密级:_____

中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

四元组权重在进化网络重建中的应用

作者姓名：吴培

指导教师：Stefan Grünewald 研究员

中国科学院上海生命科学研究院

学位类别：理学博士

学科专业：计算生物学

培养单位：中国科学院上海生命科学研究院

2016 年 11 月

Quartet weights in phylogenetic network reconstruction

By

Wu Pei

A Dissertation Submitted to
Institute of Computaional Biology, Academy of Life Science
in Shanghai, CAS
In partial fulfillment of the requirements
For the degree of
Ph.D.

Institute of Computaional Biology, Academy of Life Science
in Shanghai, CAS

November, 2016

Quartet weights in phylogenetic network reconstruction

Abstract

This thesis studies quartet weights in phylogenetic network reconstruction. Quartet weights are a 4-taxa analog of distances used in many phylogenetic tree and network reconstruction methods. Those data contain more information than distances and make it possible to infer complicated phylogenetic history more accurately. Phylogenetic networks are visualization methods that generalizes phylogenetic trees. They allow for displaying non-tree events, e.g. hybridization. This thesis develops a method that estimates quartet weights from biological data and discusses how to reconstruct phylogenetic networks using quartet weights.

Chapter 3 studies the problem of reconstructing phylogenetic networks using quartet weights and distances together. We propose two algorithms QuartetDecomposition and 2-NeighborNet. They are quartet-weight versions of SplitDecomposition and Neighbornet, two methods that reconstruct networks from distances alone. Compared with distance-based methods, our algorithms are consistent on split systems with weaker compatibility conditions. We use a mitochondrial sequence data set to verify our methods, showing that they are capable of identifying complicated non-tree events with high accuracy.

Chapter 4 deals with estimation of quartet weights from sequences. The existing methods to compute quartet weights from a multiple sequence alignment all have their drawbacks, especially if the goal is to construct a phylogenetic network rather than just a tree. Simple pattern-counting approaches tend to produce a too

flat distribution of the quartet weights, while sophisticated model-based methods often inherently assume that there is a correct tree. In this chapter we present a method to calculate quartet weights based on the Hadamard conjugation, including an approach for handling rate inhomogeneities. This method is consistent under the K3ST+I+Γ model, a special case of the GTR+I+Γ model. We investigate its performance on simulated sequences based on the GTR model with and without reticulations and compare it with other quartet weight methods. We found that our new method is generally performing well when the correct model is not extremely far from K3ST. We also use a real data set to verify our method, indicating that it is capable of reconstructing phylogenetic trees and networks with high accuracy.

Chapter 5 studies split systems and cluster systems with general forbidden configurations, especially their maximal cardinalities. Upper bounds of them are important for the analysis of time and space complexities of reconstruction algorithms. The asymptotic growth of maximal cardinalities of (p, q)-hierarchies are explicitly decided. Some other important result are: on a ground set of n elements, the maximal cardinality of a $(-1, 3)$-hierarchy is between $n^3/9+O(n^2)$ and $n^3/6+O(n^2)$; the maximal cardinality of a $2'$-weakly compatible split system is between $3n^2/4 + O(n)$ and $n^2 + O(n)$ and maximal cardinality of a 2-weakly compatible split system is between $3n^2/2 + O(n)$ and $O(n^{2.5})$.

Keywords: Split System, Quartet weight, Phylogenetic network

Contents

List of Figures

Chapter 1

Introduction

Phylogenetics deals with one of the most fundamental problems in biology: how to infer relationships in a set of taxa and detect historical speciation events from data of, mostly, extant species. Many discoveries are crucial in providing tools for studying phylogenetic problems. Hereby we mention some remarkable events. In nineteenth century Charles Darwin proposed the theory of evolution in his book "*On the Origin of Species*", such discovery starts a new era of biology. "Nothing in Biology Makes Sense Except in the Light of Evolution", quoting from Dobzhansky [1], asserts the importance of evolution in appropriate measure. The tree structure naturally arises from proposed speciation mechanism. Another breakthrough taken place in 1953 when Watson and Crick [2] resolved the structure of DNA, which has been identified as genetic materials earlier. Such discovery impact phylogenetics in two counts, firstly it further verifies that genetic material is encoded in sequence with alphabet $\{A, T, C, G\}$ (or $\{A, U, C, G\}$, some virus encode genetic information in RNA). Secondly, "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." in their original words. Such mechanism explains how maternal and paternal genes mixed and generate new traits at molecular level, namely, the Mendelian laws. In 1968 Kimura [3] and 1969 King and Jukes [4] independently purposed the neutral theory of molecular evolution, they found that large proportion of the mutation in DNA sequence is neutral and random (for a systematic review, see[5, 6]). That makes modeling or simulating evolution at the level of DNA sequence and computational approach for detecting historical evolutionary events possible: if most of the mutation is extremely positive or negative we can only see the impact of environment adaptation but not evolutionary path from sequence data. We should also note that

advances of sequencing technology make it possible for biologist to access genetic information in low cost. Large number of sequences are available from database like GenBank or EMBL.

The classical and most prevalent approach for modeling evolutionary is phylogenetic tree, which already appears in Darwin's manuscript (Fig. 1.1). Such model assumes that for a set of taxa, there is a root vertex that represents a taxon being most common ancestor that lived a period of time ago, and during the time taxon, more exactly a group of organisms, splits and become independent taxa. This procedure named speciation, can be occurred iteratively. This process can be drawn in branching diagram called phylogenetic tree. This approach implicitly rests on two assumptions: each speciation event is simple and evolution of each clade after speciation is independent. Such assumptions is violated if we need to model reticulate evolutions, including events like HGT (Horizontal Gene Transfer)[7–9], hybridization[10–12], recombination[13, 14], incomplete lineage sorting and complicated patterns of gene duplication and loss[15, 16]. Those events would leads to signals that cannot displayed by a single tree. Phylogenetic network is one of alternative of tree that allows for displaying those events. Such approach has been applied to studies of evolution of microbe[17], plants[18] and fish[19].

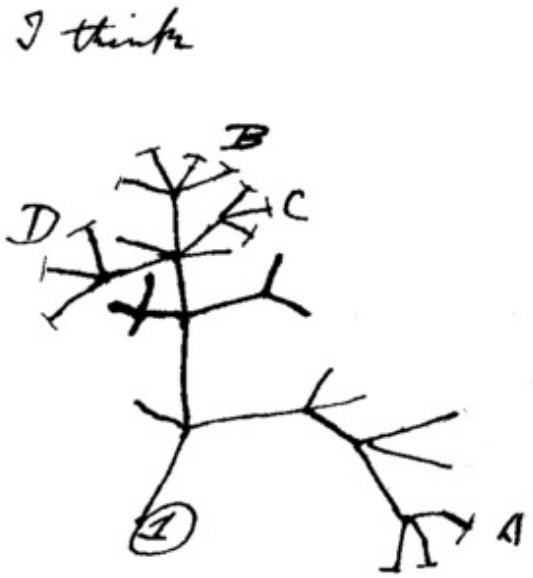


Figure 1.1: A phylogenetic tree drawn by Darwin, in *first notebook on Transmutation of Species*, 1837

One way of attacking phylogenetic problems is by using distances as an interme-

diate step, they are information extracted from pairwise comparison. Once distances were calculated, we need to find a network whose induced distances are close to them as much as possible. Distance data are more robust than sequence data since they take average across sites. We will give a brief review of network methods based on distance. The first attempts of reconstructing a network using distances trace back in 1970s. When Manfred Eigen sequenced 20 t-RNA of the *E. coli*, he calculated all distances and found that it's impossible to fit those sequences into a tree, it is even impossible for some subset of four taxa. This naturally leads to a problem, is it possible to construct a network rather than a tree? In [20] Dress *et al.* proposed *T-construction*. The method take distances as input and output a tight span, which is a complex (an object obtained by gluing polyhedra) that represents the phylogenetic relations of taxa. One of the deficit of such method is, such diagram is often too complicated and hard to analyze in general. *T-construction* is further modified into *SplitDecomposition* algorithm[21], which is the first practical network reconstruction method. The conception of split system, firstly introduced in[22], naturally arises from such algorithm. They also suggest that split system should satisfy weakly compatible condition, or two split systems could give identical distances. Circular split systems are a special class of weakly compatible split systems, they have many nice properties, Bryant *et al.* proposed *NeighborNet* method that reconstruct circular split systems. *NeighborNet* method are widely used, it is more robust and give highly resolved networks. *SplitDecomposition* give badly resolved networks for some dataset.

Distance of two taxa is obtained by pairwise scomparison. So it might lost some information when transform sequences to distances. A possible strategy that could remedy this deficit is using more global data. It has been shown that distance data is sufficient for tree reconstruction under certain model [30], however it might be beneficial if we use those higher data for network reconstruction. Latter we will show that relations of every three taxa is trivial: it contains no more information than distances data. Hence we propose methods utilizing relations of every four taxa, which can be encoded by distances and quartet weights together, we call these data 2-metric. Several methods have been proposed [31–33] using those data, in this thesis we would discuss the mathematics and applications of 2-metric.

The thesis consists of five chapters, the rest of this chapter introduces some

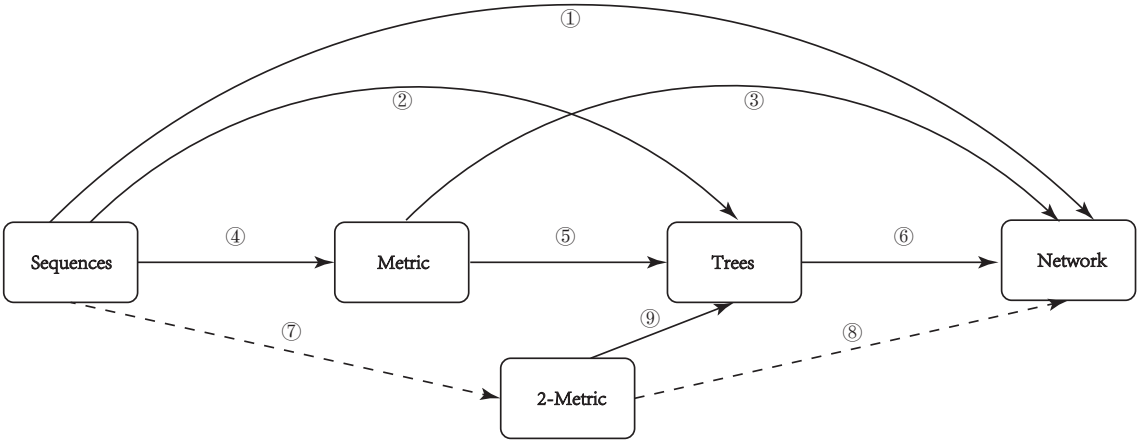


Figure 1.2: Diagrammatic view for pipeline of phylogenetic analysis, distinct methods were combined to carry out phylogenetic analysis. Some representative method of each step: 1: Median-joining network[25]; 2: Maximal parsimony[23], Maximal likelihood; 3: SplitDecomposition[21], NeighborNet[29]; 4: p-distance, logdet[34]; 5: Neighbor Joining[28], Minimal evolution[35]; 6: consensus split network[36]; 9: Quartet puzzling[37]. Dashed line indicates methods developed in this thesis.

notations and terminologies used in this thesis.

In chapter 2, we give a brief review on mathematics phylogenetic network. It has been observed that information encoded in a tree can be described by (weighted) split system satisfying certain condition called compatible condition. Phylogenetic network are visualization of split systems without those conditions, which makes displaying reticulate events possible. For clarifying the conception of induced network of a subset, the restriction operation on split system were introduced. It is also be useful in the ways that some algorithms can not distinguish certain split systems, so we need to consider families of networks without those split systems as substructures. Based on restriction operation we introduce the forbidden configuration to formalize such conditions. We also discuss weighted split system, including how distances or quartet weights were induced from a split system. The aim of reconstruction method is finding a network that fits the observed distances or quartet weights.

Chapter 3 deal with the problem of reconstructing phylogenetic networks from 2-metric. We start from study the structure of 2-metric, then developed a theory being analogue of the metric case. For 2-metric, we have the 2-very-weakly-compatible con-

dition, which corresponds to weakly compatible condition in metric case. Quartet-Decomposition method were proposed reconstructs 2-very-weakly-compatible split system from 2-metric. 2-circular split system is the analogue of circular split system in 2-metric case, they are 2-very-weakly-compatible split system and have maximal cardinality, we propose 2-NeighborNet method that reconstruct such split system. Mitochondrial sequence dataset were used to verify our method, showing that they are capable of identifying complicated non-tree events with high accuracy. At last we study the relationship of our construction and oriented matroids.

Chapter 4 propose a new method that estimates quartet weights using Hadamard conjugation. It has been noticed that almost all available quartet weight computing method are designed for reconstructing trees and not appropriate for network reconstruction. The most prevalent method for network is pattern counting, which is the analogue of Hamming distance in distance case. Such method overlooked multiple mutation, hence violates additivity assumption. Network reconstructed using pattern counting quartet weights would contains many false reticulate events. The method we propose take multiple mutation into account, which is a significant property of Hadamard conjugation. Quartet weights from such method are always consistently additive. It has been observed that different site have unequal evolutionary rates, and assuming equal rates across sites would also introduce significant error. We also consider this effect, our method includes a step estimating probabilistic distribution of rates. This method is consistent under the K3ST+I+Γ model, a special case of the GTR+I+Γ model. Simulation and real data studies are conducted, showing that unequal rate were crucial for the accuracy of quartet weight estimation. Combining with proper network reconstruction method we are able to reconstruct phylogenetic history with complicated events in high accuracy.

Chapter 5 focus on a problem under more general settings, we will consider split systems or clusters with given forbidden configurations, those compatible condition are often generated by algorithms on 2-metric or higher metric. We will study the maximal cardinality of those split system or clusters, which are useful in complexity analysis of certain algorithms. A remarkable result is maximal cardinality of split systems or clusters with any forbidden configurations is always bounded by polynomial, this is done by VC dimension argument[58]. Some special forbidden configurations were analyzed, the order of maximal cardinalities of (p, q)-hierarchies were explicitly

decided. Other important result obtained by combinatoric method is: the maximal cardinality of $(-1,3)$-hierarchy is between $n^3/9+O(n^2)$ and $n^3/6+O(n^2)$; the maximal cardinality of 2-weakly compatible split system is between $3n^2/2+O(n)$ and $O(n^{2.5})$. The last result is obtained by looking at graphs associated to split systems, such method is an powerful tool that always get finer estimation than VC dimension argument.

1.1 Notations and terminologies

Sets were always denoted by capital letters, elements were denoted by lower-case letters, vectors were denoted by bold letters. n-sets were sets with size n and n-subsets were subsets with size n. $\mathbb{Z},\mathbb{R}$ denotes integers and real numbers, $\mathbb{R}^+$ denotes non-negative real numbers. If S is a set, S^n means cartesian product of n copies of S, $\binom{S}{u}$ means the subset of S with cardinality u or cardinality satisfying condition u, for example $\binom{S}{\leq 2}$ means all subsets of S with cardinality no more than 2. If A is a set and x is an element Ax were shorthand for $A\cup\{x\}$ and $A-x$ for $A\backslash\{x\}$. The cardinality of a set A is denoted as $|A|$ or $\#A$. An caret signifies the omitted variable, for example $f(x_1,\dots,\hat{x}_i,\dots)$ is $f(x_1,\dots,x_{i-1},x_{i+1},\dots)$. $(x_1,\dots,x_n)$ signifies circular ordering (more detailed definition were in below); $[x_1,\dots,x_n]$ is linear ordering: an ordered set with $x_i<x_j$ iff $i<j$; $\{x_1,\dots,x_n\}$ is unordered set. If $f:X\to Y$ is a mapping and $A\subset Y$, $f^{-1}(A)$ is the preimage of A. If $Y\subseteq X$, define inclusion map $i:Y\to X$ being the map to identical elements.

Hereby we give a strict definition of circular ordering:

Definition 1. *A circular ordering $(x_1,\dots,x_n)$ is n marked points on circle with $x_1,\dots,x_n$ being their clockwise or anticlockwise order. Namely $(x_1,\dots,x_n)$, $(x_2,\dots,x_n,x_1)$ and $(x_n,\dots,x_1)$ are equivalent. The distance of two elements x_i and x_j are defined as $\min(|i-j|,n-|i-j|)$. two elements are neighbors iff distance of them is 1. An interval is subset like $\{x_i,x_{i+1},\dots,x_j\}$ we say a subset of $\{x_1,\dots,x_n\}$ is consists of k intervals is it can expressed as union of at least k intervals.*

Chapter 2

Background

Having introduced the motivation of phylogenetic network, we will give a formal treatment on the mathematical background of phylogenetic network in this chapter. The central idea of such formalism is, the topology of a phylogenetic network is decided by a set of bipartite, and other information were encoded by data associated with those bipartite. This chapter is a common background for the following chapters.

2.1 Phylogenetic Network and Split System

As we have discussed in previous sections, phylogenetic network is a tool that represent reticulate events by graphs instead of trees. There are various types of phylogenetic network method [38, 39] and in our discussion we will focus on split networks. There are roughly two type of network method, "abstracted" or "explicit", which corresponds to unrooted or rooted phylogenetic trees. Split network belongs to the abstract method. In the explicit method each node or edge corresponds to real phylogenetic events, while abstracted method are more focused on displaying relations of taxa. There are several reasons that abstract method are more prevalent: trees with distinct rooting and proper assigning of parameters are non-identifiable (This is the so called *pulley principle*, see Section 4.2) so biologically unrooted trees are more interesting, which makes abstract network method more widely used. Moreover the combinatorics of split networks are more simple and easy to deal with.

Definition 2. *The ground set is denoted X, which is the set of taxa we study, a split on X is a bipartition of X and denoted as $s = A|B$, with $A = X\backslash B$, and*

$A|B$ is considered to be same as $B|A$, we would assume A and B to be nonempty as default, $\emptyset|X$ count as valid split when explicitly noted. We will adopt another notation $s' = A'|B'$ for $A', B' \subseteq X$. $A' \cap B' = \emptyset$ and $A \cup B \subseteq X$. In this case we say s' is a partial split. Sometimes we will utilize the term "full split" for split in contrast of partial split. We say a full split $s = A|B$ is an k-split iff $\#A = k$ or $\#B = k$. For partial split $A|B$ we say it is $m|n$-split if $\#A = m$ and $\#B = n$, especially $2|2$-split were called quartet. For $s_1 = A_1|B_1$ and $s_2 = A_2|B_2$ if we have $A_1 \subseteq A_2$ and $B_1 \subseteq B_2$ we say s_1 is displayed by s_2 , or $s_1 \dashv s_2$. A split system S is a set of full split on X, the set of all full split on X is denoted as $\mathcal{P}(X)$, size of $\mathcal{P}(X)$ is $2^{n-1} - 1$ if $\#X = n$.

In practice we will associate each split with a positive number, for example time of evolution, in this case we call these number *weights*. We further assumes that weights are additive, which means the weight of a partial split is the sum of all full split that displays it:

$$w(s') = \sum_{s' \dashv s,\, s \in S} w(s) \tag{2.1}$$

Definition 3. *For a given weighted split system S, we define diversity of a subset $Y \subseteq X$ as:*

$$\delta(Y) = \sum_{A|B \in S\,, A \cap Y \neq \emptyset\,, B \cap Y \neq \emptyset} w(A|B)$$

The distance of to two taxa can be understood as weight of $1|1$-partial split or diversity of the 2-subset: $d(a, b) = w(a|b) = \delta(a, b)$, quartet weights were defined as the weights of some $2|2$-partial split, thus there are in all $3\binom{n}{4}$ quartet weights. Another useful formula is $w(A|B) = w(Ax|B) + w(A|xB)$ if $x \notin A \cup B$.

Then we come back to the tree case. The split formalism is helpful for understanding the information encoded in a X-tree, which is a formulation of valid phylogenetic tree.

Definition 4. *An X-tree is a pair $(\mathcal{T}, \phi)$ such that $\mathcal{T}$ is a tree and $\phi : X \to V$ is a map from taxa set X to set of vertices V with following property: if $v \in V$ has degree of connection 2, then v in the image of ϕ.*

Once an edge were removed an X-tree would break up into two connected components, thus every edge induce a bipartite of taxa set. Hence every phylogenetic

tree would induce a split system. We have such theorem characterizing split system induced by a tree.

Definition 5. *if two splits $A_1|B_1$ and $A_2|B_2$ satisfies:*

$$\emptyset \in \{A_1 \cap B_1, A_1 \cap B_2, A_2 \cap B_1, A_2 \cap B_2\} \tag{2.2}$$

we say they are compatible. A split system is compatible if the splits are pairwise compatible. An equivalent statement of compatible condition is for every four element in ground set, denoted by $1, 2, 3, 4$, $12|34$ and $13|24$ can not be displayed by splits in S simultaneously.

Theorem 1. *A split system is induced by an X-tree iff compatible. Moreover two X-tree are isomorphic iff they induce the same split system.*

A phylogenetic network is a visualization of split system, hereby we gives a formal definition.

Definition 6. *[40] A phylogenetic network on X is a finite, connected, bipartite graph $G = (V, E)$, a mapping $f : X \to V$ and a coloring $\sigma : E \to K$ which K is the set of color. The coloring should have properties below:*

1. *surjective:edge connected with one vertex should of different color.*
2. *isometric:any shortest path (we just need the edge with same color have same positive length, the assigning of length does not have impact on shortest path) between any two vertex consists of edges with different color, if the shortest path between any two vertex is not unique then the edges of them must have same set of coloring.*

The colors are in one-to-one correspondence with splits, with the properties that when removing edges with a color the phylogenetic network would break up into two connected components, which generates a bipartite of taxa set being the split, when depicting network we always draw edges with same color as parallel line segments with same length. There is a canonical method to draw phylogenetic network from splits system [41] (generally the phylogenetic network of a splits system may not be unique, and even the minimal network, in the sense that we can not remove any edge, do not necessarily exist, see Fig. 2.1), note that for compatible split system

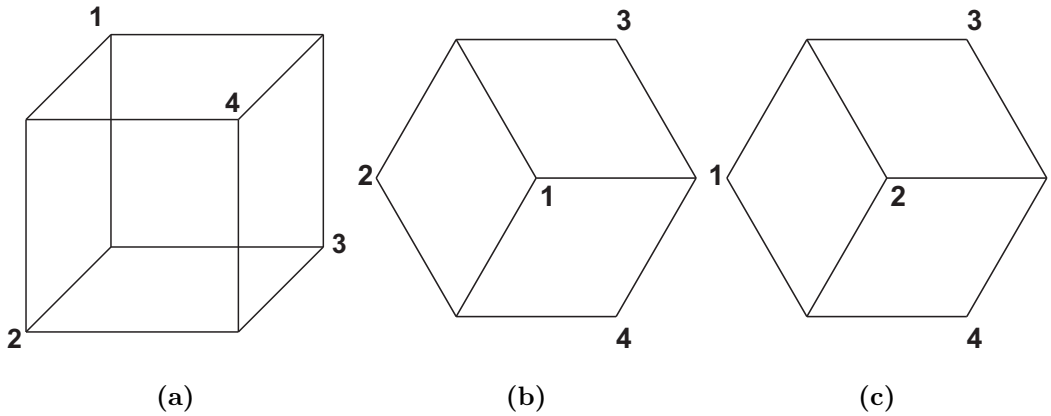


Figure 2.1: An Example of distinct phylogenetic network of same split system $\{12|34, 13|24, 14|23\}$.

the minimal phylogenetic network exist and is always the tree induce it. If we draw the graph such that the edge length is the weight of the split we have $w(A|B)$ is the shortest distance between convex hulls of taxa sets A and B (An example shown in Fig. 2.2, convex hull of a subset is defined as the union of all shortest path between any pairs in that subset).

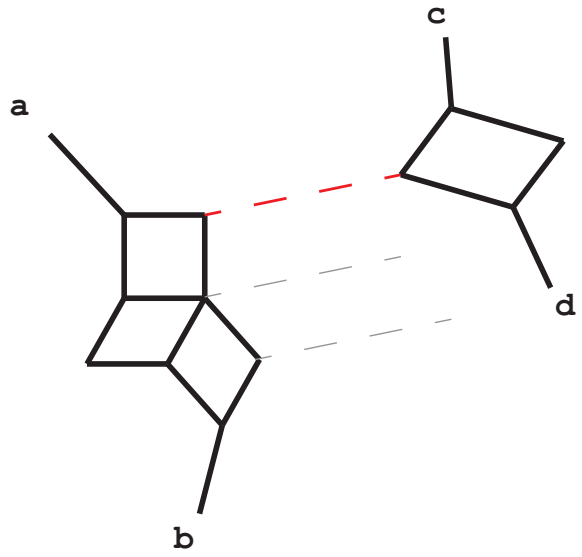


Figure 2.2: A graph illustrating the geometric meaning of $w(A|B)$. Red dash line segment indicates the distances between convex hulls of taxa sets $A = \{a, b\}$ and $B = \{c, d\}$. Grey dashed line indicates omitted parts of the network.

2.2 Restriction, Compatible condition and Forbidden Configuration Formalism

One of the radical problem for such approach is, once look at a subset of taxa set, how would global structure induce substructure of them, more explicitly, if we do the same analysis on a subset of taxa, how would the network of the subset related with the network of whole taxa set. For example, in distance-based reconstruction method, we firstly transform sequences or other biological data into distances and try to find a tree or network such that induced distances close to observed distances as much as possible. The distance can be viewed as quantity that encodes the information of substructure of 2-subsets, so it is crucial to define what do we exactly mean by "substructures". In this section we will explicitly define the conception of substructure using the restriction operation.

Definition 7. *Given a mapping $f : X \to Y$ and a split $s = A|B$ on Y. The pullback $f^\star s$ were defined as split on X: $f^\star s = f^{-1}(A)|f^{-1}(B)$. And for split system S on Y we define pullback $f^\star S = \{f^\star s | s \in S\}$. If the S is weighted, we use this formula to decide the weight of the splits in pullback: $w(s') = \sum_{f^\star s' = s,\, s \in S} w(s)$. For $X \subset Y$ we define the split and split system restricted on X as the pullback of inclusion map i, denoted as $s|_Y$ and $S|_Y$ respectively.*

A worthwhile mentioning property of restriction is such action remain distances or weighted of partial splits unchanged. More exactly $w_S(A'|B') = w_{S|_Y}(A'|B')$ for $A' \cup B' \subseteq Y$, w_S indicates weight calculated from weighted split system S.

In many cases we want to look at split system that do not have certain substructure. For simplicity we will use forbidden configuration to characterize compatible condition:

Definition 8. *A forbidden configuration is a split system on $\{1, \ldots, n\}$. And we say a split system S is with (or have) forbidden configuration S' iff there are no injective mapping $f : \{1, \ldots, n\} \to X$ such that the forbidden configuration S' is subset of $f^\star S$. If there exist a mapping f such that $S' \subseteq f^\star S$ we will write $S' \trianglelefteq S$.*

For example, the forbidden configuration of compatible split system is $\{12|34, 13|24\}$. Here we introduce two notations for convenience, the split system $\mathcal{P}(\{1, \ldots, n\})$ were denoted as F_n, the split system of all possible m-splits on

$\{1, \ldots, n\}$ were denoted F_n^m, split system with those forbidden configuration were of special interest.

Chapter 3

Applications of 2-metrics in phylogenetics

This chapter mainly develops the theory of 2-metric, which is the term refer to structure of distances and quartet weights on a set. There has been several attempts for building theory for describing the structures beyond metric in [42, 43], this chapter will provide a new view point based on split systems. We start from reviewing several canonical constructions that reconstructs network from distances, especially *T-theory* and two algorithms in Section 3.1 and 3.2. Then in Section 3.3 we generalize the SplitDecomposition algorithm to 2-metric case, namely the QuartetDecomposition algorithm. In Section 3.4 we access the possibility of generalizing T-theory into 2-metric case and only partial results were obtained. In Section 3.5 we found the proper generalization of circular split system and an algorithm were proposed to reconstruct such split system. In the last section connection with oriented matroids were discussed.

3.1 SplitDecomposition and NeighborNet

In this section we will review the SplitDecomposition and NeighborNet algorithm which both reconstruct a network from a metric space.

Historically SplitDecomposition is an algorithm derived from T-theory, however in this section we will direct present this algorithm from algebraic aspect. A similar argument can be found in [44, 45], in which the SplitDecomposition algorithm is generalized into group-valued metric.

A split system S can have size up to $2^{n-1}-1$ on a taxa set $\#X = n$, but only have $\binom{n}{2}$ distance. So there exist two nonidentical splits system induce same metric, it is impossible to reconstruct arbitrary split system from distance data, we have to assume some constraint on the split system.

Firstly we consider a metric as a $\binom{n}{2}$-dimensional vector. Then the space of all metrics on n points is a subset of $\mathbb{R}^{\binom{n}{2}}$, which is a polyhedral cone, namely the metric cone. In this way we can write out the vector of (pseudo-)metric induced by a single split.

$$\boldsymbol{d}_{a,b}(s) = \begin{cases} 0 & a|b \not\dashv s \\ 1 & a|b \dashv s \end{cases} \tag{3.1}$$

Our notation were a bit different from usual, $\boldsymbol{d}(A|B)$ stands for metric induced by single split, a, b is the index, our aim is to decompose a metric into sum of several metrics induced by splits. We start our analyze from small point number case. For 3-point case it's trivial because we have 3 distances and 3 splits which makes a bijection from the space of metrics and space of weights. We have:

$$w(a|bc) = \frac{d(a,b) + d(a,c) - d(b,c)}{2} \tag{3.2}$$

Other two weights is decided in same way. Note that $w(a|bc)$ is always non-negative which is manifested by the triangle inequality of metric. For 4 point we have 6 distance but 7 splits, thus we have to make constraints on that split system. We observe that:

$$\boldsymbol{d}(a|bcd) + \boldsymbol{d}(b|acd) + \boldsymbol{d}(c|abd) + \boldsymbol{d}(d|abc) = \boldsymbol{d}(ab|cd) + \boldsymbol{d}(ac|bd) + \boldsymbol{d}(ad|bc) \tag{3.3}$$

We can keep the metric unchanged if we add same number to all the weight of 2-split and take away a same number from trivial splits. So if we ask the minim one in three 2-splits to be zero then all split weights is fixed, such a condition is called weakly compatible condition. We have:

$$w(ab|cd) = \frac{1}{2} * (\max \begin{Bmatrix} d(a,b) + d(c,d) \\ d(a,c) + d(b,d) \\ d(a,d) + d(b,c) \end{Bmatrix} - (d(a,b) + d(c,d))) \tag{3.4}$$

From such an idea we can write out the weakly compatible condition:

Definition 9. *A split system S is weakly compatible iff having forbidden configuration F_4^2.*

For $n > 4$ case we can prove for arbitrary split $s = A|B$ in a weakly compatible split system S there exist $a_1, a_2 \in A$, $b_1, b_2 \in B$ such that only s displays $a_1a_2|b_1b_2$[21], we will generalize this property in Section 5.1.3. So we have:

$$w_d(s) = \min_{s \vdash a_1a_2|b_1b_2} w(a_1a_2|b_1b_2) \tag{3.5}$$

This is how SplitDecomposition works, we can show that method is consistent, namely if the input distance is induced from weakly compatible split system we can exactly recover the weighted split system. This also give a proof that size of a weakly compatible split system can not exceed $\binom{n}{2}$.

Then we come to the discussion of NeighborNet method. Consider taxa arranged in a circular manner $C = (x_1, \ldots, x_n)$, we can sect them with straight lines and get $\binom{n}{2}$ splits: $S(C) = \{x_i, \ldots, x_j | x_1, \ldots, x_{i-1}, x_{j+1}, \ldots, x_n\}$. A split system S is called circular split system if there exist a circular ordering C that $S \subseteq S(C)$.

The circular split system has many desirable property:

1. Circular split system is weakly compatible
2. A weakly compatible split system with cardinality $\binom{n}{2}$ is circular
3. Every compatible split system is circular
4. Network of a circular split system can always been drawn in a planar way and outer labeled

Neighbornet algorithm is a consistent method that reconstructs circular split system from metric. The first step is reconstruction of circular ordering. This is done by agglomerative process. One can compare it with Neighbor-joining method. Like Neighbor-joining[28, 46] method, starting with a set of nodes, initially we construct a score on pairs of nodes for identifying neighbors, then connect those elements into a circular ordering. Note that there are extra difficulties when reconstructing the circular ordering compared with tree case. We first need to identify neighbors (cherries in tree case), secondly when merging clusters, we have to decide which end to join, for clusters are linear orderings but not unordered sets in Neighbornet case. For example, if we merge two cluster $[x_1, x_2]$ and $[x_3]$, we may get $[x_1, x_2, x_3]$ or $[x_3, x_1, x_2]$, those cluster would lead to distinct split systems.

Neighbornet were first proposed in [29]. Later explained in [47] that such method can be understand as a greedy algorithm for the traveling salesman problem. The details of implementation in this two paper is slightly different. This difference of this two method can be illustrated in these pseudo-codes:

Data: a metric d on $X = \{1, \ldots, n\}$
Result: circular ordering $C = (x_1, \ldots, x_n)$
$Y = \{1, \ldots, n\}$ is a set of active elements;
while $\#Y > 3$ **do**
  for $\{i, j\} \in \binom{Y}{2}$ **do**
    calculate $q(i, j)$ using some neighboring score;
  end
  Choose a pair r, s that minimize $q(i, j)$, label them as neighbors.;
  while *the exist an element* $y \in Y$ *having two neighbors* x *and* z **do**
    Replace interval x, y, z by u and v ;
  end
  Update metric using some weighting method;
end
;
Using the substitution information to rebuild the circular ordering.

Algorithm 1: The Neighbornet algorithm in [29]

The performance of two implementation has no big difference, the [29] version is more robust since they take average when replacing vertices and in [47] version every step is an heuristic process on the traveling salesman problem.

The neighboring score is the same as the one of Neighbor-joining method:

$$q(i, j) = (n - 2)d(i, j) - \sum_{x \neq i} d(i, x) - \sum_{x \neq j} d(j, x)$$

One would ask why should we take the coefficient $n - 2$. There are three explanations on how this score function works:

1. $q(i, j)$ such that all trivial split have same coefficient respectively.[29]
2. $q(i, j)$ be the average of $s(C) = \sum_{0<i\leq n}^{n} d(c_i, c_{i+1})$ which C being all possible circular orderings with i, j neighbors.[47]

Data: a metric d on $X = \{1, \dots, n\}$

Result: circular ordering $C = (x_1, \dots, x_n)$

$C = \{[1], \dots, [n]\}$ is a collection of linear orderings;

while $\#C > 3$ **do**

  for $\{i, j\} \in \binom{C}{2}$ **do**

    calculate $q(C_i, C_j)$ using some neighboring score;

  end

  Choose a pair C_r, C_s that minimize $q(C_i, C_j)$.;

  for *i be marginal elements of C_r, j be marginal elements of C_s* **do**

    calculate $q'(i, j)$ using some neighboring score;

  end

  Choose a pair i, j that minimize $s'_q(i, j)$.;

  Let C_n be the new cluster that join the C_r, C_s at the i, j.

  $C = C \backslash \{C_r, C_s\} \cap \{C_n\}$;

end

;

Algorithm 2: The Neighbornet algorithm in [47]

3. $s(i, j) = \sum_{x,y \neq i,j} w(ij|xy) - w(ix|jy)$[31]

this three ways give the same neighboring score up to a factor. In the original version, the u, v should be superposition of x, y, z, distances were updated using following formula:

$$\begin{aligned} d(u, i) &= (\alpha + \beta)d(x, i) + \gamma d(y, i) \\ d(v, i) &= \alpha d(y, i) + (\beta + \gamma)d(z, i) \\ d(u, v) &= \alpha d(x, y) + \beta d(x, z) + \gamma d(y, z) \end{aligned}$$

With $\alpha + \beta + \gamma = 1$ and all positive (in practice we take $\alpha = \beta = \gamma = \frac{1}{3}$).

Remark 1. *We should note that after such replacing the circular structure is maintained. Consider the circular order $(\dots, x, y, z, \dots)$, when $\alpha = 1$ and $\beta = \gamma = 0$, it's equivalent with removing z and replace x with u and y with v, the resulting metric is induced by some weighted circular split system with circular order $(\dots, u, v, \dots)$.*

For $\alpha = 0$, $\beta = 1$ and $\gamma = 0$ the corresponding circular order is $(\dots, u, v, \dots)$, For $\alpha = \beta = 0$ and $\gamma = 1$ the corresponding circular order is still $(\dots, u, v, \dots)$. The resulting metric with general α,β and γ is linear combination of this three case, thus is induced by some weighted circular split system of circular order $(\dots, u, v, \dots)$.

For the second implementation the distance of clusters were decided by averages:

$$d(C_i, C_j) = \frac{1}{\#C_i \cdot \#C_j} \sum_{x_i \in C_i, x_j \in C_j} q(x_i, x_j)$$

Finally the split weights were estimated using non-negative least square optimization method: once we've find the circular ordering, the split system is fixed and defines a full rank linear mapping from weights to metric: $A : \boldsymbol{w} \to \boldsymbol{d}$, A can be written as matrix. Then the weights were decided by non-negative least squares procedure:

$$\begin{aligned} &\min \ ||\boldsymbol{d} - A\boldsymbol{w}||_2 \\ &s.t. \ \boldsymbol{w} \geq 0 \end{aligned} \tag{3.6}$$

Compared with SplitDecomposition method Neighbornet algorithm is more robust and tend to have more splits hence being more informative and more widely used.

3.2 T-Theory

In the previous section when we introduce the SplitDecomposition method we seemed to have the arbitrariness to choose forbidden configuration $\{12|34, 13|24, 14|23\}$ or $\{1|234, 2|134, 3|124, 4|123\}$. It's partially from biological reason that trivial splits should always be in a split system, but this is stemming from the T-theory, or say, the properties of metric space [48], here we gives a brief review, mostly following [49].

Consider the category $\mathcal{M}$ with objects being metric spaces and morphism being non-expansive mappings (the mapping $f : M \to N$ such that $d_N(f(x), f(y)) \leq d_M(x, y)$ is called non-expansive, those are always continuous.) Such category have the injective hull property, namely, for every object M there is a injective object $\bar{M}$ and monomorphism $M \to \bar{M}$ being universal, note that in $\mathcal{M}$ every monomorphism is a embedding.

