-negative, non-increasing, function on the positive integers such that $\sum_{i=1}^{n-1} \omega_i = 1$. The weight ω induces the distance between genomes S and T as follows:*

$$d(S,T) = 2(n-1) - \sum_{i=1}^{n-1} (n_i^S + n_i^T)\omega_i \tag{3.2.1}$$

where n_j^X is the number of one-way $(1, j)$-adjacent gene pairs, i.e. j-adjacent on genome X and 1-adjacent on the other genome.

Similar to definition 3.1.4, we represent the number of either-way $(1, j)$-adjacent gene pairs by $n_i' = n_i^S + n_i^T$, thus we have

$$d(S,T) = 2(n-1) - \sum_{i=1}^{n-1} n_i' \omega_i, \tag{3.2.2}$$

Given two genomes S and T with the same genes, we consider the weight ω that is allocated, inasmuch as possible, to small i-adjacencies, thus emphasizing the local similarities of the two genomes. This motivates the study of $\min_{\omega} d(S,T)$.

Theorem 3.2.2. *For genomes S and T, the weight $\boldsymbol{\omega}$ that minimizes the distance* (3.2.1) *has*

$$\omega_i = \begin{cases} \frac{1}{k^*}, & \text{if } 1 \le i \le k^* \\ \\ 0, & \text{otherwise,} \end{cases} \tag{3.2.3}$$

where k^* is a natural number and maximizes the function

$$f(k) = \frac{1}{k}\sum_{i=1}^{k}(n_i^S + n_i^T) \tag{3.2.4}$$

$$= \frac{1}{k}\sum_{i=1}^{k} n_i'. \tag{3.2.5}$$

Proof. Based on Equation (3.2.2), minimizing $d(S,T)$ is equivalent to maximizing the summation

$$R = \sum_{i=1}^{n-1} n_i'\omega_i \tag{3.2.6}$$

We first note that a uniform upper bound on ω_i is $\frac{1}{i}$. i.e. $0 \le \omega_i \le \frac{1}{i}$. This follows because if it is not true, i.e. there exists a weight $\omega_i > \frac{1}{i}$, then $\omega_j > \frac{1}{i}$ for $j < i$ since the ωs are non-increasing, so that we have $\sum_{j=1}^{i-1}\omega_j > \frac{i-1}{i} = 1 - \frac{1}{i}$, which contradicts $\sum_{j=1}^{i-1}\omega_j \le \sum_{j=1}^{n-1}\omega_j - \omega_i < 1 - \frac{1}{i}$ because the summation of all weight is 1. By the same argument, we also obtain the proposition that if $\omega_i = \frac{1}{i}$ for some value of i, then $\omega_1 = \omega_2 = \cdots = \omega_{i-1} = \omega_i = \frac{1}{i}$ and $\omega_{i+1} = \cdots = \omega_{n-1} = 0$.

Now we show that for any solution, i.e., a $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_{n-1})$ that maximizes equation (3.2.6), there must be one weight in $\boldsymbol{\omega}$ which attains this upper bound.

To prove this, let weights $\omega_1, \omega_2, \ldots, \omega_{n-1}$ maximize Equation (3.2.6) for given values of $n_1', n_2', \ldots, n_{n-1}'$, such that $\zeta = \max R$. If all the n_i''s are equal or all the ω_i

are equal, the theorem holds trivially.

For all other cases, assume that there is no weight in $\boldsymbol{\omega}$ that attains its upper bound. We define the set $\mathcal{C} = \{\ i \mid \omega_i > \omega_{i+1},\ 1 \leq i \leq n-2\} \neq \varnothing$. Let $\xi = \min_{\mathcal{C}}(\min(\omega_i - \omega_{i+1}, \frac{1}{i} - \omega_i)) > 0$, by assumption. We select two weights ω_i and ω_j where $n'_i \neq n'_j$. Without loss of generality, we fix $i < j$. Set a sign function $I = 1$, if $n'_i < n'_j$ and $I = -1$, if $n'_i > n'_j$. Then we define

$$\begin{aligned} \zeta' &= \sum_{k=1}^{i-1} n'_k\omega_k + n'_i(\omega_i - I\xi) + \sum_{k=i+1}^{j-1} n'_k\omega_k + n'_j(\omega_j + I\xi) + \sum_{k=j+1}^{n-1} n'_k\omega_k && (3.2.7)\\ &= \zeta + (n'_j - n'_i)\cdot I \cdot \xi && (3.2.8)\\ &> \zeta. && (3.2.9) \end{aligned}$$

Then ζ is not the maximal value, contradicting the assumption about $\boldsymbol{\omega}$. Hence, there must exist a weight ω_i in $\boldsymbol{\omega}$ attaining its upper-bound $\frac{1}{i}$. Then the optimal weight is $\omega_1 = \omega_2 = \cdots = \omega_{i-1} = \omega_i = \frac{1}{i}$ and $\omega_{i+1} = \cdots = \omega_{n-1} = 0$.

Substituting this $\boldsymbol{\omega}$ in (3.2.6), produces the expression of form (3.2.4). So maximizing (3.2.6) is the same as maximizing (3.2.4). □

Thus if we set $\theta = k^*$, we should find a large number of generalized adjacencies, but not at the cost of unreasonably increasing the number of potential adjacencies. The cut-off k^* differs widely of course according to the pair of genomes S and T being compared, and this variation increases with n. However, under the uniform measure on the set of permutations, $f(k)$ does not vary much, in the statistical sense, at least as n gets large. Thus we use $\mathrm{E}[k^*]$, as function of n, to find the natural value for the cut-off parameters in the uniform weight-based distance.

N.B. The cut-off k^* is not always unique, though the range of maximizing values will be very narrow. In the simulation studies in the ensuing sections, we will adopt the convention of using the minimal solution value for k^*.

Clearly, it is easy to transform the definition 3.2.1 and theorem 3.2.2 from either-way generalized adjacency gene cluster to one-way generalized adjacency gene cluster. We get a new definition and theorem as follows:

Definition 3.2.3. *Let $(1, j)$-adjacency* ***weight*** *ω_i be any non-negative, non-increasing, function on the positive integers such that $\sum_{i=1}^{n-1} \omega_i = 1$. The weight ω induces the distance between genomes S and T under the one-way $(1, \theta)$-adjacency cluster criterion as follows:*

$$d(S,T) = (n-1) - \sum_{i=1}^{n-1} n_i^S \omega_i \tag{3.2.10}$$

where n_j^X is the number of one-way $(1, j)$-adjacent gene pairs that are j-adjacent on genome X and 1-adjacent on the other genome.

Theorem 3.2.4. *For genomes S and T, the weight $\boldsymbol{\omega}$ that minimizes the distance* (3.2.10) *has*

$$\omega_i = \begin{cases} \frac{1}{k^*}, & \text{if } 1 \leq i \leq k^* \\ \\ 0, & \text{otherwise,} \end{cases} \tag{3.2.11}$$

where k^* is a natural number and maximizes the function

$$f(k) = \frac{1}{k} \sum_{i=1}^{k} n_i^S. \tag{3.2.12}$$

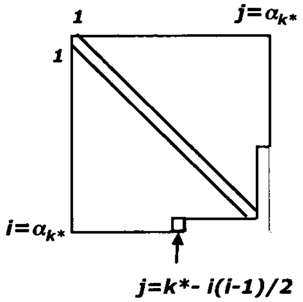


Figure 3.3: k is augmented from left to right, starting at the top row, in the lower triangle including the diagonal. Values of ω_{ij} in the upper triangle determined by symmetry.

3.2.2 The "natural" Bivariate Weights

Theorem 3.2.5. *Let* $\alpha_k = \lfloor\frac{\sqrt{1+8(k-1)}+1}{2}\rfloor$. *The (i,j)-adjacency weight* $\boldsymbol{\omega}$ *that minimizes the distance* (3.1.1) *has*

$$\omega_{ij} = \begin{cases} \frac{1}{k^*}, & \text{if } i < \alpha_{k^*},\ j \le i, \\ & \text{or } i = \alpha_{k^*},\ j \le k^* - \frac{i(i-1)}{2} \\ & \\ 0, & \text{otherwise} \end{cases} \tag{3.2.13}$$

where k^* is a natural number and maximizes the function

$$f(k) = \frac{1}{k}\left[\sum_{i=1}^{\alpha_k-1}\sum_{j=1}^{i}(n_{ij}+n_{ji}) + \sum_{j=1}^{k-\frac{1}{2}\alpha_k(\alpha_k-1)}(n_{\alpha_k j}+n_{j\alpha_k})\right] \tag{3.2.14}$$

where n_{ij} is the number of gene pairs i-adjacent on S and j-adjacent on T. (See Figure 3.3 for 2-dimensional area measured by k^*.)

Proof. Since $\omega_{ij} = \omega_{ji}$, Equation (3.1.1) is equivalent to

$$d(S,T) = 2(n-1) - \sum_{i=1}^{n-1}\sum_{j=1}^{i}(n_{ij}+n_{ji})\omega_{ij}. \tag{3.2.15}$$

So we can use the lower-triangle matrix to represent the bivariate weight, i.e.

$$\boldsymbol{\omega} = \begin{pmatrix} \omega_{11} & & & & & & \\ \omega_{21} & \omega_{22} & & & & & \\ \omega_{31} & \omega_{32} & \omega_{33} & & & & \\ & \cdots & & \cdots & & & \\ \omega_{(n-2)1} & \omega_{(n-2)2} & \omega_{(n-2)3} & \cdots & \omega_{(n-2)(n-3)} & \omega_{(n-2)(n-2)} & \\ \omega_{(n-1)1} & \omega_{(n-1)2} & \omega_{(n-1)3} & \cdots & \omega_{(n-1)(n-3)} & \omega_{(n-1)(n-2)} & \omega_{(n-1)(n-1)} \end{pmatrix} \tag{3.2.16}$$

and for each ω_{ij} $(1 \le j \le i \le n-1)$, the corresponding coefficient is $n_{ij} + n_{ji}$. Thus we transform the two-dimensional weight matrix to one-dimensional weight sequence

$$\boldsymbol{\omega} = \{\omega_{11}\ ,\ \omega_{21}\ ,\ \omega_{22}\ ,\ \cdots\ ,\ \omega_{(n-1)(n-2)}\ ,\ \omega_{(n-1)(n-1)}\} \tag{3.2.17}$$

Obviously, the $\omega s'$ in Equation (3.2.17) satisfy the weight definition in Section 3.2.1. Therefore we have that ω_{ij} satisfies a uniform distribution with a cut-off value k^* based on Theorem 3.2.2. Let ω_{ij} be the last nonzero ω. Then the two-dimensional weight matrix is

$$\boldsymbol{\omega} = \begin{pmatrix} \omega_{11} & & & & & & & \\ \omega_{21} & \omega_{22} & & & & & & \\ & \cdots & \cdots & \cdots & & & & \\ \omega_{i1} & \omega_{i2} & \cdots & \omega_{ij} & 0 \cdots 0 & & & \\ & \cdots & \cdots & \cdots & \cdots & & & \\ 0 & 0 & \cdots & 0 & 0 \cdots 0 & \cdots & 0 & \\ 0 & 0 & \cdots & 0 & 0 \cdots 0 & \cdots & 0 & 0 \end{pmatrix} \tag{3.2.18}$$

So we have

$$k^* = \frac{i(i-1)}{2} + j \tag{3.2.19}$$

Since $1 \leq j \leq i \leq n-1$, then

$$\frac{i(i-1)}{2} + 1 \leq k^* \leq \frac{i(i+1)}{2} \tag{3.2.20}$$

Solving (3.2.20) , we obtain the bounds for i,

$$\frac{\sqrt{1+8k^*}-1}{2} \leq i \leq \frac{\sqrt{1+8(k^*-1)}+1}{2} \tag{3.2.21}$$

Because i, j and k^* are all natural numbers, i must be $\lfloor\frac{\sqrt{1+8(k^*-1)}+1}{2}\rfloor$ and $j = k^* - \frac{i(i-1)}{2}$. Therefore, the conclusion holds. □

This theorem and Theorem 3.2.2 suggest that for the parameter values which are most sensitive to the similarities between two genomes, i.e., where the genome distance is minimized, the best weighting is uniform over all generalized adjacencies in a set determined by k^*. Then the value of θ and ψ should be set to $\theta = \psi = \lfloor\frac{\sqrt{1+8(k^*-1)}+1}{2}\rfloor \approx \sqrt{2k^*}$ to find the (θ, ψ) generalized adjacency gene clusters. As in the previous section we can use the $E[k^*]$ to estimate k^* due to the small variance of $f(k)$.

3.3 The Expected Number of Adjacent Gene Pairs

As mentioned in previous sections, the best weighting to minimize the genome distance is uniform over all generalized adjacencies in a set determined by k^* and the parameters should be a function of k^*. Because the formula which k^* maximizes is determined by n_{ij} (or n_j for $(1, \theta)$ case), we should study properties of the n_{ij}s (or the n_js for the $(1, \theta)$ case) to find the value of k^*.

Without loss of generality, we may always relabel the genes in one genome so that it becomes $I = (1, 2, \ldots, n)$ while the other is considered the random genome R, sampled from all $n!$ possible genomes, each with probability of $\frac{1}{n!}$. So the content of each gene in the random genome R is the position of the gene in the reference genome I.

3.3.1 The Expected Number of $(1, k)$-Adjacencies in Two Random Genomes

Let n_k^X be the total number of gene pairs (g, h), that are k-adjacent in genome X and 1-adjacent in the other.

Theorem 3.3.1. *For two random genomes S and T with n genes, the expected value of the random variable n_k^S (or n_k^T) is $\frac{2(n-k)}{n}$, i.e.*

$$\mathrm{E}(n_k^S) = \mathrm{E}(n_k^T) = \frac{2(n-k)}{n} \tag{3.3.1}$$

Proof. Considering $E(n_k^S)$, the event $\{x_i \overset{k}{\sim} x_j$ in genome S and $x_i \sim x_j$ in genome $T\}$ is equivalent to the event $\{i \overset{k}{\sim} i+1$ in the random genome R$\}$.

We define y_i^k as follows:

$$y_i^k = \begin{cases} 1, & \text{if } i \overset{k}{\sim} i+1 \text{ in } R \\ \\ 0, & \text{otherwise} \end{cases} \tag{3.3.2}$$

Then

$$n_k^S = \sum_{i=1}^{n-1} y_i^k. \tag{3.3.3}$$

Clearly, the number of gene pairs in a genome with n genes is $n(n-1)$. Now we are going to calculate the number of gene pairs in the random genome R which satisfies $i \overset{k}{\sim} i+1$, i.e., the two genes are 1-adjacent in the genome R and k-adjacent in the genome I. For two genes x and y which are strictly adjacent in the genome R, if the location of one gene, say x, is in the interval $[k, n-k]$ of I, then there are two positions to choose for the other gene y in I; while if the location of the gene x is out of the interval $[k, n-k]$ of I, we only have one position to choose for the gene y in I. So the number of gene pairs in the random genome R that satisfy $i \overset{k}{\sim} i+1$ is $2(n-k)$. Thus we obtain

$$P_n(y_i^k = 1) = P_n(i \overset{k}{\sim} i+1) = \frac{2(n-k)}{n(n-1)} \tag{3.3.4}$$

and also

$$E(n_k) = \sum_{i=1}^{n-1} E(y_i^k) = \sum_{i=1}^{n-1} P_n(y_i^k = 1) = \frac{2(n-k)}{n}. \tag{3.3.5}$$

□

Theorem 3.3.2. *Let n_k^S be the total number of gene pairs which are k-adjacent in genome S and 1-adjacent in the other genome, then n_k^S converges in distribution to the Poisson with parameter $\frac{2(n-k)}{n}$.*

Proof. As we know:

Theorem 3.3.3. (Theorem 30.1 in [2]). *Let μ 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 μ is the only probability measure with the moments α_1, $\alpha_2, \ldots$.*

Theorem 3.3.4. (Theorem 30.2 in [2]). *Suppose that the distribution of X is determined by its moments*[1]*, 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 3.3.3 and Theorem 3.3.4, we derive the following theorem

[1] *A probability measure is called determined by its moments if it satisfies the conclusion of Theorem 3.3.3.*

Theorem 3.3.5. (Theorem 2 in [37]). *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 λ^k, then their probability distributions converge to the Poisson distribution with mean λ.*

So using the same definition of y_i^k as (3.3.2), we consider the t^{th} factorial moment of n_k,

$$E[n_k(n_k-1)\dots(n_k-t+1)] = \binom{n-1}{t} t! E[y_{i_1}^k y_{i_2}^k \cdots y_{i_t}^k] \tag{3.3.6}$$

what we need to prove is $E[n_k(n_k-1)\dots(n_k-t+1)]$ converges to $[\frac{2(n-k)}{n}]^t$ for $k > 1$. Equation (3.3.6) holds because both sides of it represent the expectation of the number of ways to choose t non-zero elements from $\{y_1^k, \dots, y_{n-1}^k\}$. Since all y_i^k can only take the value 1 or 0, we have

$$\begin{aligned} E[y_{i_1}^k y_{i_2}^k \cdots y_{i_t}^k] &= P_n(y_{i_1}^k = 1, \dots, y_{i_t}^k = 1) && (3.3.7) \\ &= \prod_{j=1}^{t} P_n(y_{i_j}^k = 1) + O(\frac{1}{n^{t+1}}) \\ &= \left[\frac{2(n-k)}{n(n-1)}\right]^t + O(\frac{1}{n^{t+1}}) && (3.3.8) \end{aligned}$$

Then,

$$\begin{aligned} &\lim_{n\to+\infty} E[y(y-1)\dots(y-t+1)] \\ = &\lim_{n\to+\infty} \binom{n-1}{t} t! E[y_{i_1}^k y_{i_2}^k \cdots y_{i_t}^k] \\ = &\left[\frac{2(n-k)}{n}\right]^t && (3.3.9) \end{aligned}$$

Therefore, we conclude n_k^S converges in distribution to the Poisson with parameter $\frac{2(n-k)}{n}$. □

Wolfowitz [35] proved (cf. Xu *et al.* [36]) that n_1^S converges in distribution to the Poisson distribution with parameter 2. This is a special case of our result when n gets large.

