代数幾何学シンポジウム記録
2007年度 pp.28-41

Tropical Grassmannian and construction of phylogenetic trees

前野 俊昭
京都大学大学院工学研究科

概 要

Phylogenetic tree is a basic object of the evolutionary study. In the mathematical language, phylogenetic tree is a metrized finite graph with labelled leaves. It is known that the space of phylogenetic trees is realized as a tropical Grassmannian. A generalization of phylogenetic tree called phylogenetic network, which permits a particular kind of loops, is used by researchers of evolution. We discuss how the embedding of the tropical Grassmannian into the ambient affine space is interpreted in terms of phylogenetic networks and trees.

系統樹 (phylogenetic tree) は進化生物学の研究において中心的な興味の対象である．系統樹の典型例としては，素朴には図 1 のように種の進化を表したものが思い浮かぶだろう．このようなグラフの各頂点を種の分化が起こった時点に対応させ，時間の経過

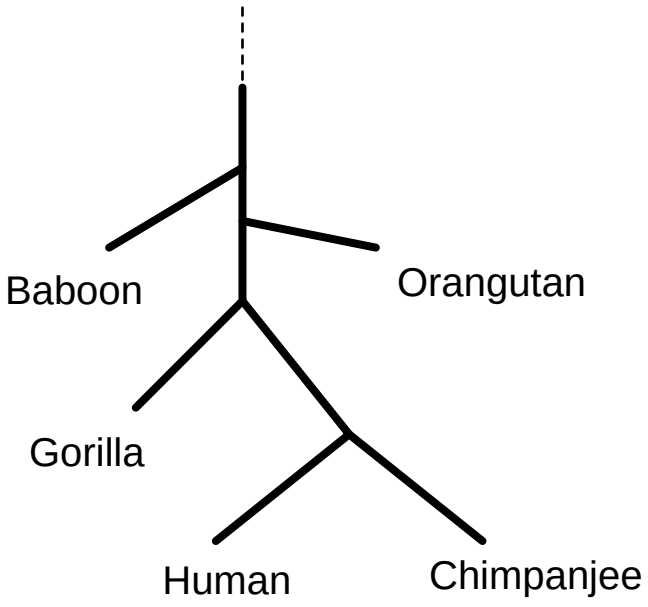

図 1: Phylogenetic tree

を各辺の長さに反映させれば，このグラフは種がどのように分化してきたかの歴史を図式的に表していることになる．しかし当然のことながら，このようなグラフの各頂点は遠い過去の時点に属していて現在の我々には見ることができず，我々が知ることができるのは「葉」の部分のデータのみである．従って，葉の部分に現れたデータから系統樹を推定・再構成することが問題となる．

現代の分子進化学では遺伝子の塩基配列やタンパク質のアミノ酸配列のデータに基いて系統樹を構成する．塩基配列のデータは $ACCTGATAA\cdots$ のように 4 種類の文字からなる有限長さの語 (word) である．従ってこれら配列相互の間の距離を Hamming 距離で測ることができる.[1] これは突然変異による塩基置換を測ることに相当する．こうしたデータ間の距離を復元するような距離付き木を見つけることが問題となっていることである．つまり系統樹の推定問題は，ある有限データの位置関係をグラフとして視覚化する問題として捉えることができるだろう．

ところで，種の分化の時間経過を表す系統樹は「歴史的に」唯一つの二分木として存在するはずである（期待距離系統樹と言う）が，配列データの距離に基く系統樹は時間経過ではなくデータ間の位置関係を反映するものであって意味合いが異なる（実現距離系統樹と言う）ので注意が必要である．さらに，配列データは平行置換や遺伝子組換え (recombination) などの生物学的な撹乱のために，それらの位置関係を木として実現することが不可能になっている場合がある．後に見る incompatible data というものがその状況を表している．このような場合に木としての視覚化をあきらめ，ループを許したグラフによってデータの位置関係を表すという手法が進化学者によって用いられている．このようなグラフを系統ネットワークという．このノートの前半では [5] に従い，系統樹と系統ネットワークの構成について簡単にまとめておく．

進化学のコンテクストを離れて数学的に考えると，今問題となっている系統樹は距離付きの木であるが，このような対象全体のなす空間がどのようなトポロジーを持っているかについては [2] で研究された．葉の個数を止めて考えると系統樹全体のなす空間は polyhedral fan としての構造を持っている．さらに，この空間は tropical Grassmannian $\mathscr{G}_{2,n}$ としてトロピカル幾何的に捉えられることが [7] で示されている．このような枠組みの中で系統樹と系統ネットワークの構成問題を理解することが，このノートの主な目的である．

トロピカル幾何学は，言わば tropical semi-ring 上の代数幾何学である．Tropical semi-ring は実数体に備わっている和と積の演算をそれぞれ max と $+$ で置き換えて，max を「和」，$+$ を「積」と見なしてできる代数系である．こうすると,「和」に関して単位元や逆元の存在は保証されなくなる．このアイデアは，元々は冪等解析の研究においてオートマトンの行列表示を与えるために [6] で導入された．トロピカル幾何学は（実）代数幾何学やトーリック幾何学の超離散化を与えるような新しい幾何学として注目されている．詳しくは [4], [8] 等を参照されたい．