Hereby we introduce a equivalent characterization of injective space.

Definition 10. *Denote the close ball with center x and radius r as $B(x,r)$. A metric space is hyperconvex iff:*

1. $r_1 + r_2 \geq d(x_1, x_2) \Rightarrow B(x_1, r_1) \cap B(x_2, r_2) \neq \emptyset$.
2. *If a finite set of closed balls are pairwise intersecting, then the intersection of them are non-empty.*

One could verify that L^∞ space is hyperconvex, which is an important prototype we'll use.

Theorem 2. *[49] A metric space is hyperconvex iff injective.*

This instantly leads to the *T-construction* that construct the injective hull of every metric space (we only consider the finite case, note that there is no big difference for infinite metric space). The idea is to make a space hyperconvex one need to adjoint points to manifest two conditions in definition 9. For every finite metric space (X, D), $X = \{x_1, \ldots, x_n\}$ we define an unbounded polytope $P(X, D)$:

$$P(X,D) = \left\{ v = \begin{pmatrix} v_1 \\ \ldots \\ v_n \end{pmatrix} | v_i + v_j \leq d(x_i, x_j) \right\} \tag{3.7}$$

Definition 11. *Given a finite metric space (X, D). We define the tight span [20] $T(X, D)$ to be the (component-wise) minimal elements.*

A equivalent characterization is

$$T(X,D) = \left\{ v = \begin{pmatrix} v_1 \\ \ldots \\ v_n \end{pmatrix} | v_i = \sup_j (d(x_i, x_j) - v_j) \right\}$$

.

$T(X, D)$ were naturally endowed with the L^∞-metric. We have a canonical isometric embedment, the Kuratowski mapping $f : X \to T(X, D)$,
$x_i \mapsto (d(x_1, x_i), \ldots, 0, \ldots, d(x_n, x_i))$. The $T(X, D)$ was exactly the injective hull we mentioned before [20].

Another way to think of T-theory is: consider the single extension y of the space X, such extension can be parameterized by $d(y, x_i)$ of all $x_i \in X$, this then we denote such number v_i, thus $v_i + v_j \geq d(x_i, x_j)$ and $v_i - v_j \leq d(x_i, x_j)$. When we consider the minimal elements of them the second inequality in automatically satisfied. In other words, tight span is moduli space of minimal single extensions, in this aspect Kuratowski mapping is natural.

One of the nice property of tight span is if the metric is a tree metric, namely the metric induced by a tree with leaf being the points of the metric space the tight span is exactly the tree. If the metric is non-tree, there are parallel edges that divides the taxa set into two clusters, however length of such edge are generally hard to compute[50], and in general, tight span is not a phylogenetic network. A more strict condition is introduced and give rise to more computable method.

Fix a ground set X the summation of two metric $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ are defined as the sum pointwise as before. If $\boldsymbol{d} = \boldsymbol{d}_1 + \boldsymbol{d}_2$ we always have $P(X, \boldsymbol{d}_1) + P(X, \boldsymbol{d}_2) \subseteq P(X, \boldsymbol{d})$(the sum is Minkowski sum). If $P(X, \boldsymbol{d}_1) + P(X, \boldsymbol{d}_2) = P(X, \boldsymbol{d})$ we say the decomposition $\boldsymbol{d} = \boldsymbol{d}_1 + \boldsymbol{d}_2$ is coherent, an more useful equivalent condition is $T(X, \boldsymbol{d}_1) + T(X, \boldsymbol{d}_2) \supseteq T(X, \boldsymbol{d})$. Our aim is, decompose a given metric $\boldsymbol{d}$ as the sum of several split metric and a residue $\boldsymbol{d}'$ being small as much as possible $\delta_1 \boldsymbol{d}_1 + \cdots + \delta_m \boldsymbol{d}_m + \boldsymbol{d}'$. This is exactly what SplitDecomposition doing.

Then we will show that consider the metric $d(i, j) = 2(i, j \in \{1, 2, 3, 4\})$. The decomposition $\boldsymbol{d} = \boldsymbol{d}(1|234) + \boldsymbol{d}(2|134) + \boldsymbol{d}(3|124) + \boldsymbol{d}(4|123)$ is coherent while $\boldsymbol{d} = \boldsymbol{d}(12|34) + \boldsymbol{d}(13|24) + \boldsymbol{d}(14|23)$ not.

Lemma 1. *$T(X, \boldsymbol{d}(A|B))$ is a line segment from $(v_1^1, \ldots, v_n^1)$ and $(v_1^2, \ldots, v_n^2)$. $v_i^1 = 0, v_i^2 = 1$ iff $x_i \in A$ and $v_i^1 = 1, v_i^2 = 0$ iff $x_i \in B$.*

For $\boldsymbol{d} = \boldsymbol{d}(12|34) + \boldsymbol{d}(13|24) + \boldsymbol{d}(14|23)$, $d(i, j) = 2$ for $i, j \in \{1, 2, 3, 4\}$ and $i \neq j$. We can show that the $T(X, \boldsymbol{d})$ is the union of line segment of $(0, 2, 2, 2)$ to $(1, 1, 1, 1)$ $(2, 0, 2, 2)$ to $(1, 1, 1, 1)$ $(2, 2, 0, 2)$ to $(1, 1, 1, 1)$ and $(2, 2, 2, 0)$ to $(1, 1, 1, 1)$. $T(X, \boldsymbol{d}(1|234)) + T(X, \boldsymbol{d}(2|134)) + T(X, \boldsymbol{d}(3|124)) + T(X, \boldsymbol{d}(4|123))$ is a cube that contains these line segments, while $T(X, \boldsymbol{d}(12|34)) + T(X, \boldsymbol{d}(13|24)) + T(X, \boldsymbol{d}(14|23))$ is a subset of hyperplane $x_1 + \cdots + x_4 = 6$, hence $(1, 1, 1, 1)$ do not lies on it.

3.3 2-metric and QuartetDecomposition

After those preparations we will introduce the main object in this paper: 2-metric, which is the term for metric and quartet weight structure on the ground set. In this section we will look at how to generalize the theory of SplitDecomposition into the 2-metric cases.

3.3.1 2-metric on Small Set

Like what we do in the section explaining the SplitDecomposition. We start from analyzing the small taxa number case. We will also clarify the condition for a 2-metric being proper such conditions were modeled on 2-metric induced from weighted split system but choose a condition on subsets with minimal cardinalities.

3.3.1.1 4-taxa

From the previous section for 4 taxa we can show that:

$$w(ab|cd) - w(ac|bd) = \frac{d(a,c) + d(b,d) - d(a,b) - d(c,d)}{2} \tag{3.8}$$

If given all the distance we can always infer all 3 quartet weight from one for every four element. So we have at most $\binom{n}{4} + \binom{n}{2}$ independent variables, latter we will show that this bound is tight.

Remark 2. *The 2-metric can also be cryptomorphically encoded using 2- and 4-diversities described in [43]. One could verify that:*

$$w(ab|cd) = \delta(a,b,c,d) - \frac{1}{2} * (\delta(a,c) + \delta(b,c) + \delta(a,d) + \delta(b,d))$$

We will show there is no linear dependencies of those diversities.

3.3.1.2 5-taxa

For 5 point we have $2^{5-1} - 1 = 15$ split which is equal to $\binom{5}{4} + \binom{5}{2}$. So we are still able to calculate all the split weight. By solving the linear equation we have:

$$w(ab|cde) = \frac{1}{2} * (w(ab|de) - w(ac|de) + w(ac|bd) - w(bd|ce) + w(ab|ce)) \tag{3.9}$$

The same argument can be found in [33]. Noted that if equation (3.8) holds the result is not affected by the order of a, b or c, d, e. Like metric on arbitrary set

should be symmetric and subadditive. We also assume quartet weights satisfies some condition:

Definition 12. *Suppose X is a metric space and $w : X^4 \to \mathbb{R}^+$ is a quartet weight on X iff:*

1. $w(a,b,c,d) = w(b,a,c,d) = w(c,d,a,b)$
2. $w(a,b,c,d) - w(a,c,b,d) = \frac{1}{2}(d(a,c) + d(b,d) - d(a,b) - d(c,d))$
3. $w(a,b,d,e) - w(a,c,d,e) + w(a,c,b,d) - w(b,d,c,e) + w(a,b,c,e) \geq 0$
4. $w(a,b,d,e) + w(a,c,d,e) + w(a,c,b,d) - w(b,d,c,e) + w(a,b,c,e) \leq d(a,d) + d(a,e) - d(d,e)$

Given a weighted split system, $w(a,b,c,d) = w(ab|cd)$ is the induced quartet weight. A 2-metric on a set X is defined as all the distances and quartet weight on X satisfying those condition. We will use the prefix 2- for 2-metric analogies. We use $\boldsymbol{q}$ to denote the vector of 2-metric:

$$\boldsymbol{q} = (d(a,b), d(a,c), d(b,c), \ldots, w(a,b,c,d), \ldots)^T \tag{3.10}$$

Like before we use $\boldsymbol{q}(A|B)$ be $\boldsymbol{q}$ induced by a split system of a single split $A|B$ with weight 1. A 2-metric is L_1 iff induced by some weighted split system namely it's linear combination of $\boldsymbol{q}(s)$ with positive coefficients. Let Q is the space of all $\boldsymbol{q}$ induced by weighted split system. $\bar{Q}$ be its linear closure.

Remark 3. *The condition 4. is equivalent with $w(a|bcde) \geq 0$. The whole condition is the same as saying, for every subset of 5 elements the induced 2-metric is L_1.*

3.3.1.3 6-taxa

If we are given 2-metric on a 6-set, we can calculate arbitrary partial split weight except full split from arguments in previous section. By simple calculation we can observe that:

$$\boldsymbol{q}(ab|cdef) + \cdots + \boldsymbol{q}(ef|abcd) \tag{3.11}$$

$$= \boldsymbol{q}(a|bcedf) + \cdots + \boldsymbol{q}(f|abcde) + \boldsymbol{q}(abc|def) + \cdots + \boldsymbol{q}(aef|bcd) \tag{3.12}$$

This is the unique linear combination of splits on 6 points that vanishes up to a factor. If we add the same number on weights of set $\{ab|cdef, \ldots, ef|abcd\}$ and

subtract the same number on weights of set $\{a|bcdef, \ldots, f|abcde,$ $abc|def, \ldots, aef|bcd\}$ the 2-metric remain unchanged. So by fixing the lowest weights in $\{abc|def, \ldots, aef|bcd\}$ to be zero the weight of all splits were fixed (setting one of the weights in $\{ab|cdef, \ldots, ef|abcd\}$ will also gives a unique result, we will explain why we choose this condition later, we also use the assumption that the trivial split were always represented). So we have:

$$w(abc|def) = \max_{\{u,v,w\}=\{a,b,c\},\{x,y,z\}=\{d,e,f\}} \{0, w(uv|xyz) - w(uv|wyz)\} \tag{3.13}$$

From the condition from 6-point case comes the idea of 2-very-weakly compatible condition:

Definition 13. *A 2-very-weakly compatible split system is a split system with forbidden configuration* F_6^3.

Theorem 3. *If* S *is 2-very-weakly compatible*

1. $\{\boldsymbol{q}(s_i)|s_i \in S\}$ *is linearly independent.*
2. $\#S \leq \binom{n}{4} + \binom{n}{2}$

Proof:

1. [51]. If $\sum a_i \boldsymbol{q}(s_i) = 0$, think of a weighted split system with split system S and weights a_i. It's distances and quartet weights vanishes, we need to show that $a_i = 0$ for all i. We will prove that from induction, the $n = 4, 5, 6$ case has been discussed in previous sections. If $n > 6$, we could assumes that this property holds for all $n-1$ case, so we have for all $x \in X$, weights of all $S|_{X-x}$ vanishes. Then if $A|Bx \in S$, we have $Ax|B \in S$ and weight being opposite number of weight of $A|Bx$ by thinking of restriction on set $X - x$. Hence S would contain all possible split if one of the coefficients a_i are nonzero, here comes the contradiction.
2. Since s_i is linearly independent, $\#S \leq dim(\bar{Q}) \leq \binom{n}{4} + \binom{n}{2}$

□

Corollary 1. *A 2-very-weakly split system* S *with weights in* $\mathbb{R}$ *(hence can be negative) having all weights of all* $3|3$*-splits being zero then it's only consists of trivial and* 2*-splits.*

Later we will show that the bound is tight, and 2-very-weakly-compatible split system with maximal cardinality have nice structure. If we choose the forbidden configuration to be $12|3456, \ldots, 56|1234$ that the linearly independence and bound property also hold however using computer search one can easily verify that in this case one can not reach the bound $\binom{n}{4} + \binom{n}{2}$ for $n > 6$.

3.3.2 QuartetDecomposition algorithm

We implemented a network reconstruction method QuartetDecomposition being the quartet weight analog of SplitDecomposition. Like SplitDecomposition, we want $\boldsymbol{q} = w_1\boldsymbol{q}(s_1) + w_2\boldsymbol{q}(s_2) + \cdots + \boldsymbol{q}'$ which $\{s_i\}$ is 2-very-weakly-compatible and $\boldsymbol{q}'$ is split-prime [20, 21].

Definition 14. *A 2-metric is called split prime, if for any split $A|B$ there is $a_1, a_2, a_3 \in A$, $b_1, b_2, b_3 \in B$ s.t. $w(a_1a_2a_3|b_1b_2b_3) = 0$.*

Remind that the split prime is sufficient for defining a unique solution in metric case. However we do not have a unique solution for 2-metric. Let $s_1 = 124|3567$, $s_2 = 134|2567$, $s_3 = 123|4567$ and $s_4 = 234|1567$ and $s_5 = 1234|567$ we have $\boldsymbol{q}(s_1) + \boldsymbol{q}(s_2) + \boldsymbol{q}(s_3) + \boldsymbol{q}(s_4) = \boldsymbol{q}(s_5) + \boldsymbol{q}'$, $\boldsymbol{q}'$ being split prime.

The 2-very weakly compatible split system do not have the finite closure property like compatible split system do, namely there are no bounded number m such that for any $s = A|B \in S$ there always exist A' and B' such that $\#A', \#B' \leq m$ and only $s \vdash A'|B'$ for all $s \in S$. This can be shown by example of 2-circular split system. We have to use insinuate method to decide split weights.

For a input we can use the method described in to decide a list of partial 3|3 splits which have weight zero. Then to get a list of full splits which could be non-zero: all the full splits that do not displays those partial splits, then we exclude trivial or 2-split. We write $\boldsymbol{t} = (w(p_1), w(p_2), \ldots)^T$($p_i$ is all 3|3 partial splits) and $\boldsymbol{w} = (w(s_1), w(s_2), \ldots)^T$($s_i$ is all none zeros splits except for trivial and 2-split). We define matrix A as:

$$A_{ij} = \begin{cases} 0 & p_i \not\dashv s_j \\ 1 & p_i \dashv s_j \end{cases} \tag{3.14}$$

In perfect case we will have $\boldsymbol{t} = A\boldsymbol{w}$. However in most case the system of linear equations are overdetermined and do not have solution. So we will calculate the

weight by minimizing $\|\boldsymbol{t} - A\boldsymbol{w}\|$:

$$\begin{aligned} &\min \|\boldsymbol{t} - A\boldsymbol{w}\| \\ &s.t.\, \boldsymbol{w} \geq 0,\, \boldsymbol{t} \geq Aw \end{aligned} \tag{3.15}$$

This can always be done from the linear independency condition. Weights of 2-split and trivial split can be calculated by:

$$w(ab|X-ab) = \min_{c,d,e \in X} (w(ab|cde) - \sum_{s \vdash ab|cde} w(s)) \tag{3.16}$$

and

$$w(a|X-a) = \min_{b,c \in X} (w(a|bc) - \sum_{s \vdash a|bc} w(s)) \tag{3.17}$$

Theorem 4. *The QuartetDecomposition algorithm is consistent on 2-very-weakly compatible split system. And the residue is split prime.*

Proof:

1. We firstly look at the consistency problem. The process of calculating 3|3-split weight is always correct. And from Corollary 1 we know that $A\boldsymbol{w} = A\boldsymbol{w}'$ iff $\boldsymbol{w} = A\boldsymbol{w}'$. So when $\|\boldsymbol{t} - A\boldsymbol{w}\| = 0$, $\boldsymbol{w}$ is the correct weight. So this is the only case $\min \|\boldsymbol{t} - A\boldsymbol{w}\|$ is reached. The consistency of (3.16) and (3.17) is self evident.

2. Then we consider the split prime part. Consider the decomposition $\boldsymbol{q} = \sum w_i \boldsymbol{q}(s_i) + \boldsymbol{q}'$ calculated from this method. If there exist split s s.t. every $a_1, a_2, a_3 \in A$, $b_1, b_2, b_3 \in B$ have $w_{\boldsymbol{q}'}(a_1a_2a_3|b_1b_2b_3) > 0$. Firstly notice that s must be in the candidate list since $w_{\boldsymbol{q}}(a_1a_2a_3|b_1b_2b_3) \geq w_{\boldsymbol{q}'}(a_1a_2a_3|b_1b_2b_3) > 0$. Then we can add the weight of s a small number to get a smaller value of $\|t - Aw\|$. So w_i cannot be a optimal solution for $\min \|t - Aw\|$.

□

We could compare this method with *Quartetnet*, which can be illustrated by Fig. 3.1. Consider a 2-metric on 6-set, we can decide weights of all the 2|3-splits. Thus the sum of weights of two split connected by one edge is fixed. The Quartetnet algorithm adjust only split weights of a subgraph in the dashed rectangle, QuartetDecomposition considers the whole graph.

We apply this method and Quartetnet to the squamata dataset[52] (Fig. 4.15). Compared with Quartetnet, the QuartetDecomposition is capable of identifying more sophisticated evolution events.

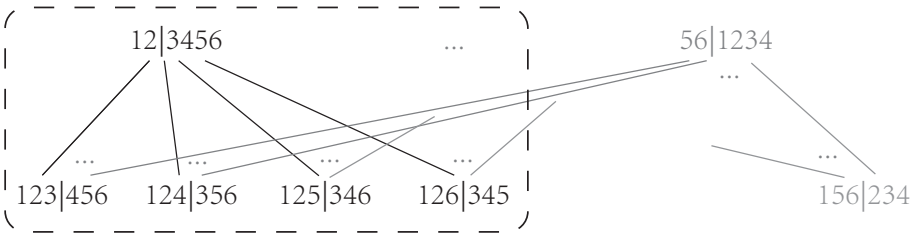
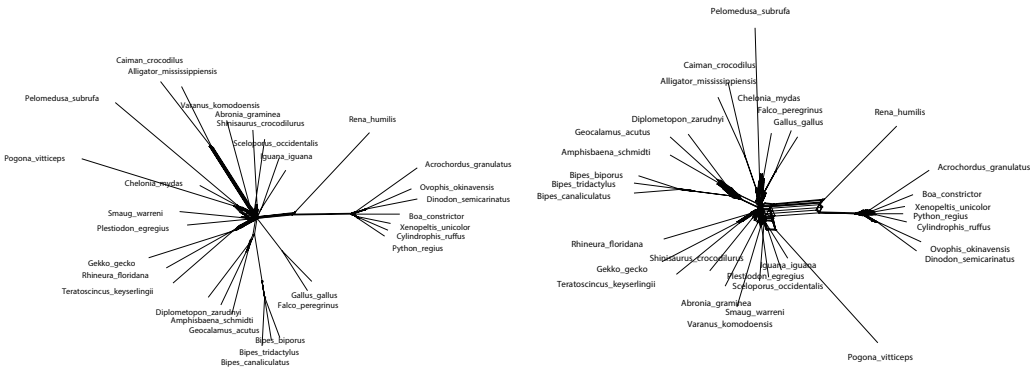


Figure 3.1: The difference between QuartetDecomposition and Quartetnet, QuartetDecomposition considers the whole graph but Quartetnet considers only the subgraph in the dashed rectangle



(a) QuartetNet **(b)** QuartetDecomposition

Figure 3.2: Reconstructed network of squamata dataset using Quartetnet and QuartetDecomposition, quartet weights were generated using Hadamard conjugation.

3.4 T-theory for 2-metric

In this section we focus on the problem that whether T-theory could be modified into 2-metric case to explain the compatible condition, and further leads to a unique decomposition method. One of possible ways for generalize the concept of coherent decomposition has been constructed in [43]. The deficit of the diversity decomposition can be illustrated using this example (Fig. 3.3). Consider a diversity on four point $1, 2, 3, 4$ induced by two splits $12|34$ and $13|24$ with weight 1. We would guess that the tight span should be a square with four vertices $1, 2, 4, 3$ in counter-clockwise order. Consider a point x in the square, let distance of x to line 12 be a and distance to line 24 be b. If x is on tight span, we have $\delta(x, 1, 3) = \min(\delta(x, 1, 2, 3) - \delta(x, 2), \delta(x, 1, 3, 4) - \delta(x, 4), \delta(x, 1, 2, 3, 4) - \delta(x, 2, 4))$ but left hand side is $1 - b$ and right hand side is $\min(2 - a - b, 1 + a - b, 2 - b)$, which always greater than left hand side, thus we can not derive a decomposition theory using diversities.

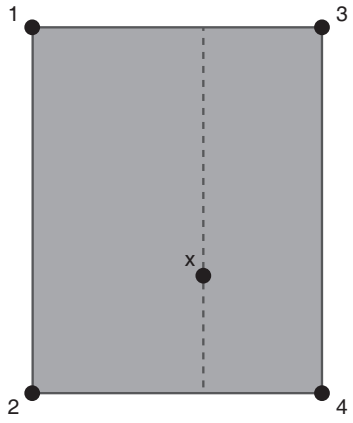


Figure 3.3: An example showing why diversity decomposition is not a proper theory

This example inspire us to use quartet weights instead of diversities to build a decomposition theory. Consider the category $\mathcal{Q}$ with objects being spaces with metric and quartet weights. The morphism are being non-expansive ones: $q(i, j, k, l) \geq q(f(i), f(j), f(k), f(l))$ and $d(i, j) \geq d(f(i), f(j))$. We could guess that in this category there are injective hulls and there are similar construction like tight span: for every spaces (X, Q) we can construct the 2-tight span $T_Q(X, Q)$.

$$T_Q(X,Q) = \left\{ \boldsymbol{v} = \begin{pmatrix} v_{x1|23} \\ \cdots \\ d_1 \\ \cdots \end{pmatrix} \mid x \text{ is a proper single extension and } \boldsymbol{v} \text{ is minimal} \right\}$$

The proper single element extension means on $X \cup \{x\}$ with the quartet weights $q(x,i,j,k) = v_{xi|jk}$ and metric $d(i,x) = d_i$ satisfies definition 11. We will use a simple example illustrating how this tight span method works.

Consider four point $1,2,3,4$ with $d(i,j) = 4$ and $w(i,j,k,l) = 1$ for all $\{i,j,k,l\} = \{1,2,3,4\}$, namely induced by the split system $\mathcal{P}(\{1,2,3,4\})$ with all weight 1.

First note that there is redundancy describing the metric and quartet weights on $\{1,2,3,4,x\}$. Thus we parameterize this by split weights $w(x|1234), \ldots,$ $w(x234|1)$, those should be positive following conditions in definition 11. In all there are 15 variables, given equalities like $1 = w(ij|kl) = w(ijx|lk) + w(ij|xkl)$, there are only 8 independent variables. Firstly we could notice that $w(x|1234) = 0$, because it's an independent variable and decreasing it would decrease the distances and quartet weights or leave them unchanged. Then only 7 variables left, we denote weights like $w(xi|jkl)$ as w_i and $w(xij|kl)$ as w_{ij}. The independent variables are $w_1, w_2, w_3, w_4, w_{12}, w_{13}, w_{14},$ variables w_{23}, w_{24}, w_{34} were still used, note that $w_{12} + w_{34} = w_{13}+w_{24} = w_{14}+w_{23} = 1$. So $q(x,1,2,3) = w_1+w_{14}$, $d(x,1) = 4-w_1+w_2+w_3+w_4-w_{12}-w_{13}-w_{14}$. $d(x,1)$ is the only variable that have negative w_1 coefficient. So if $w_1 > 0$, $w_2 = w_3 = w_4 = 0$, or else we could decrease w_1 and w_i with a same number to get a strictly smaller metric without effecting quartet weights. We could not have any other constraints, consider v with $w_2 = w_3 = w_4 = 0$, if there is $v' < v$, we could assume v' is on tight span hence $w'_2 = w'_3 = w'_4 = 0$, $q'(x,2,3,4) < q(x,2,3,4)$ so $w'_{12} < w_{12}$. Based on symmetry, $w'_{13} < w_{13}$ and $w'_{14} < w_{14}$.

$$d'(x,1) < d(x,1) \tag{3.18}$$

$$\Rightarrow \quad 4 - w'_1 - w'_{12} - w'_{13} - w'_{14} < 4 - w_1 - w_{12} - w_{13} - w_{14} \tag{3.19}$$

$$\Rightarrow \quad w'_1 + w'_{12} + w'_{13} + w'_{14} > w_1 + w_{12} + w_{13} + w_{14} \tag{3.20}$$

However we already have $w'_1+w'_{12} = q'(x,1,2,3) < q(x,1,2,3) = w_1+w_{12}$, this leads to contradiction. This argument also applies for the case of $w_1 = w_2 = w_3 = w_4 = 0$.

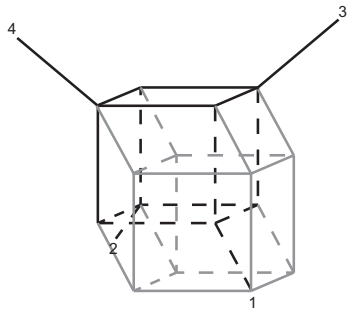


Figure 3.4: A graph demonstrating the tight span, only one of four 4−cube were shown (in grey).

So we can conclude that the tight span is the Cartesian product of a star tree and a cube(Fig. 3.4).

This argument applies to arbitrary split weight case, for this split system with arbitrary split weight, the tight span have the same topology. So for metric and quartet weights induced by a single split, we can deduce it to a 1|3 or 2|2-split case. Hence we've showed that the tight span is a line segment.

Then we consider the case we've interested in, the metric is $d(i,j) = 8$ and $q(i,j,k,l) = 2$ for all $\{i,j,k,l\} \subset \{1,2,3,4,5,6\}$. We have two possible decomposition:

$$\boldsymbol{q} = \boldsymbol{q}(1|23456) + \cdots + \boldsymbol{q}(6|12345) + \boldsymbol{q}(123|456) + \cdots + \boldsymbol{q}(156|234) \quad (3.21)$$

$$= \boldsymbol{q}(12|3456) + \cdots + \boldsymbol{q}(56|1234) \quad (3.22)$$

Then we consider the symmetric single element extension that $d(i,x) = 5$ and $q(x,i,j,k) = 1$. For the second decomposition x lies in the Minkowski sum of tight span. We can take x to be "at the $klmn$ end" of the line segment for every split $ij|klmn$, for which $d(i,x) = 1$, $d(n,x) = 0$, $q(i,x,m,n) = 0$ etc. But for the first decomposition we always have $\sum q(x,i,j,k) \geq 90$ so the first decomposition is not coherent.

However more comprehensive computation showed that the second decomposition are also not coherent. This is done by following procedure: Firstly compute the vertices of Minkowski sum of those line segments, secondly compute the minimal vertices in those, then to check whether it's in $T_Q(X,Q)$.

In the first decomposition (3.21) there are $2^{16} = 65536$ vertices, in which 7168 vertices are minimal and 4698 lies on tight span; in the second one (3.22) there are $2^{15} = 32768$ vertices, in which 7915 vertices are minimal and 7036 lies on tight span. While thinking in term of dimension this is not surprising, the tight span lies in $\mathbb{R}^{26}$, which one may expect that it has dimension 26 (actually it is according to our calculation), but the dimension of those minkowski sum are 16 and 15 respectively, this might suggest that an analog of T-theory, do not exist for 2-metric or higher case.

3.5 2-circular Split System

In this section we construct an example of 2-very-weakly-compatible split system that having cardinality $\binom{n}{4} + \binom{n}{2}$, which plays a similar role of circular split system in metric case. We also proposed an algorithm that reconstruct such split system from 2-metric.

Definition 15. *For a circular ordering $C = (x_1, \ldots, x_n)$ on a set X, we define full 2-circular split system to be*

$$\begin{aligned} S^2(C) \;=\; & \{x_i, \ldots, x_j, x_k, \ldots, x_l | x_1, \ldots, x_{i-1}, x_{j+1}, \ldots, x_{k-1}, x_{l+1}, \ldots, \\ & x_n : 1 \le i < j < k < l \le n\} \cup \{x_i, \ldots, x_j | x_1, \ldots, x_{i-1}, x_{j+1}, \ldots, \\ & x_n : 1 \le i < j \le n\} \end{aligned} \tag{3.23}$$

A split system is 2-circular iff it is subset of $S^2(C)$ for some circular ordering $C = (x_1, \ldots, x_n)$. We say a split $s = A|B$ consists of k intervals in circular ordering C iff A is consists of k intervals. Namely $S^2(C)$ is splits that consists of no more than 2 intervals.

Theorem 5. *A 2-circular split system is 2-very-weakly compatible.*

Proof: For every six elements $x_1, \ldots, x_6$, suppose the ordering of this six elements in C is $(x_1, \ldots, x_6)$. Then we have $x_1x_3x_5|x_2x_4x_6$ not displayed by any split in $S^2(C)$. So it is 2-very-weakly compatible. □

The 2-circular split system is an analog of circular split system in weak-compatible split system, it's an important example for 2-very-weakly compatible with maximal

cardinality. Note that there are 2-very-weakly compatible split system having maximal cardinality but not 2-circular.

In this section we set out to find an analogue for Neighbornet method [29] in 2-metric case. Neighbornet method is a method that constructing a circular ordering from metric. This method is a agglomerative method, which start from identifying some pair to be neighbors and combine this pair to get a cluster. This step goes on and on until we have reduce the problem to trivial case. Like the Neighbornet method, we are also aim to find a circular ordering, we follows basically the same procedure.

The implementation of NeighborNet algorithm were distinct in [29] and [47]. We basically follows the method and notation in [47]. The taxa were denoted $x_1, \ldots, x_n$, For each step we have several clusters: $C_1, \ldots, C_m$; $C_i = [x_{i1}, \ldots, x_{in_i}]$. Each cluster should be think of line graph. Namely you can not permute the elements but you can flip it. The taxa at two end of cluster C_i were denoted as $\hat{C}_i$, namely $\hat{C}_i = \{x_{i1}, x_{in_i}\}$. In each step we firstly select two clusters to merge then we decide which end should we connect if one of the clusters have more than 1 elements. So we need to have two scoring function, first the cluster neighboring score $Q(C_r, C_s)$ for which $\arg\min_{i,j} Q(C_r, C_s)$ being a pair of neighbors, and a neighboring score $\hat{Q}(x_r, x_s)$ for reduction namely if $x_r \in C_r$ and $x_s \in C_s$ are neighbors then $\hat{Q}(x_r, x_s)$ is minimal in $\hat{Q}(x'_r, x'_s)$ for $x'_r \in \hat{C}_r$ and $x'_r \in \hat{C}_r$.

For looking for an analog of NeighborNet method for 2-circular split system, we start from constructing the neighboring score in the initial step, when all clusters are single elements, hence the neighboring score is a function on pairs of taxa and be minimal on a pair being neighbors. Hence We would expect that the neighboring score should be in this form

$$\begin{aligned} s(i,j) = a * d(i,j) + b * (\sum_{x \neq i} d(i,x) + \sum_{x \neq j} d(j,x)) + c * \sum_{x,y \neq i,j} q(i,j,x,y) \\ + d * \sum_{x,y \neq i,j} q(i,x,j,y) + e * (\sum_{x,y,z \neq i} q(i,x,y,z) + \sum_{x,y,z \neq j} q(j,x,y,z)) \\ + i.t. \end{aligned} \tag{3.24}$$

For which a, b, c, d, e are numbers only dependent on n, $i.t.$ stand for independent terms. Which means terms like $\sum d(x,y)$ that do not depend on the choice of i, j. In practice those terms were simply omitted.

Lemma 2. *([53])*

$$\begin{aligned}&\sum_{x,y\neq i,j} q(i,x,j,y) - 2\sum_{x,y\neq i,j} q(i,j,x,y)\\ =&\frac{n-1}{2}((n-2)d(i,j) - \sum_{x\neq i} d(i,x) + \sum_{x\neq j} d(j,x)) + i.t.\end{aligned} \tag{3.25}$$

Proof:

$$\begin{aligned}&\sum_{x,y\neq i,j} q(i,x,j,y) - 2\sum_{x,y\neq i,j} q(i,j,x,y)\\ =&\sum_{x,y\neq i,j} (q(i,x,j,y) - q(i,j,x,y))\\ =&\frac{1}{2}\sum_{x,y\neq i,j} (d(i,j) + d(x,y) - d(i,x) - d(j,y))\\ =&\binom{n-2}{2} d(i,j) + \sum_{x,y\neq i,j} d(x,y) - \frac{n-3}{2}\sum_{x\neq i,j} d(i,x) - \frac{n-3}{2}\sum_{y\neq i,j} d(j,x)\\ =&\binom{n-2}{2} d(i,j) + \sum_{x,y} d(x,y) - \sum_{x\neq i} d(i,x) - \sum_{y\neq j} d(j,x) + d(i,j)\\ &-\frac{n-3}{2}(\sum_{x\neq i} d(i,x) + \sum_{x\neq j} d(j,x)) + (n-3)d(i,j)\\ =&\frac{n-1}{2}((n-2)d(i,j) - \sum_{x\neq i} d(i,x) + \sum_{x\neq j} d(j,x)) + \sum_{x,y} d(x,y)\end{aligned}$$

□

In literatures three methods were described for constructing the neighboring score for neighbornet case. First in [29], the neighboring score defined as unique function independent of weights of trivial splits; second in [47], the neighboring score of i, j are defined as the average of a global score of all circular ordering with i, j being neighbors; third in [31], $s(i,j) = \sum_{x,y\neq i,j} w(ij|xy)$.

The first and last method can be directly generalize into 2-circular case while the second not, we have to find a global score forehand. Luckily[47] gives a way for generalizing the global score to 2-circular case, the score should be sum of all weights in the split system .

Theorem 6. *For circular ordering $C = (c_1, \ldots, c_n)$ and a weighted 2-circular split system $S^2(C)$, we have:*

$$2\sum_{s\in S^2(C)} w(s) = \sum_{0<i\leq n}^{n} d(c_i, c_{i+1}) - \sum_{0<i<j-1<n} w(c_i, c_j, c_{i+1}, c_{j+1})$$

(we always adopt the convention that $c_{n+i} = c_i$)

Proof: Denote the split $x_i, \ldots, x_{j-1}|x_1, \ldots, x_{i-1}, x_j, \ldots, x_n$ as s_{ij} and $x_i, \ldots, x_{j-1}, x_k, \ldots, x_{l-1}|x_1, \ldots, x_{i-1}, x_j, \ldots, x_{k-1}, x_l, \ldots, x_n$ as s_{ijkl}.

Based on linearity of the equation we only needs to proof that this equation holds when the split system were induced by single splits. Remind that q_s means the 2-metric induce by a single split s with weight 1.

For $q = q_{s_{ij}}$, $\sum_{0<i\leq n}^{n} d(c_i, c_{i+1}) = d(x_{i-1}, x_i) + d(x_{j-1}, x_j) = 2$, for other terms are zero in the summation, and all $w(c_u, c_v, c_{u+1}, c_{v+1}) = 0$, hence the equation holds for this case.

For $q = q_{s_{ijkl}}$, $\sum_{0<i\leq n}^{n} d(c_i, c_{i+1}) = d(x_{i-1}, x_i)+d(x_{j-1}, x_j)+d(x_{k-1}, x_k)+d(x_{l-1}, x_l) = 4$, for other terms are zero in the summation, and all $w(c_u, c_v, c_{u+1}, c_{v+1}) = 0$ except $w(c_{i-1}, c_{k-1}, c_i, c_k) = w(c_{j-1}, c_{l-1}, c_j, c_l) = 1$, hence the equation holds for this case. This completes the whole proof.

□

Corollary 2. *For a $C = (c_1, \ldots, c_n)$ and a split $s = A|B$ such that A consists of m intervals in C, with 2-metric q_s*

$$\sum_{0<i\leq n}^{n} d(c_i, c_{i+1}) - \sum_{0<i<j-1<n} w(c_i, c_j, c_{i+1}, c_{j+1}) = -m(m-3)$$

Proof: In the proof of Theorem. 6 we've learned that each interval would contribute 2 to $\sum_{0<i\leq n}^{n} d(c_i, c_{i+1})$ and each pair of interval would contribute 2 to $\sum_{0<i<j-1<n} w(c_i, c_j, c_{i+1}, c_{j+1})$. So we have the right hand side equal to $2m - 2\binom{m}{2} = -m(m-3)$. □

For a given metric and quartet weights we define the score for a circular order as

$$s(C) = \sum_{0<i\leq n}^{n} d(c_i, c_{i+1}) - \sum_{0<i<j-1<n} w(c_i, c_j, c_{i+1}, c_{j+1}) \tag{3.26}$$

Theorem 7. *For a given 2-metric induced by a 2-circular split system with circular ordering C and weights non-zero then C is the circular ordering maximize the score $s(C)$.*

Proof: For arbitrary circular ordering C'. Suppose for every $s_i = A_i|B_i \in S^2(C)$, A_i were break in m_i intervals in C'. We have $s(C') = \sum_i -m_i(m_i - 3)w(s_i) \leq \sum_i 2w(s_i) = s(C)$, so $s(C') = s(C)$ iff $S^2(C)$=$S^2(C')$. By theorem 11. $C' = C$. □

After those preparations we can set out finding the neighboring score for 2-circular split system, this three method will give a same scoring function. Remind that all clusters are single elements, this avoid the weighting problem.

