Discrete mathematical modelling of biological processes

Momoko Hayamizu

Doctor of Philosophy

Department of Statistical Science
School of Multidisciplinary Sciences
SOKENDAI (The Graduate University for Advanced Studies)

Discrete mathematical modelling of biological processes

by

Momoko Hayamizu, M.D.

A thesis submitted to the Department of Statistical Science,
The Graduate University for Advanced Studies
for the degree of *Doctor of Philosophy*

March 2017

I would like to dedicate this thesis to my father,
Koichi Takahashi, who has invested so much in me.

“ Graph theory is a branch of mathematics which has applications in many areas: anthropology, architecture, biology, chemistry, computer science, economics, environmental conservation, psychology, and telecommunications, to name a few. The list goes on and on. In a typical situation, a problem arises in a real-world subject area that can be modeled using graphs. Then existing theorems or algorithms are used or new ones are developed to solve the original problem. ”

F. Buckley & F. Harary, *Distance in Graphs [7]*, 1990

Preface

While my original background is in biology and medicine, I have been working on different things at The Institute of Statistical Mathematics (ISM) since I entered my PhD programme in the Department of Statistical Science of The Graduate University for Advanced Studies in April 2014. At the outset, I intended to study statistical methods and machine learning techniques for biological data analysis, but my mind was changed after encountering the eye-opening book *"Phylogenetics"* [28] written by Charles Semple and Mike Steel. I knew nothing about discrete mathematics at that time, but found myself fascinated by the interaction between mathematics, statistics, computer science and biology, and ended up doing research in that field.

This thesis is a compilation of several results on discrete mathematical modelling of biological processes, most of which I obtained in my second year. There are a number of people without whom this thesis might not have been completed, but first I would like to thank my supervisor Professor Kenji Fukumizu for his patience and for his trust in me to work independently in all the time of research. I am also greatly indebted to the chair of my thesis committee, Professor Satoshi Ito, Vice Director-General of ISM for his constant encouragement and the rest of the committee: Professor Satoshi Kuriki for his support and astute guidance in writing this thesis; Dr. Shuhei Mano for his very careful reading and many helpful comments; and Dr. Yasuhide Numata, Shinshu University for his active interest in my doctoral work and for his kindness in agreeing to travel a long way to assess it.

Special thanks are due to Dr. Hiroshi Endo, Center for iPS Cell Research and Application, Kyoto University, who provided me with the inspiration to conceive the work in Part II and also contributed as a co-author to the introduction of Chapter 2. We have had many discussions on modelling of cellular differentiation since he contacted me at the end of 2012. One day he sent me several papers published in Nature Biotechnology in which algorithms for computing a minimum spanning tree (MST) were used to create a data-derived cellular tree model from distance data, namely, differences between sample cells in gene expression

patterns. He also found useful datasets online for me and suggested analysing them by a similar algorithm, so I worked on this kind of data analysis projects during my first year. As I came to see it, MST-based methods practically work well for most datasets but there is no guarantee that they always return a meaningful tree that adequately explains the original information. Knowing this, I began to wonder how to measure the validity of an automatically computed tree model and reached the notion of MST-like metric spaces. I have to admit that part of my research has grown in directions we did not initially conceive, but this thesis will, I believe, be an important step towards our larger goal.

Special thanks are also due to Professor Andrew R. Francis, Centre for Research in Mathematics, Western Sydney University for introducing me to the concept of tree-based networks. Part III of this thesis would not have come into being without his encouragement to write and publish [16] when I almost abandoned it.

I am grateful to Dr. Yoshimasa Uematsu for stimulating conversation that led me to the idea of the fourth-point condition that plays a key role in Part II; Dr. Ruriko Yoshida, Naval Postgraduate School who gave me the opportunity to present main results in Part II at a mini-symposium in SIAM Conference on Applied Algebraic Geometry in 2015 (AG'15) where I talked with Andrew for the first time; Dr. Katsutoshi Shinohara, Hitotsubashi University for constructive criticism that substantially improved an earlier version of Chapter 3; Dr. Masahiro Hachimori, Tsukuba University and Professor Yasuko Matsui, Tokai University who invited me to talk about the work in this thesis at the combinatorial mathematics seminar (COMA SEMI).

My thanks are extended to everyone in ISM for making my life here an enjoyable one. A very special thank you is due to my husband Yuto for his immense support throughout my PhD journey.

Momoko Hayamizu

Tokyo, January 2017

Abstract

This thesis treats discrete mathematical problems that arise in modelling of biological processes. It consists of the following three parts: Part I General introduction (Chapter 1); Part II Cellular differentiation (Chapter 2 and 3); and Part III Reticulate evolution (Chapter 4).

Part I briefly overviews classical work in the field of mathematical phylogenetics and outlines the motivation for the work in the other parts of the thesis.

Part II introduces and investigates the concept of minimum spanning tree (MST) metric spaces in connection with computational modelling of cellular differentiation. Chapter 2 begins with some necessary biological background and goes into the description of our problem. We say that a finite metric space, M, is an MST metric space if an MST preserves all pairwise distances between points in M. The problem is a characterisation of MST metric spaces, that is, to determine a necessary and sufficient condition on M to ensure that M is an MST metric space. We solve this problem by introducing a fourth-point condition and combining it with the classical four-point condition. Chapter 3 discuss the same problem but from a different perspective to shed new light on the notion of tree-like metric spaces. Here we do not use the four-point condition but make the simple assumption, also known as the tie-breaking rule, that the pairwise distances are all distinct. Under the tie-breaking rule, we show that M is an MST metric space if and only if it satisfies the fourth-point condition. Thus we provides a simpler characterisation of MST metric spaces, from which we can derive a measure of the MST-likeness of a finite metric space.

Part III considers a combinatorial problem concerning reticulate evolution. We provide some basics on the concept of tree-based networks that was recently introduced by Francis and Steel in [12] and define universal tree-based networks to state their question formally. Francis and Steel found a universal tree-based network with three labelled leaves and asked if there exists such a network in general. We settle this problem in the affirmative by proving that there are infinitely many universal tree-based networks with n labelled leaves for all $n > 1$.

Contents

Part I

General introduction

Chapter 1

Phylogenetic trees and tree metrics

Phylogenetic trees have been used in evolutionary biology as a standard model to depict relationships between species and have also been investigated in different areas of mathematics. In this chapter, we summarise without proofs the relevant materials on phylogenetic trees from which the motivation for this thesis stems.

1.1 Terminology

Let us start with a few basic definitions although the chapters of this thesis are rendered as self-contained as possible to facilitate access to the individual topics. We adopt the terminology of Semple and Steel [28]. Throughout this chapter, X denotes a non-empty finite set of $|X|$ different species (or 'taxa').

Definition 1.1. A *partially labelled tree $\mathcal{T}$ (on X)* is an ordered pair $(T;\phi)$, where T is an unlabelled tree with vertex set V and $\phi : X \to V$ is a map with the property that, for each $v \in V$ of degree at most two, $v \in \phi(X)$.

The set X is referred to as a *label set* and the map ϕ is called a *labelling map*.

Definition 1.2. A partially labelled tree $\mathcal{T} := (T;\phi)$ on X is said to be a *phylogenetic tree (on X)* if ϕ is a bijection from X into the set of leaves of T. If, in addition, every interior vertex of T has degree three, $\mathcal{T}$ is called a *binary phylogenetic tree (on X)*.

The notion of *partially labelled trees* is a mathematically convenient generalisation of phylogenetic trees. The above definitions make sense for weighted trees as well. In what follows, we assume that $T := (V, E)$ is an edge-weighted tree associated with a positive real-valued weighting $w : E \to \mathbb{R}^+$. The distance d_T in T is defined to be the shortest path metric in T as usual, and given a partially labelled tree $\mathcal{T} := (T, \phi)$ on X, let $d_{\mathcal{T}}(x, y) := d_T(\phi(x), \phi(y))$ for all $x, y \in X$.

1.2 Fundamental theorem of phylogenetics

In a typical setting of phylogenetic inference, we are given the differences between species, which are measured in terms of genetic or morphological traits, and wish to represent the dissimilarities by some phylogenetic tree. Then, it is natural to first check whether the observed dissimilarities can be precisely realised by a phylogenetic tree. To state the problem more formally, we need some definitions as follows.

Definition 1.3. An arbitrary function $\delta : X \times X \to \mathbb{R}$ is said to be a *dissimilarity map* if for all $x, y \in X$, $\delta(x, x) = 0$ and $\delta(x, y) = \delta(y, x)$. In particular, a dissimilarity map δ on X is said to be a *pseudometric* on X if it is non-negative and satisfies the triangular inequality.

Definition 1.4. A dissimilarity map δ is a *tree metric (on X)* if there is a partially labelled tree $\mathcal{T} := (T; \phi)$ on X associated with a positive real-valued weighting $w : E(T) \to \mathbb{R}^+$ such that, for all $x, y \in X$, $\delta(x, y) = d_{\mathcal{T}}(\phi(x), \phi(y))$.

The problem is to find a necessary and sufficient condition on an arbitrary dissimilarity map δ that ensures δ is a tree metric on X. Buneman completely settled it in [9] by introducing the *four-point condition* (Definition 1.5) and Theorem 1.8.

Definition 1.5. A dissimilarity map δ on X is said to satisfy the *four-point condition* if, for every four (not necessarily distinct) elements $p, q, r, s \in X$,

$$\delta(p, q) + \delta(r, s) \leq \max\{\delta(p, r) + \delta(q, s), \delta(p, s) + \delta(q, r)\}.$$

Remark 1.6. The above is equivalent to saying that two of the three sums $\delta(p, q) + \delta(r, s)$, $\delta(p, r) + \delta(q, s)$, and $\delta(p, s) + \delta(q, r)$ are equal and not less than the third.

Remark 1.7. Because the elements $p, q, r, s \in X$ are not necessarily distinct, the four-point condition implies the triangular inequality.

Theorem 1.8 (Buneman [9])**.** *Let δ be a dissimilarity map on X. Then, δ is a tree metric on X if and only if δ satisfies the four-point condition.*

Once we have checked that a dissimilarity map δ on X is a tree metric, our next concern will be whether δ uniquely specifies its partially labelled tree representation $\mathcal{T}$. The following theorem ensures the uniqueness of $\mathcal{T}$, up to isomorphism. The interested reader may refer to Buneman's earlier paper [8] and Theorem 7.1.8 of [28] for proofs, and to [19] for a treatment of a more general case.