[1] 実際には配列相互の比較は簡単なことではない．データを取ることの技術的な困難もあるし，そもそも配列のどの位置を比較するべきか決定するのは容易な問題ではない．突然変異の過程で文字の置換だけではなく，欠失や挿入も起こりうる．このような状況での文字列のパターンマッチングを最適アライメント問題と言う．[1]

1 Construction of phylogenetic trees

次のように長さ l で $0,1$ からなる文字列 $A_1,\dots,A_n$ が与えられたとする．

$$\begin{array}{lccccc} A_1: & A_1(1) & A_1(2) & \cdots & A_1(l) & \\ \vdots & \vdots & \vdots & \vdots & \vdots & (A_i(j)\in\{0,1\}). \\ A_n: & A_n(1) & A_n(2) & \cdots & A_n(l), & \end{array}$$

このようなデータを配列データと呼ぶことにする．このとき，文字列 A_i と A_j の間の距離 d_{ij} を Hamming 距離で測ることができる．即ち，

$$d_{ij}=\#\{x \mid A_i(x)\neq A_j(x)\}$$

である．問題は，$A_1,\dots,A_n$ でラベル付けされた n 個の外頂点を持ち距離 (d_{ij}) を実現するような木 (tree) T を見つけることである．以下，T がどのように構成されるか具体例で見てみよう．

まず，$x\in\{1,\dots,l\}$ に対し，写像 p_x を次のように定める．

$$\begin{array}{lccc} p_x: & S=\{A_1,\dots,A_n\} & \to & \{0,1\} \\ & A_i & \mapsto & A_i(x) \end{array}$$

各写像 p_x に対して集合 S の分割

$$\pi_x:\ S=p_x^{-1}(0)\cup p_x^{-1}(1)$$

が定まる．簡単のため，$S=\{A_{i_1},\dots,A_{i_k}\}\cup\{A_{i_{k+1}},\dots,A_{i_n}\}$ のような分割を単に $i_1\cdots i_k-i_{k+1}\cdots i_n$ と表そう．例として次のような配列データが与えられたとする．

$$\begin{array}{lccccccccc} A_1: & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ A_2: & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ A_3: & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ A_4: & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ A_5: & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{array}$$

この配列データから定まる分割は次の通りである．

$$\pi_1=1-2345,\ \pi_2=1345-2,\ \pi_3=345-12,$$

$$\pi_4=124-35,\ \pi_5=5-1234,\ \pi_6=12-345,$$

$$\pi_7=12345-\emptyset,\ \pi_8=3-1245,\ \pi_9=1235-4.$$

ここで π_7 は $S=S\cup\emptyset$ という自明な分割であり，T の構成という目的には役に立たない分割である．また，$\pi_1,\pi_2,\pi_5,\pi_8,\pi_9$ は，$S=S_1\cup S_2$ であって，$\#S_1$ or $\#S_2=1$ という形の分割である．このような分割は non-informative な分割と呼ばれる．これ

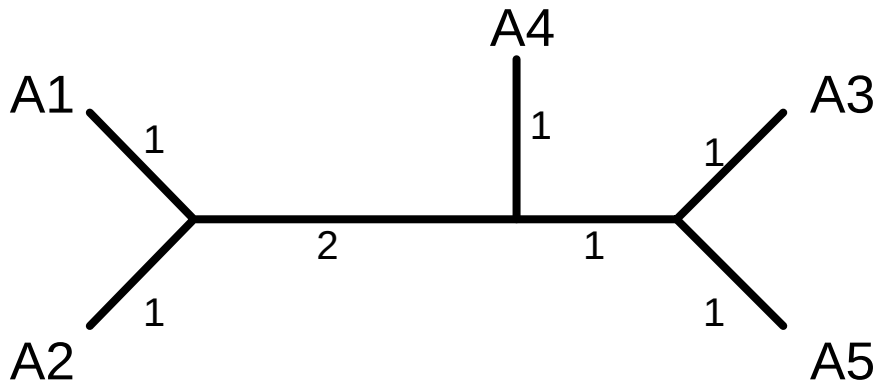

図 2: Phylogenetic tree

は，T のトポロジーを定めるのに必要な情報を持っていないという意味である．但し，このような分割も T の外辺の長さに関わっているので，距離の情報には寄与している．結局，上の配列データから決まる分割のうちで T のトポロジーを定めているのは，$\pi_3 = \pi_6 = 12 - 345$ と $\pi_4 = 124 - 35$ である．距離の情報も考慮に入れれば，T は次のような距離付きの木として定まる．

Remark 1.1 上の構成には対称群 S_l による shuffle の対称性がある．即ち，与えられた配列データ

$$\begin{array}{cccccc} A_1: & A_1(1) & A_1(2) & \cdots & A_1(l) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_n: & A_n(1) & A_n(2) & \cdots & A_n(l), \end{array}$$

の各列を入れ替えてやっても結果として得られる T は不変である．

Definition 1.1 次の条件を満たすような有限木 T を *phylogenetic tree with n operational taxonomic units (OTU,* 操作的分類単位*)* と呼ぶことにしよう．
(1) T は有限個の辺を持つ距離付き木である．
(2) T の外頂点は n 個で，$\{A_1, \ldots, A_n\}$ により一対一にラベル付けられている．
(3) 内頂点の次数は 3 以上．
また，*phylogenetic tree of n OTU's* の集合を $\mathscr{T}_n$ で表し，$T, T' \in \mathscr{T}_n$ が外辺の長さを無視して同一であるとき $T \approx T'$ と定めることで同値関係 $\approx$ を定義する．