Theorem 8. *We have three way for defining scoring functions:*

1. *If we expand $s(i,j)$ by weights of splits, all trivial and 2-split have same coefficient respectively.*

2. *$s(i,j)$ be the average of $s(C)$ which C being all possible circular orderings with i,j neighbors.*

3. *$s(i,j) = \sum_{x,y,z,w\neq i,j} w(ijx|yzw) - w(ixy|jzw)$*

this three ways give the same neighboring score up to a factor and independent terms.

Proof: We break up the proof into two parts, firstly show that condition 1 is readily a characterizing condition, then neighboring score 2 and 3 satisfies condition 1.

1. We do this by calculate out $s(i,j)$ satisfying this condition. Which is equivalent with that for every 2-metric $\boldsymbol{q}_s$, $s(i,j)$ is same for any $i,j \in X$ when s being 1- or 2-split respectively. Using lemma. 11 we can cancel out the $b*(\sum_{x\neq i} d(i,x)+\sum_{x\neq j} d(j,x))$ in 3.25. Then if $s = i|\{X-i\}$, $s(i,j) = a$ and if $s = x|X-x(x \neq i,j)$, $s(i,j) = 0$, hence $a = 0$. Next we consider the 2-splits. Let $x, y \neq i, j$. For $s = ij|\{X-ij\}$, $s(i,j) = \binom{n-2}{2}c + 2\binom{n-2}{2}e$; for $s = ix|\{X-ix\}$, $s(i,j) = (n-3)d + (\binom{n-2}{2} + (n-3))e$; for $s = ix|\{X-ix\}$, $s(i,j) = c + 2(n-3)e$. $s(i,j)$ should be same for all those cases. Solving this we have:

$$\begin{aligned} c &= 4(n-3) \\ d &= n^2 - 5n + 8 \\ e &= 2(n-1) \end{aligned} \tag{3.27}$$

2. All we needs to do is to verify for s being 1- or 2-split $s(i,j)$ is same respectively. For 2. $s(C)$ is always 2 so $s(i,j)$ is always 2 regardless of s, for 3 $s(i,j)$ always be 0.

□

Finally with all method we get:

$$\begin{aligned} s(i,j) &= 4(n-3)\sum_{x,y\neq i,j} q(i,j,x,y) + (n^2-5n+8)\sum_{x,y\neq i,j} q(i,x,j,y) \\ &\quad +2(n-1)(\sum_{x,y,z\neq i} q(i,x,y,z) + \sum_{x,y,z\neq j} q(j,x,y,z)). \end{aligned} \tag{3.28}$$

It's necessary to point out that unlike the circular case the pair maximizing this scoring function may not be neighbors: think of a circular ordering $(\dots, x', x, \dots, y', y, \dots)$, this induces a 2-circular split system. The weight of split $xyx'|\dots$, $xyy'|\dots$ are 1, and the weight of the rest are infinitesimal positive numbers. Using the third formalism of scoring function we can easily verify that x and y have maximal neighboring score. However the split system uniquely define a circular ordering which x and y are not neighbors. We would expect that this scoring function will gives a working algorithm which may not be consistent in those pathological cases, hence output reasonable results with real data.

Hereby we explicitly write out the neighboring function for clusters mimicking the neighbornet one, we use the averaged quartet weights.

$$\begin{aligned} s(C_r,C_s) =& 4(m-3)\sum_{u,v\neq r,s} q(C_r,C_s,C_u,C_v) \\ &+ (m^2-5m+8)\sum_{u,v\neq r,s} q(C_r,C_u,C_s,C_v) \\ &- 2(m-1)\sum_{u,v,w\neq r} q(C_r,C_u,C_w,C_v) \\ &- 2(m-1)\sum_{u,v,w\neq s} q(C_s,C_u,C_w,C_v) \end{aligned} \tag{3.29}$$

In which:

$$q(C_r,C_s,C_u,C_v) = \frac{1}{\#C_r\cdot\#C_s\cdot\#C_u\cdot\#C_v}\sum_{x_r\in C_r, x_s\in C_s, x_u\in C_u, x_v\in C_v} q(x_r,x_s,x_u,x_v) \tag{3.30}$$

The next step is finding a the formula for reduction, namely deciding which end to join if two clusters are decided to be neighbors. We followed the implementation of [47], two end were joined if the resulting clusters could maximize the average of possible circular ordering. This formula should also have the invariant property. Hence we can write out the formula:

$$
\begin{aligned}
s'(x_r, x_s) &= 4(\bar{m}-3)\sum_{u,v\neq r,s} q(x_r, x_s, C_u, C_v) + \\
&\quad (\bar{m}^2 - 5\bar{m} + 8)\sum_{u,v\neq r,s} q(x_r, C_u, x_s, C_v) - \\
&\quad 2(\bar{m}-1)\sum_{u,v,w\neq r} q(x_r, C_u, C_w, C_v) - \\
&\quad 2(\bar{m}-1)\sum_{u,v,w\neq s} q(x_s, C_u, C_w, C_v) \qquad (3.31)
\end{aligned}
$$

In which $\bar{m} = m + \#\hat{C}_r + \#\hat{C}_s - 2$, $C^* = C \setminus \{C_r, C_s\} \cup \hat{C}_r \cup \hat{C}_s$.

So we can write out whole the algorithm, there is one subtlety here: this algorithm is not working for $n = 6$ case. Because circular ordering $[1, 2, 3, 4, 5, 6]$ and $[1, 5, 3, 4, 2, 6]$ gives out a same 2-circular split system. So the recursion should stop at $n = 6$ case.

Data: a 2-metric $\boldsymbol{q}$ on $X = \{1, \ldots, n\}$
Result: circular ordering $C = (x_1, \ldots, x_n)$
$C = \{[1], \ldots, [n]\}$ is a collection of linear orderings;
while $\#C \geq 6$ **do**
  for $i, j \in \binom{C}{2}$ **do**
    calculate $s_q(C_i, C_j)$ using formula 3.30;
  end
  Choose a pair C_r, C_s that minimize $s_{\boldsymbol{q}}(C_i, C_j)$.;
  for $i \in \hat{C}_r$, $j \in \hat{C}_s$ **do**
    calculate $s_{\boldsymbol{q}}(i, j)$ using formula 3.32;
  end
  Choose a pair i, j that minimize $s'_{\boldsymbol{q}}(i, j)$.;
  Let C_n be the new cluster that join the C_r, C_s at the i, j.
  $C = C \backslash \{C_r, C_s\} \cap \{C_n\}$;
  Update the 2-metric $\boldsymbol{q}$ using formula 3.31;
end
Calculate the score$s(C)$ of all possible circular order being the combination of 6 clusters in C, C is set to be the one that maximize the score;

Algorithm 3: The 2-neighbornet algorithm

Remark 4. *The final step need to calculate the score for at most* $5! * 2^5 = 3840$ *times, which is not very time-consuming.*

One could notice that the calculation of (3.30) can be further simplified using lemma 2. Only one of calculation of $\sum_{u,v\neq r,s} q(C_r, C_s, C_u, C_v)$ and $\sum_{u,v\neq r,s} q(C_r, C_u, C_s, C_v)$ is necessary, the previous one needs only half calculation compared with latter one. We have:

$$\begin{aligned} s(C_r, C_s) &= (m^2 - 5m + 2)((m-2)d(C_r, C_s) + \sum_{u\neq r} d(C_r, C_u) + \sum_{u\neq s} d(C_s, C_u)) \\ &\quad +4(m-2) \sum_{u,v\neq r,s} q(C_r, C_s, C_u, C_v) \\ &\quad -2 \sum_{u,v,w\neq r} q(C_r, C_u, C_w, C_v) - 2 \sum_{u,v,w\neq s} q(C_s, C_u, C_w, C_v) \end{aligned} \tag{3.32}$$

the simplified reduction formula is similar.

$$\begin{aligned} s'(x_r, x_s) &= (\bar{m}^2 - 5\bar{m} + 2)((\bar{m}-2)d(x_r, x_s) + \sum_{u\neq r} d(x_r, C_u) + \sum_{u\neq s} d(x_s, C_u)) \\ &\quad +4(\bar{m}-2) \sum_{u,v\neq r,s} q(x_r, x_s, C_u, C_v) \\ &\quad -4 \sum_{u,v,w\neq r} q(x_r, C_u, C_w, C_v) - 4 \sum_{u,v,w\neq s} q(x_s, C_u, C_w, C_v) \end{aligned} \tag{3.33}$$

At first glance this may seems to contain more terms but actually more time-saving compared with original method.

Theorem 9. *The 2-neighbornet algorithm is consistent on such split system:*

$S = \{xy|X-xy\} \cup \{x_i, \ldots, x_j | x_1, \ldots, x_{i-1}, x_{j+1}, \ldots, x_n\}$ *with all weights positive.*

Proof: From [54] we know that for proving the consistency of this algorithm, all we need is the following inequalities for $i \leq \frac{n}{2}$:

$$\min(s(x_1, x_2), \ldots, s(x_{i-1}, x_i)) < s(x_1, x_i)$$

, $s(x_2, x_3) < s(x_1, x_3)$ and $s(x_2, x_3) < s(x_1, x_4)$. We start by doing some calculation of $s(x_r, x_s)$ when the split system consists of a single split $A|B$ with weight 1.

1. $x_r \in A$, $x_s \in B$. Denote $\#A = x$ and $\#B = y$, thus we have:

$$\begin{aligned}
\sum_{u,v\neq r,s} q(x_r, x_s, x_u, x_v) &= 0 \\
\sum_{u,v\neq r,s} q(x_r, x_u, x_s, x_v) &= (x-1)(y-1) \\
\sum_{u,v,w\neq r} q(x_r, x_u, x_w, x_v) &= \frac{1}{2}(x-1)(y-1)(y-2) \\
\sum_{u,v,w\neq s} q(x_s, x_u, x_w, x_v) &= \frac{1}{2}(y-1)(x-1)(x-2)
\end{aligned}$$

Thus

$$\begin{aligned}
s(x_r, x_s) &= (n^2 - 5n + 8) * (x-1)(y-1) - (n-1)((x-1)(y-1)(y-2) \\
&\quad +(y-1)(x-1)(x-2)) \\
&= 4(x-1)(y-1)
\end{aligned}$$

2. $x_r, x_s \in A$. Denote $\#A = x$ and $\#B = y$, thus we have:

$$\begin{aligned}
\sum_{u,v\neq r,s} q(x_r, x_s, x_u, x_v) &= \frac{1}{2}y(y-1) \\
\sum_{u,v\neq r,s} q(x_r, x_u, x_s, x_v) &= 0 \\
\sum_{u,v,w\neq r} q(x_r, x_u, x_w, x_v) &= \frac{1}{2}(x-1)(y-1)(y-2) \\
\sum_{u,v,w\neq s} q(x_s, x_u, x_w, x_v) &= \frac{1}{2}(x-1)(y-1)(y-2)
\end{aligned}$$

Thus

$$\begin{aligned}
s(x_r, x_s) &= 2(n-3)y(y-1) - 2(n-1)(x-1)(y-1)(y-2) \\
&= -2(y-1)(x^2y - 2x^2 + xy^2 - 5xy + 4x - 2y^2 + 6y - 2)
\end{aligned}$$

Note that if we can subtract same number of $s(x_i, x_j)$ without infecting the final results. In summary we have $s(x_i, x_j) = 0$ if s separates x_i and x_j, $s(x_i, x_j) = -2(y-1)(x^2y - 2x^2 + xy^2 - 5xy + 4x - 2y^2 + 6y - 2) - 4(x-1)(y-1) = -2(y-1)(y-2)(x-2)(n-1)$ if $s = A|B$, $x_r, x_s \in A$, $\#A = x$ and $\#B = y$.

Then we proceed to prove the inequalities. Again we assume the split system consists of a single split $x_p, \ldots, x_{q-1}|x_1, \ldots, x_{p-1}, x_q, \ldots, x_n$ with weight 1. For the

first one, $\min(s(x_1,x_2),\ldots,s(x_{i-1},x_i)) < s(x_1,x_i)$. Denote $M = (i-1)s(x_1,x_i) - (s(x_1,x_2)+\cdots+s(x_{i-1},x_i))$. There are essentially 3 different cases of splits, denote $p-q$ as l we have $2 < l < n/2$.

1. $i < p < q$ In this case $s(x_1,x_2) = \cdots = s(x_{i-1},x_i) = -2(n-l-1)(n-l-2)(l-2)(n-1) = s(x_1,x_i)$ hence $M = 0$

2. $p < i < q$ we have $s(x_1,x_i) = 0$ and $s(x_{j-1},x_j) = -2(n-l-1)(n-l-2)(l-2)(n-1) < 0$ if $j < p$, $s(x_{j-1},x_j) = -2(l-1)(n-l-2)(l-2)(n-1) < 0$ if $j > p$ and $s(x_{j-1},x_j) = 0$ if $j = p$ hence $M > 0$.

3. $p < q < i$ $s(x_{j-1},x_j) = s(x_1,x_i)$ unless $p \le j \le q$. Thus $M = (p-q+1)(-2(l-1)(n-l-2)(l-2)(n-1)) - (l-1)(-2(n-l-1)(n-l-2)(l-2)(n-1)) = 2(n-l-2)(l-2)(n-1)(l-1)(n-2l) > 0$

Then we consider the inequalities $s(x_2,x_3) < s(x_1,x_3)$ and $s(x_2,x_3) < s(x_1,x_4)$. If $4 < p < q$ $s(x_2,x_3) = s(x_1,x_3) = s(x_1,x_4)$ and $q > 4$ because trivial and 2-split were not considered so we are only interested in the case when $p = 2,3,4$ and $q > 4$. When $p = 2$, $s(x_2,x_3) < 0$ and $s(x_1,x_3) = s(x_1,x_4) = 0$; when $p = 3$, $s(x_2,x_3) = s(x_1,x_3) = s(x_1,x_4)$; when $p = 4$, $s(x_2,x_3) = s(x_1,x_3) < 0 = s(x_1,x_4)$. This completes the whole proof.

□

We apply 2-Neighbornet method to the squamata dataset, the results were shown in Fig. 3.5. We conclude that there is no significant difference on performance of those methods, it is worthwhile mentioning that 2-metric based method find the anomaly in order of taxa *Pogona vitticeps* (with number 9) while others do not, which has been also observed in [52].

3.5.1 A Consistent Method

Inspired by [32] we could propose a method that being consistent, though could be less effective and elegant. First observe that if x, y are neighbors then for every a_1, a_2, a_3, a_4, one of weight of $xa_ia_j|ya_ka_l$, $\{i,j,k,l\} = \{1,2,3,4\}$ must be vanish. So

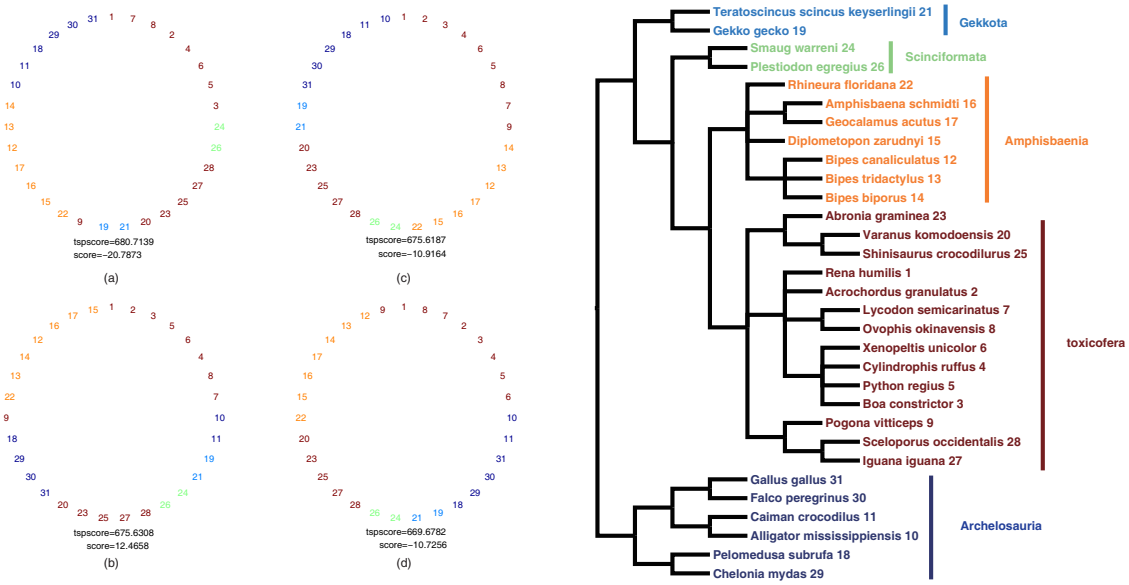


Figure 3.5: Reconstructed circular ordering of squamata dataset, Fig. (a) using method described in this section, Fig. (b) is the optimal circular ordering of score $s(C)$, Fig. (c) using Neighbornet method and Fig. (d) is the optimal circular ordering of traveling salesman problem. The optimization is done by simulated annealing.

we define

$$\sigma_1(x,y) = \sum_{a_1,a_2,a_3,a_4 \in X-\{x,y\}} \min_{\{i,j,k,l\}=\{1,2,3,4\}} w(xa_ia_j|ya_ka_l) \tag{3.34}$$

$$\sigma_2(x,y) = \sum_{a_1,a_2,a_3,a_4 \in X-\{x,y\}} w(xya_i|a_ja_ka_l) \tag{3.35}$$

The weight of $3|3$-split were calculated by formula (3.13). The x, y minimizing $\sigma_1(x,y)$ and further maximizing $\sigma_2(x,y)$ must be neighbors. This claim give raise to a consistent algorithm: just change the selection criteria in algorithm 3 with σ_1 and σ_2.

We apply this method to the Squamata dataset and reconstruct such circular ordering. We would conclude that such method do not construct more accurate network but would take more time compared with 2-neighbornet method in our case.

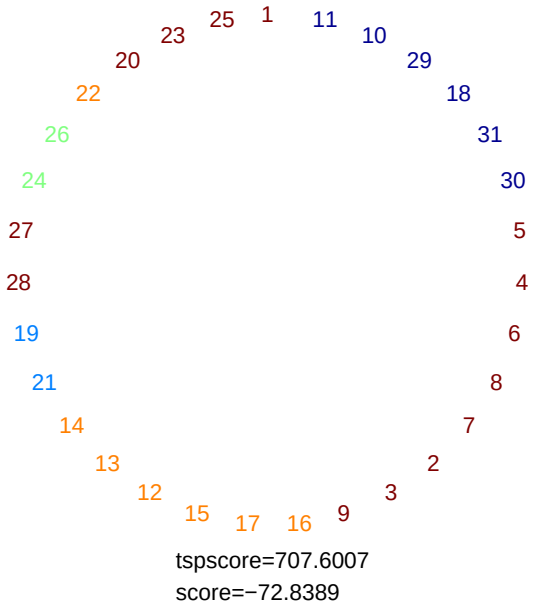


Figure 3.6: Reconstructed circular ordering of squamata dataset using the method in section 3.5.1.

3.6 Connections with Oriented Matroids

The connections of oriented matroid [55] and split system has been first noticed in [56], latter in [57] it's shown that topes of rank 3 acyclic oriented matroid can be drawn in a planar way, in [32] a practical algorithm has been proposed to for reconstructing a special class of flat split system: the neighborly flat split system. In our research we found further relations of oriented matroid and 2-very-weakly-compatible split system. 2-circular split system is an 2-very-weakly-compatible split system with maximal cardinality, however the inverse is not true: not all 2-very-weakly-compatible split system with maximal cardinality is 2-circular, in this section we identify those objects with topes of neighborly 4-polytope, which has been studied in context of oriented matroid theory. We further consider its relationship with flat split system.

We firstly gives a brief review on notions of oriented matroid theory of point configurations, for more detailed exposition one could refer to related chapters of [55]. In this case oriented matroid encodes the convexity information of point configurations. For example one could consider 4 points on a plane at general position, there are 2 non-isomorphic configurations, the first case is the 4 points are vertices

of a convex quadrilateral, the second case is 3 points spans a rectangle and 1 lies in interior. Those information can be properly encoded by oriented matroids.

An oriented matroid can be described using circuits, vectors, cocircuits, covectors, topes or chirotopes. A signed vector X is an mapping $X : E \to \{-1, 0, 1\}$, $X(v_i)$ were denoted X_i, $X^{-1}(1)$ and $X^{-1}(-1)$ were denoted as X^+ and X^- respectively, size of X were denoted as $\#X$ and defined as $\#X^+ + \#X^-$. There is an natural partial ordering: $X \prec X'$ iff $X^+ \subset X'^+$ and $X^- \subset X'^-$. Circuits, vectors, cocircuits, covectors and topes are all signed vectors. Consider a finite set of points $E = \{\mathbf{v_1}, \ldots, \mathbf{v_n}\}$ in affine plane $\mathbb{A}^{r-1}$. Each affine dependencies $\sum \lambda_i \mathbf{v_i} = \mathbf{0}$, $\sum \lambda_i = 0$ defines a vector X such that $X^+ = \{\mathbf{v_i} | \lambda_\mathbf{i} > \mathbf{0}\}$,$X^- = \{\mathbf{v_i} | \lambda_\mathbf{i} < \mathbf{0}\}$. Geometrically this implies the convex hull of X^+ and X^- are intersecting at interior points. And for each $\mathbf{w} \in \mathbb{A}^{*\mathbf{r-1}}$ and scalar ρ defines a covector X such that $X^+ = \{\mathbf{v_i} | \mathbf{v_i} * \mathbf{w} > \rho\}$,$X^- = \{\mathbf{v_i} * \mathbf{w} < \rho\}$. Which means there exist a hyperplane that X^+ and X^- lies at two side of hyperplane respectively. The circuits are minimal vectors and cocircuits are minimal covectors, topes are maximal covectors. Chirotopes are alternating functions on E^r, $\chi : E^r \to \{-1, 0, 1\}$, which were defined as $\chi(\mathbf{v_{i_1}}, \ldots, \mathbf{v_{i_r}}) = \text{sign}(\det(\mathbf{v_{i_1}} - \mathbf{v_{i_r}}, \ldots, \mathbf{v_{i_{r-1}}} - \mathbf{v_{i_r}}))$. We mainly consider points that are in general position, namely the uniform oriented matroid. In which the size of circuits are always $r + 1$ and size of cocircuits are always $\#E - r$. The value of chirotopes are always non-zero. We build the correspondence between acyclic oriented matroid and split system that split system of an oriented matroid is the set of split of topes: $X^+|X^-$. Such split system are naturally have forbidden configuration F_{r+1}, this is also called VC dimension property[58].

Definition 16. *An n-pseudoaffine split system is defined as topes of an rank $n + 1$ acyclic uniform oriented matroid. And a split system is affine if it's pseudoaffine and the oriented matroid is realizable.*

The neighbornet algorithm aims at constructing the circular split system, which is the weakly compatible split system with maximum cardinality. This naturally leads to a problem: How to characterize the 2-very-weakly compatible split system with maximal cardinality?

If we think the split system as maximal covectors (or say, topes) of oriented matroid, the rich theory of oriented matroids can be applied for further understanding the nature of maximal split system. The 2-very-weakly compatible split system

with maximal cardinality corresponds to neighborly oriented matroid of rank 5. The applications of rank 3 case in phylogenetic has been firstly introduced in [56] and further studied in [32].

Theorem 10. *A 2-very-weakly compatible split system having maximal cardinality is the topes of some neighborly oriented matroid [59, 60] of rank 5.*

Proof: The main ingredient in the proof is the theorem in [58]. Rewriting in the language of split system: a split system with forbidden configuration F_d on an n-set and with maximal cardinality (which is $\Phi_{d-2}(n-1) = \sum_{i=1}^{d-2}\binom{n-1}{i}$, see Section 5.1) is topes of a acyclic oriented matroid of rank $d-1$. In our example we have the forbidden configuration being a subset of F_6 and cardinality of the split system being $\binom{n}{4}+\binom{n}{2} = \sum_{i=1}^{4}\binom{n-1}{i}$ hence all the condition is satisfied. And all the 1-split and 2-split should be in the split system or we can add them into the split system without breaking the compatible condition and contradicts with the maximal property. Hence for every pair x, y splits $xy|X-xy$, $x|X-x$ and $y|X-y$ are elements of the split system so xy is a face of the oriented matroid. This completes the proof of neighborly part. □

The 2-circular split system indeed corresponds to alternating oriented matroid, or say, cyclic polytope, using the theory of oriented matroid we can give a constructive proof that the 2-circular split system corresponds to a unique circular ordering:

Theorem 11. *If $S^2(C_1) = S^2(C_2)$ then $C_1 = C_2$ for $|C_1| > 6$*

Proof: In term of polytope, $x, y \in C$ are neighbors in a circular ordering iff xy is a universal edge. In this way one can uniquely determine the circular ordering from the 2-circular split system. □

We also know that there are neighborly oriented matroid being non-cyclic. Thus we would guess that using topes of neighborly oriented matroid instead of 2-circular split system in the 2-neighbornet might be helpful in developing better algorithm.

3.6.1 Flat split system

In [57] flat split system were introduced, it has the property that the network can be drawn in plane without intersecting edges. Later [61] flat split system were characterized to be subset of topes of a rank 3 acyclic oriented matroid. An agglomerative

method has be raised for reconstructing a special class of flat split system: the neighborly flat split system, which is the topes of oriented matroids the can be constructed by stepwise lexicographic extension [55]. In this section we will elaborate some relations between 2-very-weakly compatible and flat split system, some of the mathematical problem is discussed.

Theorem 12. *A planar split system is 2-very-weakly compatible.*

Proof: Note that restriction of F_6^3 of set $\{1,2,3,4\}$ is F_4. A planar split system have forbidden configuration F_4, so it must be 2-very-weakly compatible. □

Remark 5. *The QuartetDecomposition method is automatically a consistent method for reconstructing flat split system, even might not very useful in practice.*

We propose the following approach for reconstructing planar split system, we first construct a 2-very-weakly compatible and further manage to filter the split system to be flat. Suppose we are able to construct 2-very-weakly compatible with maximal cardinality, namely topes of rank 5 neighborly oriented matroid. This naturally leads to a problem: is every planar split system a subset of topes of a neighborly rank 5 oriented matroid? Or, consider a stronger problem, is every planar point set be projections of vertices of a neighborly 4-polytope? (we ignore the realization problem, another problem we ignore is, being subset do not equivalent with being projection, which is the strong map conjecture, has been falsified in [62]). If yes then any planar split system is reconstructible by such method. We tend to believe that the answer is negative. Since it has already been pointed out that the dimension of ambient space of 2-neighborly 2-manifold is at least 6 [63].

We have only concrete results related with cyclic polytope.

Lemma 3. *A planar split system is 2-circular split system iff there is a circular ordering s.t. if we draw line interval between neighbors in circular ordering then every line in the plane would intersect with them at no more than* 4 *times.*

Theorem 13. *Not every planar split system is 2-circular split system.*

Proof: Consider the grid point in a coordinated plane, namely points with coordinates $(i,j), i,j \in 1,2,\ldots,n$. Consider the lines $x = 1.5,\ldots,x = n-0.5$, $y = 0.5,\ldots,y = n-1.5$, every edge must intersect with one line at least once so we

have $n^2 < 4 * 2(n-1)$ if such cyclic ordering exist. This cannot be true for n big enough. □

Remark 6. *This conclusion also holds if we use grid with arbitrary coordinates, namely* $(i,j), i,j \in \{t_1, t_2, \ldots, t_n\}$. *So we can get a neighborly point configurations by setting coordinates* $t_i = 1 + \epsilon + \cdots + \epsilon^{t-1}$, $0 < \epsilon \ll 1/n$.

3.7 Final Remark

The motivation of this section is to detect more detailed phylogenetic structure by introducing more global data. Theory of 2-metric were studied, which is turned out to be much more complicated than metric case. We deduce the 2-very-weakly compatible condition being the most general compatible condition that identifiable from 2-metric. In general direct generalization of metric method would not work for some of the nice properties of weakly compatible split system do not generalize to 2-very-weakly compatible split system. We have proposed QuartetDecomposition and 2-NeighborNet algorithm, they are quartet weight version of SplitDecomposition and NeighborNet algorithm, those method have been proved to be useful in practice. Applying our method to the squamata dataset, the reconstructed phylogenetic network displays more detailed phylogenetic events compared with conventional quartet-weight based method(Fig. 4.15). We have also showed that 2-very-weakly compatible condition has rich mathematical structure, especially related with oriented matroid theory.

One of the deficit of quartet weight approach is, one have as much as $O(n^4)$ inputs for n taxa, in practice our algorithm can handle data set with around 30 taxa. There are several methods that might could improve this problem.

1. We could start from distance data and construct a tree or network as a guidance, then we could check for every taxa quartet whether subnetwork could describe the data well, if yes we could ignore the weights from those quartet. This leads to the problem to reconstruct networks from incomplete quartet weight data.

2. Decompose the taxa set into several group such that taxa of each group should be similar enough, then we could divide the reconstruction problem into two

smaller ones[64, 65]. An interesting task is, how to identity the splits that lies "between" these two levels. For example we have identified 2 clusters A, B, can we reconstruct the split that incompatible with $A|B$? We would expect there are not too many such splits, which might accelerate calculation.

3. There might be tricks to speed up certain algorithm. For example we need to do linear optimization in QuartetDecomposition. If we could foretell which split weights lower than some threshold or which constrains is not effective, those variables and constrains can be ignored and linear optimization process could accelerate significantly.

Chapter 4

Calculating quartet weights using Hadamard conjugation

In the previous chapters we have studied a theory of 2-metric. The method for calculating distances from biological data has been extensively studied. However, the biological meaning of quartet weight are not explicitly defined in most cases and the calculation of quartet weight has been a long-existing problem[33, 66]. Most existing methods are not based on proper models thus can not generate quartet weights with high accuracy. In this chapter, we introduced a method that calculate 2-metric using Hadamard conjugation, which could improve performance by taking multi-mutation in one site into account. This assertion is verified by simulation and real data study.

4.1 Summary of existing quartet weight calculation and file format

Before proceeding to introducing the Hadamard conjugation method we'll stop for a while to make a simple review on the existing quartet weight calculating method. The most prevalent method is direct or variant of pattern counting. Consider an aligned DNA sequences of four taxa, columns were called sites. Regardless of gaps in alignment characters of each site would correspond to an element in $\{A, T, G, C\}^4$. Hence all the information in the alignment can be encoded by a 256-dimensional vector consists of frequency of site patterns. The sites with gap is omitted in most method, and one could use methods like expectation maximization to remedy this

deficit [67]. The pattern counting method were based on a model with a oversimplified assumption: mutation events were rare that occurs on each site at most one time in whole history. The most naive pattern counting method just look at the proportion of site supporting a quartet. For an explicit example:

$$w(12|34) = p_{AATT} + \cdots + p_{CCGG}$$

The sum were taken with the frequencies of string $s_1s_2s_3s_4$ that $s_1 = s_2 \neq s_3 = s_4$. A more refined version of this is

$$w(12|34) = \frac{p_{AATT} + \cdots + p_{CCGG}}{p_{AAAA} + p_{AAAT} + \cdots + P_{CCCC}}$$

The numerator is all the strings that consists of no more than 2 characters. This is the method used by QuartetNet.

Statistical geometry can be think of a variant of pattern counting which take the site that have more than 2 characters into account. For example the pattern $AAGT$ would support split 12|34 but not as strong as $AAGG$ do. This site would contribute to the split only a half of sites like $AAGG$. More exactly, if a site like $AACC$ would contribute 1 for split 12|34, then a site like $AACG$ would contribute 0.5 for split 12|34, 3|124 and 4|123, moreover a site like $ATCG$ would contribute 0.5 for split 1|234, 2|134, 3|124 and 4|123. In this method the formula determine quartet weight is:

$$w(12|34) = (p_{AATT} + \cdots + p_{CCGG}) + \frac{1}{2}(p_{AATG} + \cdots + p_{CCTG} + p_{AGTT} + \cdots + p_{CTGG})$$

This is the method used by QNet and FlatNJ method, which satisfies definition 11. It's worthwhile mentioning that in the QNet implementation the quartet weight were normalized that $w(12|34) + w(13|24) + w(14|23) = 1$.

Another class of method is under from Maximal Likelihood framework, which is also implemented in QNet. For every quartet, there are three possible tree topologies. We compute the maximized likelihood with respect to those trees. Suppose the likelihood is l_1, l_2, l_3 and length of internal edge is e_1, e_2, e_3 thus the quartet weight is decided using following formula:

$$w_i = e_i * \frac{l_i}{l_1 + l_2 + l_3} \tag{4.1}$$

Another problem is which file format can be utilized to store the quartet weight data. The nexus file format supports quartet weights. Here is an example for quartet block:

```
BEGIN Quartets;
DIMENSIONS [NTAX=number-of-taxa] NQUARTETS=number-of-quartets;
[FORMAT
[LABELS={LEFT|NO}]
[WEIGHTS={YES|NO}]
;]
MATRIX
[label1] [weight1] a1 b1 : c1 d1,
...
[labeln] [weightn] an bn : cn dn,
;
END;
```

Another format is *Qweight*, which list quartet weight of same quartet in one line for shortening file. A sample file reads:

```
taxanumber: 5;
description: artificial data;
sense: max;
taxon: 001 name: a;
taxon: 002 name: b;
taxon: 003 name: c;
taxon: 004 name: d;
taxon: 005 name: e;
quartet: 001 002 003 004 weights: 200 0 200;
quartet: 001 002 003 005 weights: 200 0 200;
quartet: 001 002 004 005 weights: 210 0 210;
quartet: 001 003 004 005 weights: 10 0 10;
quartet: 002 003 004 005 weights: 10 0 10;
```

This file format do not contain the distance information, hence QuartetNet do not construct trivial splits if using Qweight file as input. The file format used by FlatNJ[32] extended the nexus format by introducing *QUADRUPLES* block in-

cludes the weight of 1|3−splits. The file format were shown below:

```
BEGIN QUADRUPLES
DIMENSIONS NTAX=number-of-taxa NQUADRUPLES=number-of-quadruples;
[FORMAT [LABELS={LEFT|NO}] [WEIGHTS={YES|NO}];]
MATRIX
[label_1] : a1 b1 c1 d1 : [weight_a1|b1c1d1 weight_b1|a1c1d1
 weight_c1|a1b1d1 weight_d1|a1b1c1 weight_a1b1|c1d1 weight_a1c1|b1d1
 weight_a1d1|b1c1],
[label_2] : a2 b2 c2 d2 : [weight_a2|b2c2d2 weight_b2|a2c2d2
weight_c2|a2b2d2 weight_d2|a2b2c2 weight_a2b2|c2d2 weight_a2c2|b2d2
weight_a2d2|b2c2],
...
[label_n] : an bn cn dn : [weight_an|bncndn weight_bn|ancndn
weight_cn|anbndn weight_dn|anbncn weight_anbn|cndn weight_ancn|bndn
weight_andn|bncn],
;
END;
```

4.2 Markov model on phylogenetic tree

In this section we give a exposition on the canonical model describing the sequence evolution with time and speciation events, the Markov model (see [68] for a detailed discussion). From such model we can apply ML-based approach to estimate parameters.

We first consider the process of evolution of a single taxon without speciation. And the inhomogeneous inside the taxon group will always be ignored, another assumption we made is, each site is sampling from a independent identical distribution (*i.i.d*). Suppose at $t = 0$, the base frequency is $\rho = (\rho_A, \rho_T, \rho_G, \rho_C)$. And at each short time interval the expectation of total amount of transition events from character c_1 to c_2 is in direct proportion with the frequency of c_1 and unrelated with other parameters, the ratio is denoted as $Q_{c_1c_2}$ and collect as matrix Q. The dynamic of sequence evolution is controlled by a first-order differential equation.

$$\frac{d\rho}{dt} = Q\rho$$

with solution

$$\rho(t) = exp(tQ)\rho(0)$$

Denote $exp(tQ)$ as T, the matrix T is called transition matrix, with property $T_{c_1c_2} = P(E_t = c_1|E_0 = c_2)$($E_t$ means the character in time t). In practice we would assume some constraints on the substitution matrix. The general Markov model have too many parameters and lead to over-fitting problem. There are models with various number of parameters, for sequences of different length or with different properties one need to choose a proper model. The most widely used is GTR(general time reversible) model, a base distribution $\rho = (\rho_A, \rho_T, \rho_G, \rho_C)$ is called stable iff $Q\rho = 0$, a GTR model is a markov model with property $\rho_{c_1}Q_{c_1c_2} = \rho_{c_2}Q_{c_2c_1}$ for ρ being stable distribution. Equivalently, the stable distribution is always detailed balance.

The data of a Markov model is a triplet $(\mathcal{T}, \rho_r, T^i)$, $\mathcal{T}$ being a rooted X-tree, $\rho_r = (\rho_A, \rho_T, \rho_G, \rho_C)$ being base frequencies of root vertex and T^i transition matrices associated with each edge. Each vertex was associated with a random variable taking values in the characters, for example $\{A, T, G, C\}$ in DNA case. Each random value of vertex was only depend on parent vertex. Transition matrix T^i is a 4×4 matrix, for example, an edge e connects x and y and x is parental vertex of y, then T^e is defined as matrix of conditional probability $T^e_{ij} = P(y = s_i|x = s_j)$. In practice we can only observe the states of all leaf nodes, it's important that from all such data we can write out probabilities of joint distribution of leaf vertices. For example, in the tree shown in 4.1, the joint distribution can be written as follows, the sum were taken with the states of internal nodes.

$$\begin{aligned} p_{ATC} =& \rho_A T^1_{AA} T^2_{AA} T^3_{AT} T^4_{AC} + \\ & \cdots \\ & + \rho_C T^1_{CA} T^2_{CC} T^3_{CT} T^4_{CC} \end{aligned}$$

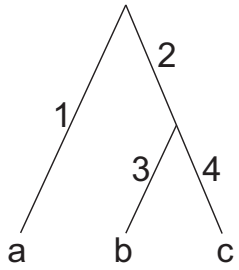


Figure 4.1: Example of Markov model on tree

An further observation is, for distinct rooting of an unrooted tree, proper assigning of transition matrix will give same joint distribution. This is the so called *pulley principle*. Consider a tree of two vertices x, y, we can take x as root and the transition matrix is $T_{ij} = P(y = s_i|x = s_j) = P(y = s_i, x = s_j)/P(x = s_j)$ or take y as root with transition matrix $\bar{T}_{ij} = P(x = s_j|y = s_i) = P(y = s_i, x = s_j)/P(y = s_i)$, this two assigning will give identical distribution. This means we can arbitrarily move the root vertex without infecting the output. In practice, we can only reconstruct an unrooted tree from sequences.