Let n_k be the total number of either-way $(1,k)$-adjacent gene pairs in two genomes, e.g. S and T. It is clear that $n_k = n_k^S + n_k^T$ as $k \neq 1$, since the two events {the two genes are k-adjacent in genome S, 1-adjacent in genome T} and {the two genes are 1-adjacent in genome S, k-adjacent in genome T} are independent except for $k = 1$. If $k = 1$, i.e. for $(1,1)$-adjacent gene pairs, then $n_k = n_k^S = n_k^T$.

Thus we have

Theorem 3.3.6. *Let n_k be the total number of either-way $(1,k)$-adjacent gene pairs in two genomes. Then n_1 is an even number and $\frac{1}{2}n_1$ converges in distribution to the Poisson with parameter $\frac{2(n-1)}{n}$. For $k > 1$, the n_k all converge, independently and independent of $\frac{1}{2}n$, in distribution to the Poisson with parameter $\frac{4(n-k)}{n}$.*

Figure 3.4 shows the comparison of the simulated values of n_k versus the Poisson, for $k = 10$ and $k = 500$.

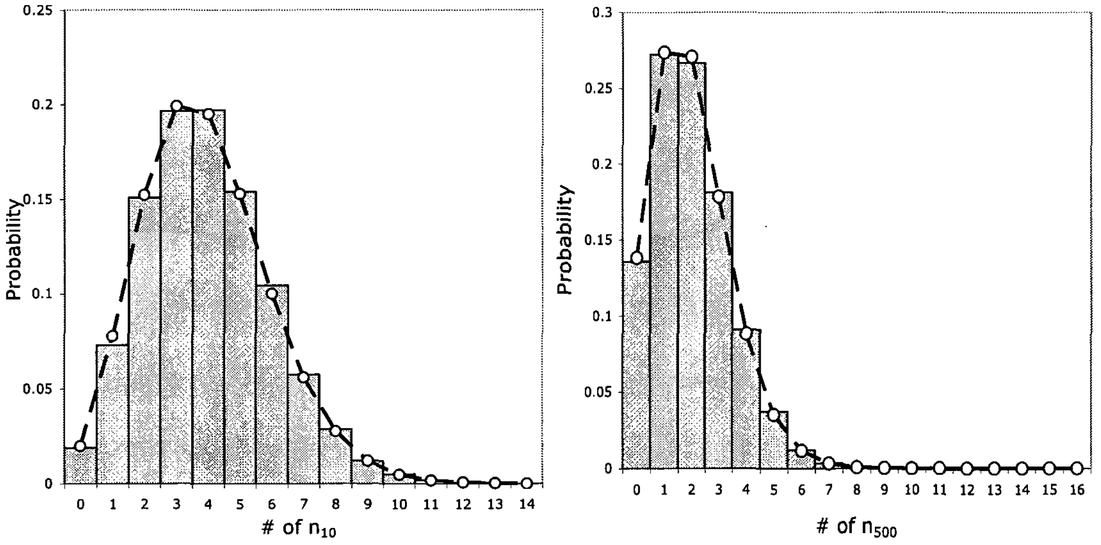


Figure 3.4: The distribution of n_k (histogram) compared to the Poisson distribution with parameter $\frac{4(n-k)}{n}$(dash line), where $k = 10$ (left) and $k = 500$ (right) Genome size $n = 1,000$, sample size $= 50,000$.

Given the limiting result in Theorem 3.3.6, we undertake to see the implications, for finite n, of the Poisson approximation. We simulated the expected value of k^* for genome sizes $n = 1000, 2000, \ldots, 100000$ by calculating k^* for each of 50,000 random permutations as in Theorem 3.2.2, and simply taking the mean. For comparison, we find k^* by generating $2Poisson(\frac{2(n-1)}{n})$ for $i = 1$ and $Poisson(\frac{4(n-1)}{n})$ for $i = 2, \ldots, n-1$, calculating mean of k^* on the basis of these values. The results are shown in Figure 3.5. So we can using the independent Poisson distribution with corresponding parameters to estimate the number of either-way $(1, k)$-adjacent gene pairs, i.e. n_k, instead of calculating from a random genome.

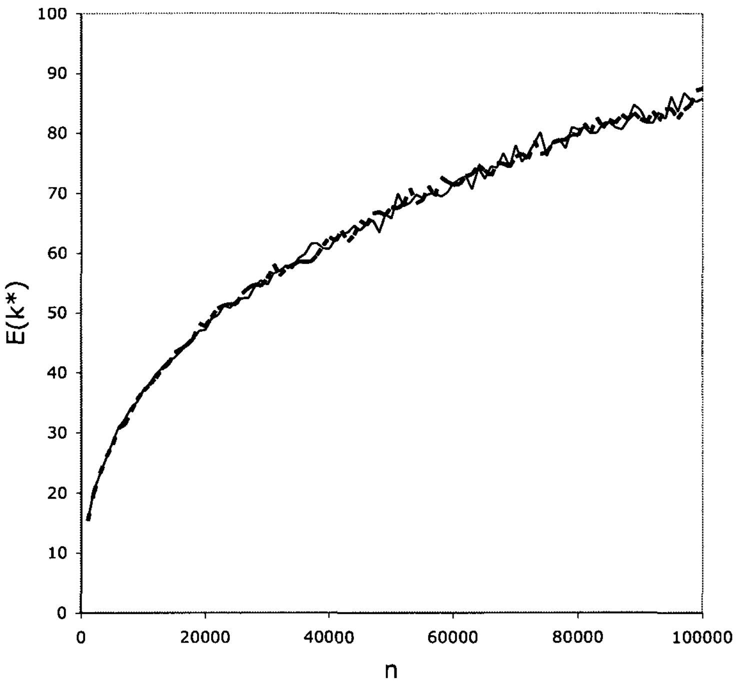


Figure 3.5: The expectation of k^* simulated by random permutation and by Poisson variables. Dotted line=Poisson.

3.3.2 The Expected Number of (j, k)-Adjacencies in Two Random Genomes

We previously calculated the expected number of both one-way and either-way $(1, k)$-adjacent gene pairs, i.e., 1-adjacent in one (or either one) genome and k-adjacent in the other. Here, we study the number of (j, k)-adjacent gene pairs. Similar to Section 3.3.1, we first consider n_{jk}, the total number of one-way (j, k)-adjacent gene pairs, (g, h), in two genomes, S and T, that are j-adjacent in genome S and k-adjacent in genome T.

Theorem 3.3.7. *For two random genomes, S and T, with n genes, let n_{jk} be the total number of one-way (j, k)-adjacent gene pairs, (x, y), which are j-adjacent in genome S and k-adjacent in genome T. Then n_{jk} converges in distribution to the Poisson with parameter $\frac{2(n-k)(n-j)}{n(n-1)}$*

Proof. The event $\{x \overset{j}{\sim} y$ in genome S , $x \overset{k}{\sim} y$ in genome T $\}$ is equivalent to the event $\{i \overset{k}{\sim} i + j$ in the random genome R$\}$. We define $y_i^{(j,k)}$ as:

$$y_i^{(j,k)} = \begin{cases} 1, & if\ i \overset{k}{\sim} i + j \text{ in } R \\ \\ 0, & \text{otherwise.} \end{cases} \tag{3.3.10}$$

Based on a similar explanation in the proof of Theorem 3.3.6, we have

$$P_n(i \overset{k}{\sim} i + j) = \begin{cases} \frac{2(n-k)}{n(n-1)}, & i = 1, 2, \dots, n - j \\ \\ 0, & \text{otherwise,} \end{cases} \tag{3.3.11}$$

So

$$\begin{aligned} P_n(y_i^{(j,k)} = 1) &= P_n(i \overset{k}{\sim} i+j,\ 1 \leq i \leq n-j) \\ &= P_n(i \overset{k}{\sim} i+j | 1 \leq i \leq n-j) P_n(1 \leq i \leq n-j) \\ &= \frac{2(n-k)(n-j)}{n(n-1)^2} \qquad (3.3.12) \end{aligned}$$

Considering the t^{th} factorial moment of n_{jk},

$$E[n_{jk}(n_{jk}-1)\ldots(n_{jk}-t+1)] = \binom{n-1}{t} t! E[y_{i_1}^{(j,k)} y_{i_2}^{(j,k)} \ldots y_{i_t}^{(j,k)}] \qquad (3.3.13)$$

Equation (3.3.13) holds because both sides of it represent the expectation of the number of ways to choose t non-zero elements from $\{y_1^{(j,k)}, \ldots, y_{n-1}^{(j,k)}\}$. Since all $y_i^{(j,k)}$ can only take the value 1 or 0, we have

$$\begin{aligned} E[y_{i_1}^{(j,k)} y_{i_2}^{(j,k)} \ldots y_{i_t}^{(j,k)}] &= P_n(y_{i_1}^{(j,k)} = 1, \ldots, y_{i_t}^{(j,k)} = 1) \qquad (3.3.14) \\ &= \prod_{r=1}^{t} P_n(y_{i_r}^{(j,k)} = 1) + O(\frac{1}{n^{t+1}}) \\ &= \left[\frac{2(n-k)(n-j)}{n(n-1)^2}\right]^t + O(\frac{1}{n^{t+1}}) \qquad (3.3.15) \end{aligned}$$

Therefore,

$$\begin{aligned} &\lim_{n\to+\infty} E[y(y-1)\ldots(y-k+1)] \\ = &\lim_{n\to+\infty} \binom{n-1}{t} t! E[y_{i_1}^{(j,k)} y_{i_2}^{(j,k)} \ldots y_{i_t}^{(j,k)}] \\ = &\left[\frac{2(n-k)(n-j)}{n(n-1)}\right]^t \qquad (3.3.16) \end{aligned}$$

Based on Theorem 2 in [37], we conclude n_{jk} converges in distribution to the Poisson with parameter $\frac{2(n-k)(n-j)}{n(n-1)}$. □

Let n'_{jk} $(1 \leq k \leq j \leq n-1)$ be the total number of either-way (j,k)-adjacent gene pairs of two random genomes. Then

Theorem 3.3.8. *If $j = k$, n'_{kk} is even and $\frac{1}{2}n'_{kk}$ converges in distribution to the Poisson with parameter $\frac{2(n-k)^2}{n(n-1)}$; if $k < j$, the n'_{jk} all converge, independently and independent of $\frac{1}{2}n'_{kk}$, in distribution to the Poisson with parameter $\frac{4(n-j)(n-k)}{n(n-1)}$*

As with Figure 3.6, we compared n'_{jk} based on random permutations to the Poisson distribution. Given the limiting result in Theorem 3.3.7, we undertake to see the implications, for finite n, of the Poisson approximation. We simulated the expectation of k^* for selected genome sizes $n = 100, 200, \ldots, 100,000$ by calculating k^* for each of 50,000 random permutations as in Theorem 3.2.5, and taking the mean. For comparison, we also estimated k^* by generating $2Poisson\left(\frac{2(n-j)(n-k)}{n(n-1)}\right)$ for $j = k \in \{1, 2, \ldots, n-1\}$ and $Poisson\left(\frac{4(n-j)(n-k)}{n(n-1)}\right)$ for $j \neq k \in \{1, 2, \ldots, n-1\}$, calculating k^* on the basis of these values. The results are shown in Figure 3.7. Note that in this bivariate case k^* grows much more rapidly than with $(1, k)$-adjacencies.

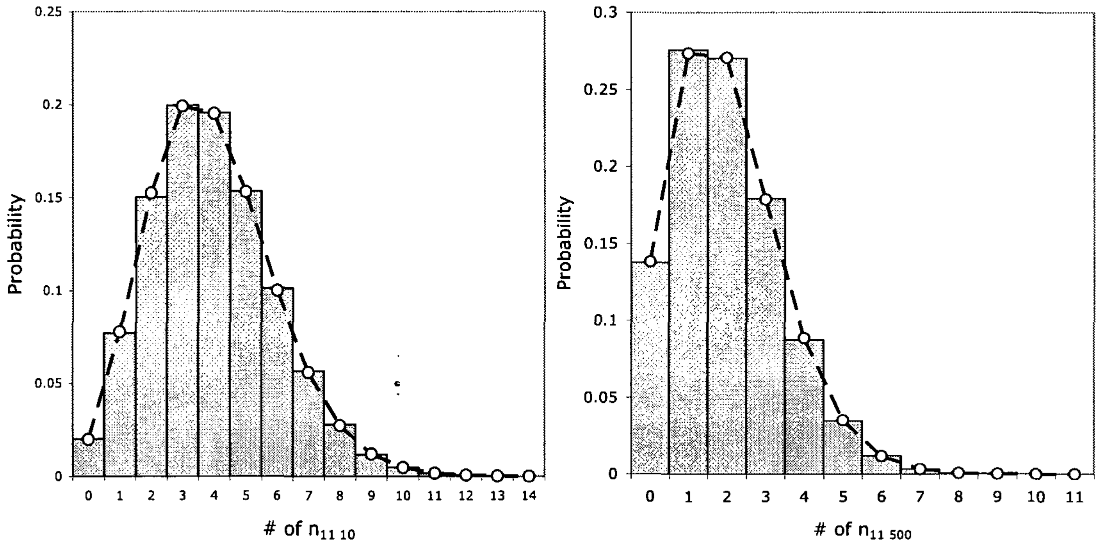


Figure 3.6: The distribution of n_{jk}(histogram) compared to the Poisson distribution(dash line) with parameter $\frac{4(n-j)(n-k)}{n(n-1)}$, where $n = 1,000$, $j = 11$ and $k = 10$ (left) or 500 (right). Sample size=50,000.

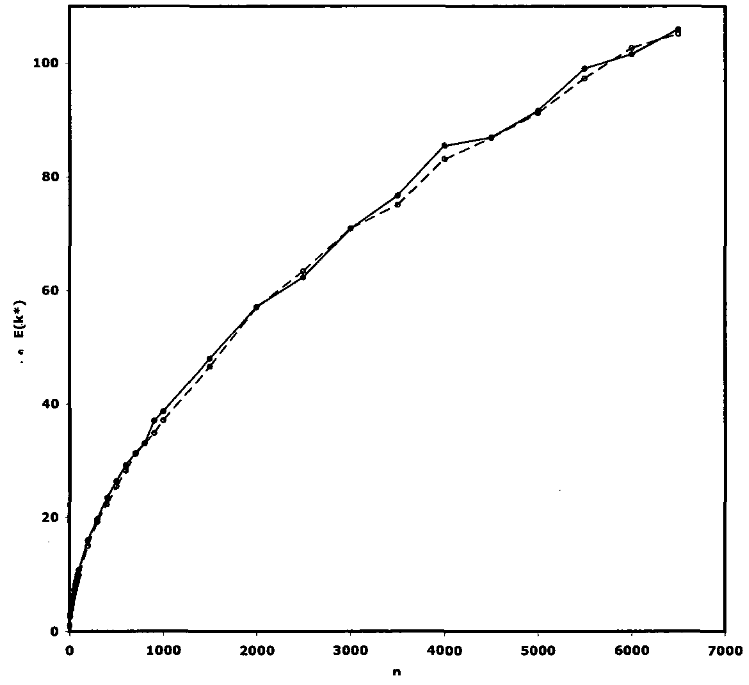


Figure 3.7: The expectation of k^* simulated by random permutation and by Poisson variables. Dotted line=Poisson.

3.4 The Expectation Value and Variance of $f(k)$

Since the number of one-way (or either-way) (i, j)-adjacent gene pairs can be represented approximately as independent Poisson distributions, we can calculate the expected value and variance of $f(k)$ in Theorem 3.2.5

Theorem 3.4.1. *Let the n_{ij}'s $(1 \leq i,\ j \leq n-1)$ satisfy Theorem 3.2.5. Then the expectation and variance of Equation* (3.2.14) *converge to* $(2-\frac{\alpha}{n})^2$ *and* $\frac{8}{\alpha^2}(1+\frac{2}{\alpha}-\frac{2}{\alpha^2})$ *as $n \longrightarrow \infty$, where* $\alpha = \lfloor\frac{\sqrt{1+8(k-1)}+1}{2}\rfloor$.