Remark 1.2 外辺の長さは 0 になってもよい．この場合は，ある外頂点と内頂点が重なった状態にあると見なすので，対応するラベル A_i が内頂点上に乗ることになる．

2 Construction of phylogenetic networks

第 1 節で構成した phylogenetic tree T では，T の各辺が分割 $\pi_1, \ldots, \pi_l$ に対応している．つまり，分割 $\pi_1, \ldots, \pi_l$ は T の辺のカットにより実現されている．しかしながら，

ある有限集合 S の分割たち $\pi_1, \ldots, \pi_l$ が与えられたとき，これら全てを辺のカットとして実現するような木は常に存在するわけではない．

Definition 2.1 集合 S の分割 $\pi : S = S_0 \cup S_1$, $\pi' : S = S'_0 \cup S'_1$ が *compatible* であるとは，$S_0 \subset S'_0$, $S_0 \subset S'_1$, $S_1 \subset S'_0$, $S_1 \subset S'_0$ のいずれかが満たされていることである．この条件が満たされていないとき，*incompatible* であるという．

次の補題は良く知られている．

Lemma 2.1 集合 S の分割たち $\pi_1, \ldots, \pi_l$ が，ある木 T の辺のカットとして実現されるための必要十分条件は，$\pi_1, \ldots, \pi_l$ が互いに *compatible* であることである．

Definition 2.2 配列データ $A_1, \ldots, A_n$ から定まる分割 $\pi_1, \ldots, \pi_l$ が互いに *compatible* であるとき，配列データ $A_1, \ldots, A_n$ は *compatible* であるという．

上の補題から，与えられた配列データから常に対応する phylogenetic tree が存在するわけではなく，compatible なデータに対してのみ構成することが可能である．それでは，incompatible なデータ $A_1, \ldots, A_n$ の位置関係を図式的に表すためにはどうすれば良いだろうか？ ここで，木を用いることをあきらめ，ある種のループを許したグラフを考えれば incompatible なデータも図式的にうまく整理することができる．それが phylogenetic network のアイデアである．

まず，集合 S の非自明な分割全体のなす集合を Π とし，non-informative (resp. informative) な分割のなす Π の部分集合を Π_0 (resp. Π_1) とする．$\#\Pi_1 = 2^{n-1} - n - 1$ である．分割 $\pi \in \Pi$ に対し，$p_\pi : S \to \{0, 1\}$ を $\pi = \{p^{-1}(0), p^{-1}(1)\}$ となるように定める．但し，$p(A_1) = 0$ となるように選んでおく．与えられた配列データ $S = \{A_1, \ldots, A_n\}$ が定める l 個の分割のうちに，$\pi \in \Pi_1$ が δ_π 個表れ，non-informative な分割 $(\{A_i\}, S \setminus \{A_i\})$ が δ'_i 個現れているとする．線型空間 $\mathbf{R}^{\Pi_1}$ の中で, δ_π を π 成分に持つような点 $(\delta_\pi)_{\pi \in \Pi_1}$ と原点とを結ぶ線を対角線とするような直方体を考え，その頂点と辺からできるグラフを Γ とする．以下，Γ の部分グラフの辺のカットとは，ある辺とその辺に平行な全ての辺を同時に取り除くことを意味する．

Definition 2.3 次の条件を満たすような Γ の部分グラフ Γ_0 を考える．
(0) Γ_0 は連結．
(1) Γ_0 は Γ の頂点 $P_{A_i} := (\delta_\pi \cdot p_\pi(A_i))_{\pi \in \Pi_1}$ を全て含む．
(2) Γ_0 は，任意の i, j に対し P_{A_i} と P_{A_j} を結ぶ Γ の中での最短路を含む．
(3) Γ_0 の辺のカットにより，$\delta_\pi \neq 0$ であるような全ての分割 $\pi \in \Pi_1$ が実現される．
(4) Γ_0 は (0) から (3) をみたす連結部分グラフのうちで極小である．つまり，Γ_0 の真部分グラフは (0) から (3) のいずれかを満たさない．

このような Γ_0 の頂点 P_{A_i} に，ラベル $i \in S$ が付いた長さ δ'_i の外辺を付け加えたものを *phylogenetic network (with n OTU's)* と言う．

Example 2.1

$$\begin{array}{lcccccccc} A_1: & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ A_2: & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ A_3: & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ A_4: & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ A_5: & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}$$

この配列データから定まる分割は

$$\pi_1 = 1235 - 4,\ \pi_2 = 2345 - 1,\ \pi_3 = 2 - 1345,$$

$$\pi_4 = 1234 - 5,\ \pi_5 = 2345 - 1,\ \pi_6 = 345 - 12,$$

$$\pi_7 = 235 - 14,\ \pi_8 = 1234 - 5$$

である．このうち，$\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_8$ は *non-informative* な分割である．残りの π_6, π_7 は互いに *incompatible* な分割であるため，これらを一つの木の辺のカットとして実現することはできない．ここで，図 *3* のように長方形状のループ（トルソー）を導入することを考える．