Theorem 1.9. *Let δ be a tree metric on X. Then, up to isomorphism, there is a unique partially labelled tree $\mathcal{T}$ that realises δ.*

Although Theorem 1.9 was originally due to Buneman [8], it seems to be a folklore fact in theoretical evolutionary biology today. In fact, Theorem 1.8, together with Theorem 1.9, is sometimes referred to as the 'fundamental theorem' of phylogenetics.

1.3 Geometric measure of tree-likeness

In the previous section we have considered when a dissimilarity map δ on X can be precisely realised by a partially labelled tree on X. However, dissimilarity maps coming from real-world data do not exactly satisfy the four-point condition usually, so in realistic situations, we need to approximate δ by a partially labelled tree on X.

Then, it would be natural to discuss a method to evaluate the tree-likeness of δ. Interestingly, this question is not merely important for biologists but also mathematically worthwhile; in fact, tree-like metrics have been well studied in pure and applied geometry since the pioneering work of Gromov [13]. Although a full discussion on his theory is, of course, beyond the scope of this thesis, we would like to touch on an illuminating result that answers the above question. We refer the reader to Chapter 6 of his original paper [13] for details and to [6] (Proposition 7.3.1) for a recent exposition. See also p. 178–9 of [28] on which the content of this section is based.

For an arbitrary dissimilarity map δ on X, the *hyperbolicity* of δ is defined to be the largest violation of the four-point condition and is denoted by $hyp(\delta)$. Namely,

$$hyp(\delta) := \max_{p,q,r,s \in X} \{\delta(p,q) + \delta(r,s) - \max\{\delta(p,r) + \delta(q,s), \delta(p,s) + \delta(q,r)\}\}.$$

We can compute the $hyp(\delta)$ in $O(|X|^4)$ time by comparing the left-hand side with the right-hand side of the inequality in Definition 1.5 for all quartets. Surprisingly, as the next theorem says, the value represents more than how it is defined:

Theorem 1.10 (Gromov [13])**.** *For any pseudometric δ on X, there is a tree metric $d_{\mathcal{T}}$ on X with*

$$d_{\mathcal{T}} \leq \delta \leq d_{\mathcal{T}} + (1 + \log_2 |X|) hyp(\delta).$$

In summary, $hyp(\delta)$ can be computed in polynomial time, and it tells us how well δ can be approximated by tree metrics when δ is a pseudometric. This is why we can use it as a quantitative measure of the tree-likeness of δ. We note, however, that it remains challenging to compute $hyp(\delta)$ for large $|X|$ efficiently (see, *e.g.*, [11]), which makes it difficult to be applied to large-scale graphs.

1.4 The motivation for this thesis

We have reviewed discrete mathematical results that had a major impact over theoretical evolutionary biology and its spillover effect on pure mathematics. However, the existing concepts and theorems are not adequate when we wish to describe various structures by using other classes of graphs than phylogenetic trees. We therefore wish to develop new ones and help to lay the foundations for discrete mathematical modelling of various biological processes.

In Part II, we consider modelling of cellular differentiation. In contrast to Part I where we have dealt with partially labelled trees, we need to use a *fully* labelled tree to describe the pairwise distances between cells. We discuss the concept of 'tree-like' metric spaces differently from the way we have done thus far and provide a new condition other than the classical four-point condition.

Part III focuses on *reticulate evolution*, *i.e.*, a complex evolutionary process that cannot be adequately represented by a phylogenetic tree. It is well known that this kind of evolution commonly occurs in large groups organisms including bacteria, fungi and plants, yet mathematical studies for modelling reticulate evolution is still at an early stage [20]. We examine an interesting idea called *tree-based networks* that was recently introduced in [12]. Although tree-based networks are interesting in that they are able to describe more complicated evolutionary relationships than phylogenetic trees can, little is known about their mathematical nature. In order to better understand it, we closely look at a combinatorial problem regarding tree-based networks.

Part II

Cellular differentiation

Chapter 2

A characterisation of minimum spanning tree metric spaces

In this chapter, we look at an emerging issue in computational cell biology and then examine a discrete mathematical problem underlying it. Evolution and cellular differentiation seem alike in that both are modelled by trees, but there is an important difference. In phylogenetic inference, a partially labelled tree serves as a reasonable model because we can only have data of extant species and have to hypothesise about extinct ones. By contrast, a model of cellular differentiation should be *fully labelled* because data are collected from all cells of interest. We therefore restrict our attention to finite metric spaces that can be realised by fully labelled trees, which we call *minimum spanning tree metric spaces*. Our main question is to characterise this type of metric spaces. The four-point condition is obviously no longer sufficient as fully labelled trees comprise a special case of partially labelled ones, so we introduce a *fourth-point condition* to complement with it.

2.1 Introduction

Classical methods for the minimum spanning tree (MST) problem have gained increasing popularity as a data analysis tool across different disciplines of biology. In fact, algorithms such as Kruskal's and Prim's have been frequently used in molecular epidemiology to elucidate genetic relationships among bacteria [27], and more recently have also attracted much attention for their potential to revolutionize the current understanding of cellular differentiation, as we now explain.

Cellular differentiation refers to the process by which a less specialized cell becomes a more specialized one. As illustrated in Figure 2.1, stem cells are capable of differentiating into any type of cells, but once a stem cell has begun to differentiate, it gradually loses this ability and proceeds through intermediate stages, and ends up becoming a terminally differentiated cell type.

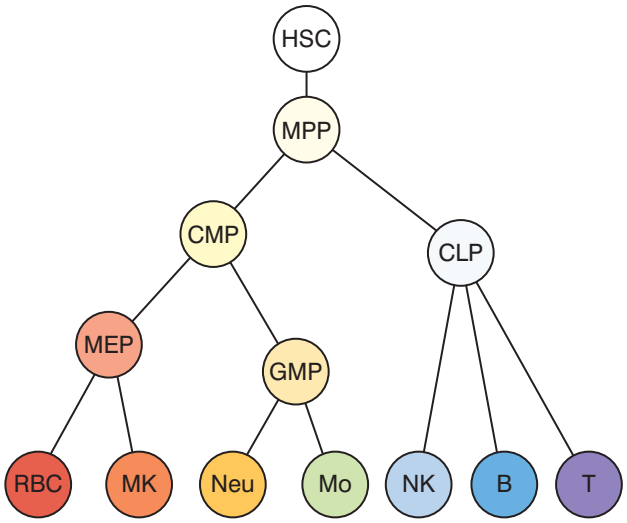

Figure 2.1: The traditional model for the differentiation of blood-related cells [1]. A hematopoietic stem cell (HSC) is placed at the apex for its potential to differentiate into any other cell type. The internal vertices of the tree signify cells at intermediate stages of differentiation, and the seven leaves represent terminally differentiated cells. MPP: multipotent progenitor; CMP: common myeloid progenitor; CLP: common lymphoid progenitor; MEP: megakaryocyte-erythroid progenitor; GMP: granulocyte-macrophage progenitor; RBC: red blood cell; MK: megakaryocyte; Neu: neutrophil; Mo: monocyte; NK: natural killer cell; B: B-lymphocyte; T: T-lymphocyte.

Although the essence of the phenomenon can be described by a tree, research on distance-based cellular tree construction is still at a very early stage because it has only recently become possible to calculate cell-to-cell distances. Unlike the process of evolution of organisms, cellular differentiation does not involve a change in the genome of a cell. Therefore, the differentiation status of a cell (*i.e.*, the cell type it is becoming and the degree of its maturity) is defined by factors other than the genome, such as the transcriptome, epigenome, and proteome, but such "omics" data of an individual cell have never been available until the recent emergence of single-cell transcriptome profiling technology. Since then, it has been feasible to measure the expression of thousands of genes in each

cell [25], and this has finally enabled us to quantify distances between cells based on differences in gene expression patterns.

Thus, algorithms for the MST problem have naturally found their applications in stem cell biology. For m genes and n individual cells, the gene expression profile of the i-th cell is represented by an m-dimensional vector x_i ($i = 1, \cdots, n$), and the pairwise distances between expression profiles are calculated using a distance function of choice and are stored in an $n \times n$ distance matrix D. Given D as an input, solving the MST problem yields a spanning tree T that extracts the $n - 1$ closest pairs of cells. It then makes sense to use MSTs for the purpose of data-driven cellular tree construction (*e.g.*, [24]). In fact, MST-based methods are not only plausible but already revealing biologically intriguing insights (*e.g.*, [14, 26, 32]).

However, a fundamental issue to be clarified is how to judge whether T is a good model to represent D. The answer to this question is not always straightforward, since there is no criterion for measuring the goodness-of-fit between D and T. Although the four-point condition, which we will discuss in Section 2.4.1, is a well-known characterisation for when D can be represented by a tree, it does not tell us whether D can be represented by a *spanning* tree. Also, one can create a distance matrix D_T from T by using the shortest path metric in T and calculate $\|D - D_T\|_p$ to compare the matrices D and D_T, but a larger discrepancy between D and D_T measured in L_p norm does not imply a greater deviation of T from the data; the value of $\|D - D_T\|_p$ overestimates differences in weights of internal edges compared to those of terminal edges of T.

Motivated by this—and inspired by the central role the four-point condition plays in the theory of δ-hyperbolic metric spaces [13, 28]— we seek for a mathematical expression presented as an equality or an inequality that could lead to criteria for measuring the 'spanning tree-likeness' of a finite metric space. Therefore, our primary goal here is to determine when a distance matrix of size n can be represented by a fully labelled tree on n vertices (Problem 2.8). We will provide an answer to this question by proving Theorem 2.18, and also show how the result is related to the MST problem in Section 2.6.

2.2 Definitions and notation

Throughout this chapter, X denotes a finite set $\{x_1, \cdots, x_n\}$ of n distinct elements, which is called a *label set*. A label set X may consist of any kind of objects. For example, suppose an element x_i of X is an m-dimensional vector that represents expression measurements of m genes within an individual cell i.

2.2.1 Metric spaces

Definition 2.1. Given a set S, a function $d_M : S \times S \to \mathbb{R}$ is said to be a *metric* on S if, for all $x, y, z \in S$, the following conditions hold:

1. $d_M(x, y) \geq 0$ (non-negativity);
2. $d_M(x, y) = 0 \Leftrightarrow x = y$ (identity of indiscernibles);
3. $d_M(x, y) = d_M(y, x)$ (symmetry);
4. $d_M(x, y) \leq d_M(x, z) + d_M(z, y)$ (triangule inequality).