Proof. Let $n'_{ij} = n_{ij} + n_{ji}$, $(1 \leq j \leq i \leq n-1)$. Then

$$\begin{aligned} f(k) &= \frac{1}{k}\left[\sum_{i=1}^{\alpha_k-1}\sum_{j=1}^{i}(n_{ij}+n_{ji}) + \sum_{j=1}^{k-\frac{1}{2}\alpha_k(\alpha_k-1)}(n_{\alpha_k j}+n_{j\alpha_k})\right] \\ &= \frac{1}{k}\left(\sum_{i=1}^{\alpha_k-1}\sum_{j=1}^{i}n'_{ij} + \sum_{j=1}^{k-\frac{1}{2}\alpha_k(\alpha_k-1)}n'_{\alpha_k j}\right) \qquad (3.4.1) \end{aligned}$$

Since n_{ij} independently converges to *Poisson* $\left(\frac{2(n-i)(n-j)}{n(n-1)}\right)$ based on Theorem 3.3.7, so we get the expectation of n'_{ij} is $\frac{4(n-i)(n-j)}{n(n-1)}$. Let $\alpha_k = \alpha$, the expectation and variance of $f(k)$ are

$$\begin{aligned} E[f(k)] &= \frac{1}{k}\left(\sum_{i=1}^{\alpha-1}\sum_{j=1}^{i}\frac{4(n-i)(n-j)}{n(n-1)} + \sum_{j=1}^{k-\frac{1}{2}\alpha(\alpha-1)}\frac{4(n-\alpha)(n-j)}{n(n-1)}\right) \qquad (3.4.2) \\ &= \frac{4kn^2-[2k^2-2(\alpha^2-3\alpha-1)k-\frac{1}{2}\alpha(\alpha^2-1)(\alpha-2)]n}{kn(n-1)} \\ &\quad + \frac{2\alpha k^2-2\alpha(\alpha^2-\alpha-1)k+\frac{1}{2}\alpha(\alpha^2-1)(\alpha^2-\alpha-\frac{2}{3})n}{kn(n-1)} \\ &\quad + \frac{\frac{1}{6}\alpha(\alpha+1)(\alpha+2)(3\alpha^2+1)}{kn(n-1)} \\ &= \frac{(2n-\alpha)^2}{n(n-1)} + \frac{\alpha}{3n(n-1)} - \frac{1}{n-1} + \frac{1}{n(n-1)} - \frac{2}{\alpha(n-1)} + \frac{2}{3\alpha n(n-1)} \\ &= (2-\frac{\alpha}{n})^2 + o(\frac{1}{n}) \quad \textit{as } k \sim \frac{1}{2}\alpha^2 \textit{ and } \alpha \in \{1,\ 2,\ \cdots,\ n-1\} \qquad (3.4.3) \end{aligned}$$

$$
\begin{aligned}
Var[f(k)] &= \frac{1}{k^2}\left(\sum_{i=1}^{\alpha-1}\sum_{j=1}^{i-1}\frac{4(n-i)(n-j)}{n(n-1)} + 4\sum_{i=1}^{\alpha-1}\frac{2(n-i)^2}{n(n-1)} + \sum_{j=1}^{k-\frac{1}{2}\alpha(\alpha-1)}\frac{4(n-\alpha)(n-j)}{n(n-1)}\right)\\
&= \frac{4(k+\alpha-1)}{k^2} + o(\frac{1}{n})\\
&= \frac{8}{\alpha^2}(1+\frac{2}{\alpha}-\frac{2}{\alpha^2}) + o(\frac{1}{n}) \quad as\ k \sim \frac{1}{2}\alpha^2 \qquad (3.4.4)
\end{aligned}
$$

□

So the expected value of $f(k)$ should be decreasing from 4 to 1 and the variance of $f(k)$ should be decreasing from 8 to 0. Figure 3.8 shows the expectation value and variance of $f(k)$ over 50,000 pairs of random genomes with 300 (left) and 1,000 (right) genes. We can see that not only is the variance of $f(k)$ decreasing, but also the range of the variance of $f(k)$ is decreasing as genome size n is increasing. Thus we can use the $E(k^*)$ to estimate k^* approximately.

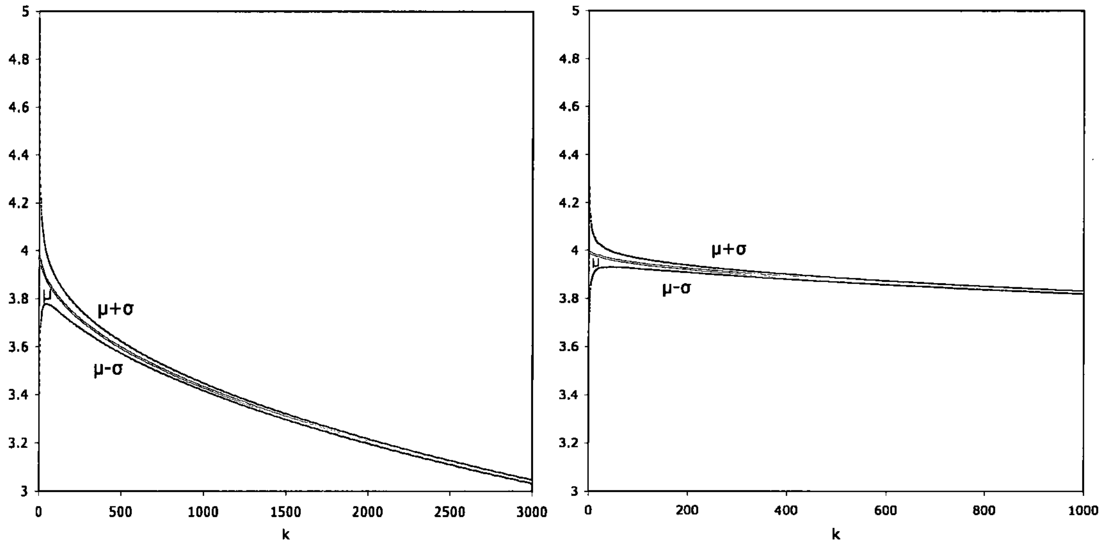


Figure 3.8: The mean and variance of $f(k)$, where genome size $n = 300$ (left) and $1,000$ (right), sample size $= 50,000$.

3.5 The Theory of Record Times

Because k^* is a maximum value of $f(k)$, looking for k^* is similar to the *upper record problem*, i.e., for a series of random variables $X_1, X_2, \ldots$, we consider the new sequence $L(m)$, $(m = 1, 2, \ldots)$, defined in the following manner:

$$L(1) = 1;\ L(m) = \min\{j : X_j > X_{L(m-1)}\}\ (m \geq 2) \tag{3.5.1}$$

where $L(m)$ is the index of the m^{th} upper record (or m^{th} record time), while the corresponding r.v. $X_{L(m)}$ is the value of the m^{th} record (or m^{th} record value).

Well-known properties of record times for i.i.d. random variables are:

- The probability that the i^{th} random variable attains a record is $\frac{1}{i}$.
- The expected number of records up to the i^{th} random variable is $\log i$.
- the average time at the record for n random variables is $\frac{n}{2}$.

The quantity k^* in Theorem 3.2.5 is a record time over $\frac{n(n-1)}{2}$ values of $f(k)$, though these are clearly neither identical nor independent random variables. That both the mean and variance of $f(k)$ are decreasing functions of n means that records become increasingly harder to attain.

This is illustrated in Figure 3.9, which compares the proportion of record values at each (i, j) in 50,000 pairs of random genomes of size $n = 10, 30, 100, 300, 1000$, and 3000, and the accumulated number of record values up to this point, with the corresponding values of i.i.d. random variables. Note that the horizontal axis is k, which maps to $i = \sqrt{2k}$ as a position on the genome.

More important for our purposes is that the average record time is nowhere near half the number of random variables ($\frac{n^2}{4}$ in our case). Figure 3.10 clearly shows that

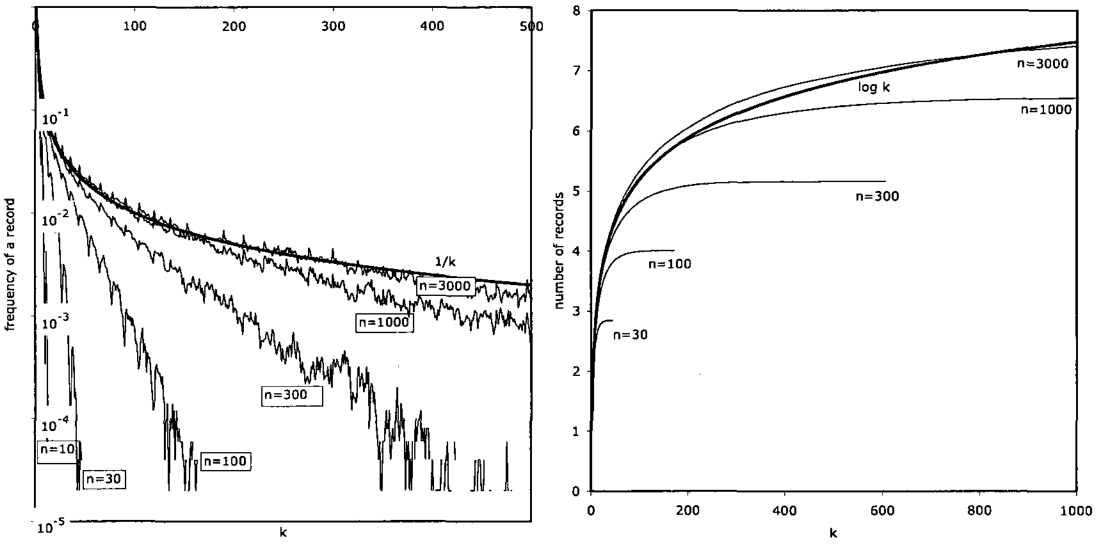


Figure 3.9: Comparison of mean optimal k values, over $50,000$ pairs of random genomes, with the record behaviour of i.i.d. random variables. Proportion of cases where k is optimal (left) and number of records attained (right), for (i, j) adjacencies as a function of genome size n. As $n \to \infty$, for any k', all curves approach the record time curves for all $k < k'$, but even at $n = 3000$, there is an eventual drop off, due to the declining mean expectations and variances of the n_{ij}.

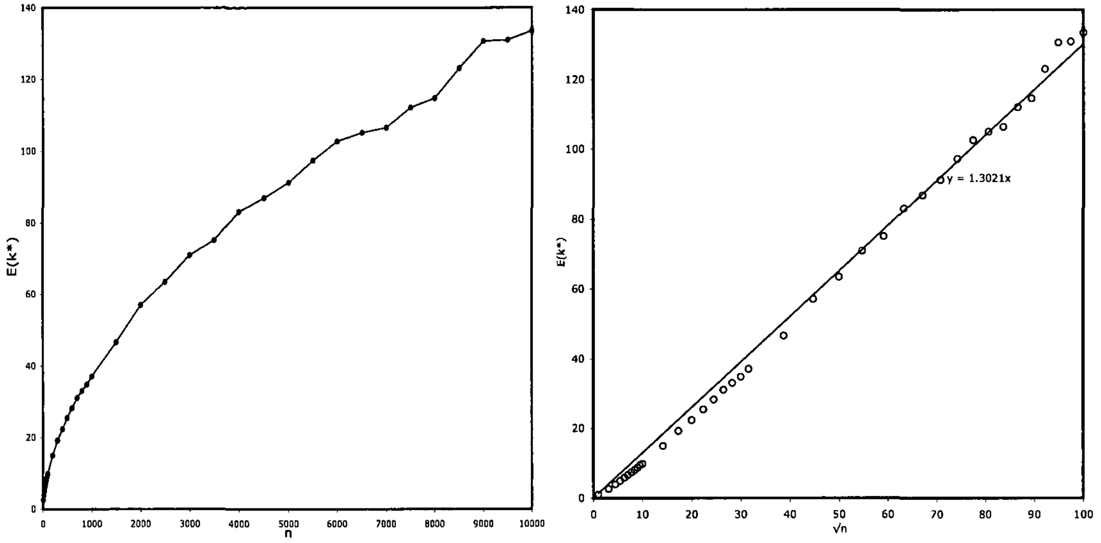


Figure 3.10: Average record time as a function of genome length.

k^* is approximately $\sqrt{n}$, so that the cut-off position on the genome of the maximizing weight system will be $o(\sqrt[4]{n})$, actually about $\sqrt[4]{4n}$. For genomes of size $n = 12,000$, the expected value of k^* is around 110, so that the cut-off θ for generalized adjacency need not be greater than 15.

Chapter 4

A "Natural" Value Based on the Cluster Significant

Theorems 3.2.2 and 3.2.5 show how to find the function which trades off the expected number, across all pairs of genomes, of generalized adjacencies against the parameters θ and ψ, with lower parameter values considered more desirable, i.e., it is better to find a large number of generalized adjacencies, but not at the cost of unreasonably increasing the number of potential adjacencies.

In this chapter, I shift attention to the other set of "natural" parameter values determined by the percolation behaviour of the (θ, ψ)-adjacency clusters. Beyond certain values of these parameters, tests of significance are no longer meaningful because all clusters rapidly coalesce together. To identify the threshold for this phenomenon, I first introduce the expression of the expected number of (i, j) adjacencies with $i \leq \theta$ and $j \leq \psi$. Then I insert this expression into connectivity formulae from the Erdős-Rényi theory of random graphs to find the corresponding percolation threshold, and compare this to simulations of random genomes.

As before, without loss of generality, we may always relabel the genes in one genome so that it becomes $I = (1, 2, \ldots, n)$ while the other is considered the random genome R, sampled from all $n!$ possible genomes, each with probability of $\frac{1}{n!}$. So the label of each gene in the random genome R is the position of the gene in the reference genome I.

4.1 The Expected Number of Adjacent Gene Pairs in a Generalized Adjacency Graph

4.1.1 One-way (θ, ψ)-Generalized Adjacency Gene Cluster

Theorem 4.1.1. *Let $N_2(n, \theta, \psi)$ be the number of adjacent gene pairs in two genomes with n genes, where the distances are no larger than θ in one genome and no larger than ψ in the other, then*

$$\lim_{n \to \infty} E[N_2(n, \theta, \psi)] = 2\psi\theta. \tag{4.1.1}$$

Proof. Based on Theorems 3.3.7 and 3.3.8, we know the number of one-way (i, j)-adjacent gene pairs independently converges to the Poisson distribution with parameter $\frac{2(n-i)(n-j)}{n(n-1)}$. Then

$$\begin{aligned} E[N_2(n, \theta, \psi)] &= \sum_{i=1}^{\theta} \sum_{j=1}^{\psi} E(n_{ij}) \\ &= \sum_{i=1}^{\theta} \sum_{j=1}^{\psi} \frac{2(n-i)(n-j)}{n(n-1)} \\ &= \frac{[(2n-1)\theta - \theta^2][(2n-1)\psi - \psi^2]}{2n(n-1)} \\ &= 2\psi\theta - \frac{\theta\psi(\theta + \psi)}{n} + \frac{\theta\psi(\theta - 1)(\psi - 1)}{2n(n-1)} \end{aligned} \tag{4.1.2}$$

where n_{ij} is the total number of one-way (i, j)-adjacent gene pairs, i.e. i-adjacent in one genome, j-adjacent in the other genome.

Therefore as $\theta,\ \psi \ll n$

$$\begin{aligned}
&\lim_{n\to\infty} E[N_2(n,\theta,\psi)] \\
= &\lim_{n\to\infty}\left(2\psi\theta - \frac{\theta\psi(\theta+\psi)}{n} + \frac{\theta\psi(\theta-1)(\psi-1)}{2n(n-1)}\right) \\
= &\ 2\psi\theta
\end{aligned} \tag{4.1.3}$$

□

4.1.2 Either-way (θ,ψ)-Generalized Adjacency Gene Cluster

Theorem 4.1.2. *Let $N_2'(n,\theta,\psi)$ be the number of gene pairs in two genomes with n genes, where the distances are no larger than θ in either one of genomes and no larger than ψ in the other, then*

$$\lim_{n\to\infty} E[N_2'(n,\theta,\psi)] = 4\psi\theta - 2\theta^2, \quad \textit{where } 1 \le \theta \le \psi \ll n. \tag{4.1.4}$$

Proof. Based on Theorem 3.3.7 and 3.3.8, we know the number of (i,j)-adjacent gene pairs independently converges to the Poisson distribution. Then for $1 \le \theta \le \psi \le n-1$,

$$\begin{aligned}
&E[N_2'(n,\theta,\psi)] \\
= &\sum_{i=1}^{\theta}\sum_{j=1}^{\psi} E(n_{ij}+n_{ji}) - \sum_{j=2}^{min(\theta,\psi)}\sum_{i=1}^{j-1} E(n_{ij}+n_{ji}) - \sum_{i=1}^{min(\theta,\psi)} E(n_{ii}) \\
= &\sum_{i=1}^{\theta}\sum_{j=1}^{\psi}\frac{4(n-i)(n-j)}{n(n-1)} - \sum_{j=2}^{\theta}\sum_{i=1}^{j-1}\frac{4(n-i)(n-j)}{n(n-1)} - \sum_{i=1}^{\theta}\frac{2(n-i)^2}{n(n-1)} \\
= &\ (4\psi\theta - 2\theta^2) - \frac{2\theta(\psi^2-\theta^2+\psi\theta)}{n} + \frac{\theta(\theta-1)(2\psi^2-2\psi-\theta^2+\theta)}{2n(n-1)},
\end{aligned} \tag{4.1.5}$$

where n_{ij} is the total number of one-way (i,j)-adjacent gene pairs, i.e. i-adjacent in one genome, j-adjacent in the other genome.