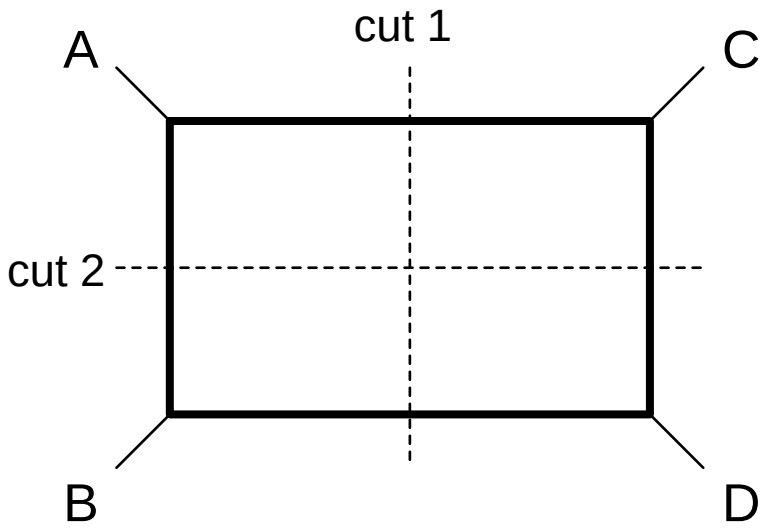

図 3: Cuts of torso

すると，今の例に対応するネットワークは図 *4* のようになる．ここで現れたトルソーの *2* 通りの切断により，*incompatible* な *2* つの分割 π_6, π_7 が実現されている．

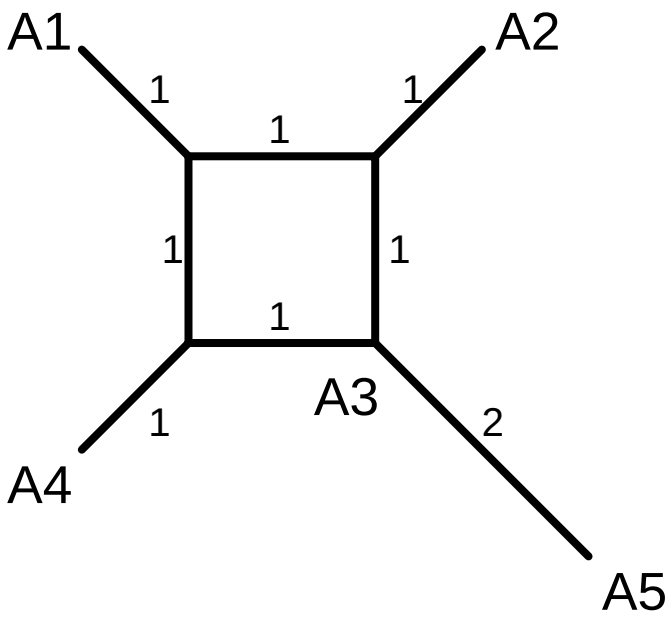

図 4: Phylogenetic network

3 Continuous model

ここまでで構成した phylogenetic tree, phylogenetic networkは，離散データと Hamming 距離に基いているため，得られたグラフの各辺の長さは常に整数値をとっている．ここではより一般に辺の長さが実数値をとるようなグラフを得るために連続的な配列データの概念を導入しておく．任意の正の実数 $\lambda > 0$ に対し，次のような n 個の写像が与えられたとする．

$$\begin{array}{rccc} A_1: & [0,\lambda] & \to & \{0,1\} \\ \vdots & \vdots & & \vdots \\ A_n: & [0,\lambda] & \to & \{0,1\} \end{array}$$

ここで，各 A_i は区間 $[0,\lambda]$ 上有限個の点を除いて連続な写像である．また，ここでは二つの写像 $f, g : [0,\lambda] \to \{0,1\}$ が $[0,\lambda]$ の有限個の点を除いて一致しているときに f と g は同一視されるものとする．このようなデータを離散的な配列データの連続類似とみなすことにする．連続的な配列データ $A_1, \ldots, A_n$ が与えられた時も，$x \in [0,\lambda]$ に対して写像 $p_x : [0,\lambda] \to \{0,1\}$ を考えることにより，$S = \{A_1, \ldots, A_n\}$ の分割 $\pi_x : S = \pi_x^{-1}(0) \cup \pi_x^{-1}(1)$ が得られる．

Definition 3.1 連続的な配列データ $(A_i : [0,\lambda] \to \{0,1\})_{i=1}^n$ が次の性質 $(*)$ を持つような配列データ $(A_i' : [0,\lambda] \to \{0,1\})_{i=1}^n$ と有限個の点での値を除いて一致するとき *compatible* なデータであると言う．

$(*)$: 任意の $x \in [0,\lambda]$ について，$(A'_i : [0,\lambda] \to \{0,1\})_{i=1}^n$ から得られる分割 π_x が *Definition 2.1* の意味で *compatible* である．

A_i と A_j の間の距離 d_{ij} は，Hamming 距離のかわりに A_i と A_j を分離するような分割を生じるセグメントの長さ（測度）m で測ることにする．即ち

$$d_{ij} = m(\{x \in [0,\lambda] \mid \pi_x \text{ separates } A_i \text{ and } A_j\})$$

連続モデルに対しても前節と同様にして phylogenetic network が構成できる．こうして得られた phylogenetic network with n OTU's の集合を $\mathscr{N}$ で表し，外辺の長さを無視する同値関係をやはり $\approx$ で表すことにする．