A finite set X equipped with a metric d_M is said to be a *finite metric space*, and is denoted by (X, d_M). Once we have chosen a metric d_M on X, we can measure the pairwise distance $d_M(x_i, x_j)$ between gene expression profiles of cell i and cell j. The square matrix D of order n with $D(i, j) := d_M(x_i, x_j)$ is called a *distance matrix*.

Definition 2.2. Given two distinct points x and x' in a finite metric space (X, d_M), the *closed metric interval* $I(x, x')$ between them is defined to be the set

$$I(x, x') := \{x_i \in X : d_M(x, x') = d_M(x, x_i) + d_M(x_i, x')\}.$$

2.2.2 Graphs

All graphs in this chapter are finite, simple, connected, and undirected, and positive weighted. An edge of a graph that joins two vertices x and y is denoted by xy. Given a graph G, the sets of vertices and edges are denoted by $V(G)$ and $E(G)$, respectively. Given a label set X and an unlabelled graph U, a vertex labelling of U is specified by a map $\phi : X \to V(U)$. The map ϕ is called a *labelling map*, and the resulting labelled graph is said to be a graph (*on* $V(U)$) *labelled by* X. A graph labelled by X is denoted by $(V, E; X, \phi, w)$ for a set V of unlabelled vertices, a set E of edges, a vertex-labelling map $\phi : X \to V$, and an edge-weighting function $w : E \to \mathbb{R}^+$. Note that ϕ is not necessarily surjective (*i.e*, some vertices are labelled, but not necessarily all) and that w is strictly positive. The distance in G is defined to be the shortest path metric in G, and is denoted by d_G.

A graph is called a *tree* if it is connected and it has no cycle. All trees considered here are unrooted. If a graph G is a tree, there is a unique path that joins two vertices x and y in G, which is represented using $[x, \cdots, y]$; in particular, we use $[x, i, \cdots, y]$ to mean that a vertex i is contained in the path and that i is adjacent to x.

Definition 2.3. Assume X is a label set. Two graphs $G_i := (V_i, E_i; X, \phi_i, w_i)$ $(i = 1, 2)$ labelled by X are said to be *isomorphic* (*as vertex-labelled, edge-weighted graphs*) if there is a one-to-one correspondence $f : V_1 \to V_2$ that satisfies the following:

- for any two distinct vertices $x, y \in V_1$, $xy \in E_1$ if and only if $f(x)f(y) \in E_2$;
- for any $xy \in E_1$, $w_1(xy) = w_2(f(x)f(y))$;
- $\phi_2 = f \circ \phi_1$.

Definition 2.4. Assume $M := (X, d_M)$ is a finite metric space, and suppose $G := (V, E; X, \phi, w)$ is a graph.

- The labelling map $\phi : X \to V$ is said to be *distance-preserving* if, for all $x, y \in X$,
$$d_G(\phi(x), \phi(y)) = d_M(x, y).$$
- The graph G is said to be a *fully labelled graph representation of* M if both of the following conditions hold:
  1. ϕ is a distance-preserving labelling map;
  2. $\phi : X \to V$ is bijective.

Remark 2.5. The condition 1 in Definition 2.4 implies that $\phi : X \to V$ is injective (otherwise, the identity of indiscernibles in Definition 2.1 would not hold).

Definition 2.6. Given a finite metric space M, a *complete graph representation* K_M *of* M is defined to be a complete graph that is a fully labelled graph representation of M.

Definition 2.7. Given a finite metric space M, a *fully labelled tree representation* T *of* M is defined to be a tree that is a fully labelled graph representation of M.

2.3 Problem description

Although every finite metric space M has its unique complete graph representation K_M, a fully labelled tree representation T of M does not necessarily exist for all M. This naturally leads to the following problem.

Problem 2.8. Given a finite metric space M, provide a necessary and sufficient condition to ensure that there is a fully labelled tree representation T of M.

2.4 Preliminaries

In this section, we describe two constituents of Theorem 2.18.

2.4.1 Four-point condition

We briefly recall the notion of partially labelled trees (see [28] for full details). Note that we focus on metrics rather than arbitrary dissimilarity maps in this chapter.

Definition 2.9. Given a finite metric space $M := (X, d_M)$, a tree $\mathcal{T} := (V, E; X, \phi, w)$ is said to be a *partially labelled tree representation of* M if it satisfies the following conditions:

1. ϕ is a distance-preserving labelling map;
2. $\{v \in V \mid deg(v) \leq 2\} \subseteq \phi(X)$.

As the condition 2 in Definition 2.9 only requires each vertex of degree at most two to be labelled with an element of X, $\mathcal{T}$ may have an unlabelled vertex (of degree at least three).

Remark 2.10. A fully labelled tree representation T of M is necessarily a partially labelled tree representation of M because the condition 2 in Definition 2.4 implies the condition 2 in Definition 2.9.

Definition 2.11. A finite metric space (X, d_M) is said to satisfy the *four-point condition* if, for every four points $q, r, s, t \in X$, the following inequality holds:

$$d_M(q, r) + d_M(s, t) \leq \max\{d_M(q, s) + d_M(r, t), d_M(r, s) + d_M(q, t)\}.$$

The following theorem, also known as the fundamental theorem of phylogenetics, characterises when a finite metric space can be represented by a partially labelled tree.

Theorem 2.12 (Buneman [9])**.** *Let $M := (X, d_M)$ be a finite metric space. Then there is a partially labelled tree representation $\mathcal{T}$ of M if and only if M satisfies the four-point condition.*

We restate Theorem 1.9 as follows.

Theorem 2.13. *Let $M := (X, d_M)$ be a finite metric space. If M satisfies the four-point condition, a partially labelled tree representation $\mathcal{T}$ of M is unique up to isomorphism.*

Remark 2.14. A graph G such that the metric space $(V(G), d_G)$ satisfies the four-point condition is also known as a *block graph*.

2.4.2 Fourth-point condition

Theorem 2.12 does not give an answer to Problem 2.8. This motivates us to introduce another condition defined as follows.

Definition 2.15 (Fig. 2.2)**.** A finite metric space (X, d_M) is said to satisfy the *fourth-point condition* if, for every three points $x, y, z \in X$, there exists a point $p^* \in X$ such that

$$d_M(x, p^*) + d_M(y, p^*) + d_M(z, p^*) = \frac{1}{2}\{d_M(x, y) + d_M(y, z) + d_M(z, x)\}.$$

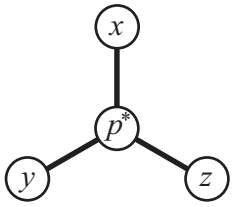

Figure 2.2: The fourth point p^* for a triplet $\{x, y, z\}$

The following result will be useful in the proof of Theorem 2.18.

Proposition 2.16. *The following is equivalent to saying that a finite metric space (X, d_M) satisfies the fourth-point condition: For every three points $x, y, z \in X$, there exists a point $p^* \in I(x, y) \cap I(y, z) \cap I(z, x) \subseteq X$.*

Proof. Because d_M is a metric, for all $x, y, z, p \in X$, we have $d_M(x, p) + d_M(y, p) + d_M(z, p) \geq \frac{1}{2}\{d_M(x, y) + d_M(y, z) + d_M(z, x)\}$. The equality holds if and only if $I(x, y) \cap I(y, z) \cap I(z, x) = \{p^*\}$. □

Remark 2.17. A graph G such that the metric space $(V(G), d_G)$ satisfies the fourth-point condition is also known as a *modular graph* . In particular, a modular graph in which each triplet of vertices has a unique median is called a *median graph*.

2.5 Results

We solve Problem 2.8 by proving the following theorem.

Theorem 2.18. *Let $M := (X, d_M)$ be a finite metric space. Then, there is a fully labelled tree representation T of M if and only if M satisfies both the four-point condition and the fourth-point condition.*

Proof. For any finite metric space of which a fully labelled tree representation exists, both the four-point condition (4PC) and the fourth-point condition (4thPC) clearly hold. Assuming M satisfies these two conditions, we prove the converse. Because M satisfies the 4PC, Theorem 2.12 and Theorem 2.13 ensure that there is a unique partially labelled tree representation $\mathcal{T}$ of M. Let $(V, E; X, \phi, w)$ denote $\mathcal{T}$. The assumption that (X, d_M) satisfies the 4thPC implies that $(\phi(X), d_{\mathcal{T}})$ also satisfies the 4thPC because $\phi : X \to V$ is a distance-preserving labelling map. Note that for any two distinct points u and v in the metric space $(V, d_{\mathcal{T}})$, the set of all vertices contained in the path $[u, \cdots, v]$ is identical to the closed metric interval $I(u, v)$ between u and v because $\mathcal{T}$ is a positive-weighted tree.

In order to obtain a contradiction, we suppose there is a vertex v of $\mathcal{T}$ such that $deg(v) \geq 3$ and $v \notin \phi(X)$. Then, there are three distinct vertices $a, b, c \in V$ that are adjacent to v. For $v_1 \in \{a, b, c\}$, we consider the following two cases:

Case 1. $v_1 \in \phi(X)$
We set $x := v_1$.

Case 2. $v_1 \notin \phi(X)$
The vertex v_1 is not a leaf of $\mathcal{T}$ by the condition 2) in Definition 2.9. Therefore, there is a vertex $v_2 (\neq v)$ of $\mathcal{T}$ that is adjacent to v_1. In the case of $v_2 \in \phi(X)$, Case 1 applies. In the case of $v_2 \notin \phi(X)$, we repeat the same process for v_2. We continue the process for $v_3, v_4, \cdots, v_i$ similarly until we find a vertex $v_i \in \phi(X)$. Note that this process ends in a finite number of steps because $\mathcal{T}$ is a finite tree. We set $x := v_i$.