Therefore, if $1 \leq \theta \leq \psi \ll n$, we have

$$\begin{aligned} & \lim_{n\to\infty} E[N_2'(n,\theta,\psi)] \\ = & \lim_{n\to\infty} \left(4\psi\theta - 2\theta^2 - \frac{2\theta(\psi^2-\theta^2+\psi\theta)}{n} + \frac{\theta(\theta-1)(2\psi^2-2\psi-\theta^2+\theta)}{2n(n-1)} \right) \\ = & \ 4\psi\theta - 2\theta^2 \end{aligned} \tag{4.1.6}$$

□

4.2 The Number of Clusters with Large Size

For the reference genome I and the random genome R with n genes, we used $N_k(n,\theta,\psi)$ to represent the number of k-tuples of genes which are in the same (θ,ψ)-adjacency gene cluster defined in Definition 3.1.5 and 3.1.6. We have already calculated the expected of $N_2(n,\theta,\psi)$ in Section 4.1. In this section, we look for analytical results for the expectation of $N_k(n,\theta,\psi)$ when $k \geq 3$. We only consider $\theta,\psi \leq \lfloor\frac{n-1}{k-1}\rfloor$, because most of our discussion is under the constraint θ and ψ are much smaller than n. We first compute the expectation of $N_3(n,\theta,\theta)$ and then extend to the larger k.

4.2.1 Three Genes in One Cluster

Considering three genes u, v and w in the genomes I and R, we define $u \overset{i_1}{\sim} v \overset{i_2}{\sim} w$ as meaning genes u and v are i_1-adjacent and v and w are i_2-adjacent in a genome.

We have

Theorem 4.2.1. *Let $N_3(n,\theta) = N_3(n,\theta,\theta)$ be the total number of 3-tuples of genes, (u,v,w), which are in the same (θ,θ)-adjacency gene cluster. If $1 \leq \theta \leq \lfloor\frac{n-1}{2}\rfloor$, then the $N_3(n,\theta)$ converges in distribution to the Poisson with parameter*

$$\frac{\theta^2(5\theta^2-2\theta-1)}{n} - \frac{\theta^2(28\theta^3-29\theta^2+2\theta+5)}{3n(n-1)} + \frac{2\theta^2(19\theta^2-16\theta-8)(\theta-1)^2}{9n(n-1)(n-2)} \tag{4.2.1}$$

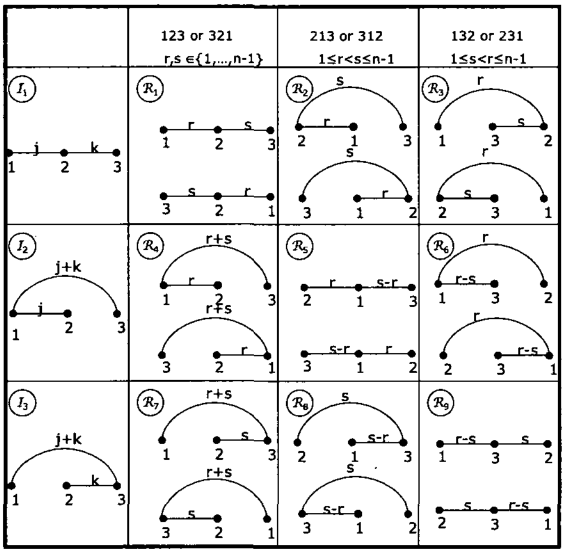


Figure 4.1: The cases to compute $N_3(n, \theta, \psi)$

Proof. Consider a 3-tuples of genes, we relabel them 1, 2, 3 from left to right in the reference genome, I, and define $1 \overset{j}{\sim} 2 \overset{k}{\sim} 3$ in genome I and $1 \overset{r}{\sim} 2 \overset{s}{\sim} 3$ in genome R. Based on Definition 3.1.6, there are eighteen configurations where genes 1, 2 and 3 are in the same (θ, θ)-adjacency gene cluster. These are distinguished by the order of the three genes in genome R and whether there are edges between them in the two genomes. These cases are shown in Figure 4.1[1].

Let $y_i^{(j,k)(r,s)}(v) = 1$ if the 3-tuples of genes, $(1, 2, 3)$ is $1 \overset{j}{\sim} 2 \overset{k}{\sim} 3$ in I, $1 \overset{r}{\sim} 2 \overset{s}{\sim} 3$ in R, the location of gene 1 in the genome I is i and the degree of gene v is 2, while other genes' degrees are 1 in Figure 4.1, while $y_i^{(j,k)(r,s)}(v) = 0$ otherwise. We define $n_{(j,k)(r,s)}^{<v>}$ the number of 3-tuples of genes satisfying $y_i^{(j,k)(r,s)}(v) = 1$, i.e. $n_{(j,k)(r,s)}^{<v>} = \sum_{i=1}^{n-2} y_i^{(j,k)(r,s)}(v)$.

[1] We combine eighteen cases to nine cases, using left to right symmetry.

Similar to the calculation of Equation 3.3.4, we get

$$P_n\left(y_i^{(j,k)(r,s)}(2) = 1|(1,2,3)\right) = \begin{cases} \frac{2(n-r-s)}{n(n-1)(n-2)}, & i = 1,2,\dots,n-j-k \\ \\ 0, & \text{otherwise,} \end{cases} \tag{4.2.2}$$

Then the expected number of 3-tuples of genes is

$$\begin{aligned} E\left(n_{(j,k)(r,s)}^{<2>}|(1,2,3)\right) &= \sum_{i=1}^{n-2}\left[P_n\left(y_i^{(j,k)(r,s)}(2) = 1|(1,2,3)\right)P_n(1 \le i \le n-j-k)\right] \\ &= \sum_{i=1}^{n-2}\left[\frac{2(n-r-s)}{n(n-1)(n-2)} \cdot \frac{n-j-k}{n-2}\right] \\ &= \frac{2(n-j-k)(n-r-s)}{n(n-1)(n-2)}. \end{aligned} \tag{4.2.3}$$

Using the same idea, we can also obtain the $E(n_{(j,k)(r,s)}^{<v>})$ in the other eight cases and find that thė general formula is

$$E\left(n_{(j,k)(r,s)}^{<v>}\right) = \begin{cases} \frac{2(n-j-k)(n-s)}{n(n-1)(n-2)}, & v = 1 \\ \frac{2(n-j-k)(n-r-s)}{n(n-1)(n-2)}, & v = 2 \\ \frac{2(n-j-k)(n-r)}{n(n-1)(n-2)}, & v = 3. \end{cases} \tag{4.2.4}$$

Hence we can calculate the expected number of 3-tuples of genes which are in the same (θ,θ)-adjacency gene cluster, using Equation (4.2.5).

$$E[N_3(n,\theta)] = \sum_{v=1}^{3}\sum_{j,k,r,s\le\theta}\sum_{(\mathcal{I},\mathcal{R})\in\{(\mathcal{I}_i,\mathcal{R}_j)|1\le i\le 3, 1\le j\le 9\}} E\left(n_{(j,k)(r,s)}^{<v>}|v,(\mathcal{I},\mathcal{R})\right), \tag{4.2.5}$$

adding all disjoint cases of genome I and R.

From Figure 4.1, we see that the number of 3-tuples of genes in the case $(\mathcal{I}_2,\mathcal{R}_4)$, which represent s the adjacency of genes in genome I, shown as $\mathcal{I}_2$, and in genome R, shown as $\mathcal{R}_4$, is already counted in the case $(\mathcal{I}_1,\mathcal{R}_1)$. So has the case $(\mathcal{I}_3,\mathcal{R}_7)$. Also for cases $(\mathcal{I}_2,\mathcal{R}_6)$ and $(\mathcal{I}_3,\mathcal{R}_8)$, their numbers have already been counted as other cases. Therefore, there are five cases left which are not included in other cases, shown

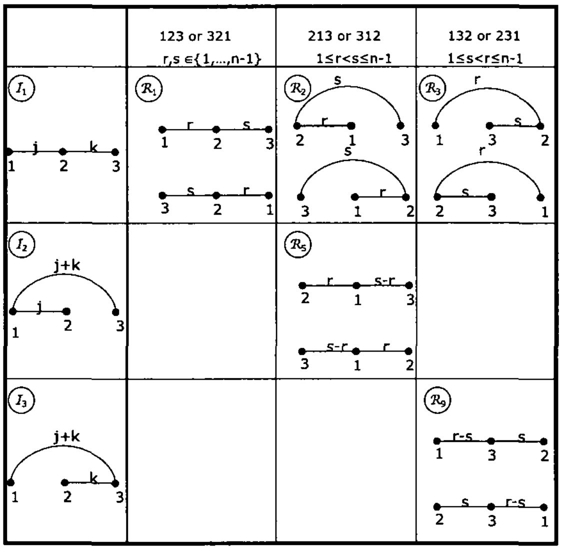


Figure 4.2: The cases to compute $N_3(n, \theta, \psi)$ which are not included the others

as Figure 4.2. Then after adding all five cases in Figure 4.2 and subtracting the overlapping for each permutation of genes in genome R, we obtain for $E[N_3(n,\theta)]$,

$$\begin{aligned}
&E[N_3(n,\theta)] \\
&= C\left(\sum_{j=1}^{\theta}\sum_{k=1}^{\theta}(n-j-k)\right)\left(\sum_{r=1}^{\theta}\sum_{s=1}^{\theta}(n-r-s)+2\sum_{r=2}^{\theta}\sum_{s=1}^{r-1}(n-r)\right) \\
&\quad +C\left(\sum_{j=1}^{\theta-1}\sum_{k=1}^{\theta-j}(n-j-k)\right)\left(\sum_{r=2}^{\theta}\sum_{s=1}^{r-1}(n-r)+\sum_{r=1}^{\theta}\sum_{s=\theta+1}^{r+\theta}(n-s)\right) \\
&= \frac{\theta^2(5\theta^2-2\theta-1)}{n}-\frac{\theta^2(28\theta^3-29\theta^2+2\theta+5)}{3n(n-1)}+\frac{2\theta^2(19\theta^2-16\theta-8)(\theta-1)^2}{9n(n-1)(n-2)}
\end{aligned} \tag{4.2.6}$$

where $C = \frac{2}{n(n-1)(n-2)}$.

Using a proof similar to Theorem 3.3.2, we can show that the $n^{<v>}_{(j,k)(r,s)}$s are approximately independent and converge to the Poisson distribution with parameter $E(n^{<v>}_{(j,k)(r,s)})$ in Equation 4.2.4. Therefore, $N_3(n,\theta)$ converges in distribution to the Poisson with parameter $E[N_3(n,\theta)]$ shown in Equation 4.2.1. □

It is straightforward to extend the $E(N_3(n,\theta,\theta))$ to $E(N_3(n,\theta,\psi))$ as in the following theorem

Theorem 4.2.2. *Let $N_3(n,\theta,\psi)$ be the total number of 3-tuples of genes, (u,v,w), which are in the same (θ,ψ) generalized adjacency gene cluster. If $1 \le \theta \le \psi \le \lfloor \frac{n-1}{2} \rfloor$, then the $N_3(n,\theta,\psi)$ converges in distribution to the Poisson with parameter*

$$\frac{\theta}{n}\left[(22\psi^3 - 36\theta\psi^2 + 26\theta^2\psi - 7\theta^3) - 6\psi^2 + 2\psi(3\theta-2) - \theta(2\theta-3)\right] + O(\frac{1}{n^2})(4.2.7)$$

Proof. Based on the proof of Theorem 4.2.1, what we need to show is the expected value of $N_3(n,\theta,\psi)$ is as in Equation 4.2.7.

Since

$$\begin{aligned}
& E(N_3(n,\theta,\psi)) \\
= & \sum_{v=1}^{3} \sum_{j,k,r,s=1}^{\psi} \sum_{\mathcal{R}\in\{\mathcal{R}_1,\mathcal{R}_2,\mathcal{R}_3\}} E\left(n_{(j,k)(r,s)}^{<v>} | v, \mathcal{R}\right) \\
& - \sum_{v=1}^{3} \sum_{j,k,r,s=\theta+1}^{\psi} \sum_{\mathcal{R}\in\{\mathcal{R}_1,\mathcal{R}_2,\mathcal{R}_3\}} E\left(n_{(j,k)(r,s)}^{<v>} | v, \mathcal{R}\right) \\
= & \quad \mathcal{C}\left(\sum_{j=1}^{\psi}\sum_{k=1}^{\psi}(n-j-k)\right)\left(\sum_{r=1}^{\psi}\sum_{s=1}^{\psi}(n-r-s) + 2\sum_{r=2}^{\psi}\sum_{s=1}^{r-1}(n-r)\right) \\
& + \mathcal{C}\left(\sum_{j=1}^{\psi-1}\sum_{k=1}^{\psi-j}(n-j-k)\right)\left(\sum_{r=2}^{\psi}\sum_{s=1}^{r-1}(n-r) + \sum_{r=1}^{\psi}\sum_{s=\psi+1}^{r+\psi}(n-s)\right) \\
& - \mathcal{C}\left(\sum_{j=\theta+1}^{\psi}\sum_{k=\theta+1}^{\psi}(n-j-k)\right)\left(\sum_{r=\theta+1}^{\psi}\sum_{s=\theta+1}^{\psi}(n-r-s) + 2\sum_{r=\theta+2}^{\psi}\sum_{s=\theta+1}^{r-1}(n-r)\right) \\
& - \mathcal{C}\left(\sum_{j=1}^{\psi-1}\sum_{k=1}^{\psi-j}(n-j-k)\right)\left(\sum_{r=\theta+2}^{\psi}\sum_{s=\theta+1}^{r-1}(n-r) + \sum_{r=\theta+1}^{\psi}\sum_{s=\psi+\theta+1}^{r+\psi}(n-s)\right) \\
= & \quad \frac{\theta}{n}\left[(22\psi^3 - 36\theta\psi^2 + 26\theta^2\psi - 7\theta^3) - 6\psi^2 + 2\psi(3\theta-2) - \theta(2\theta-3)\right] + O(\frac{1}{n^2})
\end{aligned}$$

Hence the theorem holds. □

4.2.2 $E[N_m(n,\theta)]$ Calculation for $m > 3$

It is easy to compute the expected number of three genes that are in the same cluster because we can enumerate all nine cases and simplify to five cases as shown in Figures 4.1 and 4.2. We can also find all cases of four genes in a cluster, shown in Figure 4.3 and Figure 4.4[2]. However, the larger the number of genes in one cluster, say m, is, the harder it is to enumerate the cases and compute the summation of each case because the number of summations we have to compute for m genes is $m^{m-2}(m!)^2$, which is the number of spanning tree of the complete graph multiplied by $m!$. This is too huge a number for manual enumeration. We require a general summation formula to reduce the computational effort.

From the proof of Theorem 4.2.1, we know the task of computing the $E[N_m(n,\theta)]$ is equivalent to computing the summation of all possible disjoint cases from tables such as shown in Figure 4.1, Figure 4.3 and Figure 4.4, where the first column indicates which case of m genes applies in the reference genome. Genes are labelled as an identity sequence $\{1, 2, \ldots, m\}$. The set of possibilities is the same as all spanning-trees in the complete graph with m vertices. The other structure graphs in the same row represent all possible permutations of these m genes in the random genome R with the same adjacency structure as in the reference genome. Therefore, what we must do is to find all disjoint cases for these two graphs, where one is in the reference genome, i.e. one of graphs in the first column of the table and the other one is these m genes in the other genome with the same adjacency structure as in the reference genome.

We define e_{ij} as the edge label of genes in the adjacency gene cluster graph, where i is the distance of two genes connected by the edge and j is the index of the

[2] We only drew half of the all permutations due to the symmetry of the sequence.