この場合も shuffle による対称性があることに注意する．区間 $[0,\lambda]$ の中の 4 点 a,b,c,d, $0 \le a < b < c < d \le \lambda$ に対し，写像 $\sigma(a,b;c,d) : [0,\lambda] \to [0,\lambda]$ を，

$$\sigma(a,b;c,d)(x) = \begin{cases} x, & \text{if } x < a \text{ or } x > d, \\ x-a+c, & \text{if } a \le x \le a-c+d, \\ x-a+b+c-d, & \text{if } a-c+d < x < a-b+d, \\ x+b-d, & \text{if } a-b+d \le x \le d \end{cases}$$

と定める．$\sigma(a,b;c,d)$ は 区間 $[a,b]$ と $[c,d]$ を入れ替える写像である．連続的な配列データ $A_1,\ldots,A_n$ に対し，$\sigma(a,b;c,d)$ という形の変換を有限回繰り返しても結果として得られるネットワークは変わらない．

Definition 3.2 配列データ $\alpha = (A_i : [0,\lambda] \to \{0,1\})_{i=1}^n$, $\beta = (B_i : [0,\mu] \to \{0,1\})_{i=1}^n$ が *non-informative* なセグメントの無視と，$\sigma(a,b;c,d)$ の形の変換の有限回の繰り返しで互いに移りあうとき $\alpha \sim \beta$ とすることで同値関係 $\sim$ を定義する．

Remark 3.1 $\alpha \sim \beta$ であって，α が compatile なデータならば β も compatible である．

4 Tropical Grassmannian

体 K は加法付値 $v : K^\times \to \mathbf{R}_{>0}$ の定められた付値体であるとする．即ち，v は $v(xy) = v(x)+v(y)$, $v(x+y) \le \min(v(x),v(y))$ を満たす写像である．また $v(0) = +\infty$ と定めておく．K 上の Laurent 多項式環 $K[X_1^{\pm},\ldots,X_m^{\pm}]$ の元

$$f(X) = \sum_{\underline{n} \in \mathbf{Z}^m} a_{\underline{n}} X^{\underline{n}}$$

に対し，そのトロピカル化 "$f(X)$" を

$$\text{"}f(X)\text{"} = \max_{\underline{n}=(n_1,\ldots,n_m) \in \mathbf{Z}^m} \left(A_{\underline{n}} + \sum_{i=1}^m n_i X_i \right)$$

と定める．ここでは $A_{\underline{n}} = -v(a_{\underline{n}})$ という convention を用いることにする．Laurent 多項式 $f(X) \in K[X_1^{\pm}, \ldots, X_m^{\pm}]$ に対し トロピカル超曲面 $\mathscr{T}(f) \subset \mathbf{R}^m$ は，"$f(X)$" をアフィン空間 $\mathbf{R}^m$ 上の関数と見なしたとき $\max_{\underline{n}}(A_{\underline{n}} + \sum_{i=1}^m n_i X_i)$ の少なくとも 2 つのエントリーで最大値が実現されるような点の集合である．言い換えると 区分線型関数 "$f(X)$" が微分可能でなくなるような locus を意味している．一般に，イデアル $I \subset K[X_1^{\pm}, \ldots, X_m^{\pm}]$ に対し，対応するトロピカル多様体 $\mathscr{T}(I)$ は $f \in I$ が定めるトロピカル超曲面の共通部分

$$\mathscr{T}(I) = \cap_{f \in I} \mathscr{T}(f)$$

と定められる．

Remark 4.1 通常の代数多様体と同様に，イデアル I の中の有限個の元 $f_1, \ldots, f_r$ であって $\mathscr{T}(I) = \cap_{i=1}^r \mathscr{T}(f_i)$ となるようなものが存在する ([3, Theorem 2.9])．[3] では，このような多項式 $f_1, \ldots, f_r$ の組を tropical basis と呼び，与えられたイデアル I に対してそれを計算するアルゴリズムを与えている．

以下では，tropical Grassmannian と呼ばれるトロピカル多様体について考察しよう．$K = \mathbf{R}$ とし，$v(x) = 0, \forall x \in K^{\times}$ という付値が定められているとする．アフィン空間 $\mathbf{R}^{\binom{n}{d}}$ に，$D_{i_1 \cdots i_d}$, $1 \leq i_1 < \cdots < i_d \leq n$ という座標を入れ，

$$\mathbf{R}^{\binom{n}{d}} = \operatorname{Spec} \mathbf{R}[D_{i_1 \cdots i_d} \mid 1 \leq i_1 < \cdots < i_d \leq n]$$

と考える．また，(n, n) 型正方行列のなすベクトル空間を $M_n(\mathbf{R})$ とし，その上の多項式関数のなす環を $\mathbf{R}[M_n]$ とする．各変数 $D_{i_1 \cdots i_d}$ を対応する小行列式に移すことで環準同型

$$\varphi : \mathbf{R}[D_{i_1 \cdots i_d} \mid 1 \leq i_1 < \cdots < i_d \leq n] \to \mathbf{R}[M_n]$$

が得られる．この準同型の核 $I_{d,n} := \operatorname{Ker} \varphi$ は二次の斉次式 (Plücker 関係式) で生成されている．イデアル $I_{d,n}$ を Plücker ideal と呼ぶ．Tropical Grassmannian $\mathscr{G}_{d,n}$ とは，Plücker ideal $I_{d,n}$ が定義するトロピカル多様体

$$\mathscr{G}_{d,n} = \bigcap_{f \in I_{d,n}} \mathscr{T}(f) \subset \mathbf{R}^{\binom{n}{d}}$$