There are several extensions of such model, one of the remarkable extension is rate variation across sites. For each site we associate a number called evolutionary rate r, then on this site the transition matrix is $exp(rtQ)$, the evolutionary rate r were generated from certain probability distribution $p(r)$. The most popular model for rate r is $I + \Gamma$, for which:

$$p(r|i, \alpha, \beta) = i\delta(r) + (1 - i)\frac{\beta^\alpha r^{\alpha-1} e^{-\beta r}}{\Gamma(\alpha)}$$

The inverse scale parameter β were normalized to be $\beta = (1 - i)\alpha$ such that the mean value of r equals 1.

4.3 Hadamard conjugation

Hadamard conjugation[69–71] gives a direct correspondence between pattern frequencies and tree edge length under some restricted Markov model, the initial version were only on 2-state and trees. Latter[72] generalized to K3ST model[73] and group-based model[74]. In [75] Waddell *et al.* showed that such method can cooperate with rate inhomogeneous across sites. [76] suggest that the group-based model

can be further generalized to phylogenetic network and incorporate with maximal likelihood method. Such idea were further explored in [77, 78]. It's relationship with group representation were further explored in [79]. We would also note that there are alternative models suggested for phylogenetic network like [80], but unlike Hadamard conjugation we have little idea on properties of those models thus hard to develop algorithms based on these models.

Spectronet[81] is an implementation of Hadamard conjugation, which take alignments as input and output a split network. The Hadamard conjugation have deficit that it needs exponentially time and memory with respect to taxa number. Which implies such method is not suitable for global analysis, in this paper we introduce a method that applies Hadamard conjugation on every four taxa in the taxa set to calculate the quartet weights.

Here we give a brief explanation on Hadamard conjugation and group-based model, we would focus on 4-taxa and DNA (or RNA) sequence case because we are aiming at calculating quartet weight in this special case.

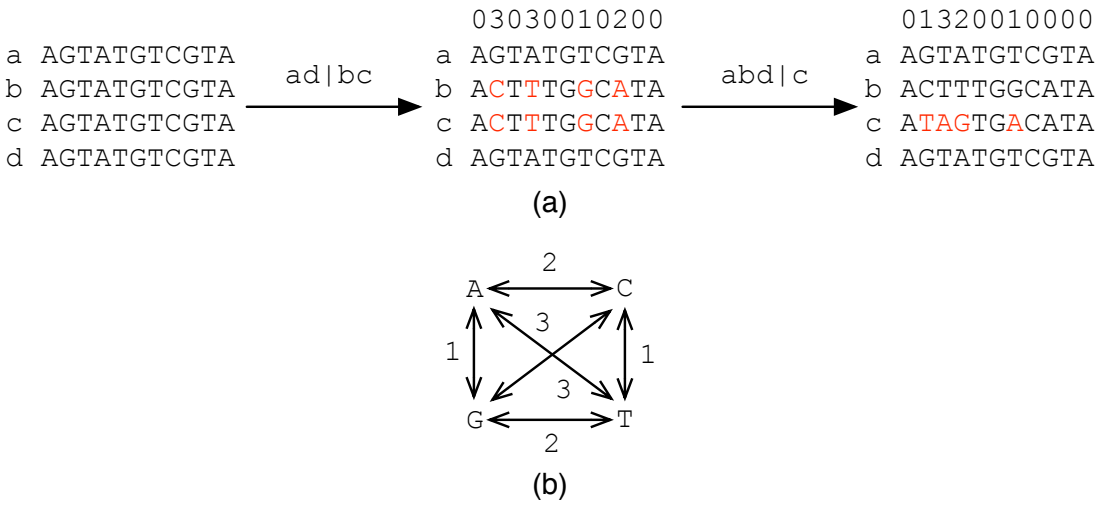


Figure 4.2: A graph illustrating Group-based Model. Fig a. illustrate how sequences changes when new splits were introduced in the system. Fig b. described the group action on the bases in K3ST model. Notice that the red letters shows that changed bases when each split were introduced in.

Definition 17. *Given a set of character $\mathcal{C}$, identify those with a abelian group $\mathcal{G}$, this induce a group action on $\mathcal{C}$. In a group-based phylogenetic process we assign a group element g in $\mathcal{G}$ to every site with a given probability distribution on group*

elements and if the original character at a site i is ξ_i and the group element at this site is g_i the character would change to $\xi_i * g_i$ after this process.

In this paper DNA sequences were considered, the set of character $\mathcal{C} = \{\mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}\}$. and the Klein group $\mathcal{V} = \mathbb{Z}_2 \oplus \mathbb{Z}_2, \mathcal{V} = \{(0,0), (0,1), (1,0), (1,1)\}$ were used, we denote these elements as $0, 1, 2, 3$ respectively.(Fig 4.2 .b shows the group action of Klein group on bases). If the distribution is $(\alpha, \beta, \gamma, \delta)$ the corresponding substitution matrix in this process is:

$$T = \begin{pmatrix} \alpha & \delta & \gamma & \beta \\ \delta & \alpha & \beta & \gamma \\ \gamma & \beta & \alpha & \delta \\ \beta & \gamma & \delta & \alpha \end{pmatrix}$$

with $\alpha + \beta + \gamma + \delta = 1$ which is the K3ST model.

Definition 18. *Given a split system S on taxa set X, we assign each split $s_i \in S$ a random distribution $P_i = (\alpha_i, \beta_i, \gamma_i, \delta_i)$ on the Klein group $\mathcal{V}$. Suppose each site is indexed by j. In this way for each split $s_i \in S$ we can generate a group element g_{ij} following distribution P_i on ever site k_j. And we pick a taxon $a \in X$ as root vertex and for split $s_i = A_i|B_i$ we always assume $a \in A_i$. Firstly we generate the characters of the root vertex by uniform distribution. Using this formula we can generate all the character state ξ_{tj} of any taxa $t \in X$*

$$\xi_{tj} = \xi_{aj} * \prod_{t \in B_i} g_{ij}$$

On one site there is $4^4 = 256$ possible site pattern. The genotypes of taxa, always a multiple alignment, can be encoded by frequencies of those site patterns, thus summarized in a 256-dimensional vector: $p = (p_{AAAA}, \ldots, p_{CCCC})$. The vector $\bar{p}$ with 64 index denote the distribution of "relative" site patterns such that: $\bar{p}_{g_1 g_2 g_3} = \sum_{s_0 * g_i = s_i, i=\{1,2,3\}} P(s_0 s_1 s_2 s_3)$ (s_i are states of taxa). Splits were denoted s_i. For each s_i there is a substitution matrix T_i, we denote its logarithm as Q_i. It must have this form:

$$log(T_i) = Q_i = \begin{pmatrix} \phi_i & \phi_{i1} & \phi_{i2} & \phi_{i3} \\ \phi_{i1} & \phi_i & \phi_{i3} & \phi_{i2} \\ \phi_{i2} & \phi_{i3} & \phi_i & \phi_{i1} \\ \phi_{i3} & \phi_{i2} & \phi_{i1} & \phi_i \end{pmatrix}$$

and $\phi_i = -\phi_{i1} - \phi_{i2} - \phi_{i3}$.

The theorem of Hadamard conjugation are as follows, a proof could refer to [77]:

Theorem 14. $s = H^{-1}(log(H\bar{p}))$

H is the Hadamard matrix of order 64. s is called spectrum which encodes the transition matrix for each splits. For our case we have:

$$s_{g_1g_2g_3} = \begin{cases} \phi_{ij} & g_t = 0 \text{ if taxon } t \text{ and 0 are not separated in split } i,\ g_t = j \text{ if not.} \\ \sum_i \phi_i & \text{all } g_t = 0 \\ 0 & \text{otherwise} \end{cases}$$

Note that some of the elements in the spectrum are always zero, this is crucial for estimating parameters in following section.

It has been found that the rate of evolution is non-constant across sites[82]. And Hadamard conjugation method can be applied to data with rate variation across sites after a slight modification[75]. For every site j we assign a rate j subject to certain distribution $f(r)$. The substitution model of split s_i on this site is $exp(Q_i r_j)$.In this case we have:

Theorem 15. *The relationship between pattern frequency and spectrum is.* $s = H^{-1}(M^{-1}(H\bar{p}))$ *in which:*

$$M(t) = \int_0^\infty e^{-rt} f(r) dr$$

This is a rather straight-forward result considering the linearity of $H\bar{p}$. In our method we consider Gamma+I distribution for convenient, namely a certain proportion of sites are invariant and substitution rate in other sites follows gamma distribution. In this case we have:

$$M(t) = (\frac{(1-i)\alpha}{(1-i)\alpha + t})^\alpha * (1-i) + i \tag{4.2}$$

$$M^{-1}(s) = (1-i)\alpha((\frac{s-i}{1-i})^{-1/\alpha} - 1) \tag{4.3}$$

in the formula i is the proportion of invariant site and α is the shape parameter in gamma distribution.

4.3.1 A proof of Hadamard conjugation

In this section we will give a proof of Hadamard conjugation, the idea is each evolution process is a convolution by certain function. Firstly we explicitly calculate an example of Hadamard conjugation of two-state four-taxa case. Denote the taxa set $X = \{1, 2, 3, 4\}$. The split system $s_1 = 1|234, \ldots, s_4 = 4|123, s_5 = 12|34, s_6 = 13|24, s_7 = 14|23$. For each s_i the transition matrix is

$$T_i = \begin{pmatrix} 1-\alpha_i & \alpha_i \\ \alpha_i & 1-\alpha_i \end{pmatrix}$$

We further assume $\phi_i = -\frac{1}{2}\ln(1-2\alpha_i)$, so

$$\ln(T_i) = \begin{pmatrix} -\phi_i & \phi_i \\ \phi_i & -\phi_i \end{pmatrix}$$

Suppose the joint distribution of four taxa is $p = (p_{0000}, \ldots, p_{1111})^T$, $\bar{p} = (\bar{p}_{000}, \ldots, \bar{p}_{111})^T$ defined as $(p_{0000} + p_{1111}, \ldots, p_{0111} + p_{1000})^T$, we introduce in a new phylogenetic process with s_i and transition matrix of α_i. Take $s_1 = 1|234$ as an example, with probability $1 - \alpha_i$ characters in a site remain unchanged and with probability α_i characters in a site of taxa $2, 3, 4$ flip to another state. Suppose after this process the pattern probability change to p' (and $\bar{p}'$) we have:

$$\begin{aligned} p'_{0000} &= (1-\alpha)p_{0000} + \alpha p_{0111} \\ p'_{0111} &= \alpha p_{0000} + (1-\alpha)p_{0111} \\ &\cdots \\ p'_{1000} &= (1-\alpha)p_{1000} + \alpha p_{1111} \\ p'_{1111} &= \alpha p_{1000} + (1-\alpha)p_{1111} \end{aligned}$$

Take sum of corresponding term:

$$\begin{aligned} \bar{p}'_{000} &= (1-\alpha)\bar{p}_{000} + \alpha\bar{p}_{111} \\ \bar{p}'_{111} &= \alpha\bar{p}_{000} + (1-\alpha)\bar{p}_{111} \\ &\cdots \\ \bar{p}'_{011} &= (1-\alpha)\bar{p}_{011} + \alpha\bar{p}_{100} \\ \bar{p}'_{100} &= \alpha\bar{p}_{011} + (1-\alpha)\bar{p}_{100} \end{aligned}$$

Thus

$$\begin{aligned}
\bar{p}'_{000} + \bar{p}'_{111} &= \bar{p}_{000} + \bar{p}_{111} \\
\bar{p}'_{000} - \bar{p}'_{111} &= (1-2\alpha_1)(\bar{p}_{000} - \bar{p}_{111}) \\
&\cdots \\
\bar{p}'_{011} + \bar{p}'_{100} &= \bar{p}_{011} + \bar{p}_{100} \\
\bar{p}'_{011} - \bar{p}'_{100} &= (1-2\alpha_1)(\bar{p}_{011} - \bar{p}_{100})
\end{aligned}$$

denote $H\bar{p}$ by h and $H\bar{p}'$ by h'

$$h' = \begin{pmatrix}
\bar{p}'_{000} + \bar{p}'_{001} + \bar{p}'_{010} + \bar{p}'_{011} + \bar{p}'_{100} + \bar{p}'_{101} + \bar{p}'_{110} + \bar{p}'_{111} \\
\bar{p}'_{000} - \bar{p}'_{001} + \bar{p}'_{010} - \bar{p}'_{011} + \bar{p}'_{100} - \bar{p}'_{101} + \bar{p}'_{110} - \bar{p}'_{111} \\
\bar{p}'_{000} + \bar{p}'_{001} - \bar{p}'_{010} - \bar{p}'_{011} + \bar{p}'_{100} + \bar{p}'_{101} - \bar{p}'_{110} - \bar{p}'_{111} \\
\bar{p}'_{000} - \bar{p}'_{001} - \bar{p}'_{010} + \bar{p}'_{011} + \bar{p}'_{100} - \bar{p}'_{101} - \bar{p}'_{110} + \bar{p}'_{111} \\
\bar{p}'_{000} + \bar{p}'_{001} + \bar{p}'_{010} + \bar{p}'_{011} - \bar{p}'_{100} - \bar{p}'_{101} - \bar{p}'_{110} - \bar{p}'_{111} \\
\bar{p}'_{000} - \bar{p}'_{001} + \bar{p}'_{010} - \bar{p}'_{011} - \bar{p}'_{100} + \bar{p}'_{101} - \bar{p}'_{110} + \bar{p}'_{111} \\
\bar{p}'_{000} + \bar{p}'_{001} - \bar{p}'_{010} - \bar{p}'_{011} - \bar{p}'_{100} - \bar{p}'_{101} + \bar{p}'_{110} + \bar{p}'_{111} \\
\bar{p}'_{000} - \bar{p}'_{001} - \bar{p}'_{010} + \bar{p}'_{011} - \bar{p}'_{100} + \bar{p}'_{101} + \bar{p}'_{110} - \bar{p}'_{111}
\end{pmatrix} = \begin{pmatrix}
h_{000} \\
(1-2\alpha_1)h_{001} \\
(1-2\alpha_1)h_{010} \\
h_{011} \\
(1-2\alpha_1)h_{100} \\
h_{101} \\
h_{110} \\
(1-2\alpha_1)h_{111}
\end{pmatrix} \tag{4.4}$$

$$\log(h') - \log(h) = \begin{pmatrix}
0 \\
-2\phi_1 \\
-2\phi_1 \\
0 \\
-2\phi_1 \\
0 \\
0 \\
-2\phi_1
\end{pmatrix} \tag{4.5}$$

Remember that $\phi_1 = -\frac{1}{2}\ln(1-2\alpha_1)$

$$H^{-1}(\log(h') - \log(h)) = \begin{pmatrix} -\phi_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \phi_1 \end{pmatrix} \tag{4.6}$$

Which coincides conclusion from Hadamard conjugation: in spectrum s, s_{000} would decrease by ϕ_1 and s_{111} would increase by ϕ_1. This calculation give us some insight on how and what spectrum changes after introducing a new split. For the most general case of arbitrary abelian group, one need to express and prove Hadamard conjugation using Fourier transform.

Hereby we review some results from Fourier transform of finite group[83, 84], in our case the group $\mathcal{G}$ is always abelian and finite, hence without integrability problem. We define its dual $\hat{\mathcal{G}} = \mathrm{Hom}(\mathcal{G}, \mathbb{C}^*)$, an element $\chi \in \hat{\mathcal{G}}$ is called character, $\chi(g)$ is always written as χg, $\hat{\mathcal{G}}$ naturally endowed with abelian group structure.

Given a locally compact abelian group $\mathcal{G}$ and Haar measure, the Fourier transform defines an operator $\mathcal{F} : L^1(\mathcal{G}) \to L^\infty(\hat{\mathcal{G}})$, $f \mapsto \hat{f}$. For $\mathcal{G}$ finite, Fourier transform is:

$$\hat{f}(\chi) = \sum_{g \in \mathcal{G}} \chi g f(g)$$

It's inverse being:

$$f(g) = \frac{1}{|\mathcal{G}|} \sum_{\chi \in \mathcal{G}} (\chi g)^{-1} \hat{f}(\chi)$$

The most useful property of our application is, the Fourier transform of convolution is product of transform, more exactly for $f, h \in L(\mathcal{G})$(for short we write complex-valued function on $\mathcal{G}$ as $L(\mathcal{G})$) define their convolution:

$$(f \star h)(g) = \sum_{g'} f(g') h(g g'^{-1})$$

we have:$\widehat{f \star h} = \hat{f}\hat{h}$.

Then we come back to our application, let the taxa set be $X = \{1, \ldots, n\}$ and 1 be root vertex, each split is represented by the cluster that do not contain root vertex. Relative pattern frequency $\bar{p}$ can be understood as an function on $\oplus\mathcal{G}^{n-1}$, in which $\mathcal{G}$ is the group in group-based method. $\oplus\mathcal{G}^{n-1}$ is denoted $[\mathcal{G}]$, its characters denoted $\widehat{[\mathcal{G}]}$, etc. Note that $\widehat{[\mathcal{G}]}$ is naturally isomorphic with $\oplus\hat{\mathcal{G}}^{n-1}$. $[g]$ is a string of g with length $n-1$ and written as $g_2 \ldots g_n$, $[\chi]$ is a string of χ with length $n-1$ and written as $\chi^2 \ldots \chi^n$. For convenience, given a set $A \subseteq X - 1$ we define the $[g]_A^{g'} = g_2 \ldots g_n$ being a string that $g_i = g'$ if $i \in B$ and $g_i = 0$ if not. The Fourier transform is a linear operator, thus written as a matrix H. Remember that in group-based model one associates each split a transition matrix T, in which $T_{g_i g_j}$ is only dependent of $g_i g_j^{-1}$, we write $T_{g_i g_j} = \alpha_{g_i g_j^{-1}}$. We further define

$$\phi_g = -\frac{1}{|G|} \sum_{\chi \in \hat{\mathcal{G}}} (\chi g)^{-1} (\ln(\sum_{g \in \mathcal{G}} (\chi g) \alpha_g)) \tag{4.7}$$

Once think ϕ and α being elements in $L(\mathcal{G})$, we have $\hat{\phi} = \log(\hat{\alpha})$, And if $Q = \log(T)$, then $Q_{g_i g_j} = \phi_{g_i g_j^{-1}}$. Now we can state the most generalized form of Hadamard conjugation:

Theorem 16. *For a group-based model, the splits system is $\{s_i : s_i = A_i|B_i\}$(we always assume taxa 1, the root taxa is in A_i). The transition matrix is T_i and ϕ_{ig} were calculated using formula 4.7. We have $s = H^{-1}(M^{-1}(H\bar{p}))$, in which s is defined as follows:*

$$s_{[g]} = \begin{cases} \phi_{ig'} & \text{if } [g] = [g]_{B_i}^{g'},\ g' \neq 0 \\ \sum_i \phi_{i0} & \text{if all } g_t = 0 \\ 0 & \text{otherwise} \end{cases}$$

Hereby we stop a while to review the 2-state case, in which $\mathcal{G} = \mathbb{Z}_2 = \{0, 1\}$ and $\hat{\mathcal{G}} = \{\chi_1, \chi_2\}$: $\chi_1(0) = \chi_1(1) = \chi_2(0) = 1$ and $\chi_2(1) = -1$. Every element $[g]$ in $[\mathcal{G}]$ corresponds to A that for every $i \in A$, $g_i = 1$ and $g_i = 0$ otherwise, every element $[\chi]$ in $[\hat{\mathcal{G}}]$ corresponds to B that for every $i \in A$, $\chi^i = \chi_2$ and $\chi^i = \chi_1$ otherwise. We have $[\chi][g] = \prod_{i \in A \cap B}(-1) = (-1)^{|A \cap B|}$, coincides with the definition of Hadamard matrix, and $\phi_0 = \frac{1}{2}\ln(1 - 2\alpha_i)\, \phi_1 = -\frac{1}{2}\ln(1 - 2\alpha_i)$, agrees with our previous definitions.

Then we proceed to proof of Hadamard conjugation, we will do this by introducing splits one by one.

Lemma 4. *If we have taxa set X with relative pattern probability $\bar{p}$, after undergoing a group-based phylogenetic process of split $s = A|B$ and transition matrix with ϕ_g. More exactly, for each site we generate an element in $\mathcal{G}$ with probability with α_g. Characters of taxa in B remain unchanged and characters of taxa in A got multiplied by g of that site. We assume after this the relative pattern probability $\bar{p}'$. Then we have $s = H^{-1}(\log(H\bar{p}')) - H^{-1}(\log(H\bar{p}))$, in which $s_{[g]^{g'}_{B_i}} = \phi_{ig'}$ and zero otherwise.*

Proof: If a site have relative pattern $[g]$ before the process and the randomly generated group element is g' then after this the relative pattern become $[g] * [g]^{g'}_{B_i}$. Thus:

$$\bar{p}'_{[g]} = \sum_{g' \in \mathcal{G}} \bar{p}_{[g]*[g]^{g'^{-1}}_{B_i}} \alpha_{ig'}$$

Thus $\bar{p}' = \bar{p} \star \bar{p}^B$, $\bar{p}^B$ is an function on $[\mathcal{G}]$ with

$$\bar{p}^B_{[g]} = \begin{cases} \alpha_{g'} & \text{if } [g] = [g]^{g'}_B \\ 0 & \text{otherwise} \end{cases}$$

Thus

$$\begin{aligned} s &= H^{-1}(\log(H\bar{p}')) - H^{-1}(\log(H\bar{p})) && (4.8) \\ &= H^{-1}(\log(\hat{\bar{p}} * \hat{\bar{p}}^B) - (\log(\hat{\bar{p}})) && (4.9) \\ &= H^{-1}(\log(\hat{\bar{p}}^B)) && (4.10) \end{aligned}$$

Then we calculate the Fourier transform of $\bar{p}^B$:

$$\begin{aligned} \hat{\bar{p}}^B([\chi]) &= \sum_{g \in \mathcal{G}} [\chi][g]^{g'}_B \alpha_g && (4.11) \\ &= \sum_{g \in \mathcal{G}} (\prod_{i \in B} \chi_i) g' \alpha_g && (4.12) \\ \log(\hat{\bar{p}}^B([\chi])) &= \log(\sum_{g \in \mathcal{G}} (\prod_{i \in B} \chi_i) g' \alpha_g) && (4.13) \\ &= \sum_{g \in \mathcal{G}} (\prod_{i \in B} \chi_i) g' \phi_g && (4.14) \end{aligned}$$

and s being the function on $[G]$ which $s_{[g]^{g'}_B} = \phi_{ig'}$ and zero elsewhere. We have $\hat{s}([\chi]) = \sum_{g \in \mathcal{G}} (\prod_{i \in B} \chi_i) g' \phi_g = \log(\hat{\bar{p}}^B([\chi]))$. Then we only need to verify the null case, when all taxa have same character in every site. Then $\bar{p}_{[g]} = 1$ when $[g] = 0 \ldots 0$ and $\bar{p}_{[g]} = 0$ else where. Hence $H\bar{p}_{[\chi]} = 1$ for every $[\chi]$ $\log(H\bar{p}) = 0$, so $H^{-1} \log(H\bar{p}) = 0$, this completes the whole proof. □

Remark 7. *The last step of calculation can be simplified using morphisms of abelian groups, note that an morphism of finite abelian group $f : \mathcal{G} \to \mathcal{H}$ induce an morphism of characters: $\hat{f} : \hat{\mathcal{H}} \to \hat{\mathcal{G}}$, and we could define push forward $f_* : L(\mathcal{G}) \to L(\mathcal{H})$: for $l \in L(\mathcal{G})$, $f_* l(h) = \sum_{f(g)=h} l(g)$. And pull back $f^{-1} : L(\mathcal{H}) \to L(\mathcal{G})$: for $l \in L(\mathcal{H})$, $f^{-1} l(g) = l(f(g))$. Let m be $\#coker(f)$. We have the following diagram being commutative:*

$$\begin{array}{ccc} L(\mathcal{G}) & \xrightarrow{f_*} & L(\mathcal{H}) \\ \downarrow{\scriptstyle \mathcal{F}|_{\mathcal{G}}} & & \downarrow{\scriptstyle \mathcal{F}|_{\mathcal{H}}} \\ L(\hat{\mathcal{G}}) & \xrightarrow{m \circ \hat{f}^{-1}} & L(\hat{\mathcal{H}}) \end{array}$$

Consider the mapping $f : \mathcal{G} \to [\mathcal{G}]$ by $f(g') = [g]_B^{g'}$. Thus $\bar{p}^B = f_*(\alpha)$, in this case $m = 1$, hence

$$\begin{aligned} s &= \mathcal{F}^{-1}(\log(\mathcal{F}(f_*(\alpha)))) && (4.15) \\ &= \mathcal{F}^{-1}(\log(\hat{f}^{-1}(\mathcal{F}(\alpha)))) && (4.16) \\ &= \mathcal{F}^{-1}(\hat{f}^{-1}(\log(\mathcal{F}(\alpha)))) && (4.17) \\ &= f^*(\mathcal{F}^{-1}(\log(\mathcal{F}(\alpha)))) && (4.18) \\ &= f^*(\phi) && (4.19) \end{aligned}$$

as desired.

4.4 A method calculating quartet weights using Hadamard conjugation

Having extensively discussed the model that sequences generating from, we will propose a method calculating quartet weights using Hadamard conjugation. Like all model-based method, we calculate parameters by selecting those that mostly fits the input. We introduce a novel method that could estimate parameters $I + \Gamma$, which would significantly improve performance.

4.4.1 Parameter estimation for $I + \Gamma$

In this section we introduce a method with estimations of parameters $I + \Gamma$. There exists several methods[85–87] aiming at solely deciding parameters especially invariant site. However, these estimations do not work for our case: the proportion of

invariant site should be less than any element in the vector $H\bar{p}$ which not manifested generally if we use parameters from these method. We proposed a method that estimating parameters in cooperate with Hadamard conjugate method. Observing that some element in the spectrum s with certain index should always be zero, in our case index the vanishing element are: $012, 013, \ldots, 332$. Every split corresponds to 3 non-zero elements and an extra s_{000} being non-zero, hence $64 - 3 * 7 - 1 = 42$ elements should be zero if the parameters are perfect, we denote this index set $\mathcal{U}$. More exactly $\mathcal{U}$ are the set of strings of character $0, 1, 2, 3$ with length 3 and have more than one kind of character except 0. This gives a way for estimating parameters. The idea is, we can adjust parameters to make those elements in s close to zero as much as possible. However, in general we should not compare s from distinct parameters, but we can compare $\bar{p}$. The method is described below.

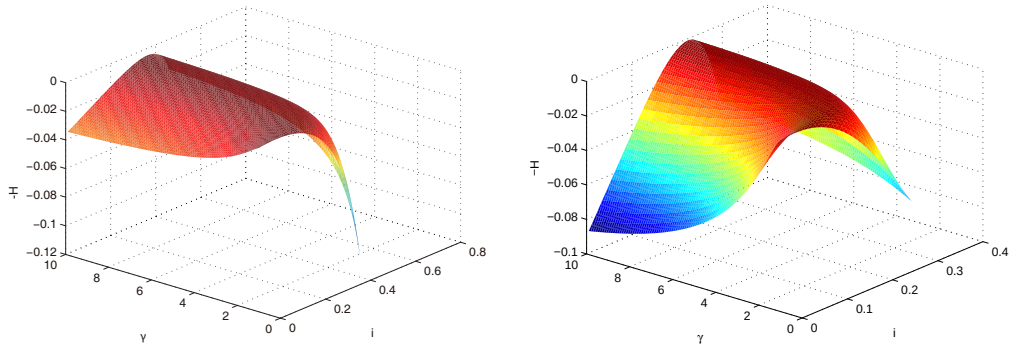
Starting from a vector $\bar{p}$ and parameters i, α we can calculate a spectrum s using formula $s = H^{-1}(M^{-1}(H\bar{p}))$, in general the $s_i, i \in \mathcal{U}$ do not vanish. We have a new spectrum s' such that set those elements 0.

$$s'_i = \begin{cases} s_i & i \notin \mathcal{U} \\ 0 & i \in \mathcal{U} \text{ and } i \neq 000 \\ -\sum_{i \notin \mathcal{U}} s_i & i = 000 \end{cases}$$

Using formula $\bar{p}' = H^{-1}(M(Hs'))$ we can recover the pattern frequency vector $\bar{p}'$, for the correct parameter $\bar{p}'$ and $\bar{p}$ should be very close. Under the framework of Maximum likelihood we seeking for maximize the likelihood $L(i, \alpha) = \prod_{i=000}^{333} \bar{p}_i'^{N * \bar{p}_i}$ (N be the number of sites, namely, the length of sequence). For simplicity we would prefer to work on an equivalent condition: minimize the relative entropy $H(\bar{p}||\bar{p}') = \sum_{i=000}^{333} log(\bar{p}_i/\bar{p}'_i) * \bar{p}_i$, H obtain its minimal when $p' = p$. Finally note that we can do this on every 4-subset of X, for convenience the $-H$ is used as score, we should maximize on the sum of $-H$ of every 4-subset of X with respect to a set of global parameters. The optimization was done by using Nelder-Mead simplex algorithm.

We use $I + \Gamma$ distribution for rate variation. For verifying whether this optimization approach would work, namely there really exists a single maximum that optimization method would converges to, we apply to real and simulated sequences, the results were shown in Fig. 4.3. both case have single maximum.

We use the *logdet* metric[34] for the edge length. So once we've calculated the



(a) **(b)**

Figure 4.3: Score $-H$ with i and γ of real and simulated sequences, (a) is real and (b) is simulated sequences, both case have single maximum.

transition matrix T under K3ST model we would define corresponding split weight to be:

$$-\log(\det T) = -tr(Q)$$

Explicitly writing out $w(12|34) = s_{011} + s_{022} + s_{033}$, $w(12|34) = s_{101} + s_{202} + s_{303}$ and $w(14|23) = s_{110} + s_{220} + s_{330}$.

4.4.2 Generating quartet weights for tree reconstruction

Some reconstruction methods uses quartet weight generated from ML method[31], the idea is fix the tree topology and using the maximized likelihood of every topology as quartet weights, it is taking long time to since it needs to analysis every four taxa. And we notice that Hadamard conjugation can be applied into such cases if we change the assumption from GTR to K3ST model. Firstly notice that for most method the if we plus the 3 quartet weight of certain four taxa by the same number the output tree topology would not change, thus using quartet weight from Hadamard conjugate directly will not provide any new insight than distance-based method. We proposed a method like this: take weight of quartet 12|34 as example, firstly estimate the parameters i, α using methods described above, then take the relative pattern frequency $\bar{p}$ of taxa 1234, then we can calculate the spectrum $s = H^{-1}(M^{-1}(H\bar{p}))$. If the sequences were from a single tree the elements in s vanishes except for those corresponds to splits of the quartet tree. We denote the index set of vanishing elements by $\mathcal{U}_{12|34}$, namely $\mathcal{U}_{12|34} = \mathcal{U} \cup \{101, 202, 303, 110, 220, 330\}$.

Then we have a new spectrum s' such that

$$s'_i = \begin{cases} s_i & i \notin \mathcal{U}_{12|34} \\ 0 & i \in \mathcal{U}_{12|34} \text{ and } i \neq 000 \\ -\sum_{i \notin \mathcal{U}_{12|34}} s_i & i = 000 \end{cases}$$

Thus as the previous section we can calculate the $\bar{p}'$ from s' and define the score for this quartet as $-H(\bar{p}||\bar{p}')$, and same as network method, sum over all 4-subset as global score and optimize using Nelder-Mead method.

4.5 Simulation

There are several problem we are interested:

1. We have introduced a method that estimates i and γ, we also have known that theoretically such method is asymptotically consistent. Thus it's crucial whether such estimation would introduce significant variance. We will study this problem by simulation.

2. The group-based model is GTR model satisfying several conditions: 1. the stationary base composition is uniform 2.the transition rate corresponds same group element should be same. We are curious to know whether Hadamard conjugate method is still performing well if these conditions are not satisfied.

3. Many method has been raised to represent the quartet weight, including a ML-based method(implemented in [31]), Statistical Geometry[88](SG) and Squangle[89]. The group-based model can generate artificial sequences from a network. We know that the Hadamard conjugation method would giving correct result in such case, how would other quartet reconstruction methods behave under such a setting?

4. The group-based model has been criticized for not biological meaningful. The mixed-tree model are more widely used in simulating reticulate evolution. We want to know whether the error is significant if we apply Hadamard conjugation method to sequences from mixed tree.

4.5.1 Stability of $I + \Gamma$ estimation

In this section we will look at the problem of stability of $I + \Gamma$ estimation, namely how would sampling error affect the accuracy of estimated parameter. We generate 4 sequences from network with pending edges length 0.6 and quartet weights 0.09, 0.18 and 0.27 respectively, all substitution models are JC model, the parameter of rate variation is $i = 0.2$ and $\gamma = 3$, sequences length were set as $50K$. We apply our method to those sequences. The results were shown in Fig. 4.4. From this study we can conclude that even we might not estimates parameters in high accuracy (Fig (a)), the curve $M^{-1}(s)$ is close to correct one (Fig (b)). For further single out the random effect, we consider the true probability distribution p, then do the PCA analysis of $M^{-1}(H\bar{p})$ with respect to parameters, the error were mainly in PCA 1, which corresponds to a global factor (Fig (c)), hence this estimation method will not introduce too much bias and variation into the final results (Fig (d)).

4.5.2 Single tree case

DNA evolution is always modeled as a continuous-time Markov model. The rate matrix of GTR model is determined of two group of parameters: i)steady-state base distribution: $\rho = (\rho_A, \rho_G, \rho_T, \rho_C)$. ii)mutation rates: $(x_{AG}, x_{AT}, x_{AC}, x_{GT}, x_{GC}, x_{TC})$. In all we have:

$$Q = \begin{pmatrix} -\frac{x_{AG}+x_{AT}+x_{AC}}{\rho_A} & \frac{x_{AG}}{\rho_A} & \frac{x_{AT}}{\rho_A} & \frac{x_{AC}}{\rho_A} \\ \frac{x_{AG}}{\rho_G} & -\frac{x_{AG}+x_{GT}+x_{GC}}{\rho_G} & \frac{x_{GT}}{\rho_G} & \frac{x_{GC}}{\rho_G} \\ \frac{x_{AT}}{\rho_T} & \frac{x_{GT}}{\rho_T} & -\frac{x_{AT}+x_{GT}+x_{TC}}{\rho_T} & \frac{x_{TC}}{\rho_T} \\ \frac{x_{AC}}{\rho_C} & \frac{x_{GC}}{\rho_C} & \frac{x_{TC}}{\rho_C} & -\frac{x_{AC}+x_{GC}+x_{TC}}{\rho_C} \end{pmatrix} \tag{4.20}$$

A GTR model is K3ST model if the following condition is satisfied: i)the steady-state is uniform: $\rho = (0.25, 0.25, 0.25, 0.25)$. ii)The mutation rate corresponds to same group element is equivalent respectively: $x_{AG} = x_{TC}, x_{AT} = x_{GC}, x_{AC} = x_{GT}$. So we will look at the performance of Hadamard conjugation method when these conditions are violated.

We would plan using these conditions to test the performance of Hadamard conjugation method:

1. Different mutation rate corresponds to one group element.

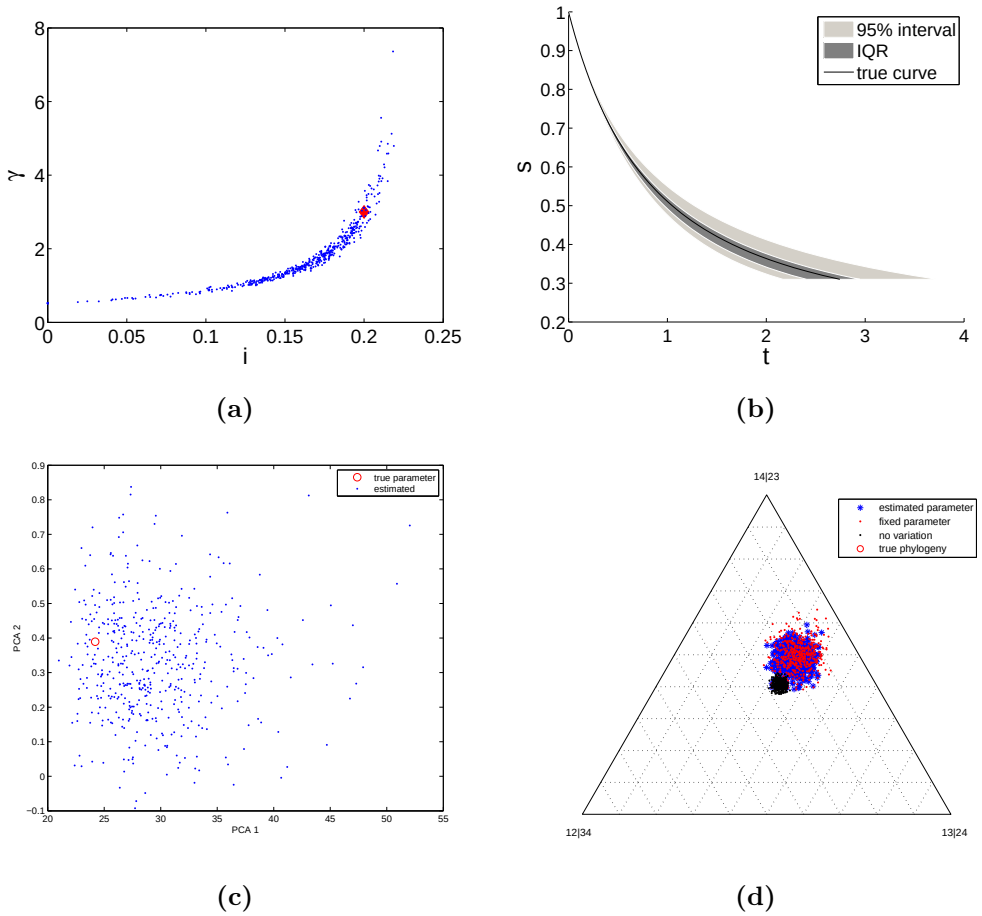


Figure 4.4: Simulation results on stability of $I + \Gamma$ estimation. In (a) the estimated i and γ value are plotted, the red circle indicates the true parameter, in (b) we plotted the function $M^{-1}(s)$(formula 4.3) with simulated i and γ, 95% interval, IQR (the interval of first and third quantile), and curve with true parameter were also plotted, in (c) we plot the first two PCA components of $M^{-1}(H\bar{p})$ with p being the true joint distribution, in (d) we plotted reconstructed quartet weights using estimated parameter, true parameter and naive method without rate variation (with legend "no variation").