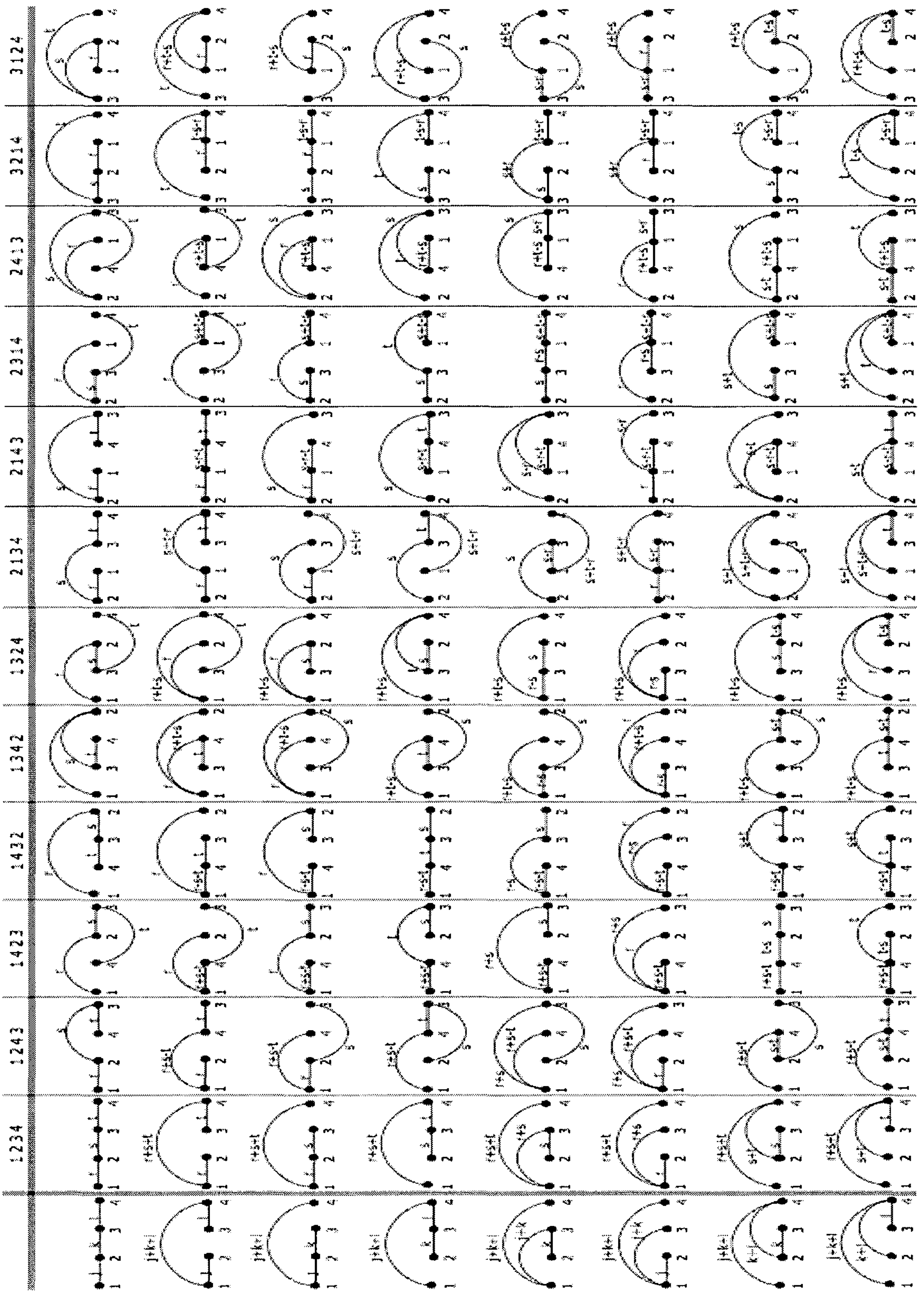


Figure 4.3: Cases to compute $N_4(n, \theta, \psi)$

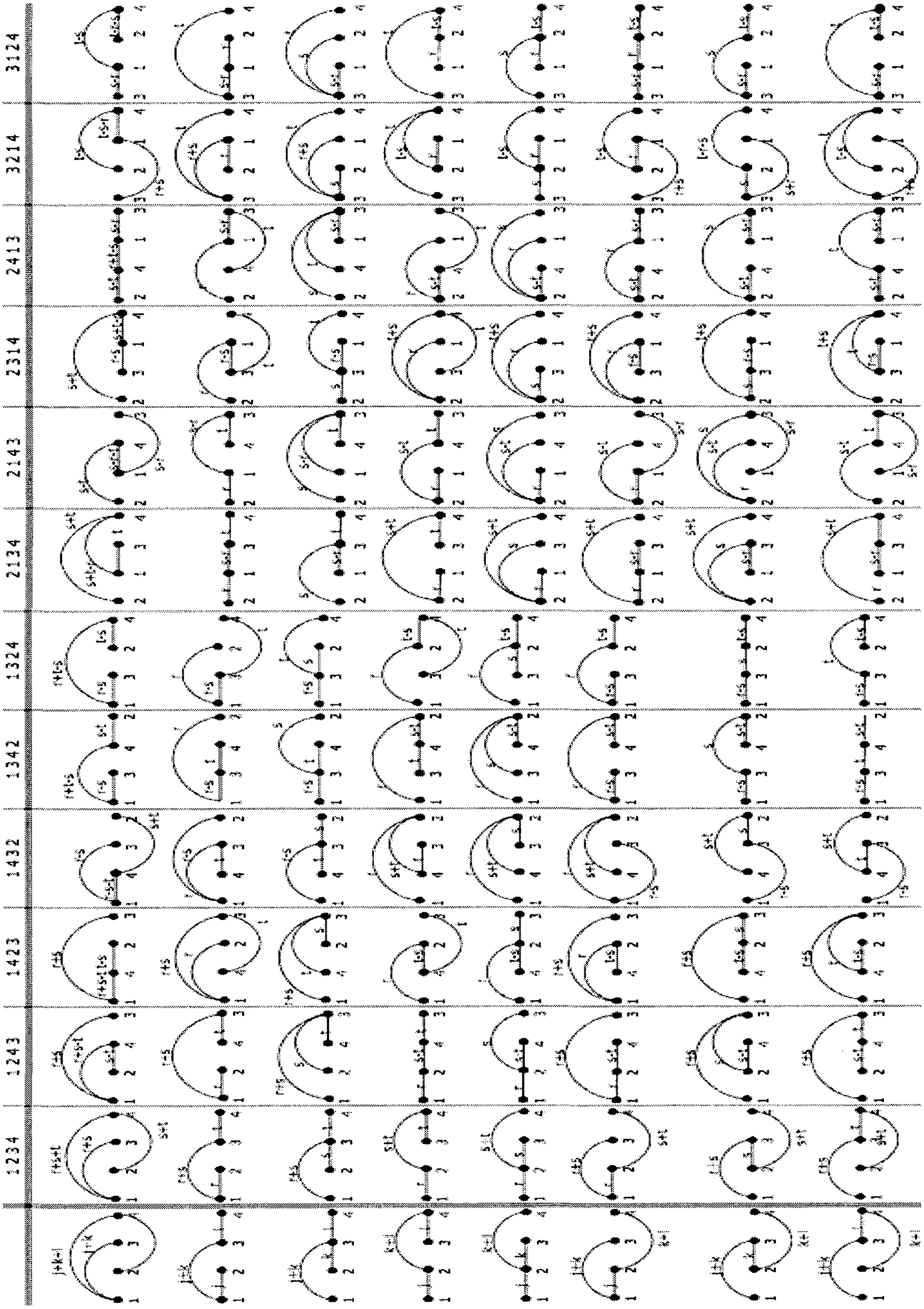


Figure 4.4: Cases to compute $N_4(n,\theta,\psi)$ (continued)

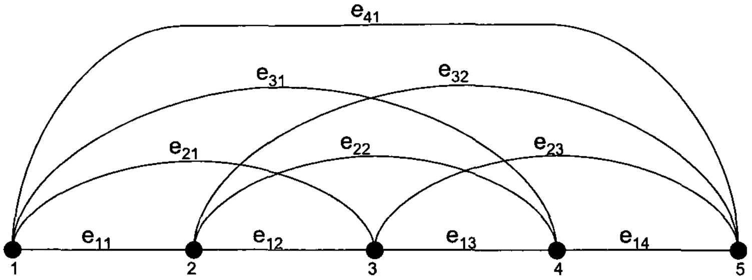


Figure 4.5: The index of edges in the adjacency graph

leftmost gene in the cluster. Figure 4.5 illustrates the e_{ij} definition for five genes in a cluster. We also represent by $|e_{ij}|$ the length of the edge e_{ij} which is the number of genes between the two genes connected by an edge, plus 1.

Obviously, there are constraints on $|e_{ij}|$ within a cluster. We observe in Figure 4.5 that:

1. $|e_{1k}| \geq 1,\ k \in \{1, 2, \ldots, m-1\}$
2. $|e_{(m-1)1}| \leq n-1,$
3. $|e_{ij}| = \sum_{k=j}^{i+j-1} |e_{1k}|$

where n is the number of genes in a genome and m is the number of genes in a cluster.

These motivate a general formula to compute each summation term in the formula of $E[N_m(n,\theta)]$, i.e.

$$\sum_{|e_{11}|=1}^{\theta-\mathcal{B}_1} \sum_{|e_{12}|=1}^{\theta-\mathcal{A}_2-\mathcal{B}_2} \cdots \sum_{|e_{1k}|=1}^{\theta-\mathcal{A}_k-\mathcal{B}_k} \cdots \sum_{|e_{1(m-1)}|=1}^{\theta-\mathcal{A}_{m-1}} \left(n - \sum_{k=1}^{m-1} |e_{1k}|\right) \tag{4.2.8}$$

where $\mathcal{A}_k$ and $\mathcal{B}_k$ are determined by the longest edge, say e_{ij}, covering the edge e_{1k}. $\mathcal{B}_k$ is the number of edges e_{1t} where $k < t < i+j$, i.e. e_{1t} is covered by e_{ij} and at the right of the edge e_{1k}. While $\mathcal{A}_k$ is the sum of the length of all edges which are

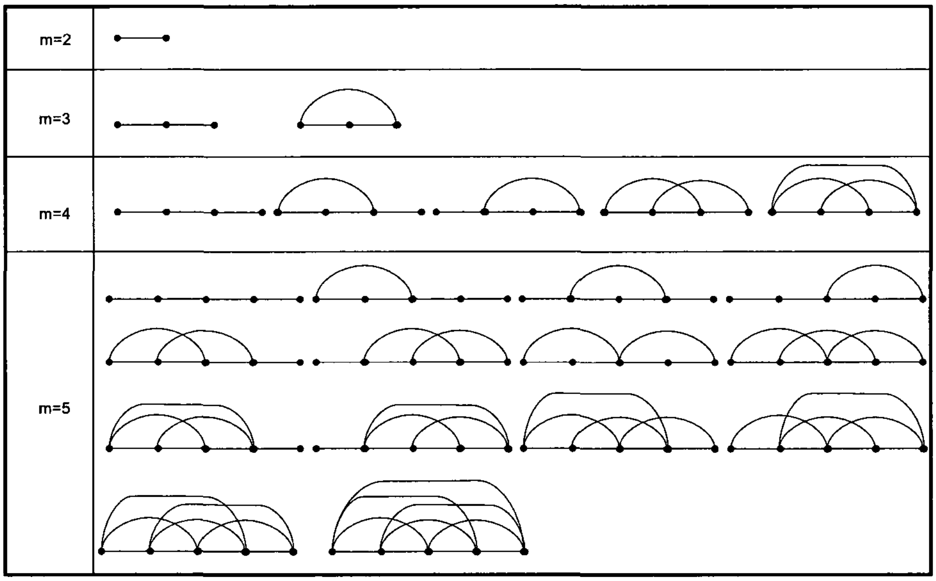


Figure 4.6: The cases used to compute the summation in $E[N_m(n,\theta)]$ calculation, where $m = 2,\ 3,\ 4,\ 5$

covered by the edge e_{ij} and at the left of the edge e_{1k}, i.e.

$$\mathcal{A}_k = \sum_{t=|e_{1j}|}^{|e_{1(k-1)}|} t. \tag{4.2.9}$$

For example, we can use the formula

$$\sum_{|e_{11}|=1}^{\theta-1} \sum_{|e_{12}|=1}^{\theta-|e_{11}|} (n - |e_{11}| - |e_{12}|) \tag{4.2.10}$$

instead of

$$\sum_{j=1}^{\theta-1} \sum_{k=1}^{\theta-j} (n - j - k) \tag{4.2.11}$$

to compute the summation of case $\mathcal{I}_2$ of Figure 4.1.

Hence, we can categorize all spanning tree of m genes based on the expression of formula (4.2.8). Figure 4.6 enumerates classes for $m = 2,\ 3,\ 4,\ 5$. So $E[N_m(n,\theta)]$ is the linear combination of choosing any two graphs, taking into account repeated occurrences of pairs of graph. So we reduce the summation computation from 9,216 cases to 25 cases for 4 genes in a cluster and from 1,800,000 cases to 196 cases for 5 genes in a cluster to get the $E[N_m(n,\theta)]$ as follows:

$$\begin{aligned} E[N_4(n,\theta,\theta)] &= \frac{\theta^2}{n(n-1)}(\frac{185}{9}\theta^4 - \frac{118}{3}\theta^3 - \frac{451}{18}\theta^2 + \frac{7}{18}\theta - \frac{14}{3}) + O(\frac{1}{n^3}) \\ E[N_5(n,\theta,\theta)] &= \frac{\theta^2}{n(n-1)(n-2)}(\frac{19337}{288}\theta^6 - \frac{13477}{72}\theta^5 + \frac{2389}{16}\theta^4 + \frac{406}{9}\theta^3 \\ &\quad -\frac{3519}{32}\theta^2 + \frac{3101}{72}\theta - \frac{395}{72}) + O(\frac{1}{n^4}) \end{aligned}$$

4.3 Percolation Threshold

Clustering procedures based on parametrised adjacency criteria, e.g., as in Refs. [16, 37], can have pathological behaviour as the criteria become less restrictive. At some point, called a *percolation threshold*, instead of large clusters being rare, they suddenly start to predominate and it becomes unusual *not* to find a large cluster.

To interpret this pathological behaviour, we simulated the probability that the size of the largest (θ,ψ)-adjacency gene cluster is larger than half of the genome size in 50,000 random pairs of genomes with size 100. Figure 4.7 shows these probabilities as a function of θ and ψ. Then beyond some contour on the (θ, ψ) plane, it becomes meaningless to test that the numbers or sizes of clusters exceed those predicted by the null hypothesis of random genomes.

It was established by Erdős and Rényi [8, 9, 10] that for random graphs where edges are independently present between pairs of the n vertices with probability p, the percolation threshold is $p = \frac{1}{n}$.

We note in Figure 4.8 that the percolation of the generalized adjacency graph is delayed considerably compared to unconstrained Erdős-Rényi graphs with the same

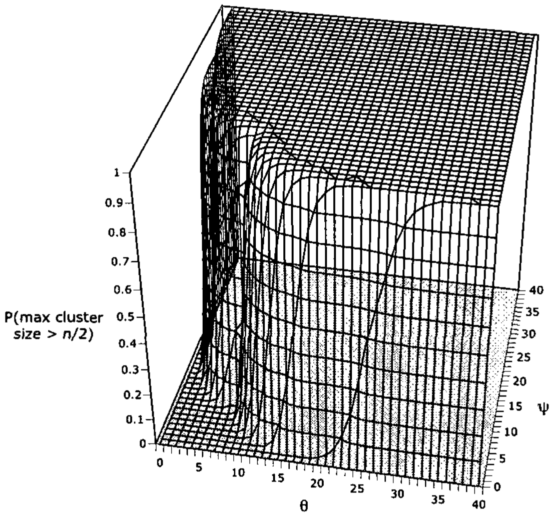


Figure 4.7: Proportion of simulations where size of the largest cluster $> \frac{n}{2}$, based on a samples of $50,000$ random permutations for $\theta, \psi = 1, 2, \ldots, 99$ and genome size $n = 100$

number of edges. To understand what aspect of the generalized adjacency graphs is responsible for this delay, we also simulated random graphs of bandwidth $\leq \theta$, since this bandwidth constraint is a property of generalized adjacency. It can be seen in Figure 4.8 that the limited bandwidth graphs also show delay in the percolation, but less than half that of generalized adjacency graphs.

As a control on our simulations, it is known (cf. [6]) that Erdős-Rényi graphs with rn edges, with r somewhat larger than $\frac{1}{2}$ have a cluster of size $(4r - 2)n$. Our percolation criterion is that one cluster must have at least $\frac{n}{2}$ vertices. Solving this, we get $r = 0.625$. This means that the $2\theta^2$ edges we use in each of our simulated graphs must be the same as $0.625n$, suggesting that $\theta = 0.56\sqrt{n}$, not far from the $0.61\sqrt{n}$ we found in our simulations.

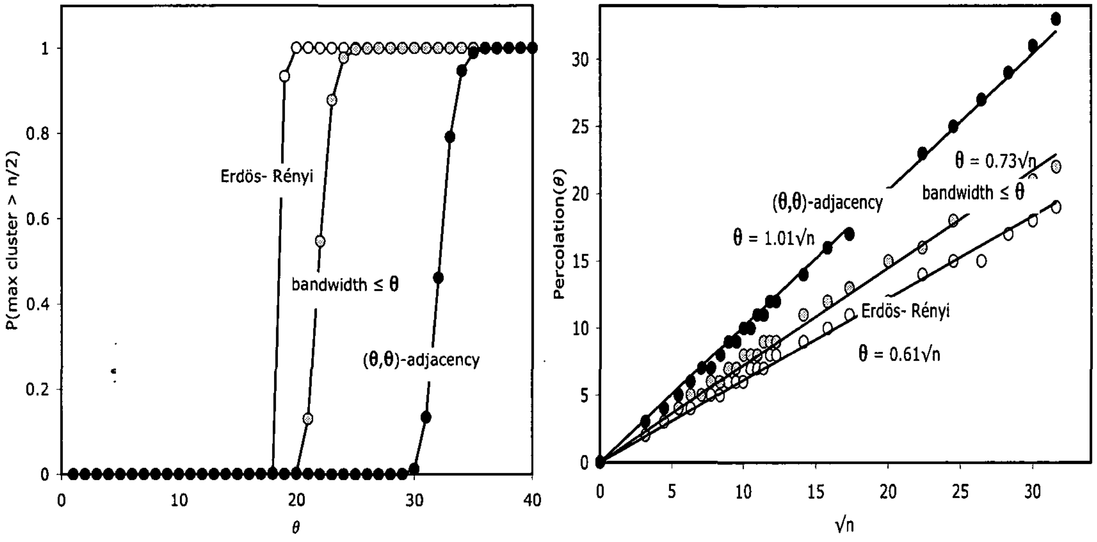


Figure 4.8: (left) Simulation with genome length $n = 1000$, with $2\theta^2$ edges in each graph, showing delayed percolation of generalized adjacency graphs with respect to Erdős-Rényi graphs. Bandwidth-limited graphs are also delayed but much less so. (right) Percolation point as a function of $\sqrt{n}$, again with $2\theta^2$ edges per graph. Delay measured by coefficient of $\sqrt{n}$ in equation for trend line.