のことである．特に $d = 2$ のとき，$\mathscr{G}_{2,n}$ は Plücker 関係式のトロピカル化

$$"D_{ij}D_{kl} - D_{ik}D_{jl} + D_{il}D_{jk}" = \max(D_{ij} + D_{kl}, D_{ik} + D_{jl}, D_{il} + D_{jk})$$

により定義されたトロピカル超局面 $\mathscr{T}_{ijkl}$ たちの交わりである．トロピカル超局面 $\mathscr{T}_{ijkl} \subset \mathbf{R}^{\binom{n}{d}}$ は，定義により 3 つの線型形式 $D_{ij} + D_{kl}, D_{ik} + D_{jl}, D_{il} + D_{jk}$ のうちの少なくとも 2 つで $\max(D_{ij} + D_{kl}, D_{ik} + D_{jl}, D_{il} + D_{jk})$ の値が実現されている点の集合を意味している．

Tropical Grassmannian $\mathscr{G}_{d,n}$ そのものは我々の問題の観点からすると redundant な情報を含んでいるので，そのような部分を除いた reduction を考えよう．線型写像

$$\begin{array}{rccc} \phi: & \mathbf{R}^n & \to & \mathbf{R}^{\binom{n}{d}} \\ & (a_1,\ldots,a_n) & \mapsto & (a_{i_1}+\cdots+a_{i_d})_{i_1\cdots i_d} \end{array}$$

を導入し，

$$\mathscr{G}''_{d,n} = \text{image of } \mathscr{G}_{d,n} \text{ in } \mathbf{R}^{\binom{n}{d}}/\mathrm{Im}\,\phi$$

と定義する．ここでは [7] と同じ記号を用いている．このように modulo $\mathrm{Im}\,\phi$ をとることは phylogenetic tree の外辺の長さを無視することに相当し，配列データのレベルでは non-informative なデータの寄与を切り落とすことを意味している．$d=2$ の場合，tropical Grassmannian $\mathscr{G}_{2,n}$ と，[2] で調べられた polyhedral complex $\mathscr{T}_n$ との関係が Speyer と Sturmfels [7] によって示された．

Theorem 4.1 (Speyer-Sturmfels [7])

$$\mathscr{T}_n/\approx \;\cong \mathscr{G}''_{2,n}$$

配列データ $(A_i : [0,\lambda] \to \{0,1\})_{i=1}^n$ のなす空間を $\mathscr{A}_n$ とし，そのうち compatible なもののなす部分空間を $\mathscr{C}_n$ で表す．また，$A_1,\ldots,A_n$ の添字 $1,\ldots,n$ を $\mathbf{Z}/n\mathbf{Z}$ の元とみなし，分割 $\pi(i)$, $i \in \mathbf{Z}/n\mathbf{Z}$ を

$$\pi(i) := (\{i, i+1, \ldots, i+[n/2]\}, S \setminus \{i, i+1, \ldots, i+[n/2]\})$$

と定める．配列データの空間 $\mathscr{A}_n$ の中で，条件「$\exists w \in S_n$, $\forall i \in \mathbf{Z}/n\mathbf{Z}$, $\delta_{w(\pi(i))} = 0$」を満たすようなものたちのなす部分空間を $\overline{\mathscr{A}}_n$, このような配列データから構成されるネットワークのなす空間を $\overline{\mathscr{N}}_n$ と表す．

我々の主結果は次の定理である．

Theorem 4.2 次のような図式が存在する．

$$\begin{array}{ccccc} \mathscr{A}_n/\sim & \cong & \mathscr{N}_n/\approx & \cong & \mathbf{R}^{2^{n-1}-n-1} \\ \cup & & \cup & & \cup \\ \overline{\mathscr{A}}_n/\sim & \cong & \overline{\mathscr{N}}_n/\approx & \cong & \mathbf{R}^{\binom{n}{2}}/\mathrm{Im}\,\phi \\ \cup & & \cup & & \cup \\ \mathscr{C}_n/\sim & \cong & \mathscr{T}_n/\approx & \cong & \mathscr{G}''_{2,n} \end{array}$$

Phylogenetic network あるいは tree の推定問題とは，上の定理での配列データの空間と $\mathscr{N}_n/\approx$ との間の同型写像を具体的に構成する問題だと理解できる．また，tropical Grassmannian $\mathscr{G}''_{2,n}$ の $\mathbf{R}^{\binom{n}{2}}/\mathrm{Im}\,\phi$ への埋め込みは配列データの空間 $\overline{\mathscr{A}}_n$ の中への compatible なデータたちの埋め込みだと思うことができる．