2. nonuniform steady-state base composition.

3. Inhomogeneous steady-state across the tree.

The underlying quartet tree were always 12|34. The length of pending edge of taxa 1, taxa 3 and taxa 2, taxa 4 were set same respectively, we denote this two number edge length 1 and edge length 2. We use average number of mutation event on each site as edge length. The test for inhomogeneous steady-state across the tree were conducted as follow: the steady-state of two internal edges were uniform, and the steady-state corresponds to the pending edge of taxa 1 and 3 were set same, so do taxa 2 and 4.

Some parameters in simulations are:

1. sequence length is infinity, namely, we always use probability as frequencies.

2. nonuniform base composition: $\rho = (0.3, 0.3, 0.2, 0.2)$ or $(0.4, 0.4, 0.1, 0.1)$.

3. Different mutation rate corresponds to one group element: $(x_{AG}, x_{AT}, x_{AC}, x_{GT}, x_{GC}, x_{TC})$ are $(1, 1, 1, 1, 1, 0)$, $(1, 1, 1, 1, 0, 0)$, $(1, 1, 1, 0, 0, 0)$ and $(1, 1, 0, 1, 0, 0)$

4. internal edge length: 0.1 or 0.05.

5. external edge length: 0 to 1.5.

6. no rate variation across sites is involved.

We also run the tree version of quartet weight for these case.

The impact of breaking the condition of K3ST but having a homogeneous steady-state across the tree is similar(case 1,2): the proportion of wrong split versus right split is high when two pending edge that not neighbors are much longer than the other two (Figure 4.5 and 4.6). While if the steady-state across the tree changes a lot the performance is getting worse while branches being longer. It's worthwhile mentioning that introducing rate variation across sites would compensate the error induced by nonuniform base composition and different mutation rate corresponds to one group element.

We also compared our results against ML method (implemented by QNet [31]) in the third case (Figure 4.9 and 4.10), ML method is also not doing well in long branch

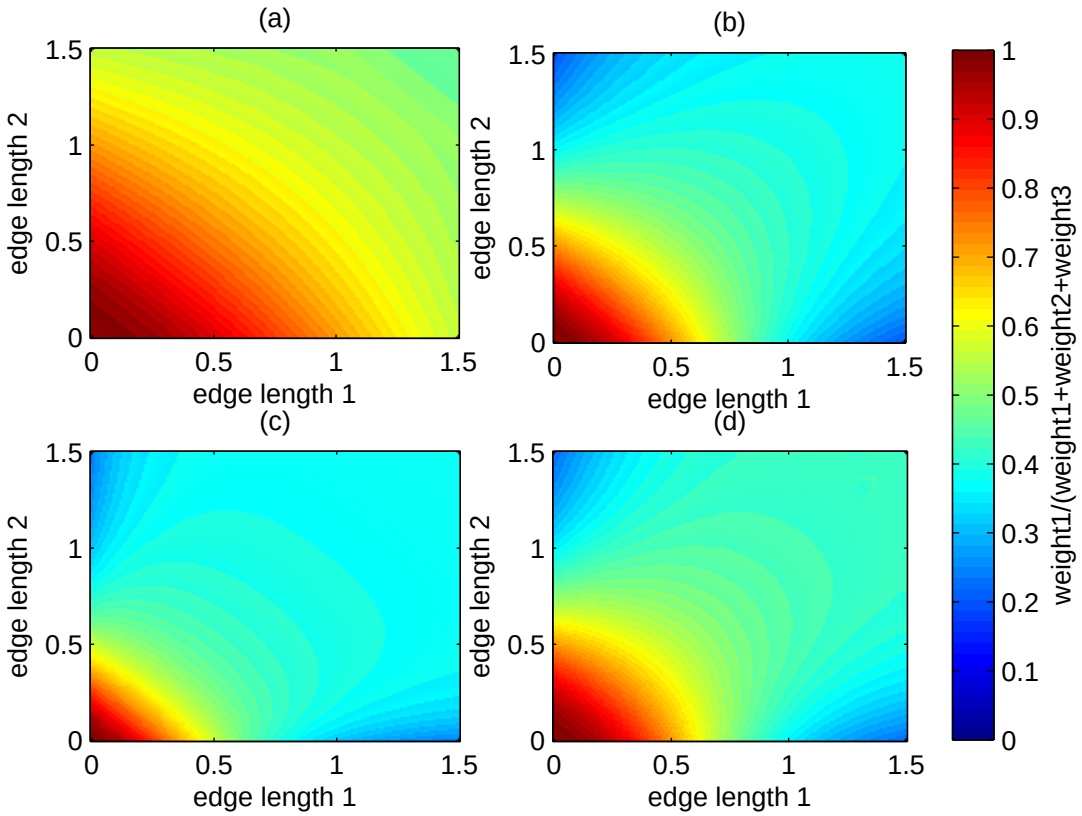


Figure 4.5: Different mutation rate corresponds to one group element. The ratio of correct quartet weight versus all quartet weight were used to show the validity of the method. It is calculated under infinite-site model, i.e. the exact probabilities not frequencies from simulation were used in Hadamard conjugation method. The base composition is uniform in all simulation and $(x_{AG}, x_{AT}, x_{AC}, x_{GT}, x_{GC}, x_{TC})$ are $(1,1,1,1,1,0)$, $(1,1,1,1,0,0)$, $(1,1,1,0,0,0)$ and $(1,1,0,1,0,0)$ respectively. Length of internal edge were all set as 0.1. We also run the tree weight method, in all those cases the likelihood of tree $12|34$ are over likelihood of tree $13|24$ and $14|23$.

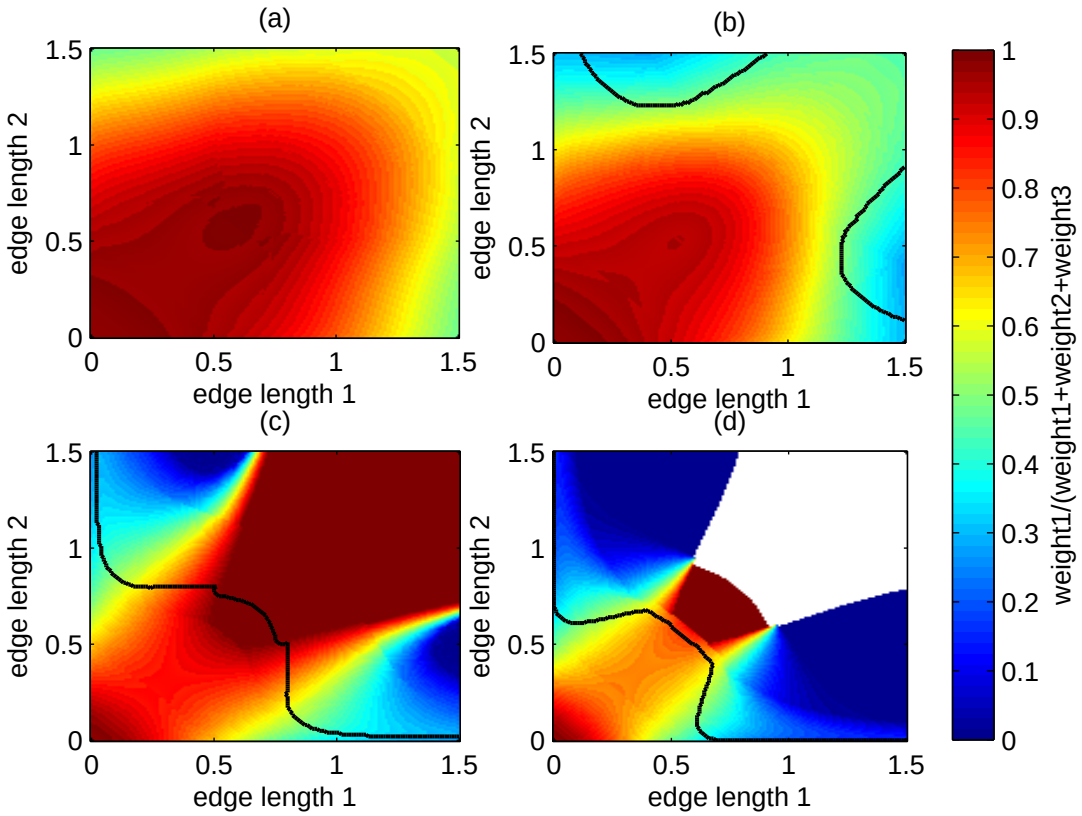


Figure 4.6: Nonuniform steady-state base composition. The ratio were also calculated under infinite site model. The base composition of (a) and (b) are $(0.3, 0.3, 0.2, 0.2)$ and (c) and (d) are $(0.4, 0.4, 0.1, 0.1)$. The length of internal edge are 0.1 for (a) (c) and 0.05 for (b) and (d) respectively. The black line is the boundary line which the likelihood of tree $12|34$ equals the greater of the two other trees. If in all those cases the likelihood of tree $12|34$ are over likelihood of the two other trees the line are not shown.

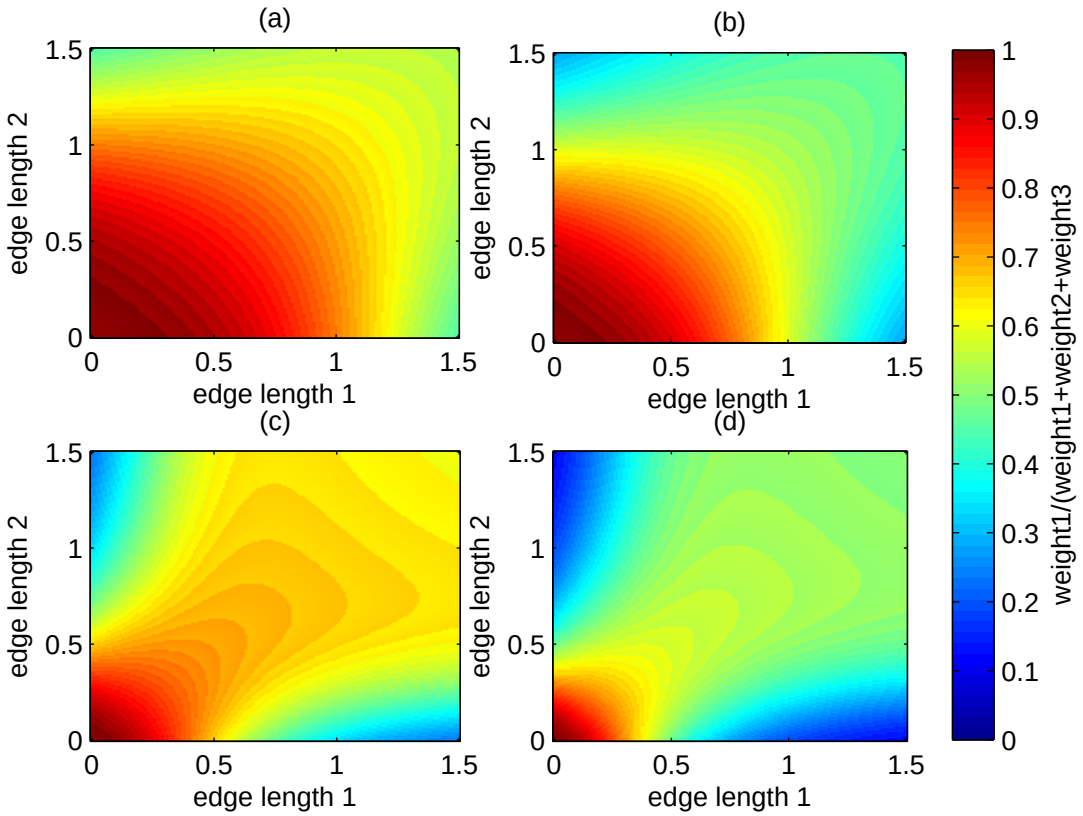


Figure 4.7: Nonuniform steady-state base composition using naive Hadamard conjugation. The settings were the same as Figure 4.6. The quartet weight method were naive Hadamard conjugation, no rate-variation across sites.

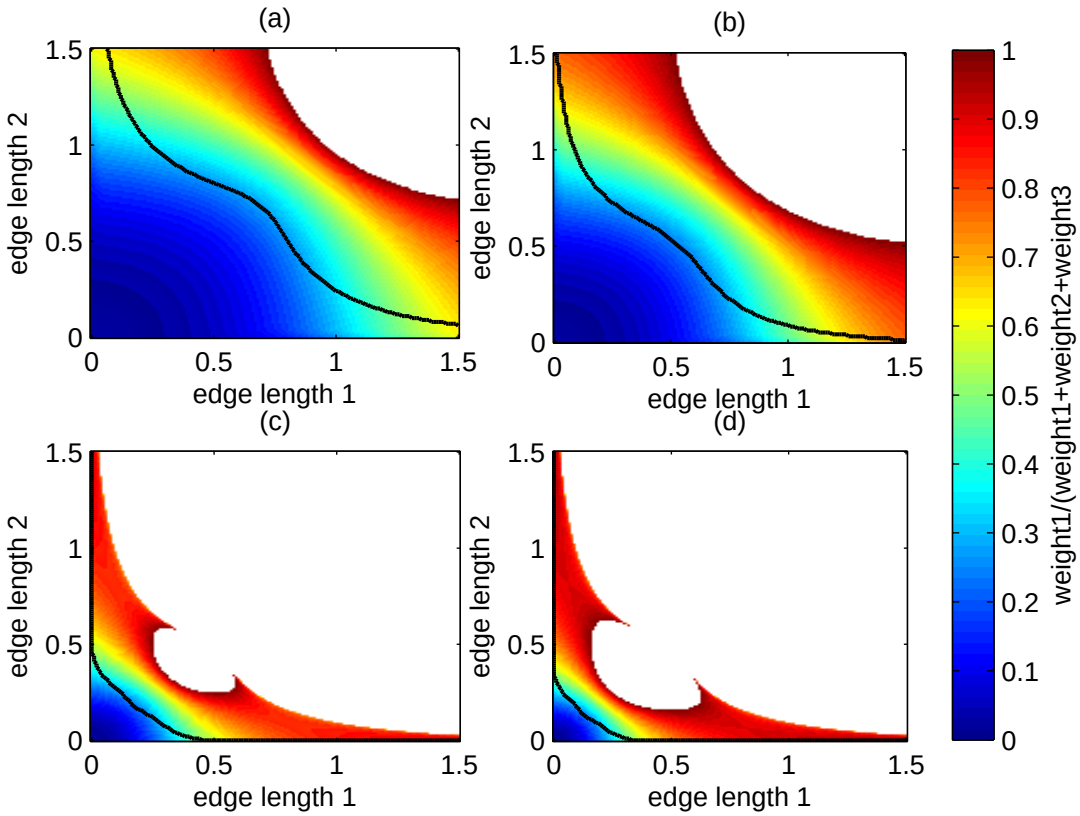


Figure 4.8: Inhomogeneous steady-state across the tree. The ratio were also calculated under infinite site model. In (a) and (b) the base composition for taxa 1 and 3 are $(0.3, 0.3, 0.2, 0.2)$ and taxa 2 and 4 are $(0.2, 0.2, 0.3, 0.3)$ and for and (c) and (d) the base composition for taxa 1 and 3 are $(0.4, 0.4, 0.1, 0.1)$ and taxa 2 and 4 are $(0.1, 0.1, 0.4, 0.4)$. The length of internal edge are 0.1 for (a) (c) and 0.05 for (b) (d) respectively. the blank zone in upright corner indicates the method fails to give a result, usually involving taking logarithm of a non-positive number.The black line is the boundary line which the likelihood of tree 12|34 are not over the two other trees.

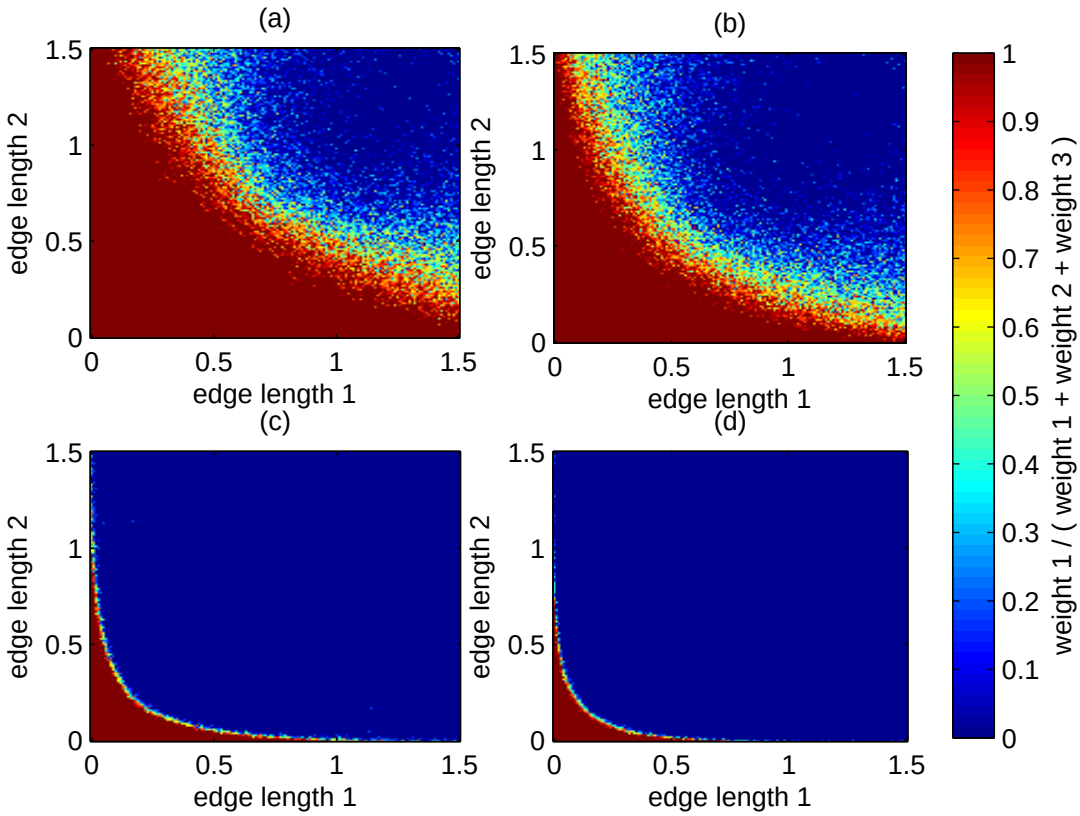


Figure 4.9: ML method on inhomogeneous steady-state across the tree. The setting of base composition and edge length was the same as Figure 4.8. And the length of sequence were set 10000. Each data point were simulated 5 times and taking average.

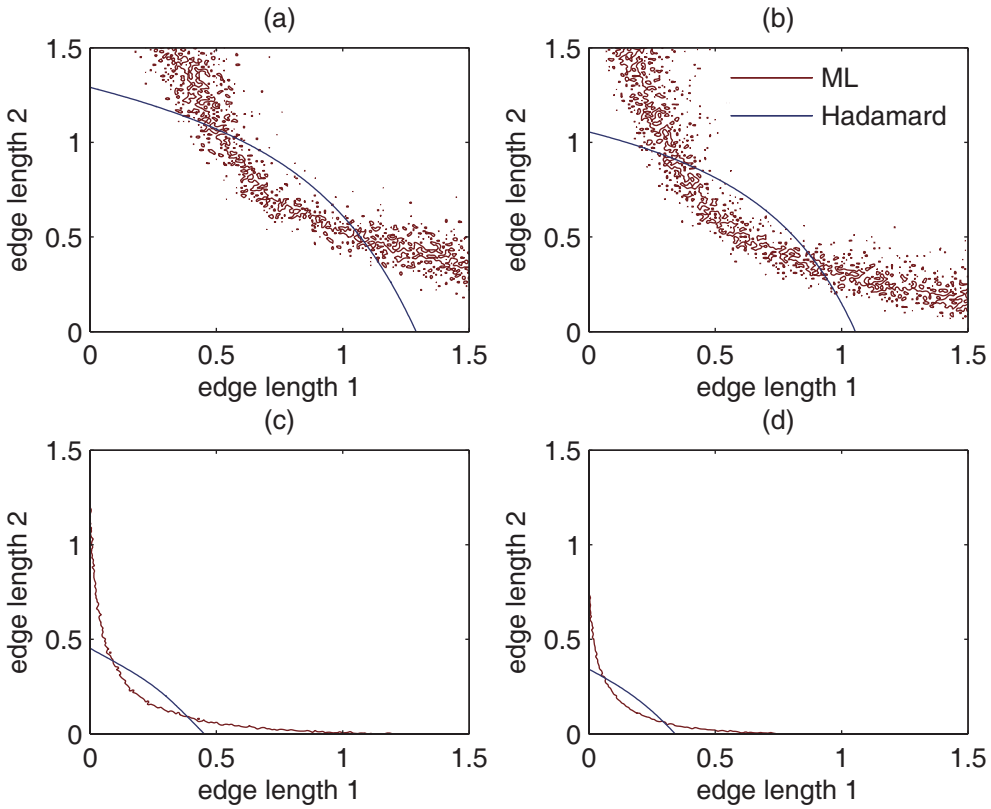


Figure 4.10: Contour line of ML and Hadamard conjugate method on inhomogeneous steady-state across the tree. The contour line were taken at 0.5, namely the sum of two wrong quartet weight equals the right weight.

case because it assumes that the substitution matrix across the tree is same. Generally the performance of ML method is no big difference than Hadamard conjugate method in these cases (Figure 4.10).

4.5.3 Network case

For we are having a model that can simulate sequences following a network and Hadamard conjugate method is consistent under such setting. We were curious to know how would other popular methods performs. Statistical Geometry[88](SG), Maximal Likelihood(ML), Squangle[89] were used for comparison. Implementation of SG and ML method were from QNet. We found that results from Hadamard conjugate method is close to real phylogeny in all these cases; both ML and Squangle tend to give extreme quartet weight while quartet weight from SG method tend to be near center Figure(4.11). This means even for a single tree the Statistical Geometry method would reconstruct a network with many reticulate events.

4.5.4 Evolution with hybrid case

In this section we simulate a more realistic scenario: sequences were generated following a tree with hybridization events. Fig. 4.12 describe the evolutionary process of five simulated taxa. The total height of tree is 0.5. Hybrid event were simulated as random mixture, with 60% from left node and 40% from right node, edge length were average number of mutation events on each site, sequence length were 10000. The quartet weight were calculated using Hadamard conjugation, statistical geometry and maximal likelihood. Result were shown in 4.13. The linear relation of quartet weight and split weights were full rank, thus reconstruction were using least square non-negative optimization. The error in this process were evaluated by cosine similarity, namely $d(\mathbf{v}_1, \mathbf{v}_2) = \arccos(\frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|})/(\frac{\pi}{2})$, $\frac{\pi}{2}$ is a normalization factor that makes the distance no more than 1. We also the prior distribution of distance, which is the distribution of distance of real split system weights and vector that randomly distributed in the positive orthant.

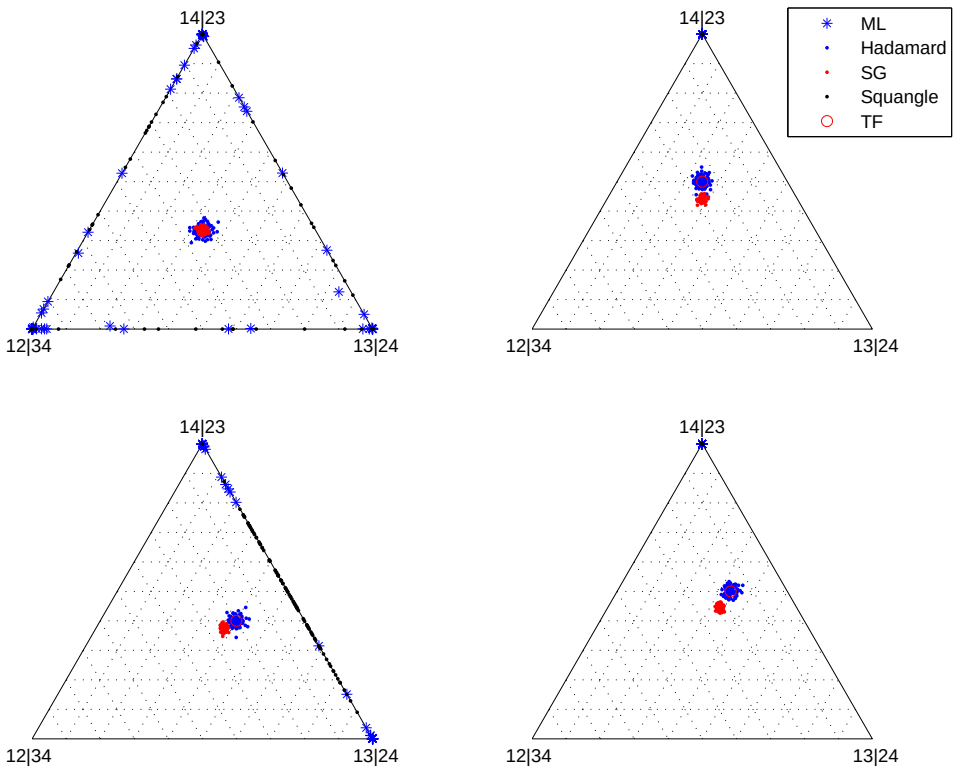


Figure 4.11: Comparison of four method in network case. The weight of three quartet weight (12|34, 13|24, 14|23) was $(0.1, 0.1, 0.1)$, $(0.1, 0.1, 0.2)$, $(0.1, 0.2, 0.2)$, $(0.1, 0.2, 0.3)$ respectively.

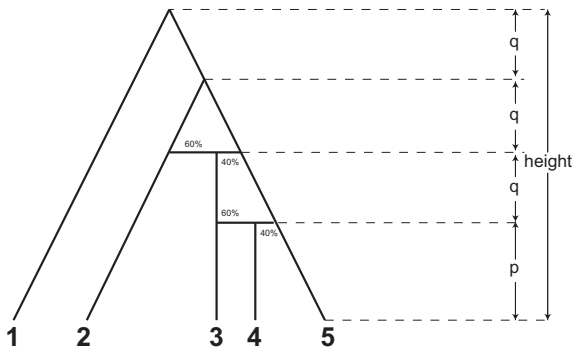


Figure 4.12: Evolution process we used for simulation, total height of tree is 0.5.

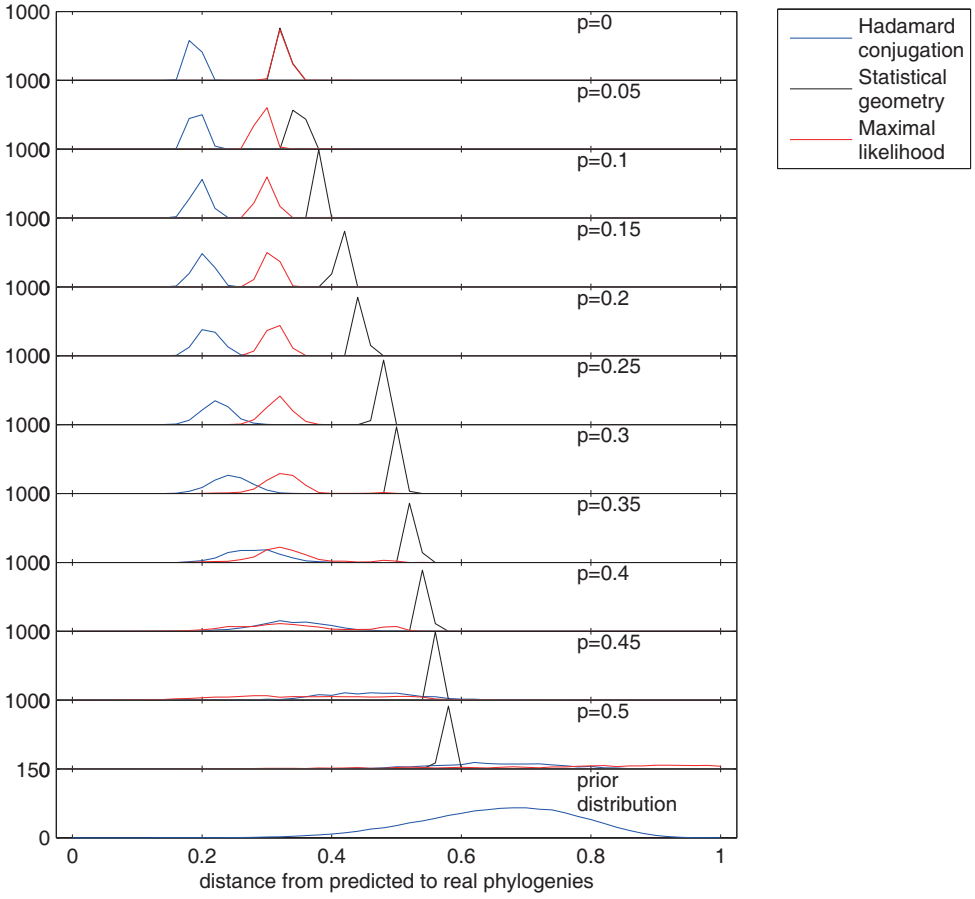


Figure 4.13: Error of prediction compared with true phylogenies.

4.5.5 Conclusion

In all, we concluded that firstly Hadamard conjugate method is the only choice if consistency is required, other method is giving incorrect results under such setting and would give false splits or reject true ones. Secondly, Hadamard conjugate method generally performs well if the base frequency is homogeneous across the tree and not fall into Felsenstein zone. The performance of Hadamard conjugation method is not satisfying when the base composition is inhomogeneous and can be affected by long-branch attraction. Hadamard conjugation method may not work well on those datasets. We also find that in general the variance of Hadamard conjugation method (Fig. 4.11) is a bit more than statistical geometry, but no much difference with maximal likelihood method (Fig. 4.13).

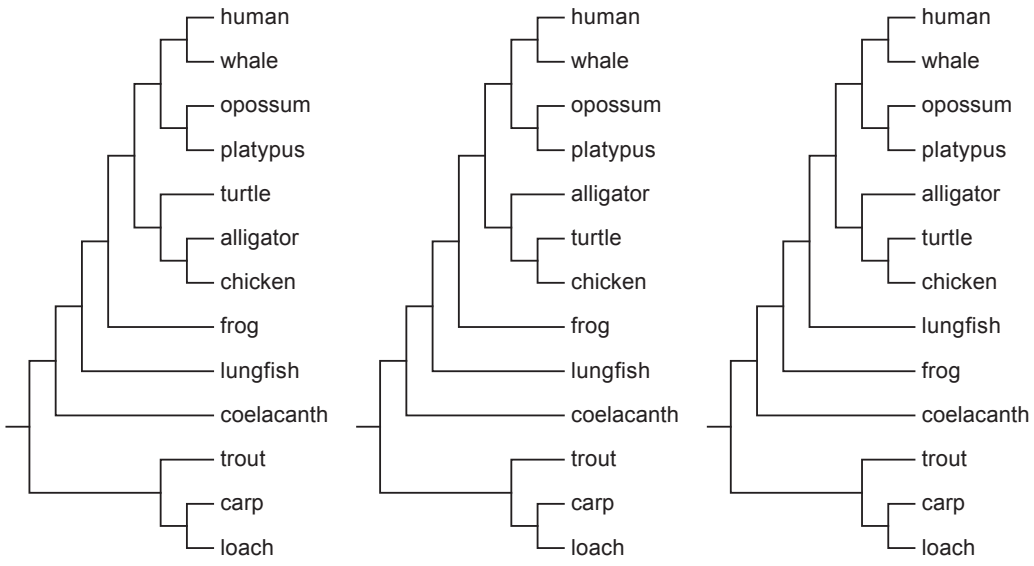
4.6 Real data

4.6.1 Zardoya dataset

In [90] Zardoya describes a dataset of 13 taxa for determining the position of turtle and use 3 teleosts as outgroup, we use this data set to verify the quality of Hadamard conjugation method. We concatenate all the genes in mitochondrial genome and align them, then we use Hadamard conjugation to get quartet weights and use Clann to get an optimized tree the meets the quartet weights as much as possible. The only difference of tree reconstructed and the one suggested in paper is the relationship of chicken, turtle and alligator. This is the most crucial part of the research yet NJ and ML method are also not reconstructing correct results (correct results requires protein sequence analysis). While there are also research[91] that gives identical results from same dataset.

4.6.2 Squamata dataset

As we've discussed in the introduction part, Many method have been developed to reconstruct phylogenetic networks using quartet weights as input, including QNet and QuartetNet. In this section we will check whether these method will give reasonable results from real biological data using quartet weights from Hadamard conjugation. We use the dataset of [52] to validate our method, this dataset consists of 31 taxa,

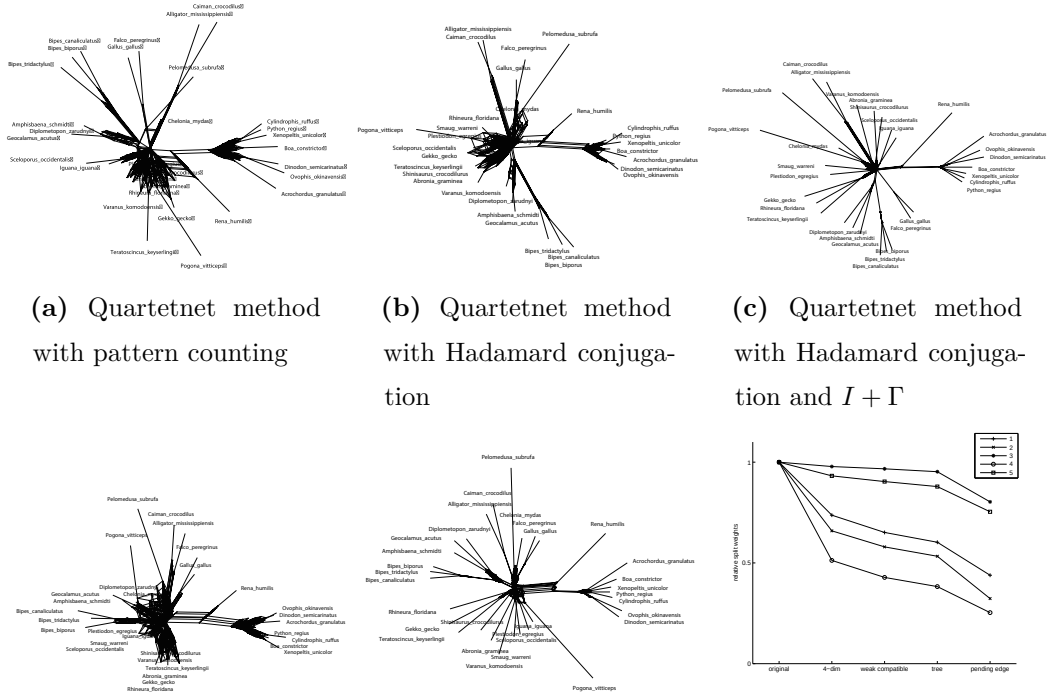


(a) proposed tree topology in [90]

(b) tree topology from Clann using Hadamard conjugate and ML quartet weight

(c) tree topology from all other method

Figure 4.14: Trees reconstructed from different quartet weight method, we use NJ,ML and subtree method using quartet weights from Hadamard conjugate, Squangle, ML, Statistical Geometry method. All method failed to reconstruct the most accepted phylogenetic order of turtle, alligator and birds.



(a) Quartetnet method with pattern counting

(b) Quartetnet method with Hadamard conjugation

(c) Quartetnet method with Hadamard conjugation and $I + \Gamma$

(d) QuartetDecomposition with naive Hadamard conjugation

(e) QuartetDecomposition with naive Hadamard conjugation and $I + \Gamma$

(f) Proportion of remaining split weight with filters, (1-5 correspond to a-e).

Figure 4.15: (a-e) are networks reconstructed from different pipelines, (f) are sum of split weights remained after different filters, normalized by the sum of split weights of original one.

including mammals, birds, lizards and squamata, which have complicated reticulate events involved and is inappropriate for ordinary approaches. It's also analyzed in [33], which using pattern counting, a method similar to Statistic geometry, as quartet weight, its accuracy were also debated in [33]. We redo the analysis using quartet weight from Hadamard conjugation to see whether we can improve the performance of Quartetnet method. We also use QuartetDecomposition method to analysis this dataset.

We use different quartet weight generating method and varied settings to reconstruct the network. The third codon in the sequence were removed in analysis, so the results in 4.15a were a bit different with results in [33]. We can conclude that Hadamard conjugation method with invariant site removed could significantly improve the performance of both Quartetnet and 2-SplitDecomposition method, i.e. reducing both the number and weights of the false splits. In a evolutionary process,

we would expect that the main part is tree like and some reticulate events exists in some parts, and the probability of certain splits exists would significantly decrease if it is too far from the tree. Such property can be utilized to check the network by putting them through different filters. For example, in the correctly reconstructed network, the split removed after greedy weak-compatible filter would have much less weights compared with those of greedy compatible filter, but for the noise signal this two part would not be significantly differed. In 4.15f, the proportion of remaining split weight with different filters of five networks are plotted. in this aspect we can conclude that the network reconstructed using quartet weight method Hadamard conjugate with invariant site is more reliable than others.