Chapter 5

The Expected Number of (i, j)-Adjacent Gene Pairs in Two Genomes with Different Gene Contents

In previous chapters, the two genomes we are comparing have the same gene content. However, for real genomes, the gene content may not be the same. Moreover, most eukaryote genomes have multiple chromosomes. Even though they have the same gene contents, the genes could be in different chromosomes. To obtain generalized adjacency gene clusters, we may compare every pair of chromosomes in the two genomes. The gene content will generally differ. In this chapter, I will consider this more general case, but I still assume every gene has at most one copy in one genome, i.e. I do not consider gene duplication.

5.1 The Expected Number of (j, k)-Adjacent Gene Pairs in Two Genomes with Partially Different Gene Sets

5.1.1 Two Genomes Have Same Size n and m Genes in Common

Since the genome size n is the same, we first match the $n - m$ unrelated genes in the two genomes arbitrarily. Then we can start with our previous results and try to modify them for the present case.

To calculate the expected number of one-way (j, k) arbitrarily adjacent gene pairs of two random genomes S and T with n genes total and m genes shared, we first calculate it for the identity genome I and random genome R with n shared genes. We then choose m genes in I to construct subsequences randomly and calculate the expected number of one-way (j, k)-adjacent gene pairs in the subsequence of R for which the genes are the same as the genes in the subsequence of I. Thus let n_{jk} be the number of one-way (j, k)-adjacent gene pairs of two genomes S and T, i.e. j-adjacent in genome S and k-adjacent in genome T.

$$\begin{aligned}
E(n_{jk}) &= \frac{1}{\binom{n}{m}} \sum_{\mathcal{A}} \sum_{\substack{i \in \mathcal{A} \\ i+j \in \mathcal{A}}} P_n(i \overset{k}{\sim} i + j) \\
&= \frac{2(n-j)}{n(n-1)\binom{n}{m}} \sum_{\mathcal{A}} \sum_{\substack{i \in \mathcal{A} \\ i+j \in \mathcal{A}}} 1 \\
&= \frac{2(n-j)}{n(n-1)\binom{n}{m}} \cdot (n-k) \binom{n-2}{m-2} \\
&= \frac{2m(m-1)(n-j)(n-k)}{n^2(n-1)^2} \qquad (5.1.1)
\end{aligned}$$

where $\mathcal{A}$ is the set of all subsequences containing m genes in the identity sequence $(1,2,3,\ldots,n)$ and $P_n(i \overset{k}{\sim} i+j)$ is defined by Equation (3.3.11).

We can use the same idea to compute the expected number of either-way (j,k)-adjacent gene pairs of two genomes S and T with n genes and m shared genes. Because the events $i \overset{k}{\sim} i+j$ and $i \overset{j}{\sim} i+k$ are independent, we have

if $j = k$,

$$\begin{aligned} E(n_{jj}) &= \frac{1}{\binom{n}{m}} \sum_{\mathcal{A}} \sum_{\substack{i \in \mathcal{A} \\ i+j \in \mathcal{A}}} P_n(i \overset{j}{\sim} i+j) \\ &= \frac{2(n-j)}{n(n-1)\binom{n}{m}} \sum_{\mathcal{A}} \sum_{\substack{i \in \mathcal{A} \\ i+j \in \mathcal{A}}} 1 \\ &= \frac{2(n-j)}{n(n-1)\binom{n}{m}} \cdot (n-j)\binom{n-2}{m-2} \\ &= \frac{2m(m-1)(n-j)^2}{n^2(n-1)^2} \end{aligned} \tag{5.1.2}$$

where $\mathcal{A}$ is the set of all subsequences containing m genes in the identity sequence $(1,2,3,\ldots,n)$ and $P_n(i \overset{j}{\sim} i+j)$ is defined by Equation (3.3.11) with $k = j$;

while if $j \neq k$,

$$\begin{aligned} E(n_{jk}) &= \frac{1}{\binom{n}{m}} \left(\sum_{\mathcal{A}} \sum_{\substack{i \in \mathcal{A} \\ i+k \in \mathcal{A}}} P_n(i \overset{j}{\sim} i+k) + \sum_{\mathcal{A}} \sum_{\substack{i \in \mathcal{A} \\ i+j \in \mathcal{A}}} P_n(i \overset{k}{\sim} i+j) \right) \\ &= \frac{2}{n(n-1)\binom{n}{m}} \left((n-j) \sum_{\mathcal{A}} \sum_{\substack{i \in \mathcal{A} \\ i+k \in \mathcal{A}}} 1 + (n-k) \sum_{\mathcal{A}} \sum_{\substack{i \in \mathcal{A} \\ i+j \in \mathcal{A}}} 1 \right) \\ &= \frac{4m(m-1)(n-j)(n-k)}{n^2(n-1)^2} \end{aligned} \tag{5.1.3}$$

where $\mathcal{A}$ is the set of all subsequences containing m genes in the identity sequence $(1,2,3,\ldots,n)$, and $P_n(i \overset{j}{\sim} i+k)$ and $P_n(i \overset{k}{\sim} i+j)$ are defined by Equation (3.3.11).

Hence

$$E(n_{jk}) = \frac{\mu m(m-1)(n-j)(n-k)}{n^2(n-1)^2}, \quad \text{where } \mu = \begin{cases} 2, & j = k \\ 4, & j \neq k. \end{cases} \tag{5.1.4}$$

5.1.2 The Size of Two Genomes S and T are Different

Let two genomes S and T have n_S and n_T genes with m genes shared. Without loss of generality, we set $m \leq n_S \leq n_T$.

First, we consider $m = n_S$.

For example, let $n_S = 5$ and $n_T = 8$. And consider genomes $S = \{1\ 2\ 3\ 4\ 5\}$ and $T = \{3\ 2\ 6\ 4\ 5\ 1\ 8\ 7\}$. We add 6 7 8 to genome S so that S and T have the same size. Using our previous approach to calculating the expected number of one-way (j, k)-adjacent gene pairs of two genomes with same genome size, we get the following result based on Equation (3.3.11).

$$E'(n_{jk}) = \sum_{i=1}^{n_S-k} \frac{2(n_T-j)}{n_T(n_T-1)} = \frac{2(n_T-j)(n_S-k)}{n_T(n_T-1)} \tag{5.1.5}$$

$E'(n_{jk})$ is the expected value which is computed by two genomes with n_T genes and n_S shared genes. Taking into account the two subsets of genes, one has n_S genes and the other one has $n_T - n_S$ genes.

Therefore the final expectation value is

$$E(n_{jk}) = \frac{1}{\binom{n_T}{n_S}} \cdot \sum_{i=1}^{n_S-k} \frac{2(n_T-j)}{n_T(n_T-1)} \tag{5.1.6}$$

$$= \frac{2(n_T-j)(n_S-k)}{n_T(n_T-1)\binom{n_T}{n_S}} \tag{5.1.7}$$

Also, we can deal with the either-way (j, k)-adjacent gene pairs as follows:

if $j \neq k$

$$\begin{aligned} E(n_{jk}) &= \frac{1}{\binom{n_T}{n_S}} \left[\sum_{i=1}^{n_S-k} \frac{2(n_T-j)}{n_T(n_T-1)} + \sum_{i=1}^{n_S-j} \frac{2(n_T-k)}{n_T(n_T-1)} \right] && (5.1.8) \\ &= \frac{2[(n_T-j)(n_S-k)+(n_S-j)(n_T-k)]}{n_T(n_T-1)\binom{n_T}{n_S}} && (5.1.9) \end{aligned}$$

and if $j = k$

$$E(n_{jj}) = \frac{1}{\binom{n_T}{n_S}} \sum_{i=1}^{n_S-j} \frac{2(n_T-j)}{n_T(n_T-1)} = \frac{2(n_T-j)(n_S-j)}{n_T(n_T-1)\binom{n_T}{n_S}} \qquad (5.1.10)$$

So combining above results, for $1 \leq m \leq n_S \leq n_T$, the expected number of either-way (j,k)-adjacent gene pairs is

$$E(n_{jk}) = \begin{cases} \frac{2m(m-1)[(n_T-j)(n_S-k)+(n_S-j)(n_T-k)]}{n_T^2(n_T-1)^2\binom{n_T}{n_S}}, & j \neq k \\ \\ \frac{2m(m-1)(n_T-j)(n_S-j)}{n_T^2(n_T-1)^2\binom{n_T}{n_S}}, & j = k \end{cases} \qquad (5.1.11)$$

Also we can get the expected number of either-way (θ, ψ)-adjacency gene clusters

$$\begin{aligned} & E[N(n_S, n_T, m, \theta, \psi)] \\ = \ & \sum_{j=1}^{\theta} \sum_{k=1}^{\psi} E(n_{jk}) - \sum_{k=2}^{min(\theta,\psi)} \sum_{j=1}^{k-1} E(n_{jk}) \\ \overset{\theta \leq \psi \leq m}{=} \ & \sum_{j=1}^{\theta} \sum_{k=1}^{\psi} \frac{2m(m-1)[(n_T-j)(n_S-k)+(n_S-j)(n_T-k)]}{n_T^2(n_T-1)^2\binom{n_T}{n_S}} \\ & - \sum_{k=2}^{min(\theta,\psi)} \sum_{j=1}^{k-1} \frac{2m(m-1)[(n_T-j)(n_S-k)+(n_S-j)(n_T-k)]}{n_T^2(n_T-1)^2\binom{n_T}{n_S}} \\ & - \sum_{j=1}^{\theta} \frac{2m(m-1)(n_T-j)(n_S-j)}{n_T^2(n_T-1)^2\binom{n_T}{n_S}} \\ = \ & \frac{2m(m-1)}{n_T^2(n_T-1)^2\binom{n_T}{n_S}} \cdot [(2\psi\theta-\theta^2)n_S n_T \\ & + \frac{1}{2}\theta(\theta^2-\psi\theta-\psi^2+\theta-2\psi)(n_T+n_S) \\ & + \frac{1}{4}\theta(\theta+1)(2\psi^2+2\psi-\theta^2-\theta)] \qquad (5.1.12) \end{aligned}$$

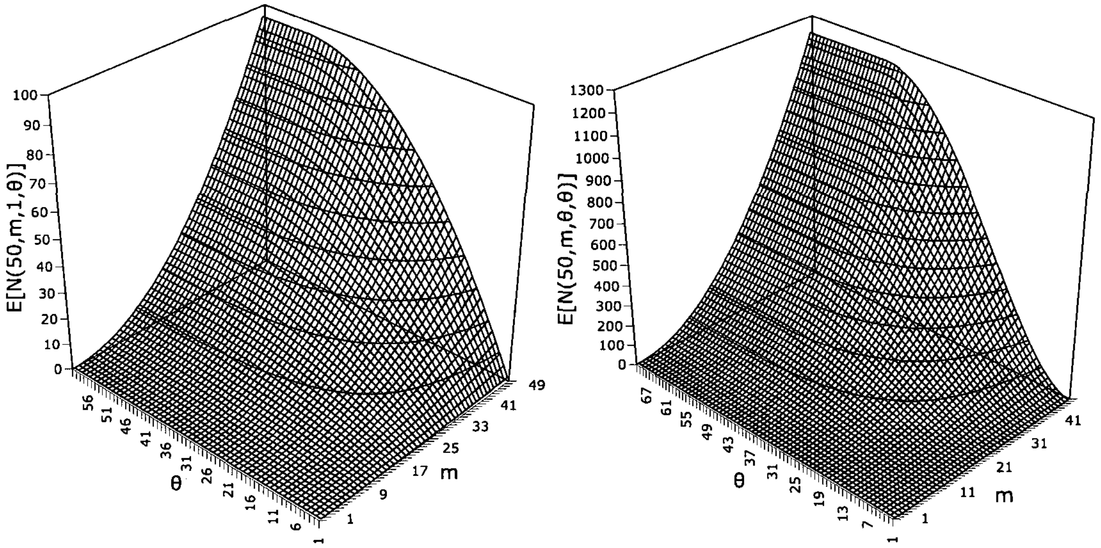


Figure 5.1: The expected number of either-way (1, θ)-adjacency and (θ,θ)-adjacency gene clusters.

If $1 \leq m \leq n_S = n_T = n$ and $1 \leq \theta = \psi \leq m$, we have

$$E[N(n,m,\theta)] = \frac{m(m-1)\theta^2(2n-\theta-1)^2}{2n^2(n-1)^2}. \tag{5.1.13}$$

If $1 < m \leq n_S = n_T = n$ and $\theta = 1,\ 1 \leq \psi \leq m$, we have

$$E[N(n,m,\psi)] = \frac{2m(m-1)[(2\psi-1)n^2 - (\psi^2+3\psi-2)n + (\psi^2+\psi-1)]}{n^2(n-1)^2}. \tag{5.1.14}$$

Figure 5.1 shows the expected number of either-way (1, θ)-adjacency gene clusters (left) and (θ,θ)-adjacency gene clusters (right) as a function of θ and shared genes number m where each genome has 50 genes in total.

Chapter 6

The Generalized Adjacency Model for Multiple Genome Comparison

In the previous chapters, the generalized adjacency model for the comparison of two genomes is studied and the "natural" parameter values were found based on the uniform weight system with cut-off point k^* and percolation threshold. In this chapter, I will extend this model to the comparison of three genomes. It is straightforward to extend the generalized gene cluster in three-genome comparison to multiple-genome comparison.

6.1 Definitions for Multiple Genome Comparison

Definition 6.1.1. *Let* Π_m *be a sequence of* m *integers, i.e.* $\Pi_m = \pi_1\pi_2\ldots\pi_m$ *and* $\mathcal{G}_m$ *be a permutation of* m *genomes,* $G_1, G_2, \ldots, G_m$. *We say two genes* x *and* y *of* $\mathcal{G}_m$ *are* ***one-way*** Π_m***-adjacent*** *in* $\mathcal{G}_m$ *if these two genes are* $x \overset{\pi_i}{\sim} y$ *in the genome* G_i *for all genomes in* $\mathcal{G}_m$. *Let* $\Pi'_m = <\pi_1\pi_2\ldots\pi_m>$[1]. *We define genes* x *and* y *of* $\mathcal{G}_m$ *to be* ***either-way*** Π'_m***-adjacent*** *in* $\mathcal{G}_m$ *if there exists a permutation* Π_m *constructed by elements* $\pi_1, \pi_2, \ldots, \pi_m$, *such that* x *and* y *are one-way* Π_m*-adjacent in* $\mathcal{G}_m$.

Definition 6.1.2. *For two sequences* S *and* T *with the same integer elements, let* S' *and* T' *be two decreasing sequences corresponding to* S *and* T. *Let* $S'[i]$ *(or* $T'[i]$*) be the* i^{th} *element of the sequence* S' *(or* T'*). We define the sequence* S *to be a* ***predecessor*** *of the sequence* T, *represent d by* $S \prec T$, *if there is an integer* i *such that* $S'[i] < T'[i]$ *and* $S'[j] = T'[j]$ *for all* $j < i$. *The sequence* T *is* ***successor*** *of the sequence* S, *represent d by* $T \succ S$, *if* $S \prec T$. *If* $S'[i] = T'[i]$ *for all* i, *then* S *and* T *are* ***equal****, represent d by* $S \asymp T$.

Definition 6.1.3. *Let* ω_{Π_m} *be the* ***weight*** *on two genes that are one-way* Π_m*-adjacent in* $\mathcal{G}_m$, *such that*

1. *all* ω_{Π_m}*s are non-negative and equal, where* Π_m *is a sequence constructed by element* $\pi_1, \pi_2, \ldots, \pi_m$
2. $\sum_{\pi_1=1}^{n-1}\sum_{\pi_2=1}^{n-1}\cdots\sum_{\pi_m=1}^{n-1}\omega_{\Pi_m} = 1$, *where* n *is the total number of genes in a genome*
3. $\omega_{\Pi_m} \geq \omega_{\Pi'_m}$ if $\Pi_m \prec \Pi'_m$

[1] "$<\pi_1\pi_2\ldots\pi_m>$" *means the decreasing sequence constructed by elements* $\pi_1, \pi_2, \ldots, \pi_m$.