Example 4.1 (Tropical Grassmannian $\mathscr{G}_{2,4}$)
Tropical Grassmannian $\mathscr{G}_{2,4} \subset \mathbf{R}^6 = \mathbf{R}^{\binom{4}{2}}$ は，Plücker 関係式のトロピカル化 $\max(D_{12}+D_{34}, D_{13}+D_{24}, D_{14}+D_{23})$ で定義されるトロピカル超曲面で，図 5 の太線部分と $\mathbf{R}^4$ との直積の構造を持っている．線型写像 $\phi : \mathbf{R}^4 \to \mathbf{R}^6$ の像で modulo をとることは，ちょうどファイバーの $\mathbf{R}^4$ をつぶすことに相当する．つまり，図 5 の太線部分の 3 本の半直線からなる集合が $\mathscr{G}''_{2,4}$ である．Tropical Grassmannian $\mathscr{G}_{2,4}$ は図 6 のような 4 つの外頂点を持つ tree をパラメトライズしている．射影 $\phi : \mathscr{G}_{2,4} \to \mathscr{G}''_{2,4}$ のファイバー $\mathbf{R}^4$ は 4 本の外辺の長さをパラメトライズしており，$\mathscr{G}''_{2,4}$ の 3 本の半直線は内辺の長さ t をパラメトライズしている．3 本の半直線はそれぞれ外辺のラベル $\{1,2,3,4\}$ の 4 通りの配置に対応している．このことから，$\mathscr{G}''_{2,4}$ は種数 0 の 4 点付き安定曲線のモジュライ空間 $\overline{\mathscr{M}}_{0,4}$ のトロピカル版 $\overline{\mathscr{M}}^{trop}_{0,4}$ と同じものと見なすことができる．

$n = 4$ のときは $\mathscr{N}_4 = \overline{\mathscr{N}}_4$ であり，$\mathscr{G}''_{2,4}$ の補集合の 3 つの連結成分は $\mathscr{N}_4 \setminus \mathscr{T}_4$ に属するネットワークの 3 通りのトポロジカルタイプに対応している．このことから，Plučker ideal $I_{2,4}$ の Gröbner fan に属する各 cone が phylogenetic network のトポロジカルタイプでラベルされていることがわかる．

Example 4.2 (Tropical Grassmannian $\mathscr{G}''_{2,5}$)
Tropical Grassmannian $\mathscr{G}''_{2,5} \subset \mathbf{R}^5 = \mathbf{R}^{\binom{5}{2}-5}$ は，Petersen graph（図 7）上の cone であることが知られている [9]. Petersen gragh の 10 個の頂点は，$ij - klm$, $\{i,j,k,l,m\} = \{1,2,3,4,5\}$, という形の分割に対応する tree たちをパラメトライズしている．今，図 7 に示された頂点 A, B, C, D, E に注目し，それぞれ $A : 13-245$, $B : 35-142$, $C : 25-134$, $D : 24-135$, $E : 14-235$ という分割に対応しているとしよう．$\mathbf{R}^5$ の原点とこれらの頂点を結んでできる半直線たちが張る cone を考えると，この cone の内点は図 8 のようなネットワークに対応している．

参考文献

[1] 阿久津達也,「バイオインフォマティクスにおける基本アルゴリズム」(山西芳裕 記), シンポジウム「第一回 数学者のための分子生物学入門」，物性研究 **81-1** (2003), 120-129.

[2] R. Billera, S. Holmes and K. Vogtman, *Geometry of the space of phylogenetic trees,* Adv. in Applied Math. **27** (2001), 733-767.

[3] T. Bogart, A. N. Jensen, D. Speyer, B. Sturmfels and R. R. Thomas, *Computing tropical varieties,* J. Symbolic Comput. **42** (2007), 54-73.

[4] J. Richter-Gebert, B. Sturmfels and Th. Theobald, *First steps in tropical geometry,* Idempotent mathematics and mathematical physics, 289-317, Contemp. Math., **377**, Amer. Math. Soc., Providence, RI, 2005.

[5] 斎藤成也,「系統関係を乱す生物学的要因と系統ネットワーク」(磯崎泰樹 記), シンポジウム「第二回 数学者のための分子生物学入門」講演記録, 27-53, 2003.

[6] I. Simon, *Limited subsets of a free monoid,* In: Proc. 19th Annual Symposium on Foundations of Computer Science, Piscataway, N.J., Institute of Electrical and Electronics Engineers, 1978, 143-150.

[7] D. Speyer and B. Sturmfels, *The tropical Grassmannian,* Adv. Geom. **4** (2004), no.3, 389-411.

[8] D. Speyer and B. Sturmfels, *Tropical Mathematics,* preprint, math.CO/0408099.

[9] B. Sturmfels, *Solving Systems of Polynomial Equations,* CBMS Regional Conference Series in Math., **97**, AMS, 2002.