4.7 Final Remark

Firstly we have proposed a method that calculates quartet weights using Hadamard conjugation. Then we verified the performance of our method using both simulations and real data. The method is asymptotically consistent on group-based model while all other method do not, it is in advantage in phylogenetic network reconstruction and the real data study have shown that it's also capable of tree reconstruction. We would also note that compared with pattern counting method the Hadamard conjugation has higher variance (Fig. 4.11), so from bias-variance trade-off principle[92], we would recommend to use on sequences longer than 5K bp, and the base frequency should be close to uniform, as indicated by simulation.

Since the Hadamard conjugation method is only consistent on group-based model, one would naturally ask the question why do not use more general model. The answer is, there is no network model that gives explicit formula of pattern frequencies like group-based model do. This is even impossible for general Markov model of tree. Even models have been proposed to generalize Markov model to network[80], its properties (identifiability, stability, etc.) is unknown. It's even harder to cooperate with rate variation across sites, which has been shown to be crucial for accuracy of our method. Compared with those models the group-based model have smaller number of parameters, which would reduce over-fitting.

In all we would conclude that the Hadamard conjugation method is a novel method to calculate quartet weights from sequences, two version were designed for

tree and network construction. Which provides an alternative of pattern counting method. Cooperating with quartet weight-based reconstruction method it would be helpful to analyze evolutionary process with complicated reticulate events.

Chapter 5

Epilogue

In this chapter we present some result that do not fit in the 2-metric context of this thesis, most results are related with attempts of building a theory for combinatorics of split systems with general forbidden configuration, especially deciding the order of maximal cardinalities, which is rather an undeveloped field and only partial results were known.

5.1 A general theory of split system

The central object of this chapter is split system, which can be think of projective analog of cluster theory, or hypergraph theory[93]. We focus on the case of split system with some forbidden configuration, we should note that the operation of taking minors in hypergraph theory and restriction is distinct, this leads to different methods. We first introduce the concept of closed family of split systems, which summarizes the property of split systems with certain forbidden configuration.

Definition 19. *A family F of split systems is said to be closed if $S' \in F$ and $S \trianglelefteq S'$ implies $S \in F$, it's called full if it contains all possible split systems. Split systems with certain forbidden configuration(s) is closed.*

Theorem 17. *A closed family of split system can be characterized by (might be infinite numbers of) forbidden configurations.*

Proof: Denote the family of all split system F'. We will prove the F equals the split system with forbidden configuration $F' \backslash F$. Denote such family $\bar{F}$, it's not hard to show $\bar{F} \subseteq F$. We will prove $F \subseteq \bar{F}$. If $S \in F$, from definition we know that for any $S' \in F' \backslash F$, $S' \not\trianglelefteq S$ thus $S \in \bar{F}$. □

We call the set of forbidden configuration of a closed family of split systems obstruction set. A closed family of split systems do not necessarily characterized by finite number of forbidden configurations, in [94] it has been point out that topes of realizable rank 3 oriented matroids have infinite number of independent forbidden configurations. We conjecture that topes of oriented matroids with certain rank, and circular splits system also have infinite number of independent forbidden configurations.

Definition 20. *The VC dimension for a split system is the size of the largest set Y such that $S|_Y = \mathcal{P}(Y)$.*

Here we introduce a notation: $\Phi_d(n) \doteq \sum_{i=0}^{d} \binom{n}{i}$

Proposition 1. *[58] If S has VC dimension d then $|S| \leq \Phi_{d-1}(n-1)$.*

In the field of machine learning, we know that the VC dimension of concept class is directly related with learnability[92, 95, 96]. We expect to give a better bound of error in terms of the exact number of concept class.

Proposition 2. *[97] Let $\Pi_H(m)$ be the maximal ways of separating m points. Then if*

$$\delta > 2\Pi_H(2m)2^{-\epsilon m/2}$$

Any consistent learn function can have error more than ϵ of probability less δ.

Hereby we list some known result:

1. For obstruction set with only one obstruction: F_d. We can reach the upper bound $\Phi_{d-2}(n-1)$. More over, if the forbidden configuration is $F_n^{\lfloor \frac{n}{2} \rfloor}$ we can also reach the upper bound (topes of uniform positively oriented matroid). It's known that every big enough point set in an n dimensional affine space must contain a subset being non-trivially neighbourly[98, 99], hence if a split system with forbidden configuration on d-set have maximal cardinality being $\Phi_{d-2}(n-1)$, the compatible condition must be weaker than $F_n^{\lfloor \frac{n}{2} \rfloor}$.

2. If we have an upper bound of different order from $O(n^{d-2})$ we call them non-trivial upper bound. the first example of non-trivial upper bound is tree, namely the obstruction set $\{\{12|34, 13|24\}\}$. Another example is the obstruction set $\{\{12|345, 13|245, 14|235, 15|234\}\}$. We proved its upper bound is $O(n^2)$. This is essentially all nontrivial examples we have.

We consider the *X-pairs* techniques first introduced in [100], which will be heavily used in this chapter. The central idea is to observe the difference of $\#S$ and $\#S|_{X-x}$. Hereby we reformulate it using notations of forbidden configurations. Note that in the following texts of this section $\emptyset|X$ would always count as valid split.

Definition 21. *For a split system S on X and $x \in X$ we define $\partial_x S$ being the split system $\{A|B : A \cup \{x\}|B,\ A|\{x\} \cup B \in S\}$ over $X - x$.*

Proposition 3. *If S is a split system on X and $x \in X$, we have: $\#S = \#S|_{X-x} + \#\partial_x S$.*

Proposition 4. *If the cardinality of split systems with obstruction $h|_{X_h-\{x\}}$ have upper bound $O(n^{k-1})$ then the cardinality of split systems with obstruction h have upper bound $O(n^k)$. (X_h being ground set of h)*

Proposition 5. *If split system S has VC-dim n, $\partial_x S$ has VC-dim no more than $n - 1$.*

Combining all propositions together we can give a proof for theorem 2. This leads to a generic method: if you want to look at the upper bound of split systems with obstruction set H you can look at the split systems with obstruction set $\{h|_{X_h-\{x\}} : h \in H\ ,\ x \in X_h\}$. (Again X_h being ground set of h), and this can be done recursively. For example the forbidden configuration of weakly compatible is $12|34, 13|24, 14|23$ the forbidden configuration after taking ∂_x is $1|23, 2|13, 3|12$, thus be induced by a linear ordering[100], so with cardinality no more than $n - 1$, so the cardinality of weakly compatible is no more than $\binom{n}{2}$.

We could further conclude that if for a closed family of split system F. Any $S \in F$ the is $x \in X$ that $\#\partial_x S \leq O(n^k)$, $\#S \leq O(n^{k+1})$. However one do not necessarily have if there always exist x, y that $\#\partial_x \partial_y S \leq O(n^k)$, $\#S \leq O(n^{k+2})$, one way for remedy this is, taking average over $x, y \in X$.

5.1.1 Linear independency

In this section we attempt to see how to combine theories of linear algebra and combinatorics of split systems. One of the clue is, for weakly compatible split system, mapping from weights to distances is injective. We first generalize this to diversities[43, 101].

Definition 22. *For a given weighted split system (S, w), we define diversity of a subset $Y \subseteq X$ as:*

$$\delta(Y) = \sum_{A|B \in S, A \cap Y \neq \emptyset, B \cap Y \neq \emptyset} w(A|B)$$

For a split $A|B$ we define vector $\boldsymbol{\delta}_{A|B}$ as

$$\boldsymbol{\delta}_{A|B}(Y) = \begin{cases} 1 & A \cap Y \neq \emptyset, B \cap Y \neq \emptyset \\ 0 & otherwise \end{cases}$$

, Y is all subsets of X, $\boldsymbol{\delta}_{A|B}^{(2k)}$ is when Y is all subsets of X with cardinality $2, 4, \ldots, 2k$.

Proposition 6. *[102] For a split system S having forbidden configuration F_{2k+2}, vector set $\{\boldsymbol{\delta}_{A|B}^{(2k)} : A|B \in S\}$ is linearly independent.*

Proof: We prove this by induction on $n = |X|$. For $n \leq 2k$, we can show this by Möbius transform [101], for $n > 2k$ assume we have a set of coefficients $w_{A|B}$ such that $\sum_{A|B \in S} w_{A|B} \boldsymbol{\delta}_{A|B}(S) = 0$ for every $S \subseteq X$, $|S| \leq 2k$. Then for every $x \in X$ we consider the S's restriction on $X - \{x\}$, we have $\sum_{A|B \in S|_{X-\{x\}}} w'_{A|B} \boldsymbol{\delta}_{A|B}(S) = 0$ for every $S \subseteq X - \{x\}$, $|S| \leq 2k$. In which:

$$w'_{A|B} = \begin{cases} w(Ax|B) & Ax|B \in S, A|Bx \notin S \\ w(A|Bx) & A|Bx \in S, Ax|B \notin S \\ w(Ax|B) + w(A|Bx) & Ax|B, A|Bx \in S \end{cases}$$

Since $S|_{X-\{x\}}$ also have the same obstruction set, from induction hypothesis we have $\boldsymbol{\delta}_{A|B}$ being linearly independent, thus all $w'_{A|B} = 0$. If we have one $A|B \in S$ such that $w(A|B)$ being non-zero, all $A - x|Bx \in S$ for any $x \in A$. Thus S would contain all the possible splits, this is in contradiction with the obstruction set for $n \geq 2k+2$.

If $n = 2k + 1$, for any split $A|B = \{x_1, \ldots, x_i\}|\{x_{i+1}, \ldots, x_n\}$ we have

$$\begin{aligned} w(A|B) &= w(\{x_1, \ldots, x_i\}|\{x_{i+1}, \ldots, x_n\}) \\ &= (-1)w(\{x_1, \ldots, x_{i-1}\}|\{x_i, \ldots, x_n\}) \\ &\ldots \\ &= (-1)^{i-1}w(\{x_1\}|\{x_2, \ldots, x_n\}) \\ &= (-1)^{i}w(\{x_1, x_n\}|\{x_2, \ldots, x_{n-1}\}) \\ &= (-1)^{i+1}w(\{x_n\}|\{x_1, \ldots, x_{n-1}\}) \\ &\ldots \\ &= (-1)^{n}w(\{x_{i+1}, \ldots, x_n\}|\{x_1, \ldots, x_i\}) \\ &= -w(B|A) \end{aligned}$$

hence $w(A|B) = 0$. □

Based on the idea we would introduce a new method to look at the compatible split system. The obstruction for compatible split system is $12|34, 13|24$. So for any four element $a, b, c, d \in X$, if quartet $ab|cd$ is represented we have $d(a, c) + d(b, d) = d(b, c) + d(a, d)$. Thus a quartet system would translate into a set of linear equations.

Lemma 5.

$$rank \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} \geq rank \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}$$

Proposition 7. *Given a quartet system on set X, the corresponding linear equations have null space dimension at most $2n - 3$.*

Proof: This condition is equivalent with that there exist $\binom{n-2}{2}$ linear independent ones from all $\binom{n}{4}$ linear relations since $\binom{n-2}{2} = \binom{n}{2} - (2n - 3)$. We will do this by induction, fix the ground set to be $\{1, \ldots, n\}$. Consider the quartet like $1a|bc$. By lemma 5 We only look at variables like $d(1, a)$. Hence this kind of linear relation would be $d(1, b) = d(1, c)$. If such relation for b and c holds we connect an edge with b and c. Thus for every triplet there is a pair connected so the graph has at most 2 connected components. Mind that each component have same value $d(1, a)$. Thus the null space for such equations is at most 2 and the rank of these linear relation ship is at least $n - 3$. Thus the rank of the whole linear relation is no less than $(n - 3) + \cdots + 1 = \binom{n-2}{2}$. □

This argument recover the fact that compatible split system has cardinality no more than $2n-3$. Such approach may lead to a new scheme for generating a bound for cardinality of general split system with certain forbidden configuration.

Firstly notice that any vanishing partial split would translate to a linear combination of diversities[101]:

$$w(U|Y-U) = \frac{1}{2}\sum_{Y\subseteq V\subseteq U}(-1)^{|U|+|V|+1}\delta(V) = 0$$

And we can combine this with the fact that diversities of odd sets are determined by diversities of even sets:

$$\delta(V) = \sum_{U\subseteq V}(-1)^{|V|}\delta(V)$$

And from the linear independency we know that we only need to consider the diversities of set size no greater than the ground set of forbidden configuration and the cardinality of splits will not exceed the dimension of the solution space of those linear equations. Thus if we want to estimate the maximal size of certain forbidden configuration, we can start from estimating the upper bound of such solution space.

For convenience we write out some terms in weight of a partial split explicitly:

Lemma 6. *A and B be two set that have same cardinality $k > 2$. $A \cup B = \{x, y\} \cup X$. Thus $w(A|B) = \delta(\{x, y\} \cup X) + c\delta(X) +$ diversities of set not containing X. in which $c = 0$ if $xy|Y$ is compatible with $A|B$, $c = -1/2$ if not.*

hence we have the following results:

Proposition 8. *If a split system S on a ground set $X = \{1, 2, \ldots, n\}$ has VC dimension 5 and for every $i, j, k, l \in \{3, \ldots, n\}$. One of partial splits in $\{12i|jkl, 12j|ikl, 12k|ijl, 12l|ijk\}$ and one of $\{1ij|2kl, 1ik|2jl, \ldots, 1kl|2ij\}$ not displayed. Then the cardinality of S cannot exceed $\frac{1}{3}n^3 + O(n^2)$.*

Proof: As we've explained, a set of not displayed partial splits can be translated into a set of linear equations of diversities. We'll pick some vanishing partial split: one of partial split for every 6-set not containing $\{1, 2\}$ and for every $\{1, 2, i, j, k, l\}$,

we pick one of partial splits in $\{12i|jkl, 12j|ikl, 12k|ijl, 12l|ijk\}$ and one of $\{1ij|2kl, 1ik|2jl, \ldots, 1kl|2ij\}$. Next we'll show that they are linearly independent.

For every partial split $A'|B'$ its weight can be expressed as linear combinations of diversities of subsets of $A' \cup B'$, if there is c_i which $\sum c_i * w(A'_i|B'_i)$ always be zero. For every 6-set not containing $\{1, 2\}$, it is the only equation that contains the diversities of the set, thus $c_i = 0$ for them. The only sets like $\{1, 2, i, j, k, l\}$ remains, for vanishing splits like $1ij|2kl$, its corresponding linear equation is the only equation containing $\delta(i, j, k, l)$, hence $c_i = 0$ for them. And for splits like $12i|jkl$ also be only one containing $\delta(1, 2, i, j, k, l)$ hence also have vanishing coefficient.thus $c_i = 0$ for them. This shows that they're linearly independent, thus the cardinality of S cannot exceed $\binom{n}{6} + \binom{n}{4} + \binom{n}{2} - \binom{n}{6} - \binom{n-2}{4} = \binom{n-1}{3} + \binom{n-2}{3} + \binom{n}{2}$.

□

5.1.2 Clusters

An cluster system on X is a set of subsets of X. From a split system S and an "infinite" point x in ground set X we can induce a cluster system on ground set $X - x$:$\{A : A|B \in S \,\text{and}\, x \in B\}$. For cluster system, we could similarly define the forbidden configuration condition.

Definition 23. *For a mapping $f : X \to Y$ and a cluster system C on Y. We define its pullback $f^\star C$ being a split system on X: $f^\star C = \{f^{-1}(A) : A \in C\}$. If C is weighted, we use this formula to decide the weight of the pullback: $w(A') = \sum_{f^{-1}(A')=A} w(A)$. For $X \subset Y$ we define the cluster system restricted on X as the pullback of inclusion map, denoted as $S|_Y$. A forbidden configuration is a cluster system on $\{1, \ldots, n\}$. And we say a cluster system C is with (or have) forbidden configuration iff there are no mapping $f : \{1, \ldots, n\} \to X$ such that the forbidden configuration is subset of $f^\star C$. We denote the forbidden configuration of all non-empty subset of $\{1, \ldots, i\}$ to be $\bar{F}_i$.*

Here by we introduce some construction, given an rooted tree, induced cluster system is defined as sets of descendants of vertices; given a linear ordering $[x_1, \ldots, x_n]$, induced cluster system were defined as: $\{\{x_1\}, \{x_1, x_2\}, \ldots\}$.

Definition 24. *For a cluster system C on X and $x \in X$ we define $\partial_x C$ being the set $\{A : A \cup \{x\}$ and $A \in C\}$*

Proposition 9. *If C is a cluster system on X and $x \in X$, we have: $|C| = |C|_{X-\{x\}}| + |\partial_x C|$.*

We can define similarity map induced by a weighted cluster system, as the analogue of diversities in weighted split system case:

$$s(A) = w(x|A) = \delta(x \cup A) - \delta(A) = \sum_{A \subseteq B} w(B)$$

And the inverse:

$$w(B) = \sum_{B \subseteq A} (-1)^{|A|+|B|} s(A) \tag{5.1}$$

And for convenience we define weights for partial splits:

$$w(\bar{A}B) = \sum_{A \cap C = \emptyset \; B \subseteq C} w(C)$$

Definition 25. *$s^{(k)}$ being the vector $(s_A : A \subset X, |A| \leq k)$. $s_A^{(k)}$ being the vector of similarities induced by a single cluster A with weight 1.*

We state the following theorem without proof, which is similar with Theorem. 6

.

Theorem 18. *For a cluster system C having forbidden configuration $\bar{F}_{k+1}$, vector set $\{s_A^{(k)} : A \in C\}$ is linearly independent.*

The similarity terminology gives us a characterization of (p, q)-hierarchy. The forbidden configuration of (p, q)-hierarchy is defined as such:

$$\begin{aligned} &2, \ldots, p+q+2 \\ &1, 3, \ldots, p+q+2 \\ &\ldots \\ &1. \ldots, p \hat{+} 1, \ldots, p+q+2 \end{aligned}$$

namely its restriction on any $p+q+2$-subset has at most p subset of size $p+q+1$.

Proposition 10. *For any $p+q+2$ elements $\{1, \ldots, p+q+2\}$ we have at least $p+2$ in all $p+q+2$ the similarities: $s(1, \ldots, \hat{i}, \ldots, p+q+2)$ are equal.*

Proof: for i being $1, \ldots, p+q+2$ at least $p+2$ of the following equations holds: $s(1, \ldots, p+q+2) = s(1, \ldots, \hat{i}, \ldots, p+q+2)$. Thus at least $p+2$ in all $p+q+2$ the similarities: $s(1, \ldots, \hat{i}, \ldots, p+q+2)$ are equal. □

Notice that the weight of all clusters can be determined by all similarities and for (p,q)-hierarchies we have $s(A) = \min_{B \subseteq A, |B|=p+q+1} s(B)$. Thus the similarities of all subsets is determined by similarities of all its $(p+q+1)$-subsets. So we have:

Definition 26. *Consider a graph on vertex set being the set of all $p+q+1$ subsets of an n-set. So for every $p+q+2$ elements in the n-set, there would be $p+q+2$ vertices that corresponds to its $p+q+1$-subset. If in every such set of vertices there is $p+2$ being connected, we call the graph is a (p,q)-connected graph of n-set.*

Proposition 11. *If a (p,q)-connected graph of n-set has at most m connected components. The cardinality of (p,q)-hierarchy on n-set will not exceed $m + \Phi_{p+q-1}(n+1)$.*

5.1.3 Finite closure property

For completeness of text we gather some results on finite closure property, which means for certain split system every split is the only one that extending a partial split with bounded cardinality, which is helpful in calculating weights.

Theorem 19. *[33] If a class of split system F have forbidden configuration like:*

$$1 \ldots (n-1)n | (n+1) \ldots 2n$$
$$1 \ldots (n-1)(n+1) | n(n+2) \ldots 2n$$
$$\ldots$$
$$2 \ldots (n+1) | 1(n+2) \ldots 2n$$

for all $S \in F$ and $s \in S$ there is a partial split $A'|B'$ such that s is the only split in S extending the partial split and $|A'| = |B'| = n$. Split system with such forbidden configuration is called $(n-1)$-weakly compatible split system.

The maximal cardinality of 2-weakly compatible split system is intensively discussed in section 5.3.2. For cluster the we have a simpler condition, we have

Theorem 20. *If a cluster system C is $(-1,n)$-hierarchy for all $c \in C$ there is a set U of cardinality less than k such that $U \subset c$ and every $U \subset c'$, $c' \in C$ we have $c \subseteq c'$.*

The 2-circular split system do not have finite closure property. Because consider a split $\{x_i, \ldots, x_j | x_1, \ldots, x_{i-1}, x_{j+1}, \ldots, x_n\}$ every partial split represented by it can be extended to another split.

5.2 Linearly independency over $\mathbb{Z}$

In this section we manage to generalize the results in [56]. The basic settings were as follows: for a split system S on ground set X, one could assign weight for each split. A weighted split system induces diversities of subsets[101], here we only focus on 2- and 4-diversities, as a comparison the original research were on metric, i.e. 2-diversities [56]. We further restrict the weights to take value in integers rather than real numbers. We will heavily use notions from lattice theory, for a general discussion, see [103]. Thus consider the mapping $f : \mathbb{Z}^S \to \mathbb{Z}^{\binom{n}{2}+\binom{n}{4}}$ defined by:

$$\delta(U) = \sum_{A|B \in S, U\cap A \neq \emptyset, U \cap B \neq \emptyset} w(A|B)$$

In which S is a split system, U is 2 or 4 subset of X, elements in $\mathbb{Z}^S$ we denoted as $(\ldots, w(s_i), \ldots)$ and elements in $\mathbb{Z}^{\binom{n}{2}+\binom{n}{4}}$ we denoted as $(\ldots, \delta(U), \ldots)$. We denote the vector of diversities of those sets $\boldsymbol{\delta}^4 = (\delta(a,b), \ldots, \delta(a,b,c,d), \ldots)$, $\boldsymbol{\delta}^4(A|B)$ is the diversities induced by a single split $A|B$ with weight 1, and ${\boldsymbol{\delta}^4}_U \doteq \delta(U)$. We define $\boldsymbol{\delta}^4(A)$ be the vector with component $\boldsymbol{\delta}^4(A)_A = 1$ and other components being 0.

Definition 27. *Define cut 2-lattice to be the lattice points satisfying*

$$\delta(a,b) + \delta(a,c) + \delta(b,c) \equiv 0 \pmod 2 \tag{5.2}$$

$$2(\delta(a,b,c,d) + \cdots + \delta(b,c,d,e)) \equiv \delta(a,b) + \cdots + \delta(d,e) \pmod 4 \tag{5.3}$$

denoted as Q_{cut}. $Q(S)$ being image of f, when S being the set of all possible split with ground set X, $Q(S)$ were denoted as Q'_{cut}. The sublattice of $\mathbb{Z}^{\binom{n}{2}+\binom{n}{4}}$ that $\delta(a,b) \equiv 0 \pmod 4$, $\delta(a,b,c,d) \equiv 0 \pmod 2$ were denoted as Q_2.

Theorem 21. $Q_2 \subset Q'_{cut} \subseteq Q_{cut}$

Proof: For convenience we introduce this shorthand notation: $\partial_x \boldsymbol{\delta}^4(A|B) \coloneqq \boldsymbol{\delta}^4(Ax|B) - \boldsymbol{\delta}^4(A|xB)$, such notation can be nested, like $\partial_x \partial_y \boldsymbol{\delta}^4(A|B) = \partial_x \boldsymbol{\delta}^4(Ay|B) -$

$\partial_x \boldsymbol{\delta}^4(A|yB) = \boldsymbol{\delta}^4(Axy|B) - \boldsymbol{\delta}^4(Ay|Bx) - \boldsymbol{\delta}^4(Ax|yB) + \boldsymbol{\delta}^4(A|xyB)$. Thus we have:

$$\partial_x \boldsymbol{\delta}^4(A|B)_U = \begin{cases} -1, & \text{if } x \in U \text{ and } U - x \subseteq A \\ 1, & \text{if } x \in U \text{ and } U - x \subseteq B \\ 0, & \text{other case} \end{cases} \tag{5.4}$$

$$\partial_x \partial_y \boldsymbol{\delta}^4(A|B)_U = \begin{cases} -1, & \text{if } \{x, y\} \subset U \text{ and } U - x - y \subseteq A \\ 1, & \text{if } \{x, y\} \subset U \text{ and } U - x - y \subseteq B \\ 2, & \text{if } U = \{x, y\} \\ 0, & \text{other case} \end{cases} \tag{5.5}$$

$$\partial_x \partial_y \partial_z \boldsymbol{\delta}^4(A|B)_U = \begin{cases} -1, & \text{if } \{x, y, z\} \subset U \text{ and } U - x - y - z \subseteq A \\ 1, & \text{if } \{x, y, z\} \subset U \text{ and } U - x - y - z \subseteq B \\ 0, & \text{other case} \end{cases} \tag{5.6}$$

$$\partial_x \partial_y \partial_z \partial_w \boldsymbol{\delta}^4(A|B)_U = 2\boldsymbol{\delta}(\{x, y, z, w\}) \tag{5.7}$$

For showing $Q_2 \subset Q'_{cut}$, we need to prove $2\boldsymbol{\delta}^4(x, y, z, w)$ and $4\boldsymbol{\delta}^4(x, y) \in Q'_{cut}$, by 5.7 we know $2\boldsymbol{\delta}^4(x, y, z, w) \in Q'_{cut}$, and consider $2\partial_x \partial_y \boldsymbol{\delta}^4(A|B)$, the only non-zero 2-diversity is $4\boldsymbol{\delta}^4(x, y)$ and all coefficients of 4-diversity is even thus can be eliminated, hence $4\boldsymbol{\delta}^4(x, y) \in Q'_{cut}$.

For the $Q'_{cut} \subseteq Q_{cut}$ part. Note that $\delta(a, b, c) = \frac{1}{2}(\delta(a, b) + \delta(a, c) + \delta(b, c))$ and $\delta(a, b, c, d, e) = \frac{1}{2}(\delta(a, b, c, d) + \cdots + \delta(b, c, d, e) - \frac{1}{2}(\delta(a, b) + \cdots + \delta(d, e)))$, thus condition 5.2 is equivalent with $\delta(a, b, c) \in \mathbb{Z}$ and condition 5.3 is equivalent with $\delta(a, b, c, d, e) \in \mathbb{Z}$. We will proof that $Q'_{cut} = Q_{cut}$ later. □

Theorem 22.

$$\det(Q_{cut}) = 2^{\binom{n-1}{2} + \binom{n-1}{4}}$$

Proof: We do this by counting the cardinality of Q_{cut}/Q_2. Consider the projection $p : Q_{cut}/Q_2 \to \mathbb{Z}^{\binom{n}{2}}$, $(\delta(a, b), \ldots, \delta(a, b, c, d), \ldots) \mapsto (\delta(a, b), \ldots)$. The image of p is exactly the cut lattice, which is the elements in $\mathbb{Z}_4^{\binom{n}{2}}$ satisfying 5.2. Hence have cardinality $4^{\binom{n}{2}}/2^{\binom{n-1}{2}} = 2^{\binom{n}{2} + (n-1)}$.

Then we consider $\ker p$, which is the vector $(0, 0, \ldots, \delta(a,b,c,d), \ldots)$, with $\delta(a,b,c,d) \in \mathbb{Z}_2$, $\delta(a,b,c,d)+\cdots+\delta(b,c,d,e) = 0$ for every $a,b,c,d,e \in X$. Fix an element $z \in X$, we have $\delta(a,b,c,d) = \delta(z,a,b,c)+\delta(z,a,c,d)+\delta(z,a,b,d)+\delta(z,b,c,d)$. And $\delta(z,a,b,d)$ is linearly independent variables, for $\delta(a,b,c,d)+\cdots+\delta(b,c,d,e) = 2(\delta(z,a,b,c)+\cdots+\delta(z,c,d,e)) = 0$. Hence $\ker p = 2^{\binom{n-1}{3}}$. In all we have $|Q_{cut}/Q_2| = |\ker p| \times |\text{image } p| = 2^{\binom{n-1}{3}+\binom{n}{2}+(n-1)}$. Thus $\det(Q_{cut}) = \det(Q_2)/|Q_{cut}/Q_2| = 2^{\binom{n-1}{2}+\binom{n-1}{4}}$.

□

Corollary 3. *This following complex of $\mathbb{Z}_2$ module is exact:*

$$\ldots \to C_{i+1} \xrightarrow{\partial} C_i \to \ldots$$

C_i is generated by $\delta(X_i)$, X_i is a subset of X with cardinality i and the ∂ operator is defined by

$$\partial_i : \delta(\{x_1, \ldots, x_i\}) \mapsto \sum_{j=1}^{i} \delta(\{x_1, \ldots, \hat{x}_j, \ldots, x_i\})$$

Theorem 23. *Fix an element $z \in X$, we could define the 2-Farris transform from diversities to similarities $f_z : \boldsymbol{\delta}^4 \to \boldsymbol{s}^4$, s is the vector $(s(a), \ldots, s(a,b), \ldots, s(a,b,c), \ldots, s(a,b,c,d), \ldots)$ in which $a,b,c,d \in X - \{z\}$, and f_z is defined by:*

$$\begin{aligned}
s(a) &= \delta(z,a) \\
s(a,b) &= \frac{1}{2}(\delta(z,a) + \delta(z,b) - \delta(a,b)) \\
s(a,b,c) &= \delta(z,a,b,c) \\
s(a,b,c,d) &= \frac{1}{2}(\delta(z,a,b,c) + \cdots - \delta(a,b,c,d) - \frac{1}{2}(\delta(z,a) + \cdots + \delta(a,b) + \ldots))
\end{aligned}$$

In other word $s(A) = w(z|A)$. We have Image$(f_z) = \mathbb{Z}^{\binom{n}{2}+\binom{n}{4}}$.

Proof: Firstly Image$(f_z) \subseteq \mathbb{Z}^{\binom{n}{2}+\binom{n}{4}}$ since $s(A) = w(z|A)$. Secondly $\det f_z = 2^{\binom{n-1}{2}+\binom{n-1}{4}}$ following the same argument in [56]. Thus determinant of Image (f_z) over $\mathbb{Z}^{\binom{n}{2}+\binom{n}{4}}$ is 1. Hence Image$(f_z) = \mathbb{Z}^{\binom{n}{2}+\binom{n}{4}}$. □

Then we consider the 4-affine split system, for which the $\boldsymbol{\delta}(s_i)$ is linear independent. We will show that this generate the cut 2-lattice.

Theorem 24. *If $S = \{s_i\}$ is a 4-affine split system, $\boldsymbol{\delta}(s_i)$ generates the cut 2-lattice.*

Proof: Denote the lattice generated by $\boldsymbol{\delta}(s_i)$ as Q_S. Again consider the projection p into the 2-diversity components. We start from showing Image(p) is the cut lattice, for points $x_i \in \mathbb{R}^4$, we can project them into a 2-dim plane. Every line in that plane would lift to a 3-dim hyperplane in $\mathbb{R}^4$. Thus every 4-affine split system must contain a 2-affine split system as subset, hence the cut lattice must be a subset of Image(p), remind that we have showed that Image(p) is a subset of cut lattice. This proofs that Image(p) is the cut lattice. Then we could consider the ker(p), as we have explicitly described before, it's a lattice generated by $2\boldsymbol{\delta}(x,y,z,w)$ and $\sum_{i\neq x,y,z}\boldsymbol{\delta}(x,y,z,i)$. The 4-affine split system have the 4 separation property, namely for every 4 element x,y,z,w, there is a hyperplane pass through this 4 points and under perturbation we have there exist $A, B \subset X$ such that $A\cup P|B\cup Q \in S$ for every $P \subseteq \{x,y,z,w\}\, Q = \{x,y,z,w\} - P$. Thus $2\boldsymbol{\delta}(x,y,z,w) = \partial_x\partial_y\partial_z\partial_w\boldsymbol{\delta}^4(A|B) \in Q_S$. And $\sum_{i\neq x,y,z}\boldsymbol{\delta}(x,y,z,i) = \partial_x\partial_y\partial_z\boldsymbol{\delta}^4(A|B)+2\sum_{i\in A}2\boldsymbol{\delta}(x,y,z,i)$. This completes the whole proof. □

Here we consider the split system $S = \{s_i\}$ which consists of all 2-splits and 4-splits on X, for which $\#X = n > 6$ and $n \neq 8$. We will show that $\boldsymbol{\delta}(s_i)$ is linearly independent and they do not generate the cut 2-lattice. We will explicitly calculate its determinant, here we use a argument that different with [56] since 2−metric case is more complicated, note that the results in [56] can be recovered using this method.

Label split system S by 2 or 4-subset of X, this is possible since $n > 6$ and $n \neq 8$, then we consider could the linear mapping from the space of split weights to diversities as endomorphism f on vector space $V = \mathbb{R}^{\binom{n}{2}+\binom{n}{4}}$, its elements were written as $\boldsymbol{v} = (v_{12},\dots,v_{ij},\dots,v_{1234},\dots,v_{ijkl},\dots)$ with $1 \leq i < j < k < l \leq n$. V is endowed with nature inner product, we will show that even f is not self-adjoint, its eigenspaces are almost orthogonal.

We define several subspaces of V:

Definition 28. *We consider linear space L_1', L_2', L_3', L_1, L_2, L_3, L_4, L_5. In L_i' the components v_{ijkl} always vanishes and in L_i the components v_{ij} always vanishes. L_1' is the space when there exist c' such that $v_{ij} = c'$ for all i,j; L_2' is the space when there exist c_i' with $\sum_i c_i' = 0$ such that $v_{ij} = c_i' + c_j'$ for all i,j; L_3' is the space when there exist c_{ij}' with $\sum_i c_{ij}' = 0$ for ever j such that $v_{ij} = c_{ij}'$ for all i,j. L_1 is the space when there exist c such that $v_{ijkl} = c$ for all i,j,k,l; L_2 is the space when there exist c_i with $\sum_i c_i = 0$ such that $v_{ijkl} = c_i+c_j+c_k+c_l$ for all i,j,k,l; L_3 is the space when*

there exist c_{ij} *with* $\sum_i c_{ij} = 0$ *for all* j *such that* $v_{ijkl} = c_{ij} + c_{ik} + c_{il} + c_{jk} + c_{jl} + c_{kl}$ *for all* i, j, k, l*;* L_4 *is the space when there exist* c_{ijk} *with* $\sum_i c_{ijk} = 0$ *for all* j, k *such that* $v_{ijkl} = c_{ijk} + c_{ijl} + c_{ikl} + c_{jkl}$ *for all* i, j, k, l*;* L_5 *is the space when there exist* c_{ijkl} *with* $\sum_i c_{ijkl} = 0$ *for all* j, k, l *such that* $v_{ijkl} = c_{ijkl}$ *for all* i, j, k, l*.*

Lemma 7. *Consider mapping* $f : \mathbb{R}^{\binom{X}{p}} \to \mathbb{R}^{\binom{X}{q}}$, $(\dots, v_S, \dots) \mapsto (\dots, v_T, \dots)$ *for* S *being* p *subset and* T *being* q *subset,* $v_T = \sum_{S \subset T} v_S$*.* f *is injective when* $\#X > p + q$ *and* $p < q$*.*

Proof: We exemplify by case when $p = 3$. We need to prove $\ker(f) = 0$, Namely if for every q subset T $\sum_{i,j,k \in T} v_{ijk} = 0$ then $v_{ijk} = 0$. Denote $\#X = n$

1. $\sum_T \sum_{i,j,k \in T} v_{ijk} = \binom{n-3}{q-3} \sum_{i,j,k} v_{ijk}$ hence $\sum_{i,j,k} v_{ijk} = 0$.
2. fix i', $\sum_{i' \in T} \sum_{i,j,k \in T} v_{ijk} = \binom{n-3}{q-3} \sum_{j,k} v_{i'jk} + \binom{n-4}{q-4} \sum_{i,j,k \neq i'} v_{ijk} = (\binom{n-3}{q-3} - \binom{n-4}{q-4}) \sum_{j,k} v_{i'jk}$. $\binom{n-3}{q-3} - \binom{n-4}{q-4} \neq \binom{n-4}{q-3} = 0$, thus for every i', $\sum_{j,k} v_{i'jk} = 0$.
3. fix i', j', $\sum_{i',j' \in T} \sum_{i,j,k \in T} v_{ijk} = \binom{n-3}{q-3} \sum_k v_{i'j'k}$ $+ \binom{n-4}{q-4} (\sum_{i,j,k \neq j'} v_{i'jk} + \sum_{i,j,k \neq i'} v_{j'jk}) + \binom{n-5}{q-5} \sum_{i,j,k \neq i',j'} v_{ijk}$. The coefficient before $\sum_k v_{i'j'k}$ is $\binom{n-3}{q-3} - 2\binom{n-4}{q-4} + \binom{n-5}{q-5} = \binom{n-5}{q-3}$, which must be non-zero since $n > q + 3$ thus $\sum_k v_{i'j'k} = 0$.
4. fix i', j', k' Consider $\sum_{i',j',k' \in T} \sum_{i,j,k \in T} v_{ijk}$. With our experience before we know that the coefficient before $v_{i'j'k'}$ is $\binom{n-3}{q-3} - 3\binom{n-4}{q-4} + 3\binom{n-5}{q-5} - \binom{n-6}{q-6} = \binom{n-6}{q-3}$, which must be non-zero since $n > q + 3$ thus $v_{i'j'k'} = 0$.

For the general p case we could follow this procedure, note that for step i, the coefficient is $\binom{n-p-i+1}{q-p}$, which is nonzero for $1 \leq i \leq p+1$ with $n \geq p+q$ and $p \leq q$. This completes the whole proof.

□

Corollary 4. *Dimension of* L'_i *and* L_i *are all* $\binom{n}{i-1} - \binom{n}{i-2}$ *and* $V = L'_1 \oplus L_1 \oplus L'_2 \oplus L_2 \oplus L'_3 \oplus L_3 \oplus L_4 \oplus L_5$

Proof: Using lemma 6 we know the mapping from c or c' to v is injective and the constraints on c or c' are independent. □

Then we come to the action of f on those subspaces, we will show that $L'_1 \oplus L_1$, $L'_2 \oplus L_2$, $L'_3 \oplus L_3$, L_4, L_5 is invariant and we write out the action explicitly, for convenient we denote the image of f by δ_{ij} or δ_{ijkl} as before.