Definition 6.1.4. *Let E_S^{θ} be the set of all i-adjacencies in S, where $1 \leq i \leq \theta$ and $\boldsymbol{\theta}_m$ be a sequence of θs, say $\{\theta_1, \theta_2, \dots, \theta_m\}$. We define a subset of $C \subseteq V$ to be a* ***one-way*** $\Pi_{\boldsymbol{\theta}_m}$ ***generalized adjacency gene cluster****, or* ***one-way*** $\Pi_{\boldsymbol{\theta}_m}$***-adjacency cluster****, if it consists of the vertices of a connected component of the* ***generalized adjacency graph****,* $G_{\mathcal{G}_m}^{\Pi_{\boldsymbol{\theta}_m}} = (V, \bigcap_{i=1}^{m} E_{G_i}^{\theta_i})$.

Definition 6.1.5. *Let E_S^{θ} be the set of all i-adjacencies in S, where $1 \leq i \leq \theta$ and $\boldsymbol{\theta}_m$ be a sequence of θs, say $\{\theta_1, \theta_2, \dots, \theta_m\}$. We define a subset of $C \subseteq V$ to be an* ***either-way*** $\Pi_{\boldsymbol{\theta}_m}$ ***generalized adjacency gene cluster****, or* ***either-way*** $\Pi_{\boldsymbol{\theta}_m}$***-adjacency cluster****, if it consists of the vertices of a connected component of the* ***generalized adjacency graph****,* $G_{\mathcal{G}_m}^{\Pi_{\boldsymbol{\theta}_m}} = (V, \bigcup_{\Pi_{\boldsymbol{\theta}_m}} \bigcap_{i=1}^{m} E_{G_i}^{\theta_i})$.

6.2 The Study of Generalized Adjacency Clusters for Three Genome Comparison

6.2.1 Definitions for Three Genome Comparison

Definition 6.2.1. *We say genes x and y of three genomes R, S and T are* ***one-way*** (i, j, k)***-adjacent*** *if they are i-adjacent in the genome R, j-adjacent in the genome S and k-adjacent in the genome T. Let $i \leq j \leq k$. We say x and y are* ***either-way*** (i, j, k)***-adjacent*** *if they are i-adjacent in one of three genomes, j-adjacent in another genome and k-adjacent in the remaining genome.*

Definition 6.2.2. *Let ω_{ijk} be the* ***weight*** *on two genes that are one-way (i, j, k)-adjacent, such that*

1. $0 \leq \omega_{ijk} = \omega_{ikj} = \omega_{jik} = \omega_{jki} = \omega_{kij} = \omega_{kji}, \;\; i, j, k \in \{1, 2, \dots, n-1\}$
2. $\sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \sum_{k=1}^{n-1} \omega_{ijk} = 1$

3. $\omega_{ijk} \geq \omega_{rst}$ if

(a) $i^{(1)} < r^{(1)}$

(b) $i^{(1)} = r^{(1)}$ *and* $i^{(2)} < r^{(2)}$

(c) $i^{(1)} = r^{(1)}$, $i^{(2)} = r^{(2)}$ *and* $i^{(3)} < r^{(3)}$

where $i^{(3)} \leq i^{(2)} \leq i^{(1)}$*, the set* $\{i^{(1)}, i^{(2)}, i^{(3)}\} = \{i, j, k\}$

and $r^{(3)} \leq r^{(2)} \leq r^{(1)}$*, the set* $\{r^{(1)}, r^{(2)}, r^{(3)}\} = \{r, s, t\}$.

Definition 6.2.3. *Let* E_S^θ *be the set of all i-adjacencies in* S*, where* $1 \leq i \leq \theta$*. We define a subset of* $C \subseteq V$ *to be a* **one-way** (θ, ϕ, ψ) **generalized adjacency gene cluster***, or* **one-way** (θ, ϕ, ψ)**-adjacency cluster***, if it consists of vertices of a connected component of the* **generalized adjacency graph***,* $G_{RST}^{\theta\phi\psi} = (V, E_R^\theta \cap E_S^\phi \cap E_T^\psi)$.

Figure. 6.1 illustrates how genomes $R = 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9$, $S = 2\ 1\ 5\ 7\ 8\ 3\ 6\ 4\ 9$ and $T = 1\ 2\ 6\ 7\ 8\ 4\ 9\ 5\ 3$ determine the one-way $(2, 3, 4)$-adjacency clusters, $\{1, 2\}$ and $\{3, 4, 5, 6, 7, 8\}$.

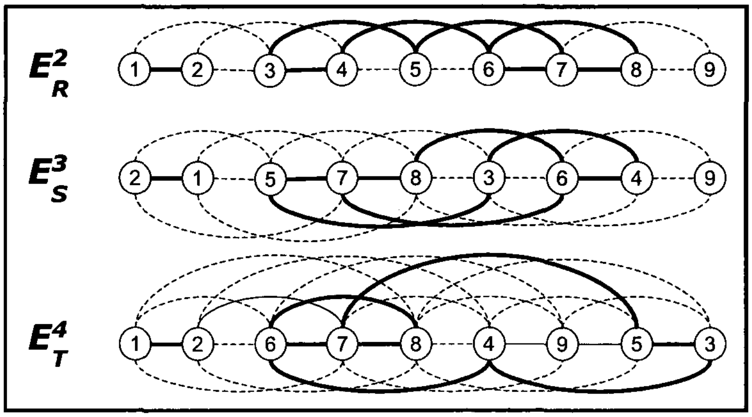


Figure 6.1: Determination of one-way $(2, 3, 4)$-adjacency clusters of genomes R, S, T.

Definition 6.2.4. *Let E^θ_S be the set of all i-adjacencies in S, where $1 \leq i \leq \theta$. We define a subset of $C \subseteq V$ to be an* **either-way (θ, ϕ, ψ) generalized adjacency gene cluster**, *or* **either-way (θ, ϕ, ψ)-adjacency cluster**, *if it consists of vertices of a connected component of the* **generalized adjacency graph**, *$G^{\theta\phi\psi}_{RST} = (V, (E^\theta_R \cap E^\phi_S \cap E^\psi_T) \cup (E^\theta_R \cap E^\psi_S \cap E^\phi_T) \cup (E^\phi_R \cap E^\theta_S \cap E^\psi_T) \cup (E^\phi_R \cap E^\psi_S \cap E^\theta_T) \cup (E^\psi_R \cap E^\theta_S \cap E^\phi_T) \cup (E^\psi_R \cap E^\phi_S \cap E^\theta_T))$.*

Figure. 6.2 illustrates how genomes $R = 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9$, $S = 2\ 1\ 5\ 7\ 8\ 3\ 6\ 4\ 9$ and $T = 1\ 2\ 6\ 7\ 8\ 4\ 9\ 5\ 3$ determine the $(2, 3, 4)$ clusters $\{1, 2\}$ and $\{3, 4, 5, 6, 7, 8, 9\}$.

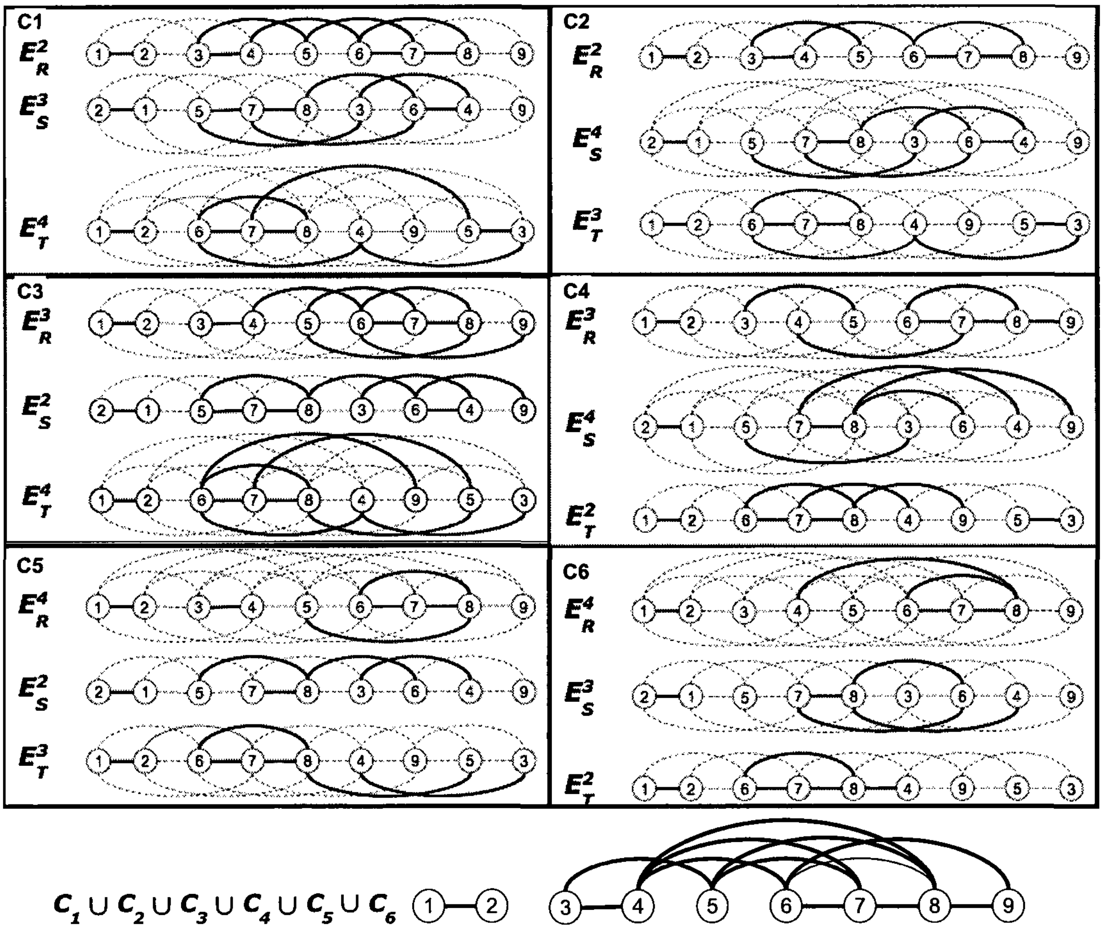


Figure 6.2: Determination of either-way $(2, 3, 4)$ clusters of genomes R,S,T.

Definition 6.2.5. *The* ***similarity function*** *of three genomes R, S and T is*

$$SIM(R,S,T) = \sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\sum_{k=1}^{n-1} n_{ijk}\omega_{ijk} - \sum_{i=1}^{n-1}\sum_{j=1}^{n-1}(n_{ijj}+n_{jij}+n_{jji})\omega_{ijj} + \sum_{i=1}^{n-1} n_{iii}\omega_{iii} \tag{6.2.1}$$

where ω_{xyz} is the weight of an (x,y,z)-adjacency; n_{xyz} is the number of gene pairs (g,h) that are x-adjacent on R, y-adjacent on S and z-adjacent on T.

6.2.2 The "Natural" Weight Function for Three Genome Comparison

Theorem 6.2.6. *Let $\alpha_k = \lfloor \frac{1}{3}(81k-1-9\sqrt{81k^2-2k})^{\frac{1}{3}} + \frac{1}{3}(81k-1-9\sqrt{81k^2-2k})^{-\frac{1}{3}} + \frac{2}{3} \rfloor \approx \lfloor \sqrt[3]{6k} \rfloor$. The either-way (i,j,k)-adjacency weight $\boldsymbol{\omega}$ that maximizes the similarity function* (6.2.1) *has*

$$\omega_{rst} = \begin{cases} \frac{1}{k^*}, & \text{if } 1 \le t \le s \le r < \alpha_{k^*}, \\ & \text{or } 1 \le s \le r = \alpha_{k^*}, 1 \le t \le s - \sqrt{s^2 - 2k^* + \frac{1}{3}r(r-1)^2} \\ & \\ 0, & \text{otherwise} \end{cases} \tag{6.2.2}$$

where k^* is a natural number and maximizes the function

$$f(k) \;=\; \frac{1}{k}\left(\sum_{r=1}^{\alpha_k-1}\sum_{s=1}^{r}\sum_{t=1}^{s} n'_{rst} + \sum_{s=1}^{\alpha_k}\sum_{t=1}^{s-\frac{1}{3}\sqrt{3s^2-6k+\alpha_k(\alpha_k-1)^2}} n'_{\alpha_k st}\right) \tag{6.2.3}$$

where $n'_{xyz} = n_{xyz} + n_{xzy} + n_{yxz} + n_{yzx} + n_{zxy} + n_{zyx}$ and n_{xyz} is the number of gene pairs x-adjacent on R, y-adjacent on S and z-adjacent on T. (See Figure 6.3 for the 3-dimensional volume measured by k^*.)

Proof. Similar to the proof of Theorem 3.2.5, we can transform the three-dimensional weight ω_{rst} to a non-increasing sequence and prove it is a uniform system. Now we will seek the cut-off point for the weights, i.e. k^*.

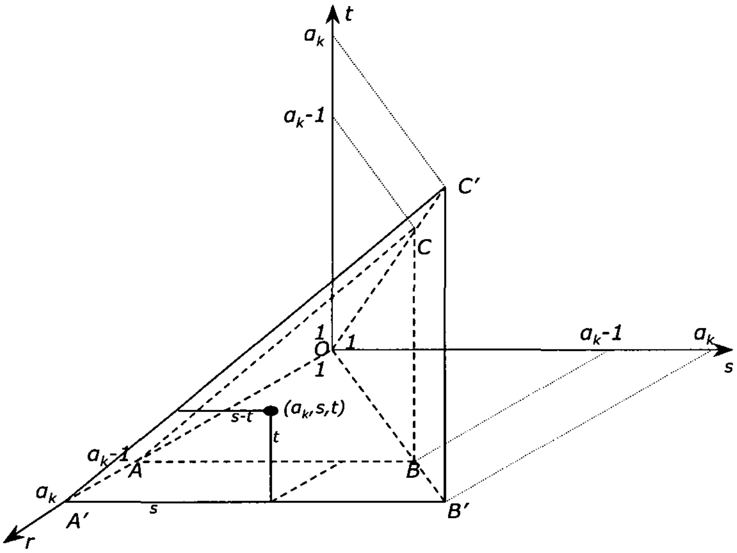


Figure 6.3: k is augmented starting at the origin, in the triangular pyramid. Values of ω_{rst} in other parts determined by symmetry.

In Figure 6.3, we obtain that the value of k is between the volume of the triangular pyramid $OABC$ and $OA'B'C'$, so we have

$$k = \frac{1}{6}r(r-1)^2 + \frac{1}{2}s^2 - \frac{1}{2}(s-t)^2,\ 1 \le t \le s \le r \le n-1 \tag{6.2.4}$$

and

$$\frac{1}{6}r(r-1)^2 \le k \le \frac{1}{6}r^2(r+1) \tag{6.2.5}$$

Solving Equation (6.2.5) , we obtain the bound of r,

$$r \ge \frac{1}{3}[(81k-1-9\sqrt{81k^2-2k})^{\frac{1}{3}} + (81k-1-9\sqrt{81k^2-2k})^{-\frac{1}{3}} - \frac{1}{3}],$$
$$r \le \frac{1}{3}[(81k-1-9\sqrt{81k^2-2k})^{\frac{1}{3}} + (81k-1-9\sqrt{81k^2-2k})^{-\frac{1}{3}} + \frac{2}{3}] \tag{6.2.6}$$

Because r and k are all natural numbers, r must be $\lfloor\frac{1}{3}(81k-1-9\sqrt{81k^2-2k})^{\frac{1}{3}} + \frac{1}{3}(81k-1-9\sqrt{81k^2-2k})^{-\frac{1}{3}} + \frac{2}{3}\rfloor$.

Consider s and t. We show by solving t in Equation (6.2.4) that

$$1 \le t \le s - \frac{1}{3}\sqrt{3s^2 - 6k + \alpha_k(\alpha_k - 1)^2} \tag{6.2.7}$$

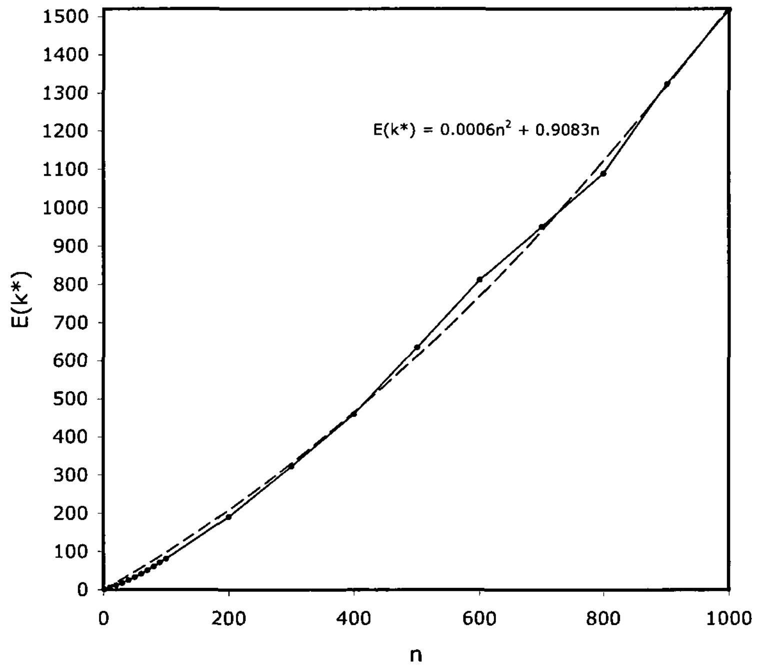


Figure 6.4: The expectation of k^* simulated by random permutation in three-genome comparison.