Therefore, regardless of whether v_1 is labelled or not, we can find a labelled vertex $x \in \phi(X)$. The vertices v and x specify the path $[v, v_1, \cdots, x]$ in $\mathcal{T}$. Applying the same argument to each of the triplet $\{a, b, c\}$, we obtain three distinct labelled vertices $x, y, z \in \phi(X)$ of $\mathcal{T}$. The vertex v is the only vertex of $\mathcal{T}$ which the three paths $[v, a, \cdots, x]$, $[v, b, \cdots, y]$ and $[v, c, \cdots, z]$ have in common (otherwise, $\mathcal{T}$ would not be a tree). This gives $I(v, x) \cap I(v, y) \cap I(v, z) = \{v\}$. Also, we have $I(x, y) \cap I(x, z) = I(x, v)$ by using $I(x, y) = I(x, v) \cup I(v, y)$ and $I(x, z) = I(x, v) \cup I(v, z)$. Then, for distinct three points $x, y, z \in \phi(X)$, $I(x, y) \cap I(y, z) \cap I(z, x) = I(x, v) \cap I(y, v) \cap I(z, v) = \{v\}$, where $v \notin \phi(X)$. Then, Proposition 2.16 states that the 4thPC does not hold for $(\phi(X), d_{\mathcal{T}})$, but this is a contradiction. Hence, if M satisfies both the 4PC and the 4thPC, every vertex of $\mathcal{T}$ is labelled with an element of X, which means that $\mathcal{T}$ is a fully labelled tree representation of M. This completes the proof. □

Theorem 2.18 can be restated as the following corollary using Remark 2.14 and Remark 2.17.

Corollary 2.19. *A finite graph is a tree if and only if it is a block graph and is also a median graph .*

2.6 Relationship to the minimum spanning tree

In this section, we only consider fully labelled graph representations. This allows us to identify a set of labelled vertices with the label set itself, so we write $(X, E; w)$ rather than $(V, E; X, \phi, w)$ for notational simplicity. Also, we may identify a label $x \in X$ with the corresponding labelled vertex $\phi(x) \in V$, and use the same symbol x for each.

The following proposition states that, if it exists, a fully labelled tree representation T of M can be found by solving the MST problem.

Proposition 2.20. *Let $M := (X, d_M)$ be a finite metric space, and $K_M := (X, \binom{X}{2}; d_M)$ be the complete graph representation of M. If there is a fully labelled tree representation T of M, then T is uniquely determined up to isomorphism. Moreover, T is isomorphic to the only MST in K_M.*

Proof. We first note that Theorem 2.13 ensures the uniqueness of a fully labelled tree representation T of M, if it exists (recall Remark 2.10).

Let $(X, E; w)$ denote T. We see that T is a spanning subtree of K_M because we have $V(T) = V(K_M)$, and as the condition 1 in Definition 2.4 implies, $w(xy) = d_M(x, y)$ holds for all $xy \in E(T)$. Let T' be an arbitrary spanning subtree of K_M with an edge set E' $(\neq E)$. To obtain a contradiction, we suppose that T' is an MST in K_M. In what follows, a path joining vertices x and y in T (or T') is represented using $[x, \cdots, y]_T$ (or $[x, \cdots, y]_{T'}$).

We claim that for any $pq \in E \setminus E'$, there exists $rs \in E([p, \cdots, q]_{T'}) \setminus E$ such that $[r, \cdots, s]_T$ contains pq. Because T' is a tree, for any $pq \in E \setminus E'$, there is a unique path $[p, \cdots, q]_{T'}$. If all edges in $[p, \cdots, q]_{T'}$ were in E, then the union of $[p, \cdots, q]_{T'}$ and pq would form a cycle C, so T would not be a tree. Then, there is an edge $rs \in E' \setminus E$ that is contained in $[p, \cdots, q]_{T'}$. Because $[r, \cdots, s]_T$ has at least one edge other than pq and all weights are strictly positive, we have $d_M(p, q) < d_M(r, s)$.

Let T'' be the spanning subtree in K_M that is obtained from T' by replacing rs with pq. The above inequality implies that the length of T'' is strictly less than that of T', but this is a contradiction. Then, T' is not an MST in K_M. Hence, we can conclude that T is a unique MST in K_M. This completes the proof. □

Proposition 2.20 gives the following corollary of Theorem 2.18.

Corollary 2.21. *Let $M := (X, d_M)$ be a finite metric space, and T_M be a minimum spanning tree in the complete graph $K_M := (X, \binom{X}{2}; d_M)$. Then, T_M and K_M are isometric if and only if M satisfies both the four-point condition and the fourth-point condition.*

2.7 Summary and discussion

Stimulated by biological applications of the MST problem, we have addressed Problem 2.8 to determine when a distance matrix of order n can be represented by a fully labelled tree on n vertices. We have settled it by proving Theorem 2.18, where our fourth-point condition is combined with Buneman's four-point condition. As we have shown in Proposition 2.20, given a finite metric space that satisfies both the four-point condition and the fourth-point condition, solving the MST problem gives a unique fully labelled tree that preserves all information about the metric space. Thus, as summarized in Corollary 2.21, we have characterised when there is an exact fit between a finite metric space and the MST.

The results in this chapter have various applications, one of which is cellular tree estimation as described in Section 2.1. We expect that they will extend the range of biological applications of the four-point condition, which has been mostly confined so far to the context of phylogenetic tree inference.

From a more general perspective, it would be interesting to discuss a quantitative measure of the MST-likeness of a finite metric space. One possible approach would be to combine two deviation measures, but another method is worth considering as the deviation from the four-point condition is hard to compute when we have thousands of sample cells (see Section 1.3). To this end, we will seek for alternative characterisation of MST metric spaces in the next chapter.

Notes

This chapter is based on the manuscript of [17] 'A characterization of minimum spanning tree-like metric spaces' (with H. Endo and K. Fukumizu), which is to appear in *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* and is available at doi: 10.1109/TCBB.2016.2550431.

Chapter 3

Alternative views on minimum spanning tree metric spaces

In the previous chapter, we introduced the fourth point condition and combined it with the classical four-point condition to give a characterisation of minimum spanning tree metric spaces. In this chapter, we attempt to shed new light on the notion of 'tree-like' metric spaces by focusing on another approach that does not use the four-point condition.

3.1 Introduction

Historically, graphs as finite metric spaces have been extensively studied [7]. Even though we approach them differently, we would like to emphasise, amongst others [22, 29, 30, 31], the classical result provided by Buneman [9]. In short, a metric on a finite set can be realised by the shortest path metric in a positive-weighted tree if and only if it satisfies the four-point condition. The four-point condition is not only frequently quoted in the context of evolutionary trees [28], but also known for its direct connection to the theory of Gromov hyperbolic metric spaces [13]. Nowadays, it is also widely known that there is a unique tree representation for every metric satisfying the four-point condition [8, 19].

Given this background, a metric space that satisfies the four-point condition is commonly considered tree-like. However, an important caveat should be addressed: the four-point condition is necessary and sufficient to ensure the existence of a *partially labelled* tree that realises a given metric [7, 19, 28]. For example, a complete graph with a uniform edge length clearly satisfies the four-point condition, but it only becomes tree-like after an extra vertex is added.

In this case, the four-point condition does not ensure that a metric is realised by a *fully labelled* tree on *the same* set. The same applies to *block graphs* (*i.e.*, unweighted graphs in which all biconnected components are complete subgraphs) [3]. Thus, it does not characterise the distance within trees but rather the shortest path metrics induced by graphs of a more general class.

This may not create an issue in the field of conventional phylogenetics, but considering the recent surge of renewed biological interest in minimum spanning tree (MST)-based tree estimation [26], determining when a metric space is realised by a positive-weighted tree on the same set is not only a natural undertaking but also a meaningful one. Thus far, this problem has not been properly recognised, much less addressed. The only two exceptions to this are the recent work provided in [2] and in [17]. It seems to be a non-trivial question not only because it cannot be answered using Buneman's theorem, but also because it is equivalent to determining a method for recognising a special case of the metric travelling salesman problem (TSP). If an input—a metric on a set of cities—is the shortest path metric in a tree on the city set, the length of the optimal tour must equal twice the length of the MST.

In this chapter, we examine the sub-type of tree metrics without relying on the four-point condition. Our work is based on three ingredients: the so-called tie-breaking assumption, which has been popular in algorithmic applications since the work provided by Kruskal in [23]; what we call the fourth-point condition, which can typically be found in the definition of median metric spaces [10]; and a simple trick for metric-preserving edge removal, which applies to any finite metric space. These concepts, which are part of our original results, are defined and discussed in Section 3.2.

As expected, if it exists, a fully labelled positive-weighted tree that realises a finite metric space is the unique MST in its associated weighted complete graph (Proposition 3.15). Our goal is to prove the following: A finite metric space under the tie-breaking rule is realised by the MST if and only if it satisfies the fourth point condition (Theorem 3.17). This implies that every finite median graph, in which the shortest path lengths between all pairs of vertices are distinct, is necessarily a tree (Corollary 3.19). This result also yields a stronger condition for understanding when a finite metric space is realised, especially by a spanning path graph (Corollary 3.21). We define and discuss the notion of a spanning tree-likeness of a finite metric space in Section 3.4.

3.2 Preliminaries

We apply the metric-related terminology provided in [10] throughout this chapter. Let (X, d_M) be a *finite metric space*, that is, a finite set, X, equipped with metric d_M. For two distinct points x and x' in X, the *closed metric interval* between them is defined to be the set

$$I(x, x') := \{i \in X : d_M(x, x') = d_M(x, i) + d_M(i, x')\}.$$

All graphs considered in this chapter will be simple, undirected, *fully labelled* (*i.e.*, each vertex is labelled), and *positive weighted* (*i.e.*, each edge has a positive length). A graph is denoted by $(V, E; w)$ for a set, V, of labelled vertices and a set, E, of edges that are associated with a positive edge-weighting function, $w : E \to \mathbb{R}^+$. Given a graph G, the sets of vertices and edges are denoted by $V(G)$ and by $E(G)$, respectively. Moreover, a graph G is said to be a graph *on* $V(G)$. Vertices may be renamed as needed, assuming no confusion arises, and a vertex labelled 'x' is referred to as vertex x. The distance in graph G is defined to be the shortest path metric and is represented using d_G.

Assume M is a finite metric space, (X, d_M). Let K_M be the associated weighted complete graph $(X, \binom{X}{2}; d_M)$ with M. An edge of K_M that joins two distinct vertices, x and x', is denoted by $e(x, x')$. This chapter uses the terms '*points*' and '*vertices*' interchangeably because there is a one-to-one correspondence between X and $V(K_M)$ for any finite metric space M.

3.2.1 Tie-breaking rule

Given a connected graph, G, a subtree that connects all the vertices of G is said to be a *spanning tree* in G. In particular, a spanning tree whose length (*i.e.*, the sum of all edge-weights) is shortest amongst all spanning trees is called a *minimum spanning tree* and abbreviated as MST. The problem of finding an MST in a connected graph is known as the MST problem, which is efficiently solved by a greedy algorithms such as Kruskal's method. In fact, one can easily find an MST in K_M by selecting edges so as not to create a cycle in ascending order of the value of d_M. Although K_M can have one or more MSTs in general, its MST is uniquely determined if the following assumption holds.