1. on $L_1' \oplus L_1$: $\delta_{ij} = 2(n-2)c' + 4(n-4)c$ and $\delta_{ijkl} = (2\binom{n-2}{3} + \binom{n-2}{2})c' + (4\binom{n-4}{3} + 6\binom{n-4}{2} + 4(n-4))c$.

2. on $L_2' \oplus L_2$:

$$\begin{aligned}
\delta_{ij} &= \sum_{k\neq j} v_{ik} + \sum_{k\neq i} v_{jk} + \sum_{k,l,m\neq j} v_{iklm} + \sum_{k,l,m\neq i} v_{jklm} && (5.8)\\
&= (n-2)(c_i' + c_j') + 2\sum_{k\neq i,j} c_k' + \binom{n-2}{3}(c_i + c_j) + 2\binom{n-3}{2}\sum_{k\neq i,j} c_k && (5.9)\\
&= (n-4)(c_i' + c_j') + (\binom{n-2}{3} - 2\binom{n-3}{2})(c_i + c_j) && (5.10)
\end{aligned}$$

$$\begin{aligned}
\delta_{ijkl} &= v_{ij} + \cdots + v_{kl} + \sum_{m\neq i,j,k,l} (v_{mi} + \cdots + v_{ml}) && (5.11)\\
&\quad + \sum_{m\neq i,j,k,l} (v_{mijk} + \cdots + v_{mjkl}) && (5.12)\\
&\quad + \sum_{m,n\neq i,j,k,l} (v_{mnij} + \cdots + v_{mnkl}) && (5.13)\\
&\quad + \sum_{m,n,o\neq i,j,k,l} (v_{mnoi} + \cdots + v_{mnol}) && (5.14)\\
&= 3(c_i' + \cdots + c_l') + (n-4)(c_i' + \cdots + c_l') + 4\sum_{m\neq i,j,k,l} c_m' && (5.15)\\
&\quad + \sum_{m\neq i,j,k,l} (4c_m + 3(c_i + \cdots + c_l)) && (5.16)\\
&\quad + \sum_{m,n\neq i,j,k,l} (6c_m + 6c_n + 3(c_i + \cdots + c_l)) && (5.17)\\
&\quad + \sum_{m,n,o\neq i,j,k,l} (4c_m + 4c_n + 4c_o + (c_i + \cdots + c_l)) && (5.18)\\
&= (n-5)(c_i' + \cdots + c_l') && (5.19)\\
&\quad + ((3(n-4) - 4) + (3\binom{n-4}{2} - 6(n-5)) && (5.20)\\
&\quad + (\binom{n-4}{3} - 4\binom{n-5}{2}))(c_i + \cdots + c_l) && (5.21)\\
&= (n-5)(c_i' + \cdots + c_l') && (5.22)\\
&\quad + 1/6(n^3 - 18n^2 + 107n - 216)(c_i + \cdots + c_l) && (5.23)
\end{aligned}$$

3. on $L_3' \oplus L_3$:

$$\begin{aligned}
\delta_{ij} &= \sum_{k\neq j} v_{ik} + \sum_{k\neq i} v_{jk} + \sum_{k,l,m\neq j} v_{iklm} + \sum_{k,l,m\neq i} v_{jklm} && (5.24)\\
&= -2c_{ij}' + \binom{n-3}{2}\sum_{k\neq i,j} (c_{ik} + c_{jk}) + 2(n-4)\sum_{k,l\neq i,j} c_{kl} && (5.25)\\
&= -2c_{ij}' - (n-4)(n-5)c_{ij} && (5.26)
\end{aligned}$$

$$\begin{aligned}
\delta_{ijkl} &= v_{ij} + \cdots + v_{kl} + \sum_{m\neq i,j,k,l} (v_{mi} + \cdots + v_{ml}) && (5.27)\\
&\quad + \sum_{m\neq i,j,k,l} (v_{mijk} + \cdots + v_{mjkl}) && (5.28)\\
&\quad + \sum_{m,n\neq i,j,k,l} (v_{mnij} + \cdots + v_{mnkl}) && (5.29)\\
&\quad + \sum_{m,n,o\neq i,j,k,l} (v_{mnoi} + \cdots + v_{mnol}) && (5.30)\\
&= c'_{ij} + \cdots + c'_{kl} - 2(c'_{ij} + \cdots + c'_{kl}) - v_{ijkl} - \sum_{m,n,o,p\neq i,j,k,l} v_{mnop} && (5.31)\\
&= -(c'_{ij} + \cdots + c'_{kl}) - (c_{ij} + \cdots + c_{kl}) - \binom{n-6}{2} \sum_{m,n\neq i,j,k,l} c_{mn} && (5.32)\\
&= -(c'_{ij} + \cdots + c'_{kl}) - (\binom{n-6}{2} + 1)(c_{ij} + \cdots + c_{kl}) && (5.33)
\end{aligned}$$

4. on L_4: based on symmetry δ_{ij} is always zero.

$$\begin{aligned}
\delta_{ijkl} &= \sum_{m\neq i,j,k,l} (v_{mijk} + \cdots + v_{mjkl}) + \sum_{m\neq i,j,k,l} (v_{mijk} && (5.34)\\
&\quad + \cdots + v_{mjkl}) && (5.35)\\
&\quad + \sum_{m,n\neq i,j,k,l} (v_{mnij} + \cdots + v_{mnkl}) && (5.36)\\
&\quad + \sum_{m,n,o\neq i,j,k,l} (v_{mnoi} + \cdots + v_{mnol}) && (5.37)\\
&= -v_{ijkl} - \sum_{m,n,o,p\neq i,j,k,l} v_{mnop} && (5.38)\\
&= -(c_{ijk} + \cdots + c_{jkl}) - (n-7) \sum_{m,n,o\neq i,j,k,l} c_{mno} && (5.39)\\
&= (n-8)(c_{ijk} + \cdots + c_{jkl}) && (5.40)
\end{aligned}$$

5. on L_5: based on symmetry δ_{ij} is always zero.

$$\begin{aligned}
\delta_{ijkl} &= \sum_{m\neq i,j,k,l} (v_{mijk} + \cdots + v_{mjkl}) + \sum_{m,n\neq i,j,k,l} (v_{mnij} + \cdots + v_{mnkl}) && (5.41)\\
&\quad + \sum_{m,n,o\neq i,j,k,l} (v_{mnoi} + \cdots + v_{mnol}) && (5.42)\\
&= -4v_{ijkl} + 6v_{ijkl} - 4v_{ijkl} && (5.43)\\
&= -2v_{ijkl} && (5.44)
\end{aligned}$$

Thus

$$\det(f|_{L'_1 \oplus L_1}) = \begin{vmatrix} 2(n-2) & 2\binom{n-2}{3} + \binom{n-2}{2} \\ 4(n-4) & 4\binom{n-4}{3} + 6\binom{n-4}{2} + 4(n-4) \end{vmatrix} = -4/3(n-6)(n-4)(n-2)$$

$$\det(f|_{L_2'\oplus L_2}) = \begin{vmatrix} n-4 & \binom{n-2}{3} - 2\binom{n-3}{2} \\ n-5 & 1/6(n^3-18n^2+107n-216) \end{vmatrix}^{n-1} \tag{5.45}$$

$$= (-1/3(n-8)(n-6)(n-4))^{n-1} \tag{5.46}$$

$$\det(f|_{L_3'\oplus L_3}) = \begin{vmatrix} -2 & (n-4)(n-5) \\ 1 & -\binom{n-6}{2} - 1 \end{vmatrix}^{1/2n(n-3)} = (-4n+24)^{1/2n(n-3)}$$

$$\det(f|_{L_4}) = (n-8)^{1/6n(n-1)(n-5)}$$

and

$$\det(f|_{L_5}) = (-2)^{n(n-1)(n-2)(n-7)/24}$$

In all

$$\begin{aligned} \det(f) &= -4/3(n-6)(n-4)(n-2)(-1/3(n-8)(n-6)(n-4))^{n-1} \\ &\quad (-4n+24)^{1/2n(n-3)}(n-8)^{1/6n(n-1)(n-5)}(-2)^{n(n-1)(n-2)(n-7)/24} \\ &= (-1)^{n(n-1)(n^2-9n+26)/8}2^{(n-1)(n-2)(n^2-7n+24)/24}(1/3)^n(n-8)^{(n-1)(n-2)(n-3)/6} \\ &\quad (n-6)^{n(n-1)/2}(n-4)^n(n-2) \end{aligned}$$

In [56] the determinant of following matrix is calculated: M of size $\binom{n}{2} \times \binom{n}{2}$ and M_{ij} be 1 iff the intersection of corresponding 2-subset has size 1. Once think of action on $L_1' \oplus L_2' \oplus L_3'$, one could see that M has eigenvalue $2(n-2)$ with multiplicity 1, $n-4$ with multiplicity $n-1$ and -2 with multiplicity $1/2n(n-3)$. In all we have $\det(M) = -(-2)^{1/2(n-1)(n-2)}(n-4)^{n-1}(n-2)$

5.3 Upper bound for some split system and cluster system

In this section we proof some inequalities on the maximal cardinality of split system with certain forbidden configurations. It has been proved that cardinality of a split system or cluster system with forbidden configuration is bounded by polynomial. The main tool utilized were combinatoric methods for getting bounds for maximal cardinality.

5.3.1 (p, q)-hierarchy

(p, q)-hierarchy were firstly introduced in [51] by Dress. In this section we will give the order of cardinality of those cluster systems.

Definition 29. *Cluster system C on X is (p, q)-hierarchy if and only if its restriction on any $(p + q + 2)$-set have $(p + q + 1)$-set cardinality no more than q.*

For example, rooted tree is (0,1)-hierarchy, linear ordering is (-1,1)-hierarchy and weak hierarchy is (-1,2)-hierarchy. An useful construction is k-point cluster system: The base set X were divided into k parts and a set in cluster system would intersect any part with exactly one element.

Theorem 25. *A $(p + 1, q - 1)$-hierarchy is a (p, q)-hierarchy, a $(p - 1, q)$-hierarchy is a (p, q)-hierarchy.*

Theorem 26. *A (p, q)-hierarchy's maximal cardinality on n-set is $O(n^{p+q+1})$ if $q > 1$, $O(n^{p+q})$ if $q \leq 1$ and $p > -1$.*

Proof:

Firstly we have the closure argument: any set in $(-1, q)$-hierarchy is a closure of q elements, i.e. given a element X' in the cluster system we can always find q elements that any set in the cluster system containing this q element must contain X'. In this way we make a injective mapping from the cluster system to all q-subsets. Thus we have $\#\mathcal{C} < O(n^{p+q+1})$. And notice that the $(p+q+1)$-point cluster system is already $(p + q - 2, 2)$-hierarchy, thus (p, q)-hierarchy, this completes the proof of $q > 1$ case.

For $q = 0$ the condition of being (p, q)-hierarchy is just the size of set in the cluster system can not exceed $p + 1$. The only case left is $q = 1$ case. In this case any set that contain a certain p-subset must be linearly ordered. we have the cardinality of m-set in the cluster system cannot exceed $\binom{n}{p}/\binom{m}{p}$. Since $\sum_{m=p}^{\infty} 1/\binom{m}{p}$ is bounded this complete the whole proof.

□

We will focus on a certain type of hierarchy: $(-1, 3)$-hierarchy, which the simplest hierarchy that we don't know the maximal cardinality. We can explicitly write out it's forbidden configuration: 123,124,134,234. The upper bound from closure

argument is $O(1/6 * n^3)$. Here we present a construction that have cardinality $O(1/9 * n^3)$.

Lemma 8. *Given n point in $\mathbb{R}^3$ with coordinates (x_i, y_i, z_i), we can construct a $(-1, 3)$-hierarchy from it: $\{c| \exists (x, y, z) \in \mathbb{R}^3\,, c = \{i|x_i < x,\, y_i < y,\, z_i < z\}\}$.*

Thus we can consider such geometric configuration: n point distributed uniformly on polyline connecting $(0, 1, 0)$,$(0.5, 0.5, 0)$ and $(1, 0, 0)$. This gives a cluster system with cardinality $1/12 * n^3 + O(n^2)$. And for each part of the line we can do the similar thing and finally have a fractal-like structure. This gives a cluster system with cardinality $1/9 * n^3 + O(n^2)$.

Here we gives a more detailed explanation: our construction take a linear ordering $l = [x_1, \ldots, x_n]$ as input and output a cluster system C, such construction is defined recursively. For $n = 1$, $C(l) = \{\{x_1\}\}$ and for $n > 1$, $C(l) = C([x_1, \ldots, x_{\lfloor \frac{n}{2} \rfloor}]) \cup C([x_n, x_{n-1} \ldots, x_{\lfloor \frac{n}{2} \rfloor + 1}]) \cup \{\{x_i, \ldots, x_j, x_{2\lfloor \frac{n}{2} \rfloor + 1 - j}, \ldots, x_k\}\}$.

For showing such cluster system is $(-1, 3)$-hierarchy, we need to show for every $a, b, c, d \in \{x_1, \ldots, x_n\}$ the restricted cluster system is $(-1, 3)$-hierarchy. For $a, b, c, d \leq \frac{n}{2}$, we first observe that $C(l)$ already includes all intervals $\{x_i, \ldots, x_j\}$ of l. $C(l)|_{\{a,b,c,d\}} = C(l)|_{\{x_1, \ldots, x_{\lfloor \frac{n}{2} \rfloor}\}}|_{\{a,b,c,d\}} = \{C([x_1, \ldots, x_{\lfloor \frac{n}{2} \rfloor}]) \cup \{\{x_i, \ldots, x_j\} | i < j \leq \lfloor \frac{n}{2} \rfloor\}\}|_{\{a,b,c,d\}} = \{C([x_1, \ldots, x_{\lfloor \frac{n}{2} \rfloor}])|_{\{a,b,c,d\}}$ hence $C(l)|_{\{a,b,c,d\}}$ is $(-1, 3)$-hierarchy. For $a < b < c \leq \frac{n}{2} < d$, $acd \notin C(l)|_{\{a,b,c,d\}}$. For $a < b \leq \frac{n}{2} < c < d$, restriction of $C([x_1, \ldots, x_{\lfloor \frac{n}{2} \rfloor}]) \cup C([x_n, x_{n-1} \ldots, x_{\lfloor \frac{n}{2} \rfloor + 1}])$ on $\{a, b, c, d\}$ is only consists of 1 or 2 set, by lemma. 6 we know that $C(l)|_{\{a,b,c,d\}}$ is $(-1, 3)$-hierarchy.

Denote $\#C(l)$ as $f(n)$. The cardinality of $\{\{x_i, \ldots, x_j, x_{2\lfloor \frac{n}{2} \rfloor + 1 - j}, \ldots, x_k\}\}$ is $\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} (i-1)^2 = 1/12 * n^3 + O(n^2)$. We have $f(n) = 2 * f(\lfloor \frac{n}{2} \rfloor) + 1/12 * n^3 + O(n^2)$, hence from arguments from asymptotic analysis we have $f(n) = 1/9 * n^3 + O(n^2)$.

5.3.2 2-weakly compatible split system

In [33] 2-weakly compatible condition was introduced, which is defined as split system with forbidden configuration $123|456, 124|356, 125|346, 126|345$, we were interested in the maximal cardinality of such split system. In the original paper an upper bound $3\binom{n}{4} + n$ was given, we have already give a better estimation $\binom{n}{4} + \binom{n}{2}$ since 2-weakly compatible is a condition stronger than 2-very-weakly compatible.

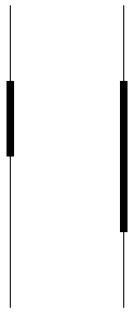


Figure 5.1: Demonstration of clusters in constructed hierarchy

We firstly manage to study the case of split system on small cardinality using computational approach. The problem of maximal cardinality can be formulated by following optimization problem:

$$\begin{array}{lccll} \max & \sum_{s_i \in \mathcal{S}} w_{s_i} & & & \\ \text{s.t.} & w_{s_i}, w_{p_j} & \in & \{0,1\} & \forall s_i \in \mathcal{S}, \forall p_j \in \mathcal{P}_{3|3} \\ & w_{abc|def} + w_{abd|cef} + w_{abe|cdf} + w_{abf|cde} & \leq & 3 & \forall a,b,c,d,e,f \in X \\ & w_{s_i} & \leq & w_{p_j} & \forall p_j \dashv s_i \end{array} \tag{5.47}$$

In which $X = \{1, \ldots, n\}$, $\mathcal{S}$ is all possible full splits on X (for convenience we could omit all 1- and 2-splits), $\mathcal{P}_{3|3}$ is all possible 3|3-partial splits on X. In all there is $2^{n-1} - \binom{n}{2} - \binom{n}{1} - 1 + \binom{n}{3}\binom{n-3}{3}$ variables and $\binom{n}{2}\binom{n-2}{4} + \binom{n}{3}\binom{n-3}{3} * 2^{n-6}$ constraints. Hereby I'll list some results from this approach.

Result 1. *There are essentially two non-isomorphic maximal 2-weakly compatible split system on* $X = \{1, \ldots, 6\}$*:* $S_1 = \{123|456, 124|356, 125|346, 134|256, 135|246, 145|236\}$ *and* $S_2 = \{123|456, 134|256, 145|236, 156|234, 126|345\}$ *(1- and 2-splits were omitted).*

Result 2. *Denote the maximal cardinality of 2-weakly compatible split system on* n *taxa by* a_n*. The optimization algorithm gives* $a_6 = 28$*,* $a_7 = 42$*,* $a_8 = 60$*,* $a_9 = 80$*.*

Theorem 27. *The maximal cardinality of 2-weakly compatible split system cannot exceed* $O(n^3)$

We will present two different proof of theorem 27, in next section we will give a better bound $O(n^{2.5})$, the aim of these two proof is to represent some fundamental techniques in study of combinatorics of split system, from now on we allows for split system to include the split $\emptyset|X$ for split system over X for convenience.

Definition 30. *$2'$-weakly compatible split system is split system with forbidden configurations* $\{12|345, 13|245, 14|235, 15|234\}$ *and* $\{123|45, 124|35, 125|34, 12|345\}$. *A split system is $2''$-weakly compatible iff with forbidden configurations* $\{1|234, 2|134, 3|124, 4|123\}$, $\{1|234, 12|34, 13|24, 14|23\}$ *and* $\{123|4, 124|3, 12|34\}$.

Such a definition derives from the fact that $\partial_x S$ is a $2'$-weakly compatible system iff S is 2-weakly compatible, and $\partial_x S$ is $2''$-compatible iff S is $2'$-weakly compatible.

Theorem 28. *For a $2''$-weakly compatible split system S on a set X with size n, $\#S \leq n$.*

Proof: For $n = 5$, one could verify that $\#S \leq 5$ by hand. For $n \geq 6$, we assume that $\#S' \leq m$ for every S' over set with cardinality $m < n$. Then take arbitrary $x \in X$, $\partial_x S$ is with forbidden configuration $\emptyset|123, 1|23$, thus we have $\#\partial_x S \leq 1$. Or take $A|B, C|D \in \partial_x S$, $A \cap C, A \cap D, B \cap C, B \cap D$ is a partition of $X - x$, by pigeonhole principle cardinality of one of those set has more than 2 elements. Suppose $1, 2 \in A \cap C$ we observe that $A \cap D$ or $B \cap C$ is not empty or $A|B = C|D$ thus we could further assume that $3 \in B \cap C$ this leads to contradiction because $\emptyset|123 \dashv C|D$ and $12|3 \dashv A|B$. Thus $\#S = \#S|_{X-x} + \#\partial_x S \leq n$.

□

Hence we have the maximal cardinality of $2'$-weakly compatible split system cannot exceed $\frac{1}{2}n^2 + O(n)$. Thus the maximal cardinality of 2-weakly compatible split system cannot exceed $\frac{1}{6}n^3 + O(n^2)$. Note that the bound of $2'$-compatible split system is tight because circular split system is $2'$-compatible and is of cardinality $\frac{1}{2}n^2 + O(n)$.

A more rough bound can be proved by a simpler argument.

Proof: Consider a split system, for a pair of element x, y we define $p(x, y)$ as the

number split that x, y belongs to one side. Thus we have:

$$\begin{aligned}
\sum_{x,y\in X} p(x,y) &= \sum_{A|B\in S} (\binom{\#A}{2} + \binom{\#B}{2}) && (5.48)\\
&= \sum_{A|B\in S} (\binom{\#A}{2} + \binom{n-\#A}{2}) && (5.49)\\
&= \sum_{A|B\in S} (\frac{1}{2}(\#A^2) + (n-\#A)^2) - n) && (5.50)\\
&\geq \sum_{A|B\in S} (\frac{1}{2}(2 * (\frac{n}{2})^2) - n) && (5.51)\\
&= \#S * (\frac{n^2}{4} - n) && (5.52)
\end{aligned}$$

Thus $\max_{x,y\in X} p(x,y) \leq \sum_{x,y\in X} p(x,y)/\binom{n}{2} \leq \#S\frac{n-4}{2n-2}$. Choose such x, y that maximize $p(x,y)$. Take the cluster system $C = \{A : A|B \in S, x, y \in B\}$. It's $(-1,3)$-hierarchy thus cardinality cannot exceed $\frac{1}{6}n^3 + O(n^2)$. Hence the cardinality of 2 compatible split system cannot exceed $(\frac{1}{6}n^3 + O(n^2)) * \frac{2n-2}{n-4} = \frac{1}{3}n^3 + O(n^2)$. □

We have proved that the maximal cardinality of 2-weakly compatible split system is between $3n^2/2 + O(n)$ and $O(n^3)$, there is no strong evidence supporting whether it's quadratic or not. To reveal the complexity of the maximal cardinality problem, we will prove the following result:

Theorem 29. *Split system with forbidden configuration* $12|345, 13|245, 14|235, 15|234$ *have maximal cardinality* $O(n^2)$ *and split system with forbidden configuration* $1234|567, 1235|467, 1236|457, 1237|456$ *have maximal cardinality* $O(n^3)$.

The first part is relatively hard, we begin by proving some lemma

Lemma 9. *If split system* S *on* $X = \{1, \ldots, n\}$ *with forbidden configuration* $1|23, 2|13, 3|12$ *has cardinality more than* 3 *(includes the* $\emptyset|X$ *split). Then there is a ordering* $[x_1, \ldots, x_n]$ *of* X *such that all* $s \in S$ *were in form of* $x_1, \ldots, x_i|x_{i+1}, \ldots, x_n$.

Proof: We firstly show that S is compatible, if not, we can find $s_1 = A_1|B_1, s_2 = A_2|B_2$ being non-compatible. Suppose $s_3 = A_3|B_3$, hereby we introduce notations like $U_{AAB} = A_1 \cap A_2 \cap B_3$ one could imagine a cube with each vertex stands for one such set and no equilateral triangle has all vertex being non-empty. We could suppose U_{AAA} is non-empty. If U_{ABB} is non-empty, then U_{BAB} should be empty then U_{BAA} should be non-empty, for the same reason U_{BBA} should be empty then

U_{BBB} should be non-empty, thus U_{ABA} is empty for U_{BAA} and U_{BBB} is non-empty, for the same reason U_{AAB} is empty, thus $s_3 = s_2$ leads to contradiction.

So U_{ABB} is empty, U_{ABA} is non-empty. For symmetry U_{BAB} is empty, U_{BAA} is non-empty, thus U_{AAB} and U_{BBB} is empty, thus s_3 must be $\emptyset|X$.

Thus S is compatible, since it has forbidden configuration $1|23, 2|13, 3|12$, the corresponding X-tree have no degree 3 vertex, so be a line graph. This completes the whole proof. □

Theorem 30. *If split system S on a set X with size $n \geq 4$ has forbidden configuration*
$1|234, 2|134, 3|124, 4|123$ *and* $1|234, 12|34, 13|24, 14|23$*, there exist an $x \in X$ such that $\#\partial_x S \leq 3$.*

Proof: For there are no more than 3 trivial splits in S and the size of x is no more than 4, we can choose an element x such that $x|X - x \notin S$.

If $\#\partial_x S \leq 3$ the proof is accomplished, if not, we pick 3 splits in $\#\partial_x S$ and denote them as $A_1|B_1, A_2|B_2, A_3|B_3$. $\#\partial_x S$ have forbidden configuration $1|23, 2|13, 3|12$. For lemma. 9, consider the symmetry we can suppose $A_1 \subset A_2 \subset A_3$ for convenient we define $A = A_1, B = A_2 - A_1, C = A_3 - A_2$ and $D = B_3$, this four set are all non-empty, non-intersecting and they forms a partition of $X - x$.

Thus we can pick an arbitrary element x' in B, for any $A'|B' \in \partial_{x'} S$, we will assume that $x \in B'$. Next we'll prove some properties of A' and use them to illustrate that there are at most 2 ways to choose A' that the compatible condition is satisfied. In the following part we'll prove that $A' - B \in \{A, C \cup D\}$

1. $A' \not\subseteq B$.

   If not, $\exists a \in A, c \in C$ thus we have

   $$A|xBCD \vdash a|xx'c, xD|ABC \vdash x|ax'c, x'A'|B' \vdash x'|axc, CD|AxB \vdash c|axx'$$

   noted that this implied that $A' \cap A$ and $A' \cap (C \cup D)$ is not empty at the same time.

2. **one of $A' \cap A$ and $A' \cap (C \cup D)$ must be empty.**

   If not, $\exists a \in A' \cap A, c \in A' \cap (C \cup D)$ thus we have

   $$x'A'|B \vdash x'ac|x, A'|x'B' \vdash ac|xx', xA|BCD \vdash ax|cx', xCD|AB \vdash ax'|cx$$

   from 1 and 2 we know that exactly one of $A' \cap A$ and $A' \cap (C \cup D)$ is empty

3. **if $A' \cap A$ is non-empty, $A \subseteq A'$.**

   If not, $\exists a' \in A - A'$ and we arbitrarily pick an element $c \in C$ thus we have

   $$A|xBCD \vdash a'|cxx', x'A'|B' \vdash x'|a'cx, xA|ABC \vdash x|a'cx', CD|xAB \vdash c|a'xx'$$

4. **if $A' \cap (C \cup D) \neq \emptyset$ $A' \cap D$ is non-empty.**

   If not, $A' \cap D = \emptyset$, and $\exists c \in C \cap A', d \in D$ thus we have

   $$xA|BCD \vdash x|x'cd, AB|xCD \vdash x'|xcd, D|xABC \vdash d|xx'c, A'|x'B' \vdash c|xx'd$$

5. **if $A' \cap D$ is non-empty, $C \subseteq A'$.**

   If not, $\exists d \in A' \cup D, c \in C$ and $c \notin A'$ thus we have

   $$AB|xCD \vdash x'|xcd, xAB|CD \vdash xx'|cd, xD|ABC \vdash xd|x'c, x'A'|B' \vdash xc|x'd$$

6. **if $C \subseteq A'$, $D \subseteq A'$**

   If not, $\exists d \in D - A'$ then we arbitrarily choose $c \in C$ thus we have

   $$xA|BCD \vdash x|x'cd, AB|xCD \vdash x'|xcd, D|xABC \vdash d|xx'c, A'|x'B' \vdash c|xx'd$$

   from 4,5 and 6. We conclude that if $A' \cap (C \cup D) \neq \emptyset$, $A' \cap (C \cup D) = C \cup D$, combine 3 we have $A' - B \in \{A, C \cup D\}$

Then we'll prove that there is only one A' that $A' - B = A$ and only one A' that $A' - B = C \cup D$, suppose if we have two different set A'_1 and A'_2 that $A'_1 - B = A$ and $A'_2 - B = A$ then consider the symmetry we can assume $b' \in A'_1$ but $b' \notin A'_2$ and we arbitrarily choose $a \in A$. Since $b' \notin A'_2$, $b' \notin A$ so $b' \in B$ thus we have.

$$A|xBCD \vdash a|b'xx', xA|BCD \vdash ax|b'x', A'_1|x'B'_1 \vdash ab'|xx', B'_2|x'A'_2 \vdash ax'|b'x$$

for the $A' - B = C \cup D$ case we can use the similar strategy: if we have two different set A'_1 and A'_2 that $A'_1 - B = C \cup D$ and $A'_2 - B = C \cup D$ then consider the symmetry we can assume $b' \in A'_1$ but $b' \notin A'_2$ and we arbitrarily choose $a \in C \cup D$. Since $b' \notin A'_2$, $b' \notin C \cup D$ so $b' \in B$, thus we have.

$$CD|xBA \vdash a|b'xx', xCD|AB \vdash ax|b'x', A'_1|x'B'_1 \vdash ab'|xx', B'_2|x'A'_2 \vdash ax'|b'x$$

This means if we put A' into two classes by $A' - B = A \, or \, C \cup D$ each class have at most one element. This means we can always find a element $x \in X$ such that $\#\partial_x S \leq 3$.

□

Following this lemma we could give a proof of theorem. 29:

Proof: From proposition 29 we know that the split system with forbidden congifuration $12|345, 13|245, 14|235, 15|234$ has maximal cardinality no more than $O(n^2)$, the circular split system are with such forbidden configuration, thus maximal cardinality of split system with such forbidden configuration is $O(n^2)$.

For split system with forbidden configuration
$1234|567, 1235|467, 1236|457, 1237|456$, with the same argument of second proof of theorem 27 we can show maximal cardinality of split system with such forbidden configuration is $O(n^3)$. Then we construct a split system with this cardinality, by considering only 3-splits, take 3-sets as a cluster system, the split system is with forbidden configuration $1234|567, 1235|467, 1236|457, 1237|456$ iff the cluster system is a $(-1, 3)$-hierarchy, the 3-point cluster system would suffice the compatible condition and have cardinality $O(n^3)$. □

Conjecture 1. *Maximal cardinality of split system with forbidden configuration*

$$\begin{aligned} 12 &\mid 3\dots m \\ &\dots \\ 1i &\mid 3\dots\hat{i}\dots m \\ &\dots \\ 1m &\mid 2\dots m-1 \end{aligned}$$

is bounded by $O(n^{\lfloor \frac{m}{2} \rfloor})$.

Hereby we propose several weaker problem that could help understand the structure of 2-weakly compatible split system:

1. What is the maximal cardinality of subset of 2-circular split system being 2-weakly compatible.

2. What is the maximal cardinality of subset of the following split system being 2-weakly compatible: Consider ground set $X = \{x_1, \dots, x_u, y_1, \dots, y_v, z_1, \dots, z_w\}$, we write split $s_{ijk} = x_1, \dots, x_i, y_1, \dots, y_j, z_1, \dots, z_k | x_{i+1}, \dots, x_u, y_{j+1}, \dots, y_v, z_{k+1}, \dots, z_w$, the split system defined as $\bar{S} = \{s_{ijk} | 1 \le i \le u, 1 \le j \le v, 1 \le k \le w\}$.

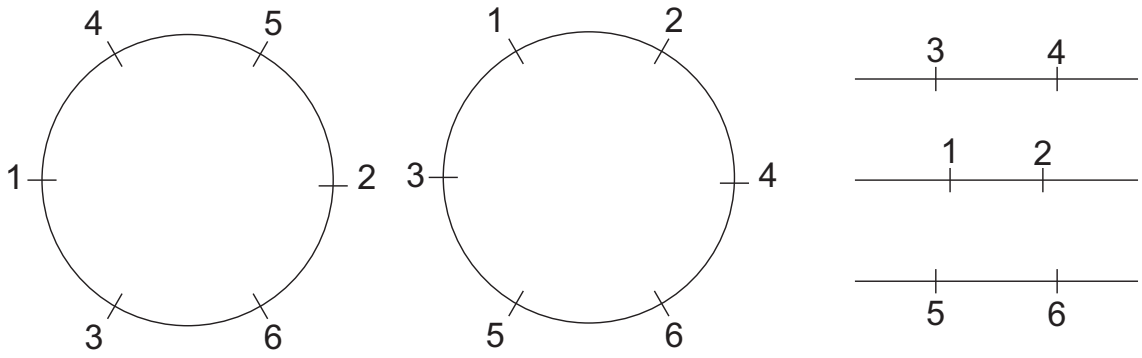


Figure 5.2: Examples of point configures that split system contain the forbidden configuration $123|456, 124|356, 125|346, 126|345$.

5.3.3 Graph Associated to a Split system

In this section we represent a powerful tool that can effectively estimate cardinality of split system with arbitrary forbidden configuration. One property of such method

is, being possible to get upper bound $O(n^\omega)$ with ω non-integer even irrational, and can always give finer estimation than VC-dim. We will use 2-weakly compatible split system as example, showing that cardinality is bounded above by $O(n^{2.5})$.

Definition 31. *For each split system we associate a graph that vertices being splits, and $(A_1|B_1, A_2|B_2)$ is an edge if there is x such that $A_1 = A_2x$ and $B_2 = B_1x$, such edge were labeled x. We say two adjacent edge $e_1 = (s_1, s_2)$ and $e_2 = (s_2, s_3)$ were parallel iff they were labeled with x, y such that $x|y$ is displayed by s_2 and antiparallel iff otherwise.*

Lemma 10. *If for a closed family of split system, the average degree of associated graph do not exceed ω, the maximal cardinality is less than $O(n^{\omega/2})$.*

Proof: Denote the maximal cardinality of split system a_n on a ground set of size n. We note that the cardinality of edge is $\sum_x \#\partial_x S$, the cardinality of edge is no more than $\omega a_n/2$ and no less than $n(a_n - a_{n-1})$. We get:

$$\omega a_n/2 \geq n(a_n - a_{n-1}) \tag{5.53}$$

$$a_n \leq \frac{n}{n - \omega/2} a_{n-1} \tag{5.54}$$

$$< (\frac{n}{n-1})^{\omega/2} a_{n-1} \tag{5.55}$$

$$\leq n^{\omega/2} a_1 \tag{5.56}$$

$$= O(n^{\omega/2}) \tag{5.57}$$

□

Lemma 11. *If S is 2-weakly compatible, we have the following property of associated graph.*

- *Each vertex v associates with non-trivial split has degree no more than 6 and at most 3 of them are pairwise anti-parallel.*
- *If there is a path $(e_1, \ldots, e_n)$ of $n \geq 2$(assume $e_i = (v_i, v_{i+1})$) that e_i and e_{i+1} being parallel for all i, and there are two edge connecting v_1 anti-parallel with e_1 then at most 1 edge connecting v_{n+1} anti-parallel with e_n.*

Proof:

- If a split $s = A|B$ is not trivial, its degree is at most 6. Suppose edges were labeled with x_i there is $x_1, \ldots, x_4 \in A$ or B, we could suppose $x_1, \ldots, x_4 \in A$, thus $A-x_1|Bx_1, \ldots, A-x_4|Bx_4 \in S$ which is against the compatible condition.

- If there is a parallel path $(e_1, \ldots, e_n)$ of $n \geq 2$ (assume $e_i = (v_i, v_{i+1})$, e_i associates with split s_i). We could have $s_1 = Ax_1 \ldots x_n|B$, $s_{n+1} = Bx_1 \ldots x_n|A$, if there are two edge connecting v_1 anti-parallel with e_1 and two edge connecting v_{n+1} anti-parallel with e_n, there is $a_1, a_2 \in A$ and $b_1, b_2 \in A$ such that split $Ax_1 \ldots x_n - a_1|Ba_1$, $Ax_1 \ldots x_n - a_2|Ba_2$, $Bx_1 \ldots x_n - b_1|Ab_1$ and $Bx_1 \ldots x_n - b_2|Ab_2$ were in the split system, here comes the contradiction(when restricting on $\{x_1, x_2, a_1, a_2, b_1, b_2\}$).

□

Theorem 31. *The maximal cardinality of 2-weakly compatible split system is bounded by* $O(n^{2.5})$.

Proof: If S is 2-weakly compatible split system on X. We first remove split $\emptyset|X$ and all trivial splits from S, note that they all have degree n, and construct associate graph $G = (V, E)$, we will prove that average degree of G do not exceed 5. Let A be the set of vertices with degree 6. We first observe that edges in the union of all parallel path start from A can be oriented simultaneously such that two edge were anti-parallel iff they are both in or out of same vertices: we always start from vertices in A, it is possible since no parallel path start and end with vertices in A, denote such set of edges E' and set of vertices involved V'. The edge not in E' must be anti-parallel with edges in E' if possible, while counting "in" and "out" degree of vertices such edge were count as "in" for both vertices it connects. The condition of lemma 11 implies that if there is a path $v_1 \to \cdots \to v_n$ and v_1 has degree 6 then $v_3, \ldots, v_n$ have in degree at most 2. We divide V into four sets: A, B, C, D. $v \in B$ iff $v \notin A$ there is an edge connecting v and some vertices in A, with in degree 3 and out degree no more than 2, let $C = V' - A - B$ and $D = V - V'$. For each set $S \subset V$ we write $c(S) = \sum_{v \in S}(d(v) - 5)(d(v)$ is degree of $v)$ and define in/out degree of a set is the cardinality of edges go in/out from that set. Firstly $c(A) = \#A$, denote such number ϵ. And a vertex with degree 6 must have at least 3 neighbors with degree less than 6 thus the out degree of A is no less than 3ϵ, denote the out degree of A by σ. Then consider set B, suppose vertices with degree $i(i = 3, 4, 5)$ were denoted

ϵ_i, $c(B) = -2\epsilon_3 - \epsilon_4$, the in degree should be no more than $3\epsilon_3 + 3\epsilon_4 + 3\epsilon_5$ and out should be $\epsilon_4 + 2\epsilon_5$(this is because no edge connecting vertices both in B can be parallel with edge from A or the end of previous edge can not have in degree 3), since $\sigma - (3\epsilon_3 + 3\epsilon_4 + 3\epsilon_5)$ being number of vertices from A to C, thus being non-negative, denoted by δ. Then for set C, the in degree is $\epsilon_4 + 2\epsilon_5 + \delta$ and vanishing out degree, we further observe that the if a vertex has degree 3, then the difference of in and out is at most 1 and contributes to $c(C)$ a -2, if a vertex has degree 2, then the difference of in and out is at most 2 and contributes to $c(C)$ a -3, if a vertex has degree 1, then the difference of in and out is at most 1 and contributes to $c(C)$ a -4, otherwise out degree is no less than in. Thus ratio of $-c(C)$ against in degree of C is greater than 1.5, $-c(C) \geq 1.5(\epsilon_4 + 2\epsilon_5 + \delta)$. And $c(D) \leq 0$. Collecting all we get:

$$
\begin{aligned}
c(V) &= c(A) + c(B) + c(C) + c(D) && (5.58)\\
&\leq \epsilon - 2\epsilon_3 - \epsilon_4 - 1.5(\epsilon_4 + 2\epsilon_5 + \delta) && (5.59)\\
&\leq \frac{1}{3}\sigma - 2\epsilon_3 - 2.5\epsilon_4 - 3\epsilon_5 - 1.5\delta && (5.60)\\
&= \frac{1}{3}(\delta + 3\epsilon_3 + 3\epsilon_4 + 3\epsilon_5) - 2\epsilon_3 - 2.5\epsilon_4 - 3\epsilon_5 - 1.5\delta && (5.61)\\
&\leq 0 && (5.62)
\end{aligned}
$$

Thus the average degree is no more than 5. Then we can put back the trivial splits and split $\emptyset|X$, we have:

$$
\begin{aligned}
(5(a_n - n - 1) + n(n+1))/2 &\geq n(a_n - a_{n-1}) && (5.63)\\
a_n &\leq \frac{n}{n-2.5}a_{n-1} + \frac{n^2 - 4n}{2n-5} && (5.64)\\
&< (\frac{n}{n-1})^{2.5}a_{n-1} + \frac{n^2 - 4n}{2n-5} && (5.65)\\
&\leq n^{2.5}((\frac{1}{5})^{2.5}a_5 + \sum_{i=6}^{n}\frac{i-4}{(2i-2.5)i^{1.5}}) && (5.66)\\
&= O(n^{2.5}) && (5.67)
\end{aligned}
$$

□

Remark 8. *If we can show that the average degree is less than $\omega > 4$, the maximal cardinality is less than $O(n^{\omega/2})$. Denote the ground set $X = \{1, \ldots, n\}$, Consider split system $S = \{i, \ldots, j|1, \ldots, i-1, j+1, \ldots, n\} \cup \{1, i, \ldots, j|2, \ldots, i-1, j+1,$*

$\ldots, n\}$. *Such split system is 2-weakly compatible and the graph associated with this split system almost all vertices have degree* 5*, which indicates the best result we can achieve using graph argument is* $O(n^{2.5})$.