Hence, we set the conclusion of the theorem. □

This theorem shows that most sensitive values of the parameters, θ, ϕ and ψ in three-genome comparison should be $\theta = \phi = \psi = \lfloor \frac{1}{3}(81k - 1 - 9\sqrt{81k^2 - 2k})^{\frac{1}{3}} + \frac{1}{3}(81k - 1 - 9\sqrt{81k^2 - 2k})^{-\frac{1}{3}} + \frac{2}{3} \rfloor \approx \lfloor \sqrt[3]{6k} \rfloor$ to find the (θ, ϕ, ψ) generalized adjacency gene clusters. We can use the $E[k^*]$ to estimate k^*. Figure 6.4 illustrates the simulation result of the expected value of k^* (solid line) for genome sizes $n = 10, 20, \ldots, 100, 200, 300, \ldots, 1000$ by calculating k^* for each of $50,000$ random permutations as in Theorem 6.2.6, and simply taking the mean. We also display the trend of the expected value of k^* as a function of genome size n (dashed line).

In order to use the same method as in two-genome comparison to obtain k^*, we need know what the n_{rst}s are and whether they satisfy the same property as n_{ij} in the previous chapter, i.e. the total number of one-way (i,j)-adjacent gene pairs in two-genome comparison.

Theorem 6.2.7. *For three random genomes, R, S and T, with n genes, let n_{rst} be the total number of one-way adjacent gene pairs, (x,y), that are r-adjacent in genome R, s-adjacent in genome S and t-adjacent in genome T. Then n_{rst}, converges in distribution to the Poisson with parameter $\frac{4(n-r)(n-s)(n-t)}{n^2(n-1)^2}$*

Proof. Events $\{x \overset{s}{\sim} y\, in\, genome\, S \mid x \overset{r}{\sim} y\, in\, genome\, R\}$ and $\{x \overset{t}{\sim} y\, in\, genome\, T \mid x \overset{r}{\sim} y\, in\, genome\, R\}$ are independent. Based on Definition (3.3.10) and Equation (3.3.11), then we get

$$\begin{aligned}
& E\left(\{x \overset{r}{\sim} y \text{ } in \text{ } R\}, \{x \overset{s}{\sim} y \text{ } in \text{ } S\}, \{x \overset{t}{\sim} y \text{ } in \text{ } T\}\right) \\
= \;& P_n\left(\{x \overset{r}{\sim} y \text{ } in \text{ } R\}, \{x \overset{s}{\sim} y \text{ } in \text{ } S\}, \{x \overset{t}{\sim} y \text{ } in \text{ } T\}\right) \\
= \;& P_n\left(\{x \overset{s}{\sim} y \text{ } in \text{ } S\} \mid \{x \overset{r}{\sim} y \text{ } in \text{ } R\}\right) \cdot P_n\left(\{x \overset{t}{\sim} y \text{ } in \text{ } T\} \mid \{x \overset{r}{\sim} y \text{ } in \text{ } R\}\right) \cdot P_n\left(\{x \overset{r}{\sim} y \text{ } in \text{ } R\}\right) \\
= \;& \frac{4(n-s)(n-t)(n-r)}{n^2(n-1)^3} \qquad (6.2.8)
\end{aligned}$$

Therefore

$$\begin{aligned}
E(n_{rst}) &= (n-1) \cdot E\left(\{x \overset{r}{\sim} y \text{ } in \text{ } R\}, \{x \overset{s}{\sim} y \text{ } in \text{ } S\}, \{x \overset{t}{\sim} y \text{ } in \text{ } T\}\right) \\
&= \frac{4(n-r)(n-s)(n-t)}{n^2(n-1)^2}. \qquad (6.2.9)
\end{aligned}$$

Using the same idea as in Theorem 3.3.7, we conclude n_{rst} converges to the Poisson distribution with parameter $\frac{4(n-r)(n-s)(n-t)}{n^2(n-1)^2}$. □

Based on the proof of Theorem 3.2.5 and Theorem 6.2.6, we know that for the parameter value selection, we consider $1 \leq j \leq i \leq n-1$ in two-genome comparison and $1 \leq t \leq s \leq r \leq n-1$ in three-genome comparison to get the formula and the range of k because of the symmetry of weights (see Equation (3.2.19), (3.2.20), (6.2.4) and (6.2.5)) and find that the parameter values should be $\lfloor\sqrt{2k}\rfloor$ in two-genome comparison and $\lfloor\sqrt[3]{6k}\rfloor$ in three-genome comparison. Using the same idea, we can obtain k in m-genome comparison where $m > 3$, by considering $1 \leq i_1 \leq i_2 \leq \cdots \leq i_m \leq n-1$ and find the parameter value to be approximatly $\lfloor \sqrt[m]{m! \cdot k}\rfloor$. Thus, the parameter values of m-genome comparison, $\theta_1, \theta_2, \ldots, \theta_m$ should be $\lfloor \sqrt[m]{m! \cdot k}\rfloor$ to find a large number of generalized adjacencies, but not at the cost of unreasonably increasing the number of potential adjacencies.

Chapter 7

Conclusion and Open Questions

7.1 Conclusion

In this thesis, I first introduced the background of gene clustering (in Chapter 1) and some previous works of gene clustering (in Chapter 2). Then I considered two-genome comparison with the same gene content and introduced the definition of a two-parameter class of gene proximity criteria (in Chapter 3), where two genes are said to be one-way (or either-way) (i, j)-adjacent if they are separated by $i - 1$ genes on a chromosome in one (or either one) of the genomes and $j - 1$ genes in the other. We define a one-way (or either-way) (θ, ψ) cluster in terms of a graph where the genes are vertices and edges are drawn between those (i, j)-adjacent gene pairs where $\min(i, j) < \min(\theta, \psi)$ and $\max(i, j) < \max(\theta, \psi)$. The connected components of the graph are then the one-way (or either-way) (θ, ψ) clusters. The virtue of generalized adjacency clusters is that they embody gene order considerations within the cluster. In contrast to r-windows [7] and max-gap clusters [1, 16], generalized adjacency cannot have two genes close together in one genome but far apart in the other, although the cluster could be very large.

As with other criteria, the quantities θ and ψ would seem arbitrary parameters in our definition of a cluster. We remove some of this arbitrariness, by finding "natural values" for θ and ψ as a function of n, the total number of genes in the genomes. We find two such functions; the first trades off the expected number, across all pairs of genomes, of generalized adjacencies against the parameters θ and ψ, with lower parameter values considered more desirable, i.e., it is good to find a large number of generalized adjacencies, but not at the cost of including unreasonably remote adjacencies (in Chapter 3).

To do this we first define a wide class of similarities (or equivalently, distances) between two genomes in terms of weights on the (i, j)-adjacencies, namely any system of fixed-sum, symmetric, non-negative weights ω non-increasing in i and j. This is the most general way of representing decreasing weight with increasing separation of the genes on the chromosome. In any pair of genomes, we then wish to maximize the sum of the weights, which essentially maximizes the sensitivity of the criterion. Our main result is a theorem showing that the solution reduces to a uniform weight on gene separations up to a certain value of both θ and ψ, and zero weight on larger separations.

If we are willing to accept the loss of sensitivity, and prefer to search for clusters more widely dispersed on chromosomes, there is a second set of "natural" parameter values that serve as an upper bound on the meaningful choices θ and ψ (in Chapter 4). These values are the percolation thresholds of the (θ, ψ) clusters. Beyond these values, tests of significance are no longer meaningful because all clusters rapidly coalesce together. It is no longer surprising, revealing or significant to find large groups of genes clustering together, even in pairs of random genomes.

Percolation has been studied for max-gap clusters [16], but the main analytical results on percolation pertain to completely random (Erdős-Rényi) graphs. The graphs associated with (θ, ψ) clusters manifest delayed percolation, so the use of Erdős-Rényi percolation values would be a "safe" but conservative way of avoiding dangerously high values of the parameters. I show how to translate known results on Erdőss-Rényi percolation back to generalized adjacency clusters. I also introduced random bandwidth-limited graphs and use simulations to compare the delays of generalized adjacency and bandwidth-limited percolation with respect to Erdős-Rényi percolation in order to understand what structural properties of generalized adjacency are responsible for the delay.

After that, I extended my research to more general situations, i.e., two-genome comparison with different gene content (in Chapter 5) and multiple genome comparison (in Chapter 6). I gave the definitions and some results from both theoretical analysis and simulation.

7.2 Open Questions

Locating k^* analytically as a function of n

In this thesis, I obtained a graph of the expected value of k^* as a function of genome size n by simulating k^* for each of $50,000$ random permutations with some genome sizes n, for example $n = 1000, 2000, \ldots, 100000$ in two-genome comparison. It remains to locate k^* analytically as a function of n.

Exploring other structural properties

It would be interesting to find other structural properties, besides bandwidth, responsible for the delayed percolation of generalized adjacency graphs. In this thesis, I showed by simulation that bandwidth is partially responsible because the limited bandwidth graphs (Figure 4.8) show delay in the percolation. However, the delay shown in limited bandwidth graphs is less than half that in generalized adjacency graphs, which means bandwidth constraint is not the only thing responsible for the delayed percolation behaviour of generalized adjacency graphs.

Publication List

Aside from the normal supervisory and collaborative participation of my thesis director David Sankoff, I was responsible for all aspects of these papers except when specified

1. Z. Yang and D. Sankoff. Natural parameter values for generalized gene adjacency. In Ciccarelli, F.D. and Miklós, I. editors, *Proceedings of RECOMB 2009 Workshop on Comparative Genomics (RECOMB CG 2009)*, volume 5187 of *Lecture Notes in Bioinformatics*, pages 13-23. Springer, 2009.

2. D. Sankoff, C. Zheng, A. Muñoz, Z. Yang, Z. Adam, R. Warren, V. Choi and Q. Zhu. Issues in the Reconstruction of Gene Order Evolution.*Journal of Computer Science and Technology.*, 25(1): 10-25 Jan. 2010

   I provided models, theorems and simulations for section 6 of this paper. I also prepared the manuscript, especially technical write-ups and graphics in this section.

3. X. Xu, Z. Yang and David Sankoff. Statistics of Generalized Adjacency Gene Clusters. *Computers in Biology and Medicine*, submitted

   Most of the work in this paper was done by Ximing Xu. I provided simulation and prepared the manuscript of the Section 6.

4. Z. Yang and D. Sankoff. Natural parameter values for generalized gene adjacency. *Journal of Computational Biology*. submitted

Bibliography

[1] A. Bergeron, S. Corteel, and M. Raffinot. The algorithmic of gene teams. In D. Gusfield and R. Guigo, editors, *2nd Workshop on Algorithms in BioInformatics(WABI 2002)*, volume 2452 of *Lecture Notes in Computer Science*, pages 464–476. Springer-Verlag, 2002.

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

[3] G. Blin, C. Chauve, and G. Fertin. Gene order and phylogenetic reconstruction: application to gamma-proteobacteria. In A. McLysaght and D. H. Huson, editors, *Proceedings of RECOMB 2005 Workshop on Comparative Genomics (Berlin Heidelberg, 2005)*, volume 3678 of *Lecture Notes in Bioinformatics*, pages 11–20. Springer-Verlag, 2005.

[4] X. Chen, Z. Su, P. Dam, B. Palenik, Y. Xu, and T. Jiang. Operon prediction by comparative genomics: an application to the synechococcus sp. wh8102 genome. *Nucleic Acids Research*, 32(7):2147–2157, 2004.

[5] G. Didier. Common intervals of two sequences. In G. Benson and R. Page, editors, *3rd Workshop on Algorithms in BioInformatics(WABI 2003)*, volume 2812 of *Lecture Notes in Computer Science*, pages 17–24, 2003.

[6] R. D'Souza, D. Achlioptas, and J. Spencer. Explosive percolation in random networks. *Science*, 323:1453–1455, 2009.

[7] D. Durand and D. Sankoff. Tests for gene clusters. *Journal of Computational Biology*, 10:453–482, 2003.

[8] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[9] P. Erdős and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[10] P. Erdős and A. Rényi. On the strength of connectedness of a random graphs. *Acta Mathematica Scientia Hungary*, 12:261–267, 1961.

[11] M. D. Eroolaeva, O. White, and S. Salzberg. Prediction of operons in microbial genomes. *Nucleic Acids Research*, 5(29):1216–1221, 2001.

[12] W. M. Fitch. Distinguishing homologous from analogous proteins. *Syst Zool*, 19(2):99–113, 1970.

[13] W. M. Fitch. Homology: a personal view on some of the problems. *Trends Genet*, 16(5):227–231, 2000.

[14] S. Heber and J. Stoye. Algorithms for finding gene clusters. In O. Gascuel and B. M. E. Moret, editors, *1st Workshop on Algorithms in BioInformatics(WABI 2001)*, volume 2149 of *Lecture Notes in Computer Science*, pages 254–265, 2001.

[15] R. Hoberman and D. Durand. The incompatible desiderata of gene cluster properties. In A. McLysaght, editor, *Comparative Genomics, RECOMB International Workshop (2005)*, volume 3678 of *Lecture Notes in Computer Science*, pages 73–87. Springer-Verlag, 2005.

[16] R. Hoberman, D. Sankoff, and D. Durand. The statistical analysis of spatially clustered genes under the maximum gap criterion. *Journal of Computational Biology*, 12:1081–1100, 2005.

[17] M. A. Huynen and P. Bork. Measuring genome evolution. *Proceedings of the National Academy of Sciences*, 95(11):5849–5856, 1998.

[18] A. B. Kolsto. Dynamic bacterial genome organization. *Molecular Microbiology*, 24:241–248, 1997.

[19] J. Lawrence and J. R. Roth. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143:1843–1860, 1996.

[20] J. Nadeau and D. Sankoff. Counting on comparative maps. *Trends Genet*, 14(12):495–501, 1998.

[21] S. J. O'Brien, J. Wienberg, and L. A. Lyons. Comparative genomics: lessons from cats. *Trends Genet*, 10(13):393–399, 1997.

[22] R. Overbeek, M. Fonstein, M. Dsouza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6):2896–2901, 1999.

[23] S. Pasek, A. Bergeron, J.L. Risler, A. Louis, E. Ollivier, and M. Raffinot. Identification of genomic features using microsyntenies of domains : domain teams. *Genome Research*, 15:867–874, 2005.

[24] P. Pevzner and G. Tesler. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Research*, 13(1):37–45, 2003.

[25] M. N. Price, K. H. Huang, E. J. Alm, and A. P. Arkin. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Research.*, 33(3):880–892, 2005.

[26] N. Raghupathy and D. Durand. Individual gene cluster statistics in noisy maps. In A. McLysaght, editor, *Proceedings of RECOMB 2005 Workshop on Comparative Genomics (Berlin Heidelberg, 2005)*, volume 3678 of *Lecture Notes in Computer Science*, pages 106–120. Springer-Verlag, 2005.

[27] D. Sankoff. Rearrangements and chromosomal evolution. *Current opinion in genetics and development*, 13(6):583–587, 2003.

[28] D. Sankoff, V. Ferretti, and J.H. Nadeau. Conserved segment identification. *Journal of Computational Biology*, 4:559–565, 1997.

[29] D. Sankoff and J. H. Nadeau. Chromosome rearrangements in evolution: From gene order to genome sequence and back. *Proceedings of the National Academy of Sciences*, 100(20):11188–11189, 2003.

[30] M. Suyama and P. Bork. Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet*, 1(17):10–13, 2001.

[31] K. M. Swenson, M. Marron, J. V. Earnest-deyoung, and B. M. E. Moret. Approximating the true evolutionary distance between two genomes. In *Proc. 7th Workshop on Algorithm Engineering and Experiments*, pages 121–129. SIAM Press, 2005.

[32] J. Tamames. Evolution of gene order conservation in prokaryotes. *Genome Biology*, 6(2):0020.1–0020.11, 2001.

[33] J. Tamames, G. Casari, C. Ouzounis, and A. Valencia. Conserved clusters of functionally related genes in two bacterial genomes. *Journal of Molecular Evolution.*, 44:66–73, 1997.

[34] T. Uno and M. Yagiura. Fast algorithms to enumerate all common intervals of two permutations. *Algorithmica*, 26(2):290–309, 2000.

[35] J. Wolfowitz. Note on runs of consecutive elements. *Annals of Mathematics and Statistics*, 15:97–98, 1944.

[36] W. Xu, B. Alain, and D. Sankoff. Poisson adjacency distributions in genome comparison: multichromosomal, circular, signed and unsigned cases. *Bioinformatics*, 24(16):i146–i152, 2008.

[37] X. Xu and D. Sankoff. Tests for gene clusters satisfying the generalized adjacency criterion. In ALC Bazzan, M. Craven, NF. Martins, and Berlin, editors, *Proceedings of the Brazilian Symposium on Bioinformatics (BSB 2008).*, volume 5167 of *Lecture Notes in Bioinformatics*, pages 152–160. Springer, 2008.

[38] I. Yanai, J. C. Mellor, and C. Delisi. Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet*, 18(4):176–179, 2002.

[39] Q. Zhu, Z. Adam, V. Choi, and D. Sankoff. Generalized gene adjacencies, graph bandwidth, and clusters in yeast evolution. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 6(2):213–220, 2009.