Definition 3.1. A finite metric space, (X, d_M), is said to satisfy the *tie-breaking rule* if the values of d_M are distinct for all pairs in X.

The tie-breaking rule has been widely known since it was introduced by Borůvka [5] (cited in [23]) and by Kruskal [23]. This assumption is strong enough

to ensure the uniqueness of the MST but it is reasonable in many practical situations and is convenient as it can be quickly checked in $O(|X|^2)$ time. The present chapter explores its another benefit through a discussion on relation between an MST and a finite metric space.

3.2.2 The fourth point condition

We first recall Definition 2.15 and Proposition 2.16.

Definition 3.2 (Figure 2.2, Chapter 2)**.** A finite metric space, (X, d_M), is said to satisfy the *fourth-point condition* if, for every (not necessarily distinct) three points $x, y, z \in X$, there exists a point, $p^* \in X$, such that

$$d_M(x, p^*) + d_M(y, p^*) + d_M(z, p^*) = \frac{1}{2}\{d_M(x, y) + d_M(y, z) + d_M(z, x)\}.$$

Proposition 3.3. *The following is equivalent to saying that finite metric space (X, d_M) satisfies the fourth-point condition: For every (not necessarily distinct) three points $x, y, z \in X$, there exists a point $p^* \in I(x, y) \cap I(y, z) \cap I(z, x)$.*

Remark 3.4. Fourth point p^* is not necessarily unique for each triplet in X (see the five-point metric space induced by the complete bipartite graph $K_{2,3}$, which can be seen in Figure 3.1).

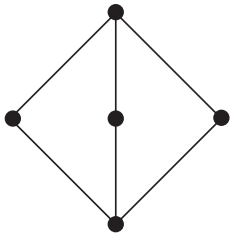

Figure 3.1: The complete bipartite graph $K_{2,3}$

Proposition 3.5. *Let (X, d_M) be a finite metric space that satisfies the fourth-point condition. Then fourth point $p^* \in X$ is unique for each triplet in X if and only if X does not contain a subset $S \subseteq X$ of five points such that (S, d_M) is realised by a weighted complete bipartite graph $K_{2,3}$ with uniform edge weights.*

Proof. Suppose there are two fourth points p_1^* and p_2^* ($p_1^* \neq p_2^*$) for a triplet $\{x, y, z\}$ in X. By the assumption that the fourth-point condition holds, there is a fourth point $x' \in X$ for $\{p_1^*, p_2^*, x\}$ (see Figure 3.2). Similarly, let y' and $z' \in X$ be fourth points for $\{p_1^*, p_2^*, y\}$ and $\{p_1^*, p_2^*, z\}$, respectively. Let $\alpha := d_M(p_1^*, x')$,

$\beta := d_M(p_1^*, y')$ and $\gamma := d_M(p_1^*, z')$. Because both p_1^* and p_2^* are fourth points for a triplet $\{x, y, z\}$, we have

$$\begin{aligned} d_M(p_2^*, x') + d_M(p_2^*, y') + d_M(p_2^*, z') &= d_M(p_1^*, x') + d_M(p_1^*, y') + d_M(p_1^*, z') \\ &= \alpha + \beta + \gamma. \end{aligned}$$

We may set $d_M(p_2^*, x') = \alpha - a$, $d_M(p_2^*, x') = \beta + b$ and $d_M(p_2^*, x') = \gamma + a - b$ with $a, b \geq 0$ (If each term in the left hand side is strictly greater or smaller than that in the right side, the sums of three distances would not be equal). By Proposition 3.3, $p_1^*, p_2^* \in I(x, y)$ holds. Then we have $\alpha+\beta = \beta+b+\alpha-a$ and hence $a = b$. Applying similar arguments to $I(y, z)$ and $I(z, x)$, we obtain $a = b = 0$. Also, we deduce $\alpha = \beta = \gamma$ from $x', y', z' \in I(p_1^*, p_2^*)$. Here α, β and γ must be strictly positive because if they were equal to zero, we would have $x' = y' = z' = p_1^* = p_2^*$ but this contradicts the assumption of $p_1^* \neq p_2^*$. Then setting $S := \{x', y', z', p_1^*, p_2^*\}$, we conclude that (S, d_M) is realised by an $K_{2,3}$ with a uniform edge length. The converse is obvious. □

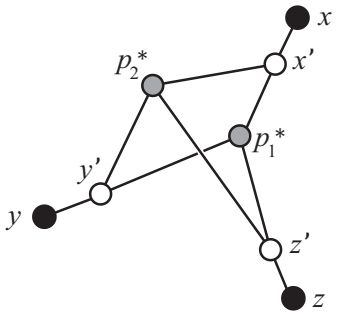

Figure 3.2: The proof of Proposition 3.5

The following is an immediate consequence of Proposition 3.5.

Proposition 3.6. *If a finite metric space, (X, d_M), under the tie-breaking rule satisfies the fourth-point condition, fourth point $p^* \in X$ is unique for each triplet in X.*

Remark 3.7. Fourth point p^* is also known as a *median* for $\{x, y, z\}$ because it minimises the sum of the distances to the three points, and a metric space in which there is a unique median for each triplet (or a graph inducing this kind of metric space) is said to be *median* [3, 10]. Although a discussion of this topic is provided in [3, 4], it should be noted that median graphs include multiple types of graphs other than trees, such as grid and square graphs.

Lemma 3.8. *Let C be a cycle graph, $(V, E; w)$, with $\sum_{e\in E} w(e) = c$. Also, let d_C be the shortest path metric in C. Given three distinct points $x, y, z \in V$ such that $d_C(x, y) + d_C(y, z) + d_C(z, x) = c$, fourth point p^* exists in V if and only if* $\max\{d_C(x, y), d_C(y, z), d_C(z, x)\} = c/2$.

Proof. We can assume $d_C(z, x) = \max\{d_C(x, y), d_C(y, z), d_C(z, x)\}$ without loss of generality. Clearly, $y \in I(x, y) \cap I(y, z)$. Therefore, $I(x, y) \cap I(y, z) \cap I(z, x) = \emptyset$ if and only if $y \notin I(z, x)$. Under the assumption that the length of C is fixed at c, this is equivalent to stating that $d_C(z, x) \neq c/2$. Thus, $I(x, y) \cap I(y, z) \cap I(z, x) = \emptyset$ if and only if $d_C(z, x) \neq c/2$. Applying Proposition 3.3 completes the proof. □

3.2.3 Basic geodesic graphs

In this subsection, we present a simple trick for metric-preserving edge removal, which can be used to represent an arbitrary finite metric space as a graph with the fewest edges. Let M be a finite metric space, (X, d_M), and assume K_M is the weighted complete graph associated with M.

Definition 3.9. Suppose G is a connected graph on finite set X with shortest path metric d_G. Graph G is said to *realise* M if $d_G(x, x') = d_M(x, x')$ for all $x, x' \in X$.

Definition 3.10. Given $x, x' \in X$, the edge, $e(x, x')$, of K_M is said to be *non-basic* if there is a permutation, $(x_1, x_2, \cdots, x_k)$, on a non-empty subset of $X \setminus \{x, x'\}$ such that cyclic permutation $(x, x_1, x_2, ..., x_k, x')$ satisfies

$$d_M(x, x') = d_M(x, x_1) + d_M(x_1, x_2) + \cdots + d_M(x_k, x').$$

The edge is called *basic* otherwise.

Proposition 3.11. *Let x, y, z be three different vertices of K_M. When the three edges, $e(x, y)$, $e(y, z)$, and $e(z, x)$, of K_M are basic, fourth point p^* does not exist for $\{x, y, z\}$. If a non-basic edge exists, say $e(x, y)$, points x and y are the only two candidates for p^*.*

The proof of this proposition is straightforward.

Definition 3.12. Assume B_M is the set of all basic edges of K_M, and suppose λ is a restriction of d_M to B_M. A subgraph, $G_M := (X, B_M; \lambda)$, in K_M is called *the basic geodesic graph in K_M*.

Lemma 3.13. *The basic geodesic graph, G_M, in K_M is a connected graph on X that realises M.*

Proof. It suffices to prove that G_M is connected. If $e(x, x') \in E(K_M)$ is basic, the vertices x and x' are obviously connected in G_M. Assuming that $e(x, x')$ is non-basic, we show that there is a path of basic edges joining x and x' in K_M.

We define C to be a cycle with the greatest number of vertices (or edges) amongst all cycles in K_M that are of length $2d_M(x, x')$ and contain $e(x, x')$. Let $V(C) = \{x, x'\} \cup Y$ where $Y := \{x_1, \cdots, x_k\}$ is a non-empty subset of $X \setminus \{x, x'\}$. We have $d_M(x, x') = d_M(x, x_1) + \sum_{i=1}^{k-1} d_M(x_i, x_{i+1}) + d_M(x_k, x')$, as in Definition 3.10. We know that any path in K_M joining two vertices x_i and x_j of C must have a length greater than or equal to $d_C(x_i, x_j)$ because $e(x, x')$ would be longer than the path connecting x and x' through x_i and x_j otherwise. We use this fact at the end of the proof.

In order to obtain a contradiction, we suppose $e(y, y') \in E(C) \setminus e(x, x')$ is non-basic. We define C' to be a cycle in K_M of overall length $2d_M(y, y')$ with $e(y, y') \in E(C')$, which is similar to our previous case except that $|V(C')|$ is unimportant. Let $V(C') = \{y, y'\} \cup Z$ with a non-empty subset $Z := \{y_1, \cdots, y_l\}$ of $X \setminus \{y, y'\}$. Again, we have $d_M(y, y') = d_M(y, y_1) + \sum_{i=1}^{l-1} d_M(y_i, y_{i+1}) + d_M(y_l, y')$. We note that $Y \cap Z$ is non-empty because otherwise K_M would contain a cycle of the same length as C but with more vertices than C (see Figure 3.3). Let $y'' \in Y \cap Z$. By our hypothesis that $e(y, y')$ is a non-basic edge in K_M, $d_M(y, y') = d_{C'}(y, y') = d_{C'}(y, y'') + d_{C'}(y', y'')$. Then, we have $d_{C'}(y, y'') < d_{C'}(y, y') = d_C(y, y')$. Moreover, we claim $d_C(y, y') < d_C(y, y'')$. First, C' does not contain $e(x, x')$ because $d_M(x, x') > d_M(y, y')$ but $e(y, y')$ must be strictly longer than any other edge in C'. Then, by assuming that y' lies in the shortest path joining y and y'' in C (note that this entails no loss of generality as the roles of y and y' can be exchanged), we see that our claim is indeed true. Thus, $d_{C'}(y, y'') < d_C(y, y'')$. It follows that K_M contains a path joining y and y'' of length less than $d_C(y, y'')$, but this is a contradiction. Hence, $e(y, y')$ is basic, which completes the proof. □

Definition 3.14. Finite metric space M is said to be a *spanning tree metric space* if the basic geodesic graph, G_M, in the weighted complete graph, K_M, is a spanning subtree in K_M. In particular, M is said to be a *spanning path metric space* if G_M is a *path graph* (*i.e.*, a tree with two vertices of degree one and remaining vertices of degree two) that spans all the vertices of K_M.