We should note that estimation using edge number can be generalized to higher dimension. In this note binomial coefficients were analytically extended to all real numbers.

Definition 32. *In the graph associated with a split system, a box(or n-box) is a subgraph isomorphic to n-hypercube and parallel edges were labeled with same taxa.*

Theorem 32. *If for a closed family of split system, the ratio of cardinality of k-boxes against cardinality of vertices do not exceed* $\omega = \binom{\alpha}{k}$ *(*$\alpha \geq k$ *and do not necessarily be integer) in associated graph, the maximal cardinality is less than* $O(n^\alpha)$.

Lemma 12. $\binom{n}{\alpha} = O(n^\alpha)$ *for fixed* $\alpha > 0$.

Lemma 13. *For a sequence of number* a_n*, if* $\sum_{i=0}^{k}(-1)^i\binom{k}{i}a_{n-i} \leq \frac{\omega}{\binom{n}{k}}a_n$*, we have* $a_n < O(n^\alpha)$.

Proof: Denote $\mathrm{d}^j a_n = \sum_{i=0}^{j}(-1)^i\binom{j}{i}a_{n-i}$, namely $\mathrm{d}^0 a_n = a_n$ and $\mathrm{d}^j a_n = \mathrm{d}^{j-1}a_n - \mathrm{d}^{j-1}a_{n-1}$. Let b_n be a series of number such that $b_i = a_i$ for $i = 1, \ldots, k+1$ and $\sum_{i=0}^{k}(-1)^i\binom{k}{i}b_{n-i} = \frac{\omega}{\binom{n}{k}}b_n$. b_n must be linear combinatorics of $\binom{n+\alpha_i-k}{\alpha_i}$, in which α_i are solutions of $\omega = \binom{\alpha}{k}$, thus $b_n = O(n^\alpha)$ since α is the greatest root($\binom{\alpha}{k}$ is strictly increasing as function of α when $\alpha > k$).

Then we prove for every $i = 0, \ldots, k$, $\mathrm{d}^i a_n \leq \mathrm{d}^i b_n$, we do this by induction. Suppose this holds for $n = k+1, \ldots, m-1$ then $\mathrm{d}^k a_m \leq \frac{\omega}{\binom{m}{k}}a_m$, $\mathrm{d}^k b_m = \frac{\omega}{\binom{m}{k}}b_m$, thus

$$(1 - \frac{\omega}{\binom{m}{k}})a_m \leq d^{k-1}a_{m-1} + \cdots + d^0 a_{m-1} \tag{5.68}$$

$$\leq d^{k-1}b_{m-1} + \cdots + d^0 b_{m-1} \tag{5.69}$$

$$= (1 - \frac{\omega}{\binom{m}{k}})b_m \tag{5.70}$$

thus $a_m \leq b_m$, $\mathrm{d}^k a_m \leq \frac{\omega}{\binom{m}{k}}a_m \leq \frac{\omega}{\binom{m}{k}}b_m \leq \mathrm{d}^k b_m$. Using induction again $\mathrm{d}^j a_m = \mathrm{d}^{j+1}a_m + \mathrm{d}^j a_{m-1} \leq \mathrm{d}^j b_m$. This completes the proof. □

Proof:[Proof of theorem 32] For $Y \subseteq X$ define $a_Y = \#S|_Y$, a_i being the average of $a_Y = \#S|_Y$ for $\#Y = i$. For $x_1, \ldots, x_k \in Y$ we have

$\#\partial_{x_1,\ldots,x_k} S|_Y = \sum_{I\subseteq\{x_1,\ldots,x_k\}} (-1)^{\#I} \#S|_{Y-I}$. On the other hand $\sum_{x_1,\ldots,x_k\in Y} \#\partial_{x_1,\ldots,x_k} S|_Y$ is the number of k-boxes thus no more than $\omega S|_Y$. Averaging this for all $\#Y = i$, we get $\binom{i}{k}\, \mathrm{d}^k a_i \leq \omega a_i$. Since for all split system a_k is bounded by 2^k, using lemma 13 $a_n = O(n^\alpha)$. □

As a summary we can use the following steps to analyze a given family of split system:

- Find the forbidden subgraph, a general method is: for a forbidden configuration $S = \{s_i : A_i|B_i\}$ on ground set $\{1,\ldots,n\}$, take network associated with dual split system with ground set S: $\{\bar{s}_i : i = 1,\ldots,n\}$, $\bar{s}_i = S_i|S_i'$, $S_i = \{s_j : i \in A_j\}$ and S_i' being its complement. See(Fig. 5.3) an example of 2−weakly compatible.

- Calculate bound of average degree, or density of k-boxes, this requires techniques from combinatorics, especially graph theory.

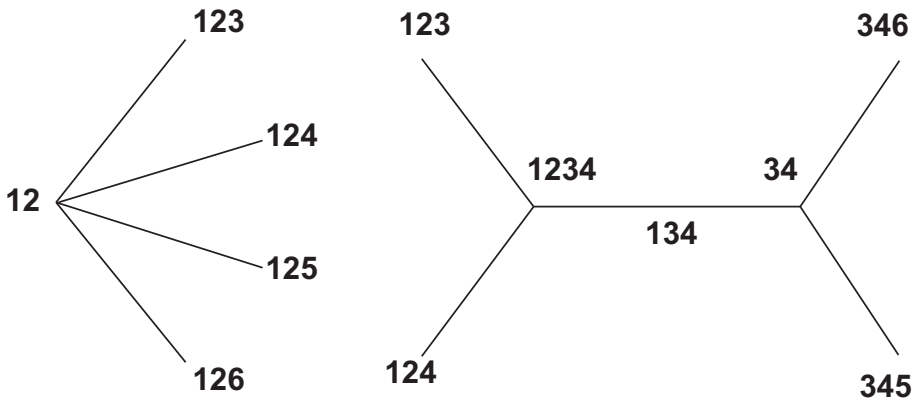


Figure 5.3: An example showing how to construct forbidden subgraph from forbidden configuration of 2−weakly compatible split system: in the left one $\{A_i\} = 123, 124, 125, 126$, in the right one $\{A_i\} = 123, 124, 345, 346$

Following these steps we can give a bound on cardinality of split systems.

5.4 Final Remark

In this section split system with general compatible condition is studied. Such problem arises not only from phylogenetics but also many fields, including multicommodity flow problem([104]), machine learning([58]), metric theory([105–107]),

lattice([107]) and oriented matroids([56, 61]). Many combinatoric problem can be reformulated using the language of split system or cluster system. For example the sunflower condition[108] can be reformulated as forbidden configuration $\{1, \ldots, n\}$ of cluster system, and if we desire the kernel to be non-empty the forbidden configuration is $\{12, \ldots, 1(n+1)\}$. As indicated by theorem 17, a family of split system (or cluster system) can be characterized by forbidden configuration iff it is closed w.r.t taking subsystem or restriction, this is the rational of ubiquitous of forbidden configuration formalism.

Here we list several open problems we were interested in related with split system.

1. *Maximal cardinality for general split system or cluster system* For a split system or cluster system with certain forbidden configuration(s), what is maximal cardinality for such system and how to characterize systems with maximal cardinality? For compatible and weakly-compatible the maximal cardinality is $2n-3$ and $\binom{n}{2}$ respectively; which corresponds to binary X-trees and circular split system. In the previous chapters we have shown that for k-very-weakly-compatible the maximal cardinality is $\sum_{i=1}^{k}\binom{n}{2i}$, realized by topes of neighborly rank $2k+1$ oriented matroid. However we have encountered compatible conditions that maximal cardinality can not be easily decided. The problem of maximal cardinality have two part: 1.deciding the lower bound, which often involves constructing a split system of cardinality as big as possible. 2.deciding the upper bound, which needs algebraic or combinatoric methods. We want to know whether non-constructive method, like probabilistic method[109, 110], could give better lower bound, or to make modification on flag algebra [111, 112] for split system, which has been proved to be powerful in other extreme problem originated from phylogenetics [113]. An problem of special interest is, is growth rate of maximal cardinality of a split system with finite forbidden configurations always be polynomial?

2. *Linear dependency* We have established some linear independency theorems (for example proposition 6). One could ask whether we could associate splits with other type of vectors that applicable for general compatible condition. Another possible approach is, using a split system with compatible condition stronger than VC-dimension less than 4 as an example, we know that the mapping from split weights to distances and 4-diversities $f : (w_1, \ldots, w_i) \mapsto$

$(\delta(a,b),\dots,\delta(a,b,c,d),\dots)$ is injective. If we could found $1,2 \in X$ such that f compounds with projection $g : (\delta(a,b),\dots,\delta(a,b,c,d),\dots)$ $\mapsto (\delta(a,b),\dots,\delta(1,2,a,b),\dots,\delta(1,a,b,c),\dots,\delta(2,a,b,c),\dots)$ is also injective. We could conclude that size of split system is bounded by $O(n^3)$. We know that 2-weakly compatible split system do not necessarily have such property ($S_1 = \{123|456, 124|356, 125|346, 134|256, 135|246, 145|236\}$ is an example). At last we want to know if there is an explanation of linear independency by introducing proper inner product: a set of vector is linearly independent if they are pairwise orthogonal ([102], pp. 77).

3. *Duality* Given a compatible condition we could ask the maximal cardinality of split system on a ground set of fixed cardinality. We could also ask for a dual problem: what is the maximal cardinality of ground set X with split system of fixed cardinality, such that for every pair of elements $a, b \in X$, $a|b$ is displayed by split system, we denote those two numbers $f(n), g(n)$ respectively. Such duality is motivated by hyperplane-point duality in projective geometry. For compatible split system $g(n) = n + 1$, for weakly compatible we conjecture that $g(n) = 2n$ realized by circular split system. A problem of special interest is, how $f(n)$ and $g(n)$ related.

4. *Tree* From the discussion above we could have a looming idea that trees, or compatible split system is a rather special object. In contrary with split system with forbidden configuration F_n or F_{2n}^n, there is no higher analog of trees. We conjecture there should be a system of families of split systems $\mathcal{F}_i$ such that it should have compatible condition stronger than F_{i+1} and have maximal cardinality $O(n^i)$. $\mathcal{F}_1$ should be family of compatible split system. It's worthwhile to mention that valuated matroids[114] (or affine buildings[42, 115], tropical grassmannians[116, 117], etc.) might be a proper generalization of trees, it's also being an nice example of system of set of linear equations with known maximal dimension of solutions, but their underlying combinatoric structure is far distinct from split system.

Bibliography

[1] Dobzhansky T. Nothing in biology makes sense except in the light of evolution[J]. The American Biology Teacher. 1973, 35(3):125–129.

[2] Watson J D, Crick F H C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid[J/OL]. Nature. apr 1953, 171(4356):737–738. http://dx.doi.org/10.1038/171737a0. DOI: 10.1038/171737a0.

[3] Kimura M. Evolutionary rate at the molecular level.[J]. Nature. 1968, 217 (5129):624–626. DOI: 10.1038/217624a0.

[4] King J L, Jukes T H. Non-Darwinian evolution.: volume 164[M]. [S.l.]: [s.n.], 1969: 788–798. DOI: 10.1126/science.164.3881.788.

[5] Nei M, Suzuki Y, Nozawa M. The neutral theory of molecular evolution in the genomic era.[J/OL]. Annual review of genomics and human genetics. jan 2010, 11:265–89. http://www.annualreviews.org/doi/abs/10.1146/annurev-genom-082908-150129. DOI: 10.1146/annurev-genom-082908-150129.

[6] Kimura M. The neutral theory of molecular evolution: a review of recent evidence.[J]. Idengaku zasshi. aug 1991, 66(4):367–86.

[7] Keeling P J, Palmer J D. Horizontal gene transfer in eukaryotic evolution.[J/OL]. Nature reviews. Genetics. aug 2008, 9(8):605–18. http://dx.doi.org/10.1038/nrg2386. DOI: 10.1038/nrg2386.

[8] Koonin E V, Makarova K S, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification.[J/OL]. Annual review of microbiology. jan 2001, 55:709–42. http://www.annualreviews.org/doi/abs/10.1146/annurev.micro.55.1.709. DOI: 10.1146/annurev.micro.55.1.709.

[9] Fraser C, Alm E J, Polz M F, et al. The bacterial species challenge: making sense of genetic and ecological diversity.[J/OL]. Science (New York, N.Y.). mar 2009, 323(5915):741–6. http://www.sciencemag.org/content/323/5915/741. DOI: 10.1126/science.1159388.

[10] Rieseberg L H. The role of hybridization in evolution: old wine in new skins[J]. American Journal of Botany. 1995, 82(7):944–953.

[11] Rieseberg L H, Carney S E. Plant hybridization[J]. New Phytologist. dec 1998, 140(4):599–624. DOI: 10.1046/j.1469-8137.1998.00315.x.

[12] Rieseberg L H, Ellstrand N C, Arnold M. What Can Molecular and Morphological Markers Tell Us About Plant Hybridization?[J]. Critical Reviews in Plant Sciences. jan 1993, 12(3):213–241. DOI: 10.1080/07352689309701902.

[13] Buckler E S, Ippolito A, Holtsford T P. The evolution of ribosomal DNA divergent paralogues and phylogenetic implications[J]. Genetics. 1997, 145 (3):821–832.

[14] Schierup M H, Hein J. Consequences of recombination on traditional phylogenetic analysis[J]. Genetics. 2000, 156(2):879–891.

[15] Planet P J, Kachlany S C, Fine D H, et al. The Widespread Colonization Island of Actinobacillus actinomycetemcomitans.[J]. Nature genetics. jun 2003, 34 (2):193–8. DOI: 10.1038/ng1154.

[16] Delwiche C F, Palmer J D. Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids[J]. Molecular Biology and Evolution. jul 1996, 13(6):873–882. DOI: 10.1093/oxfordjournals.molbev.a025647.

[17] Wu M, Sun L V, Vamathevan J, et al. Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: a streamlined genome overrun by mobile genetic elements.[J/OL]. PLoS biology. mar 2004, 2 (3):E69. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0020069. DOI: 10.1371/journal.pbio.0020069.

[18] Kilian B, Ozkan H, Deusch O, et al. Independent Wheat B and G Genome Origins in Outcrossing Aegilops Progenitor Haplotypes[J/OL]. Molecular Bi-

ology and Evolution. oct 2006, 24(1):217–227. http://mbe.oxfordjournals.org/content/24/1/217.short. DOI: 10.1093/molbev/msl151.

[19] Crow K D, Stadler P F, Lynch V J, et al. The "fish-specific" Hox cluster duplication is coincident with the origin of teleosts.[J/OL]. Molecular biology and evolution. jan 2006, 23(1):121–36. http://mbe.oxfordjournals.org/content/23/1/121.short. DOI: 10.1093/molbev/msj020.

[20] Bandelt H J, Dress A. A canonical decomposition theory for metrics on a finite set[J]. Advances in Mathematics. 1992, 92:47–105.

[21] Bandelt H J, Dress A W. Split decomposition: a new and useful approach to phylogenetic analysis of distance data.[J/OL]. Molecular Phylogenetics and Evolution. 1992, 1(3):242–252. http://www.ncbi.nlm.nih.gov/pubmed/1342941.

[22] Buneman P. The recovery of trees from measures of dissimilarity: pp[M]. [S.l.]: Edinburgh University Press, 1971: 387–395.

[23] Fitch W M, Margoliash E. Construction of phylogenetic trees[J]. Science. 1967, 155(760):279–284.

[24] Bandelt H J, Forster P, Sykes B C, et al. Mitochondrial portraits of human populations using median networks.[J/OL]. Genetics. 1995, 141(2):743–753. http://www.genetics.org/cgi/content/abstract/141/2/743.

[25] Bandelt H J, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies.[J/OL]. Molecular Biology and Evolution. 1999, 16(1): 37–48. http://www.ncbi.nlm.nih.gov/pubmed/10331250.

[26] Fitch W M. Networks and viral evolution[J/OL]. Journal of Molecular Evolution. 1997, 44(S1):S65–S75. http://www.springerlink.com/index/10.1007/PL00000059. DOI: 10.1007/PL00000059.

[27] Sokal R R, Michener C D. A statistical method for evaluating systematic relationships[J/OL]. University of Kansas Scientific Bulletin. 1958, 28(22):1409–1438. http://www.citeulike.org/user/druvus/article/1327877. DOI: 10.1111/j.1600-0447.1958.tb01740.x.

[28] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees.[J/OL]. Molecular biology and evolution. jul 1987, 4 (4):406–25. http://www.ncbi.nlm.nih.gov/pubmed/18343690.

[29] Bryant D, Moulton V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks.[J/OL]. Molecular Biology and Evolution. 2004, 21(2):255–265. http://www.ncbi.nlm.nih.gov/pubmed/14660700.

[30] Roch S. Toward extracting all phylogenetic information from matrices of evolutionary distances.[J/OL]. Science. 2010, 327(5971):1376–1379. http://www.ncbi.nlm.nih.gov/pubmed/20223986.

[31] Grünewald S, Forslund K, Dress A, et al. QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets.[J/OL]. Molecular Biology and Evolution. 2007, 24(2):532–538. http://www.ncbi.nlm.nih.gov/pubmed/17119010.

[32] Balvočiūté M, Spillner A, Moulton V. FlatNJ: A novel network-based approach to visualize evolutionary and biogeographical relationships[J]. Systematic Biology. 2014, 63(3):383–396.

[33] Yang J, Grünewald S, Wan X F. Quartet-Net: A Quartet Based Method to Reconstruct Phylogenetic Networks.[J/OL]. Molecular biology and evolution. 2013, 30(Felsenstein 2004):1206–1217. http://www.ncbi.nlm.nih.gov/pubmed/23493256. DOI: 10.1093/molbev/mst040.

[34] Barry D, Hartigan J A. Asynchronous distance between homologous DNA sequences.[J/OL]. Biometrics. 1987, 43(2):261–276. http://www.ncbi.nlm.nih.gov/pubmed/3607200.

[35] Rzhetsky A, Nei M. A simple method for estimating and testing minimum-evolution trees[J]. Mol. Biol. Evol. 1992, 9(5):945–967.

[36] Holland B, Moulton V. Consensus networks: A method for visualising incompatibilities in collections of trees[M]. [S.l.]: Springer, 2003: 165–176.

[37] Strimmer K. Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies[J]. Molecular Biology. 1994, 13(7):964–969.

[38] Huson D H, Bryant D. Application of phylogenetic networks in evolutionary studies: volume 23[M]. [S.l.]: [s.n.], 2006: 254–267. DOI: 10.1093/molbev/msj030.

[39] Huson D H, Scornavacca C. A survey of combinatorial methods for phylogenetic networks: volume 3[M]. [S.l.]: [s.n.], 2011: 23–35. DOI: 10.1093/gbe/evq077.

[40] Huson D H, Rupp R, Scornavacca C. Phylogenetic Networks: Concepts, Algorithms and Applications: volume 57[M/OL]: Cambridge University Press, 2011: 362. http://sysbio.oxfordjournals.org/cgi/doi/10.1093/sysbio/syr055. DOI: 10.1093/sysbio/syr055.

[41] Dress A, Huber K T, Moulton V. Some variations on a theme by Buneman[J/OL]. Annals of Combinatorics. 1997. http://www.springerlink.com/index/4W081T6N7832L546.pdf.

[42] Dress A, Terhalle W. The tree of life and other affine buildings.[J/OL]. Documenta Mathematica. 1998, 565–574. https://eudml.org/doc/233296.

[43] Bryant D, Tupper P F. Hyperconvexity and tight-span theory for diversities[J/OL]. Advances in Mathematics. dec 2012, 231(6):3172–3198. http://arxiv.org/abs/1006.1095. DOI: 10.1016/j.aim.2012.08.008.

[44] Dress A. Split Decomposition over an Abelian Group Part 1: Generalities[J/OL]. Annals of Combinatorics. jun 2009, 13(2):199–232. http://link.springer.com/10.1007/s00026-009-0020-2. DOI: 10.1007/s00026-009-0020-2.

[45] Dress A. Split decomposition over an abelian group, Part 2: Group-valued split systems with weakly compatible support[J/OL]. Discrete Applied Mathematics. may 2009, 157(10):2349–2360. http://www.sciencedirect.com/science/article/pii/S0166218X08003727. DOI: 10.1016/j.dam.2008.06.041.

[46] Gascuel O, Steel M. Neighbor-Joining Revealed[J/OL]. Molecular Biology and Evolution. 2006, 23(11):1997–2000. http://www.ncbi.nlm.nih.gov/pubmed/16877499. DOI: 10.1093/molbev/msl072.

[47] Levy D, Pachter L. The Neighbor-Net Algorithm[J/OL]. Advances in Applied Mathematics. 2007, 1(2):240–258. http://arxiv.org/abs/math/0702515.

[48] Isbell J R. Six theorems about injective metric spaces[J/OL]. Commentarii Mathematici Helvetici. dec 1964, 39(1):65–76. http://retro.seals.ch/digbib/view?rid=comahe-002:1964-1965:39::7. DOI: 10.1007/BF02566944.

[49] Espínola R, Khamsi M A. Introduction to hyperconvex spaces[M]. [S.l.]: Springer, 2001: 391–435.

[50] Dress A, Moulton V, Spillner A, et al. Obtaining splits from cut sets of tight spans[J/OL]. Discrete Applied Mathematics. jul 2013, 161(10-11):1409–1420. http://dl.acm.org/citation.cfm?id=2467348.2467637. DOI: 10.1016/j.dam.2013.02.001.

[51] Dress A. Towards a theory of holistic clustering[J]. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. 1997, 37:271–289.

[52] Townsend T M, Larson A, Louis E, et al. Molecular phylogenetics of Squamata: the position of snakes, amphisbaenians, and dibamids, and the root of the squamate tree[J]. Systematic Biology. 2004, 53:735–757.

[53] Pachter L, Sturmfels B. Algebraic statistics for computational biology: volume 13[M]. [S.l.]: Cambridge University Press, 2005.

[54] Bryant D, Moulton V, Spillner A. Consistency of the neighbor-net algorithm.[J/OL]. Algorithms for molecular biology : AMB. jan 2007, 2:8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1948893&tool=pmcentrez&rendertype=abstract. DOI: 10.1186/1748-7188-2-8.

[55] Björner A, Las Vergnas M, Sturmfels B, et al. Oriented Matroids: volume 1[M]. [S.l.]: Cambridge University Press, 2000.

[56] Bryant D, Dress A. Linearly independent split systems[J/OL]. European Journal of Combinatorics. 2007, 28(6):1814–1831. http://linkinghub.elsevier.com/retrieve/pii/S0195669806000874. DOI: 10.1016/j.ejc.2006.04.007.

[57] Spillner A, Nguyen B, Moulton V. Constructing and Drawing Regular Planar Split Networks[J/OL]. IEEE/ACM Transactions on Computational Biology and Bioinformatics. mar 2012, 9(2):395–407. http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=5989792. DOI: 10.1109/TCBB.2011.115.

[58] Gärtner B, Welzl E. Vapnik-Chervonenkis dimension and (pseudo-)hyperplane arrangements[J/OL]. Discrete & Computational Geometry. dec 1994, 12(1): 399–432. http://link.springer.com/10.1007/BF02574389. DOI: 10.1007/BF02574389.

[59] Shemer I. Neighborly polytopes[J/OL]. Israel Journal of Mathematics. jan 1982, 43(4):291–311. http://link.springer.com/10.1007/BF02761235. DOI: 10.1007/BF02761235.

[60] Sturmfels B. Neighborly Polytopes and Oriented Matroids[J/OL]. European Journal of Combinatorics. nov 1988, 9(6):537–546. http://www.sciencedirect.com/science/article/pii/S0195669888800507. DOI: 10.1016/S0195-6698(88)80050-7.

[61] Balvočiūtė M, Bryant D, Spillner A. When can splits be drawn in the plane?[M/OL]. sep 2015. http://arxiv.org/abs/1509.06104.

[62] Richter-Gebert J. Oriented matroids with few mutations[J/OL]. Discrete & Computational Geometry. sep 1993, 10(3):251–269. http://link.springer.com/10.1007/BF02573980. DOI: 10.1007/BF02573980.

[63] Kalai G, Wigderson A. Neighborly Embedded Manifolds[J/OL]. Discrete & Computational Geometry. feb 2008, 40(3):319–324. http://link.springer.com/10.1007/s00454-008-9065-y. DOI: 10.1007/s00454-008-9065-y.

[64] Dress A W, Huber K T, Koolen J, et al. Cut points in metric spaces[J]. Applied Mathematics Letters. jun 2008, 21(6):545–548. DOI: 10.1016/j.aml.2007.05.018.

[65] Dress A W, Huber K T, Koolen J, et al. Compatible decompositions and block realizations of finite metrics[J]. European Journal of Combinatorics. oct 2008, 29(7):1617–1633. DOI: 10.1016/j.ejc.2007.10.003.

[66] Weyer-Menkhoff J, Devauchelle C, Grossmann A, et al. Integer linear programming as a tool for constructing trees from quartet data.[J/OL]. Computational Biology and Chemistry. 2005, 29(3):196–203. http://www.ncbi.nlm.nih.gov/pubmed/15979039.

[67] Holmes I. Using evolutionary Expectation Maximization to estimate indel rates[J/OL]. Bioinformatics. feb 2005, 21(10):2294–2300. http://bioinformatics.oxfordjournals.org/content/21/10/2294.short. DOI: 10.1093/bioinformatics/bti177.

[68] Bryant D, Galtier N, Poursat M A. Likelihood calculation in molecular phylogenetics[J]. Mathematics of Evolution and Phylogeny, cáp. 2005, 2:33–62.

[69] Hendy M D, Penny D. Spectral analysis of phylogenetic data: volume 10[M]. [S.l.]: [s.n.], 1993: 5–24. DOI: 10.1007/BF02638451.

[70] Hendy M D. The relationship between simple evolutionary tree models and observable sequence data[J]. Systematic Biology. 1989, 38(4):310–321.

[71] Hendy M D, Penny D. A framework for the quantitative study of evolutionary trees[J/OL]. Systematic Zoology. 1989, 38(4):297–309. http://sysbio.oxfordjournals.org/content/38/4/297.short. DOI: 10.1093/sysbio/syp096.

[72] Hendy M D, Penny D, Steel M A. A discrete Fourier analysis for evolutionary trees.[J]. Proceedings of the National Academy of Sciences of the United States of America. 1994, 91(8):3339–3343.

[73] Kimura M. Estimation of evolutionary distances between homologous nucleotide sequences.[J/OL]. Proceedings of the National Academy of Sciences of the United States of America. jan 1981, 78(1):454–458. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=319072&tool=pmcentrez&rendertype=abstract.

[74] Szekely L, Steel M, Erdos P. Fourier Calculus on Evolutionary Trees[J/OL]. Advances in Applied Mathematics. jun 1993, 14(2):200–216. http://www.sciencedirect.com/science/article/pii/S0196885883710110. DOI: 10.1006/aama.1993.1011.

[75] Waddell P J, Penny D, Moore T. Hadamard conjugations and modeling sequence evolution with unequal rates across sites.[J/OL]. Molecular Phylogenetics and Evolution. 1997, 8(1):33–50. http://www.ncbi.nlm.nih.gov/pubmed/9242594.

[76] von Haeseler A, Churchill G A. Network models for sequence evolution[J/OL]. J Mol Evol. 1993, 37(1):77–85. http://www.ncbi.nlm.nih.gov/pubmed/8395605.

[77] Bryant D. Extending Tree Models to Split Networks[M]. Cambridge, UK: Cambridge University Press, 2005: 322–334.

[78] Bryant D. Hadamard Phylogenetic Methods and the n-taxon process[J/OL]. Bulletin of Mathematical Biology. 2008, 71(2):20. http://arxiv.org/abs/0806.1378.

[79] Sumner J G, Charleston M A, Jermiin L S, et al. Markov invariants, plethysms, and phylogenetics (the long version)[J/OL]. Arxiv preprint arXiv07113503. 2007, 39. http://arxiv.org/abs/0711.3503.

[80] Sumner J G, Holland B H, Jarvis P D. The algebra of the general Markov model on phylogenetic trees and networks[J/OL]. Bulletin of Mathematical Biology. 2010, 74:17. http://arxiv.org/abs/1012.5165.

[81] Huber K T, Langton M, Penny D, et al. Spectronet: a package for computing spectra and median networks.[J]. Applied bioinformatics. 2002, 1(3):159–161.

[82] Simon C, Nigro L, Sullivan J, et al. Large differences in substitutional pattern and evolutionary rate of 12S ribosomal RNA genes.[J/OL]. Molecular Biology and Evolution. 1996, 13(7):923–932. http://www.ncbi.nlm.nih.gov/pubmed/8752001.

[83] Folland G B. Studies in advanced mathematics A course in abstract harmonic analysis[M]. 1st ed. [S.l.]: CRC Press, 1995.

[84] Deitmar A. Universitext A First Course in Harmonic Analysis[M]. 2nd ed. [S.l.]: Springer, 2005.

[85] Steel M, Huson D, Lockhart P J. Invariable Sites Models and Their Use in Phylogeny Reconstruction[J/OL]. Systematic Biology. jun 2000, 49(2):225–232. http://sysbio.oxfordjournals.org/content/49/2/225.abstract?ijkey=c93a0b7a20b41ae18057d3f45f58599b40cb1c2f&keytype2=tf{_}ipsecsha. DOI: 10.1093/sysbio/49.2.225.

[86] Allman E S, Rhodes J a. Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites[J/OL]. Mathematical Biosciences. 2008, 211(1):18–33. http://arxiv.org/abs/q-bio/0702050. DOI: 10.1016/j.mbs.2007.09.001.

[87] Jayaswal V, Robinson J, Jermiin L. Estimation of phylogeny and invariant sites under the general Markov model of nucleotide sequence evolution.[J/OL]. Systematic Biology. 2007, 56(2):155–162. http://www.ncbi.nlm.nih.gov/pubmed/17454972.

[88] Eigen M, Winkler-Oswatitsch R, Dress A. Statistical geometry in sequence space: a method of quantitative comparative sequence analysis.[J/OL]. Proceedings of the National Academy of Sciences of the United States of America. 1988, 85(16):5913–5917. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=281875&tool=pmcentrez&rendertype=abstract.

[89] Holland B R, Jarvis P D, Sumner J G. Low-parameter phylogenetic inference under the general markov model.[J/OL]. Systematic biology. jan 2013, 62(1):78–92. http://sysbio.oxfordjournals.org/cgi/content/long/62/1/78. DOI: 10.1093/sysbio/sys072.

[90] Zardoya R, Meyer A. Complete mitochondrial genome suggests diapsid affinities of turtles.[J/OL]. Proceedings of the National Academy of Sciences of the United States of America. 1998, 95(24):14226–14231. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=24355&tool=pmcentrez&rendertype=abstract.

[91] Hedges S B, Poling L L. A molecular phylogeny of reptiles.[J/OL]. Science. 1999, 283(5404):998–1001. http://www.ncbi.nlm.nih.gov/pubmed/9974396. DOI: 10.1126/science.283.5404.998.

[92] Vapnik V N. Statistical Learning Theory[M]. [S.l.]: [s.n.], 1998: 760. DOI: 10.2307/1271368.

[93] Bretto A. Hypergraph Theory: An Introduction[J]. Mathematical Engineering. 2013, 11. DOI: 10.1007/978-3-319-00080-0.

[94] Bokowski J, Sturmfels B. An infinite family of minor-minimal nonrealizable 3-chirotopes[J/OL]. Mathematische Zeitschrift. dec 1989, 200(4):583–589. http://link.springer.com/10.1007/BF01160956. DOI: 10.1007/BF01160956.

[95] Anthony M, Shawe-Taylor J. A result of Vapnik with applications[J]. Discrete Applied Mathematics. 1993, 47(3):207–217. DOI: 10.1016/0166-218X(93)90126-9.

[96] Vapnik V N, Chervonenkis A Y. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities[J]. Theory of Probability & Its Applications. jan 1971, 16(2):264–280. DOI: 10.1137/1116025.

[97] Blumer A, Ehrenfeucht A, Haussler D, et al. Learnability and the Vapnik-Chervonenkis dimension[J]. Journal of the ACM. 1989, 36(4):929–965. DOI: 10.1145/76359.76371.

[98] Grunbaum B, Shephard G C. Convex Polytopes: volume 1[M]. [S.l.]: [s.n.], 1969: 126. DOI: 10.1112/blms/1.3.257.

[99] Suk A. A note on order-type homogeneous pointsets[J/OL]. Mathematika. dec 2013, 60(01):37–42. http://arxiv.org/abs/1305.5934. DOI: 10.1112/S0025579313000247.

[100] Grünewald S, Koolen J H, Moulton V, et al. The size of 3-compatible, weakly compatible split systems[J/OL]. Journal of Applied Mathematics and Computing. feb 2012, 40(1-2):249–259. http://link.springer.com/10.1007/s12190-012-0546-z. DOI: 10.1007/s12190-012-0546-z.

[101] Bryant D, Klaere S. The link between segregation and phylogenetic diversity[J]. Journal of Mathematical Biology. 2012, 64(1-2):149–162.

[102] Babai L, Frankl P. Linear Algebra Methods in Combinatorics: With Applications to Geometry and Computer Science[M]. [S.l.]: Department of Computer Science, univ. of Chicag, 1992.

[103] Birkhoff G, Birkhoff G, Birkhoff G, et al. Lattice theory: volume 25[M]. [S.l.]: American Mathematical Society New York, 1948.

[104] Karzanov A, Lomonosov M. Flow systems in undirected networks[J]. Mathematical Programming, Institute for System Studies. 1978, (issue 1 in Russian): 59–66.

[105] Deza M, Laurent M. Facets for the cut cone I[J]. Mathematical Programming. aug 1992, 56(1-3):121–160. DOI: 10.1007/BF01580897.

[106] Avis D, Deza M. The cut cone,L1 embeddability, complexity, and multicommodity flows[J]. Networks. oct 1991, 21(6):595–617. DOI: 10.1002/net.3230210602.

[107] Deza M, Laurent M. Geometry of cuts and metrics[J]. Book. 1997, 15:xii+587. DOI: 10.1007/978-3-642-04295-9.

[108] Erdös P, Rado R. Intersection Theorems for Systems of Sets[J/OL]. Journal of the London Mathematical Society. jan 1960, s1-35(1):85–90. http://jlms.oxfordjournals.org/content/s1-35/1/85.short. DOI: 10.1112/jlms/s1-35.1.85.

[109] Erdos P, Spencer J. Probabilistic methods in combinatorics[J]. AMC. 1974, 10:12.

[110] Alon N, Spencer J H. The probabilistic method[M]. [S.l.]: John Wiley & Sons, 2015.

[111] Razborov A A. Flag algebras[J]. The Journal of Symbolic Logic. 2007, 72 (04):1239–1282.

[112] Razborov A A. Flag algebras: an interim report[M]. [S.l.]: Springer, 2013: 207–232.

[113] Alon N, Naves H, Sudakov B. On the maximum quartet distance between phylogenetic trees[M/OL]. may 2015. http://arxiv.org/abs/1505.04344.

[114] Dress A W, Wenzel W. Valuated matroids[J/OL]. Advances in Mathematics. jun 1992, 93(2):214–250. http://www.sciencedirect.com/science/article/pii/000187089290028J. DOI: 10.1016/0001-8708(92)90028-J.

[115] Joswig M, Sturmfels B, Yu J. Affine Buildings and Tropical Convexity[M/OL]. jun 2007: 22. http://arxiv.org/abs/0706.1918.

[116] Herrmann S, Joswig M, Speyer D E. Dressians, tropical Grassmannians, and their rays[J]. Forum Mathematicum. 2014, 26(6):1853–1881. DOI: 10.1515/forum-2012-0030.

[117] Speyer D, Sturmfels B, Ziegler G M. Advances in Geometry The tropical Grassmannian[J]. Adv. Geom. 2004, 4:389–411. DOI: 10.1515/advg.2004.023.