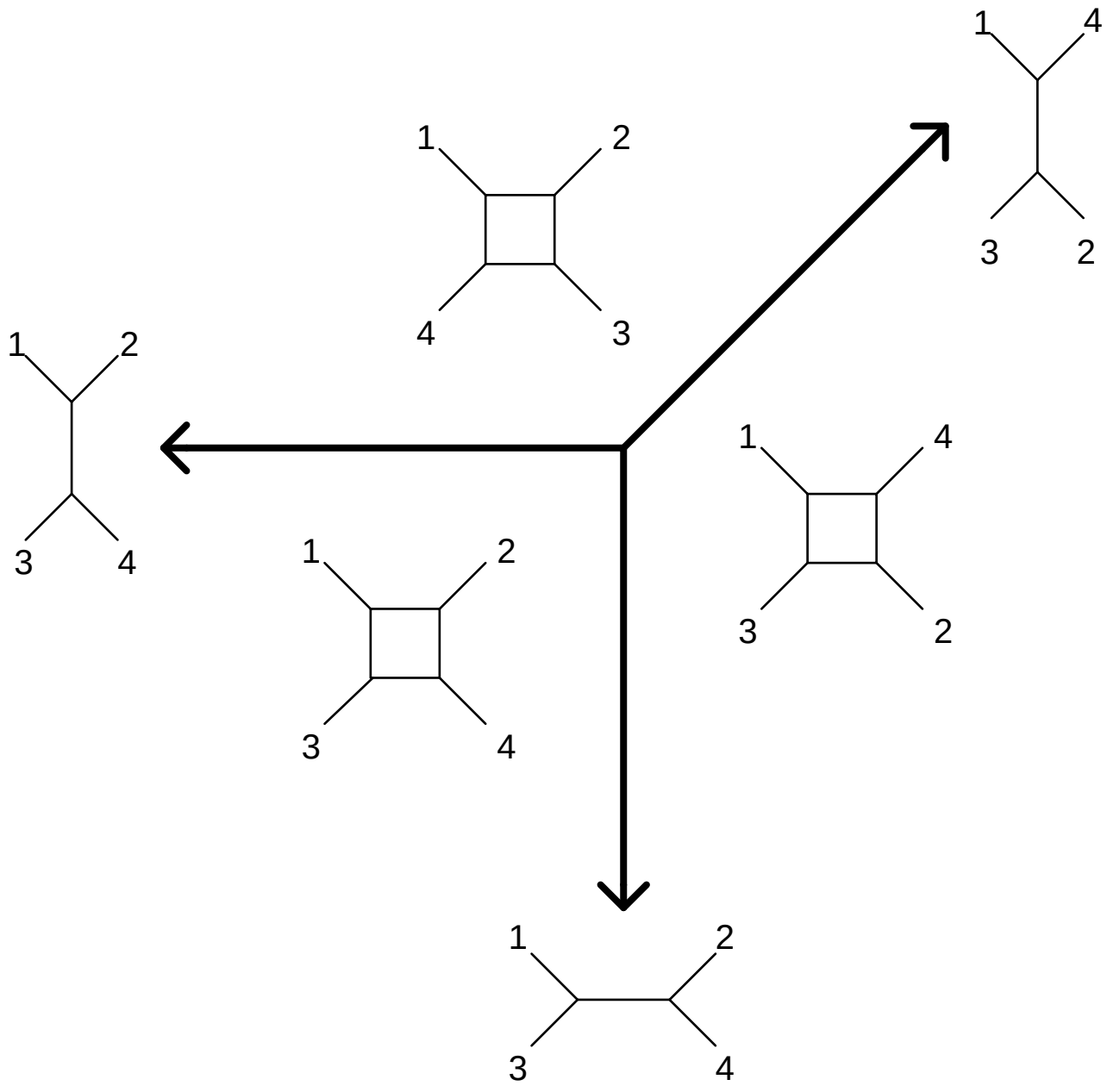

図 5: reduced tropical Grassmannian $\mathscr{G}''_{2,4}$

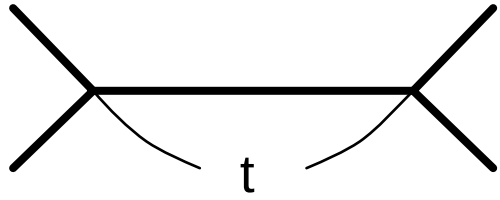

図 6: Phylogenetic tree parametrized by $\mathscr{G}_{2,4}$

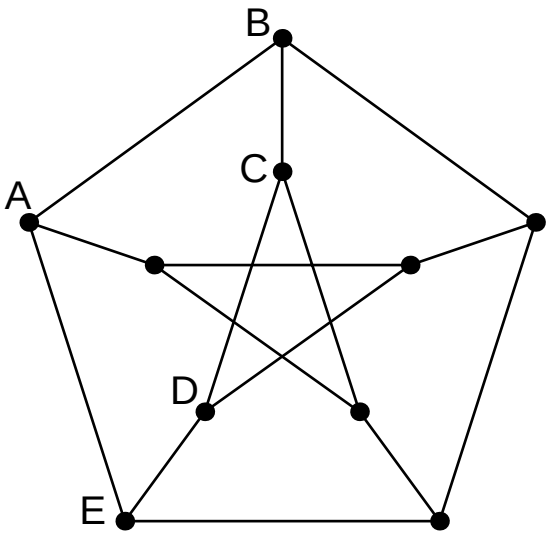

図 7: Petersen graph

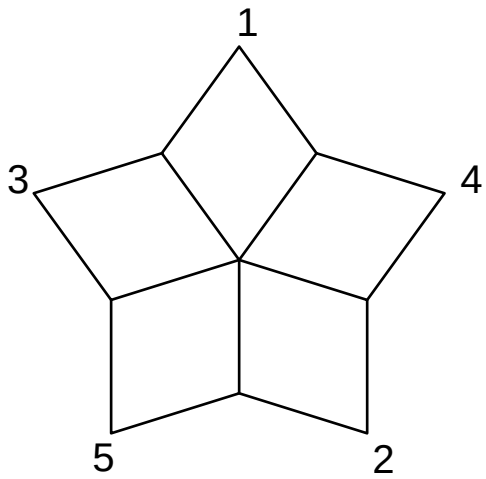

図 8: Network in $\overline{\mathscr{N}_5}$