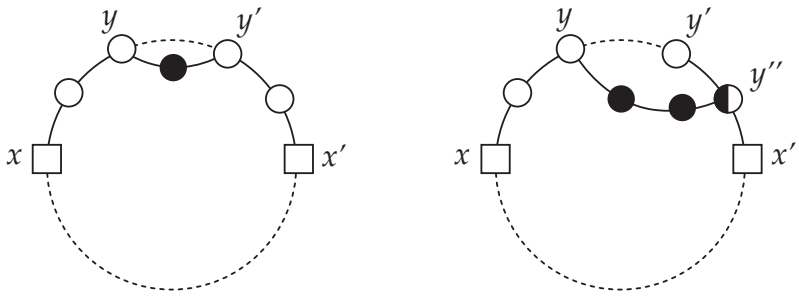

Figure 3.3: The proof of Lemma 3.13. The dashed lines are assumed to be non-basic edges. The white and black round vertices represent elements in Y and in Z, respectively. The left panel describes the case of $Y \cap Z = \emptyset$, which we can exclude (see text). The right panel illustrates the case in which there is a vertex $y'' \in Y \cap Z$.

Proposition 3.15. *Let $M := (X, d_M)$ be a spanning tree metric space and G_M be the basic geodesic graph in K_M. Then the following statements hold:*

1. *G_M is the unique minimum spanning tree in K_M;*
2. *G_M is the unique fully labelled tree on X that realises M.*

Proof. (1) Assume $B := E(G_M)$ and let $\overline{B} := E(K_M) \setminus B$. Because $|B| = |X| - 1$, G_M is the only spanning tree in K_M such that all edges are basic. In addition, let $e(x, x') \in \overline{B}$. Because G_M is a tree, there is a unique path joining x and x', denoted by P. Each edge of P must be strictly shorter than $d_M(x, x')$ for the following reasons: the length of P equals $d_M(x, x')$; the number of edges of P exceeds one; and the edge weights are all positive. Therefore, replacing an arbitrary edge of P with $e(x, x')$ results in a spanning tree in K_M of greater length. Hence, G_M is shorter than any other spanning trees in K_M. (2) Suppose that M is realised by fully labelled tree T on X. This implies that each edge of T has a positive weight. We can recover K_M from T by summing the weights along every path in T that has two or more edges. This process indicates that T is isomorphic to the basic geodesic graph in K_M. Hence, given (1), we know T is unique. □

Remark 3.16. Proposition 3.15 states that a metric space is uniquely realised by the only MST if it is a spanning tree metric space. Note that we do not need Buneman's four-point condition in the argument (*cf.* [2]).

3.3 Main results

Theorem 3.17. *Let M be a finite metric space, (X, d_M), under the tie-breaking rule. Then M is a spanning tree metric space if and only if it satisfies the fourth-point condition.*

Proof. (i) The fourth-point condition clearly holds for all spanning tree metric spaces. (ii) If d_M is not a spanning tree metric on X, then we will show that there is a triplet in X that violates the fourth-point condition. According to Lemma 3.13, our assumption implies that the basic geodesic graph, $G_M = (X, B; \lambda)$, in K_M contains at least one cycle. Suppose $C := (X_k, B_k; \lambda_k)$ is the shortest cycle in G_M, where $X_k \subseteq X$, $B_k \subseteq B$, $|X_k| = |B_k| = k$, and λ_k is the restriction of λ to B_k. Then Proposition 3.11 yields $k \geq 4$. Let c denote the sum of the λ_k over all elements in B_k. Also, assume that d_C is the shortest path metric in C. For all $i, j \in X_k$, no path in G_M joining i and j has a shorter length than $d_C(i, j)$ (otherwise, C would not be the shortest cycle in G_M). Therefore, $d_C(i, j) = \min\{a_{ij}, c - a_{ij}\}$, in which a_{ij} represents the length of the path in C that travels from i to j in a clockwise direction.

Consider a route in which we visit the points in X_k. Let $s \in X_k$ be the starting point from which we travel along the circle in a clockwise direction. We assign a label, 'L' or 'R', to every point $i \in X_k \setminus \{s\}$: label 'L' is assigned if $a_{si} < c/2$, and we use label 'R' if $a_{si} \geq c/2$. If every point in $X_k \setminus \{s\}$ was labelled 'L', the last edge we would traverse returning to s would be non-geodesic or non-basic. Therefore, there exists one and only one basic edge between vertices labelled 'L' and 'R'. Suppose that t signifies the last point with label 'L' and u indicates the first point with label 'R' as on the left in Figure 3.4. Note that $d_M(s, t) + d_M(t, u) + d_M(u, s) = c$.

We assume that p^* exists for $\{s, t, u\}$ (otherwise, the assertion of the theorem immediately follows). Lemma 3.8 gives us $\max\{d_M(s, t), d_M(t, u), d_M(u, s)\} = c/2$. Thus, $d_M(u, s) = c/2$ (the edge joining t and u is basic, and $d_M(s, t) < c/2$). Let $v(\neq u)$ be a point in X_k with label 'R' that is between u and s as on the right in Figure 3.4. We know point v exists because $e(u, s)$ would be non-basic otherwise. According to the tie-breaking rule, we note that $a_{tv} \neq c - a_{tv}$. We can also set $a_{tv} < c - a_{tv}$ in order to select $\{s, t, v\}$. Although we should select $\{t, u, v\}$ when $a_{tv} > c - a_{tv}$, we limit our consideration to the former case. Therefore, we have $d_M(s, t) + d_M(t, v) + d_M(v, s) = c$ again, but each of the three terms does not equal $c/2$ (recall that $d_M(s, u) = c/2$). Hence, Lemma 3.8 implies that p^* does not exist for $\{s, t, v\}$, and this completes the proof. □

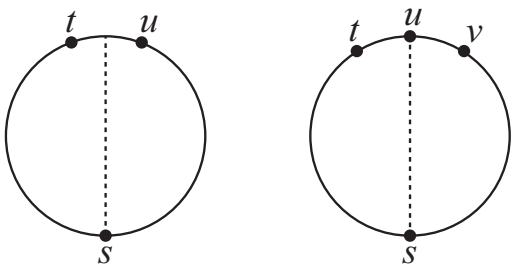

Figure 3.4: Points in the proof of Theorem 3.17

Remark 3.18. Given a finite metric space on X, we can determine in $O(|X|^4)$ time whether it is a spanning tree metric space

Corollary 3.19. *Let G be a median graph on finite set X and let d_G be the shortest path metric of G. If each pair in X has a different value for d_G, then G is a tree.*

Remark 3.20. As was mentioned in Remark 3.7, the fourth-point condition *per se* is not a sufficient condition, but it is a necessary condition in order to ensure that a finite metric space is induced by the shortest path metric in a tree (*cf.* a cycle graph on four vertices with a uniform edge length).

Corollary 3.21. *Suppose $M := (X, d_M)$ is a finite metric space under the tie-breaking rule. Then M is a spanning path metric space (Definition 3.14) if and only if it satisfies* the three-point condition*: for every (not necessarily distinct) three points $x, y, z \in X$, we have*

$$\max\{d_M(x,y), d_M(y,z), d_M(z,x)\} = \frac{1}{2}\{d_M(x,y) + d_M(y,z) + d_M(z,x)\}.$$

The condition can be confirmed in $O(|X|^3)$ time. If M is a spanning path metric space, it is realised by the unique shortest path that joins the farthest two points in X.

Proof. We only prove the first statement. The three-point condition obviously holds for all spanning path metric spaces. Therefore, we assume that the three-point condition holds and show that the basic geodesic graph, G_M, in K_M is a path graph on X. It is clear that y is a fourth point, p^*, for $\{x, y, z\}$ when the left-hand side equals $d_M(z, x)$. This means that the fourth-point condition automatically holds for any finite metric space that satisfies the three-point condition. Therefore, our assumption implies that G_M is a tree on X. The three-point condition also indicates that every vertex in G_M has a degree of one or two. In other words, if vertex x has degree three or more, then any three distinct vertices adjacent to x would violate the three-point condition. Hence, G_M is a path graph on X, which completes the proof. □

3.4 Discussion and future directions

The results in this chapter have implications for measuring the *fully labelled* tree-likeness of M, which cannot be quantified by δ-hyperbolicity. As we have seen in Section 1.3, the hyperbolicity of finite metric spaces (or graphs) is a concept provided by Gromov [13, 28] and measures the deviance of a metric space from Buneman's four-point condition. If a metric space, M, satisfies the four-point condition, then the hyperbolicity of M equals 0, and M is said to be 0-hyperbolic. As was previously discussed, any complete graph with a uniform edge length is 0-hyperbolic, and all metric triangles are also 0-hyperbolic. Therefore, although the value of hyperbolicity is usually called the 'tree-likeness' of M, a more precise interpretation refers to the *partially labelled* tree-likeness of M.

In the light of extensive use of MST algorithms for estimating a fully labelled tree to explain observed distance data, it would be important to consider how to evaluate the minimum spanning tree-likeness of a finite metric space. In Chapter 2, we have shown that a finite metric space M is a spanning tree metric space if and only if both the four point condition and the fourth-point condition holds (Theorem 2.18). Then, one possible idea is to integrate two measures, a degree of deviance from the four-point condition and that of the fourth-point condition. However, because δ-hyperbolicity is known to be hard to compute, we would like to circumvent it if possible.

Theorem 3.17 says that, whenever the tie-breaking assumption holds, we can avoid taking account of δ-hyperbolicity and focus on computing the deviation from the fourth-point condition. The assumption holds in many practical cases, and we can even make an arbitrary finite metric space (X, d_M) obey it by slightly changing the values of d_M where necessary. Then it would be interesting to explore a nice measure of how far d_M deviates from the fourth-point condition, just as Gromov excogitated δ-hyperbolicity from the four-point condition. Here, for illustrative purposes, we define ρ as follows and say that M is *ρ-roundabout*:

$$\rho := \max_{x,y,z \in X} \min_{i \in X} \{d_M(x,i) + d_M(y,i) + d_M(z,i) - \frac{1}{2}(d_M(x,y) + d_M(y,z) + d_M(z,x))\}.$$

The value of ρ indicates how far M deviates from the fourth-point condition, and under the tie-breaking rule, it can be used to quantify the spanning tree-likeness of M or the circuitousness of d_M (see Figure 3.5). As Proposition 3.15 implies, M is 0-roundabout if and only if there is an exact fit between M and the MST. The degree of violation of the three-point condition similarly provides the spanning path-likeness of M—the maximum discrepancy between the left and

right-hand sides of the triangular inequality. On the other hand, hyperbolicity does not provide any information because all metric triangles are 0-hyperbolic.

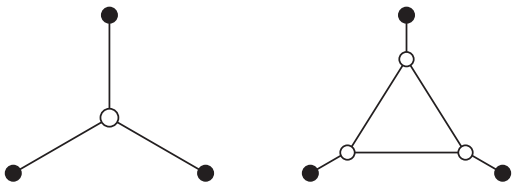

Figure 3.5: Illustrations of spanning tree-likeness ($\rho = 0$ and $\rho > 0$)

Our future work includes determining how to measure the deviation of M from the fourth-point condition. In particular, it would be nice if we could find a mathematical quantity of use in estimating the approximation error of M by the MST. Another research direction is to consider how to perturb d_M without serious change in the MST-likeness of (X, d_M).

Notes

This chapter is based on the preprint [18] 'On minimum spanning tree-like metric spaces' (with K. Fukumizu), which is available at arXiv:1505.06145.

Part III

Reticulate evolution

Chapter 4

On the existence of infinitely many universal tree-based networks

The idea of the 'tree of life' dates back to at least the early 19th century, when Charles Darwin first sketched how species might evolve in his famous notebook (*First Notebook on Transmutation of Species* 1837). While it is a useful metaphor to describe the evolutionary relationships between living and extinct organisms, we know today that a phylogenetic tree oversimplifies the interconnectedness between species on Earth and that the 'web of life' may more accurately describe the reality of evolution. In this chapter, we are interested in modelling tangled evolutionary processes called *reticulate evolution* and thereby consider mathematical aspects of *phylogenetic networks*.

4.1 Introduction

Throughout this chapter, n denotes a natural number that is greater than 1 and X represents the set $\{1, 2, \cdots, n\}$. All graphs considered here are directed acyclic graphs. A graph G' is said to be a *subdivision* of a graph G if G' can be obtained from G by inserting vertices into arcs of G zero or more times. Given a vertex v of a graph with $indeg(v) = outdeg(v) = 1$, *smoothing* (or *supressing*) v refers to removing v and then adding an arc from the parent to the child of v. Two graphs are said to be *homeomorphic* if they become isomorphic after smoothing all vertices of in-degree one and out-degree one.

In [12], Francis and Steel recently introduced the class of networks that can be created from phylogenetic trees merely by placing additional arcs (to be defined formally later) and posed interesting problems on their mathematical properties.

For the reader's convenience, we briefly recall the relevant background from [12].

Definition 4.1. A *rooted binary phylogenetic network on X* is defined to be a directed acyclic graph (V, A) with the following properties:

- $X = \{v \in V \mid indeg(v) = 1,\ outdeg(v) = 0\}$;
- there is a unique vertex $\rho \in V$ with $indeg(\rho) = 0$ and $outdeg(\rho) \in \{1, 2\}$;
- for all $v \in V \setminus \{X \cup \{\rho\}\}$, $\{indeg(v), outdeg(v)\} = \{1, 2\}$.

The vertices in X are called *leaves*, and the vertex ρ is called the *root*.

Definition 4.2. Suppose $\mathcal{T} = (V, A)$ is a rooted binary phylogenetic tree on X. A rooted binary phylogenetic network $\mathcal{N}$ on X is said to be a *tree-based network on X with base tree $\mathcal{T}$* if there are a subdivision $\mathcal{T}' = (V', A')$ of $\mathcal{T}$ and a set I of mutually vertex-disjoint arcs between vertices in $V' \setminus V$ such that $(V', A' \cup I)$ is acyclic and is homeomorphic to $\mathcal{N}$. The vertices in $V' \setminus V$ are called *attachment points*, and the arcs in I are called *linking arcs*.

Tree-based networks can represent more realistic relationships among taxa than phylogenetic trees without compromising the concept of 'underlying trees' (*cf.*, [12, 21]). They may have an important role to play in modern phylogenetic inference, but there are many open problems on their mathematical properties, one of which we would like to address here.

In order to state the problem formally, we now introduce the notion of universal tree-based networks. A tree-based network on X is said to be *universal* if any binary phylogenetic tree on X can be a base tree. We can define universal tree-based networks in a more concrete manner with the number $(2n - 3)!!$ of binary phylogenetic trees on X as follows.

Definition 4.3. A tree-based network $\mathcal{N} = (V, A)$ on X is said to be *universal* if for any binary phylogenetic tree $\mathcal{T}^{(i)}$ on X ($i \in \{1, 2, \cdots, (2n - 3)!!\}$), there is a set $I^{(i)} \subset A$ of linking arcs such that $(V, A \setminus I^{(i)})$ is homeomorphic to $\mathcal{T}^{(i)}$.

Problem 4.4 ([12])**.** Does a universal tree-based network on a set X of n leaves exist for all n?

This problem is of interest because it explores whether a phylogenetic tree on X is always reconstructable from a tree-based network on X. In [12], Francis and Steel pointed out that the answer is 'yes' for $n = 3$. In this chapter, we will completely settle their question in the affirmative and provide further insights into universal tree-based networks (Theorem 4.8).

4.2 Preliminaries

Here, we slightly generalise the concept of tree shapes. Given a tree-based network $\mathcal{N}$ on X, ignoring the labels on the leaves of $\mathcal{N}$ results in an unlabelled tree-based network N with n leaves. We use the two different types of symbols, such as N and $\mathcal{N}$, to mean unlabelled and labelled tree-based networks, respectively. Two tree-based networks $\mathcal{N}$ and $\mathcal{N}'$ on X are said to be *shape equivalent* if N and N' are isomorphic. This equivalence relation partitions a set of the tree-based networks on X into equivalence classes called *tree-based network shapes with n leaves*.

Definition 4.5. A tree-based network shape N with n leaves is said to be *universal* if for any rooted binary phylogenetic tree shape $T^{(i)}$ with n leaves ($i \in \{1, 2, \cdots, r_n\}$), there is a set $I^{(i)}$ of linking arcs such that $(V, A \setminus I^{(i)})$ is homeomorphic to $T^{(i)}$. Here, r_n denotes the number of rooted binary phylogenetic tree shapes with n leaves.

The following proposition is not directly relevant to the content of this chapter, but ideas behind it, which are summarised in Remark 4.7, will be useful in the proof of Theorem 4.8.

Proposition 4.6 ([15])**.** *Let $r_1 := 1$ and $k \in \mathbb{N}$ with $k > 1$. Then, we have the following recurrence equation:*

$$r_n = \begin{cases} 1 & \text{if } n = 2; \\ \sum_{i=1}^{k-1} r_i r_{n-i} & \text{if } n = 2k-1; \\ \frac{r_k(r_k+1)}{2} + \sum_{i=1}^{k-1} r_i r_{n-i} & \text{if } n = 2k. \end{cases}$$

Remark 4.7. We assume that T_1 represents a rooted chain shape. Any rooted binary phylogenetic tree shape T_n with n leaves can be decomposed into two first-order subshapes T_m and T_{n-m} with $m \in \mathbb{N}$. In other words, using Harding's notation [15], we can write $T_n = T_m + T_{n-m}$.

4.3 Results

Theorem 4.8. *For any natural number $n > 1$, there are infinitely many universal tree-based networks on a set X of n leaves.*

Proof. First, we will show that there is a universal tree-based network shape with n leaves. Let U_n be a rooted binary phylogenetic network shape with n leaves as illustrated in the left panel of Figure 4.1, which can be obtained by adding ($n -$

$1)(n-2)/2$ linking arcs and $(n-1)(n-2)$ attachment points to a rooted caterpillar tree shape with n leaves. By definition, U_n is a tree-based network shape with n leaves. We will prove that U_n is universal by induction. (i) It is easy to see that U_2 and U_3 are universal. (ii) Assuming U_k is universal for any $k \in \mathbb{N}$ $(2 \leq k \leq n)$, we will show that U_{n+1} is universal. We claim that any binary phylogenetic tree shape T_{n+1} with $T_{n+1} = T_n + T_1$ can be a base tree shape of U_{n+1}. Indeed, U_{n+1} contains mutually vertex-disjoint arcs whose removal turns U_{n+1} into the union of two subgraphs that are homeomorphic to U_n and T_1, respectively (see the middle panel of Figure 4.1). Because U_n is universal, our claim holds true. We next claim that any binary phylogenetic tree shape T_{n+1} with $T_{n+1} = T_k + T_{n-k+1}$ can be a base tree of U_{n+1}. The right panel of Figure 4.1 indicates that U_{n+1} contains two distinct subsets of mutually vertex-disjoint arcs, one of which delineates U_{n-k+1} (shown in thick gray line) and the other distinguishes U_k from the remainder. Because both U_{n-k+1} and U_k are universal, our claim holds true. Therefore, U_{n+1} is universal. Hence, U_n is universal for all n.

Next, we will provide a method to create infinitely many universal tree-based networks on X from U_n. Let $\mathcal{E}_n$ be a tree-based network on X obtained from U_n by specifying a permutation π_0 of X. In what follows, we use the same notation i both for a leaf labelled i and for the terminal arc incident with i. A *crossover* σ_{ij} refers to a pair of crossed additional arcs between two distinct terminal arcs i and j as described in Figure 4.2. Note that σ_{ij} can be viewed as representing the transposition $(i\ j)$ of the labels. For any permutation π_1 $(\neq \pi_0)$ of X, there is a series of adjacent crossovers that converts π_0 into π_1 and then vice versa (note that any permutation can be expressed as a product of transpositions and that the symmetric group S_n is generated by the adjacent transpositions). Then, by sequentially adding $n! - 1$ series of crossovers, we can construct a universal tree-based network $\mathcal{U}_n$ on X from $\mathcal{E}_n$. Moreover, it is possible to create infinitely many universal tree-based networks on X because we may add an arbitrary number of redundant linking arcs between the terminal arcs of $\mathcal{U}_n$. This completes the proof. □

The construction described in the proof of Theorem 4.8 adds more arcs than necessary (*cf.* Figure 1 in [12]). It would be interesting to consider how to construct universal tree-based networks on X with the smallest number of arcs.

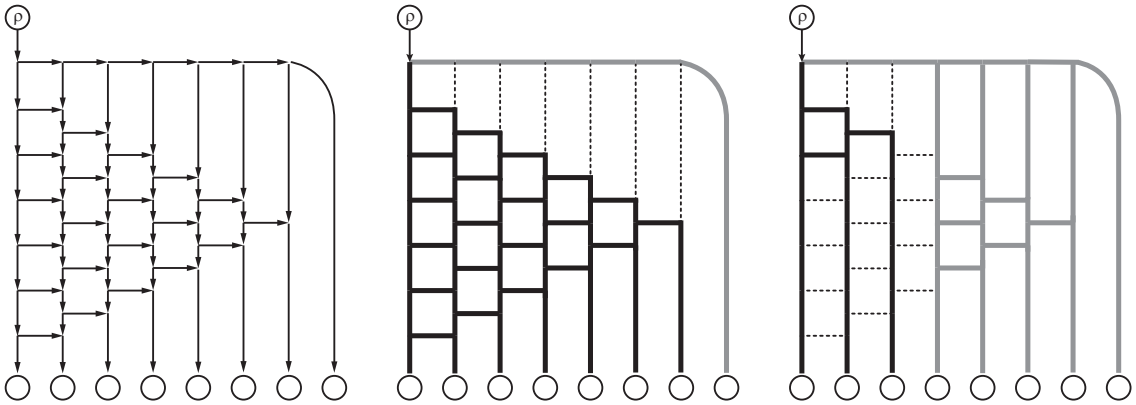

Figure 4.1: The first part of the proof of Theorem 4.8. The left panel is an illustration of U_n for $n = 8$. The other panels show examples of T_{n+1} in U_{n+1} for $n + 1 = 8$, and the right panel describes the case of $T_{n+1} = T_k + T_{n-k+1}$ with $k = 3$.

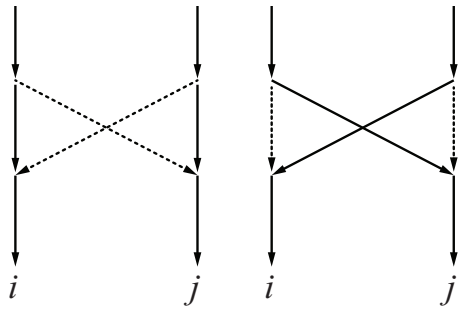

Figure 4.2: The second part of the proof of Theorem 4.8. Left: A *crossover* σ_{ij} is defined to be a pair of crossed additional arcs placed between arcs i and j ($i \neq j$) after subdividing both arcs twice. Right: When the two arcs in σ_{ij} are selected as tree arcs, σ_{ij} represents the transposition $(i\ j)$.

Notes

The original version [16] of this chapter was published in *Journal of Theoretical Biology*, Vol. 396, 7 May 2016, pp. 204–206 (doi:10.1016/j.jtbi.2016.02.023). We studied Problem 4.4 independently from Louxin Zhang [33].

Afterword

When I was in medical school and residency, I frequently thought that I had made a wrong decision by choosing medicine. I had never dreamt of working in a hospital but aspired to be a basic researcher to advance science and to be of service to mankind. Although I had no clear picture of how to achieve this, I always had a passion for integrating knowledge from different disciplines and grew up to what I am today.

Despite my initial regrets, I now appreciate my decision as I love the advantage of being able to think beyond barriers that exist between biological and mathematical sciences. It enables me to work at the intersection of those disciplines and makes me unique as a researcher. In the final year of my PhD, I received a grant to start a research project I had long envisaged, but it would not have been possible without this privilege.

Reading this thesis will remind me how I found my identity in an earlier stage of life and how I opened the door to the next steps. Therefore, even though the work in this thesis is only the first small step, it will be treasured all my life.

Momoko Hayamizu
Tokyo, January 2017

Bibliography

[1] K. Akashi, D. Traver, T. Miyamoto, and I. L. Weissman. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, Vol. 404, No. 6774, pp. 193–197, 2000.

[2] A. Baldisserri. Buneman's theorem for trees with exactly n vertices. *arXiv:1407.0048v1 [math.CO]*, 2014. preprint.

[3] H-J. Bandelt and V. Chepoi. Metric graph theory and geometry: a survey. In *Surveys on discrete and computational geometry*, Vol. 453 of *Contemporary Mathematics*, pp. 49–86. Amer. Math. Soc., Providence, RI, 2008.

[4] H-J. Bandelt, V. Chepoi, and D. Eppstein. Combinatorics and geometry of finite and infinite squaregraphs. *SIAM Journal on Discrete Mathematics*, Vol. 24, No. 4, pp. 1399–1440, 2010.

[5] O. Borůvka. O jistém problemu minimálním (About a certain minimal problem). *Práce moravské přírodovědecké společnosti (Acta Societatis Scientiarum Naturalium Moravicae)*, Vol. III, No. 3, pp. 37–58, 1926.

[6] B. H. Bowditch. Notes on Gromov's hyperbolicity criterion for path-metric spaces. In *Group theory from a geometrical viewpoint (ed. E. Ghys, A. Haefliger, A. Verjovsky)*, pp. 64–167. World Scientific, Singapore, 1991.

[7] F. Buckley and F. Harary. *Distance in Graphs*. Addison-Wesley Publishing Company, Advanced Book Program, Redwood City, CA, 1990.

[8] P. Buneman. The recovery of trees from measures of dissimilarity. *Mathematics in the Archaeological and Historical Sciences*, 1971.

[9] P. Buneman. A note on the metric properties of trees. *Journal of Combinatorial Theory, Series B*, Vol. 17, pp. 48–50, 1974.

[10] M. M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, Heidelberg, third edition, 2014.

[11] H. Fournier, A. Ismail, and A. Vigneron. Computing the Gromov hyperbolicity of a discrete metric space. *Information Processing Letters*, Vol. 115, No. 6, pp. 576–579, 2015.

[12] A. R. Francis and M. Steel. Which Phylogenetic Networks are Merely Trees with Additional Arcs? *Systematic Biology*, Vol. 64, No. 5, pp. 768–777, 2015.

[13] M. Gromov. Hyperbolic groups. In *Essays in group theory*, Vol. 8 of *Mathematical Sciences Research Institute Publications*, pp. 75–263. Springer, New York, 1987.

[14] G. Guo, S. Luc, E. Marco, TW. Lin, C. Peng, M. A. Kerenyi, S. Beyaz, W. Kim, J. Xu, P. P. Das, T. Neff, K. Zou, GC. Yuan, and S. H. Orkin. Mapping Cellular Hierarchy by Single-Cell Analysis of the Cell Surface Repertoire. *Cell stem cell*, Vol. 13, No. 4, pp. 492–505, 2013.

[15] E. F. Harding. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, Vol. 3, No. 1, pp. 44–77, 1971.

[16] M. Hayamizu. On the existence of infinitely many universal tree-based networks. *Journal of Theoretical Biology*, Vol. 396, pp. 204–206, 2016.

[17] M. Hayamizu, H. Endo, and K. Fukumizu. A characterization of minimum spanning tree-like metric spaces. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2015. in press.

[18] M. Hayamizu and K. Fukumizu. On minimum spanning tree-like metric spaces. *arXiv:1505.06145 [math.CO]*, 2015. preprint.

[19] M. D. Hendy. The path sets of weighted partially labelled trees. *The Australasian Journal of Combinatorics*, Vol. 5, pp. 277–284, 1992.

[20] A. Hernandez-Lopez. Of Trees and Bushes: Phylogenetic Networks as Tools to Detect, Visualize and Model Reticulate Evolution. In *Evolutionary Biology: Exobiology and Evolutionary Mechanisms*, pp. 145–164. Springer, 2013.

[21] D. H. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, 2010.

[22] V. Imrih and È. Stockiĭ. The optimal embeddings of metrics into graphs. *Akademija Nauk SSSR. Sibirskoe Otdelenie. Sibirskiĭ Matematičeskiĭ Žurnal*, Vol. 13, pp. 558–565, 1972.

[23] J. B. Kruskal, Jr. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, Vol. 7, pp. 48–50, 1956.

[24] P. M. Magwene, P. Lizardi, and J. Kim. Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics*, Vol. 19, No. 7, pp. 842–850, 2003.

[25] V. Moignard and B. Göttgens. Transcriptional mechanisms of cell fate decisions revealed by single cell expression profiling. *BioEssays*, Vol. 36, No. 4, pp. 419–426, 2014.

[26] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs, Jr., R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature biotechnology*, Vol. 29, No. 10, pp. 886–891, 2011.

[27] S. J. Salipante and B. G. Hall. Inadequacies of Minimum Spanning Trees in Molecular Epidemiology. *Journal of Clinical Microbiology*, Vol. 49, No. 10, pp. 3568–3575, 2011.

[28] C. Semple and M. Steel. *Phylogenetics*, Vol. 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[29] J. M. S. Simões-Pereira. A Note on the Tree Realizability of a Distance Matrix. *Journal of Combinatorial Theory*, Vol. 6, pp. 303–310, 1969.

[30] J. M. S. Simões-Pereira. An Optimality Criterion for Graph Embeddings of Metrics. *SIAM Journal on Discrete Mathematics*, Vol. 1, No. 2, pp. 223–229, 1988.

[31] È. D. Stockiĭ. On the imbedding of finite metrics into a graph. *Sibirskiĭ Matematičeskiĭ Žurnal*, Vol. 5, pp. 1203–1206, 1964.

[32] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, Vol. 32, No. 4, pp. 381–386, 2014.

[33] LX. Zhang. On Tree-Based Phylogenetic Networks. *Journal of Computational Biology*, Vol. 23, No. 7, pp. 553–565, 2016.