統計数理(2012)
第 60 巻 第 2 号 279–288

特集「多様性と進化の統計解析」
[研究詳解]

分子系統学における代数的方法

Ruriko Yoshida† (訳：間野修平)

(受付 2011 年 12 月 26 日；改訂 2012 年 2 月 9 日；採択 2 月 14 日)

要　　旨

近年，現代生物学と高等数学の間で多くの共同研究がなされている．成長しつつある代数統計学は，組み合わせ論，計算代数，多面体幾何の方法を統計的計算とモデリングに適用する分野であるが，計算生物学との間に多くの重要な関係が確立している．系統学は，系統学的不変量，樹空間の幾何，系統の再構築の解析など，代数統計学の豊富な応用対象を提供してきた．本詳解の目的は，この分野の導入，さらに知るための網羅的ではない手引き，さらに代数的手法が分子系統学の文脈で用いられている具体的な事例をいくつか与えることにある．

キーワード：代数統計学，系統学．

1. 導入

計算生物学において，微生物やヒトゲノムの進化の基礎的理解は重要であり，公衆衛生，医科学，生態学，自然保護に応用される(例えば，Pedersen, 2006; Pillay et al., 2007; Friesen et al., 2006; Craven et al., 2001; Edwards et al., 2007; Haarmann et al., 2005; Barluenga et al., 2006)．

与えられたアラインメントから系統樹を推測する手法には，最尤法，距離に基づく方法，節約法，ベイズ法など，いろいろある(Felsenstein, 2003)．

樹の再構築へのアプローチとして最も確立しているのは最尤法である．最尤法では相同なDNA 配列の進化を系統樹上の連続時間マルコフ連鎖により記述する．連続時間マルコフ連鎖は置換行列により特徴づけられ，系統樹は種の関係を枝の長さ(分岐時間)と共通祖先によって要約する．DNA 配列は葉ノードでのみ観察され，樹上の情報，置換事象(時間とタイプ)，枝の長さは欠損している．連続時間マルコフ過程の推移行列 $P(t)$ はパラメタを含む置換率行列 Q により $\exp(Qt)$ と表せる．固定した樹の置換率のパラメタと樹の枝の長さを観察データに基づいて推定するためには，EM (expectation maximization)アルゴリズムが用いられ(Hobolth and Jensen, 2005)，その更新ステップは Q の固有値と固有ベクトルにより陽に表せる．この統計モデルを理解するためには(例えば，識別性の証明)計算代数の手法が使われる．なぜなら，この進化モデル，すなわち系統樹を再構築するための置換率行列 Q をもつ連続時間マルコフ過程は，代数統計モデルだからである．

連続時間マルコフ過程は任意の配列の集合の間に距離の測度を与える．配列を関係づける系統樹を再構築するために配列の対ごとの距離を使うこともでき，すべての対の間の距離の集合から系統樹を再構築する方法である距離に基づく方法が用いられる．アラインメントから計算されたすべてのペアごとの距離の集合は，通常は樹の計量を与えない．ここで，樹の計量とは，樹の構造をもつ距離行列を指す．それゆえ，距離に基づく方法では，何らかの基準を用いてア

† Department of Statistics, University of Kentucky, Lexington, KY 4050, U.S.A.

ラインメントから計算された対ごとの距離の集合に最も近い樹の計量を探すことを目指す．距離に基づく手法は幾何と組み合わせ論に関係する．実際，樹の空間，すなわち，すべての距離行列の集合の上のすべての樹の計量の集合は多面体錐の和集合として記述できる．

生命科学における問題の増大する複雑さは，数理科学，統計科学に新しい概念と計算手法を要求する．近年，古くは応用数学の一部とは考えられていなかったいくつかの数学の分野が，様々な生物学的問題に関与している．そのような分野の一つが代数学，とくに記号計算の方法論に基づく計算的視点である．ほんの数例をあげれば，DNA とタンパク配列データ(Pachter and Sturmfels, 2004a; Levy et al., 2006)，RNA の 2 次構造(Heitsch and Condon, 2002)，ウイルスの会合(Sitharam and Agbandje-McKenna, 2006)，細胞生化学ネットワークのモデリング(Laubenbacher and Stigler, 2004; Jarrah et al., 2007)，さらに代謝ネットワークを確認する代数モデル(Mishra and Mysore, 2007)がある．代数生物学で用いられる代数的手法は，グレブナ基底，有限体，多面体幾何，特異点解消のテクニックにおよぶが，ここではこれらの応用分野を 3 つの簡単な例により紹介する．

現存の種の集合の間の祖先の関係を決定するにあたって 1 つの共通するアプローチは，モデルに基づく系統学を用いることである．この設定において，統計モデルは現存種の DNA のアラインメントの列を観察するとき期待される確率を表現するのに用いられる．代数学が関わるのは，系統学的モデルの確率はモデルのパラメタの多項式だからである．それゆえ，系統学的モデルは多次元確率単体の半代数的部分集合(本稿では不等式の解を含めるとき“半”代数的という)である．これらのモデルを解析するための代数的テクニックのひとつは，モデルの系統学的不変量(Allman and Rhodes, 2003; Sturmfels and Sullivant, 2005)，すなわち，モデルに属する任意の確率分布により満たされなくてはならない同時確率分布上の多項式の等式の研究である．これらの不変量は，系統樹モデルの適合度の検定のための統計量の構築に使うことができる(Eriksson, 2005; Casanellas and Fernández-Sánchez, 2007)．

系統学への代数的アプローチは代数統計学の一分野であり，計算生物学に多くの応用がある．これらの関係はよく引用される書籍 Algebraic Statistics for Computational Biology (Pachter and Sturmfels, 2005)に強調されている．代数統計は多くの統計モデルが(半)代数的集合であるという観察に基づいて構築されている．その幾何とこれらの代数的集合上の等式の研究は統計的推測に有用である．このアプローチは系統学に応用され，HIV における薬剤抵抗性亢進の推測(Beerenwinkel and Sullivant, 2009)，配列のアラインメントのパラメトリックな挙動の判定(Pachter and Sturmfels, 2004a)，適合度地形の幾何の研究(Beerenwinkel et al., 2007)などがある．代数統計学には生物学以外にも多くの応用があり，ほんの数例をあげると，データ開示抑制(Fienberg and Slavkovic, 2004)，実験計画(Pistone et al., 2001)，対数線形モデルの仮説検定(Diaconis and Sturmfels, 1998)，最尤推定(Buot and Richards, 2006; Catanese et al., 2006)，ベイズ統計における積分計算の近似(Watanabe, 2001; Rusakov and Geiger, 2005)などがある．Drton and Sullivant (2007)には，指数分布族の文脈で代数統計モデル研究の一般的な枠組みが提示されている．離散確率変数のほぼすべての統計モデルが対象になり，多くの連続確率変数のモデルもこの方法で扱うことができる．したがって，これらの代数統計的テクニックは計算，数理生物学のさらに多くの領域で有用と思われる．

この詳解では，さらに知るための網羅的ではない手引き，さらに代数的手法が分子系統学の文脈で用いられている具体的な事例をいくつか与える．

2. 事例研究

本節では特に 3 つの話題について詳しく述べる．1) ある進化モデルの下での系統学的不変

量，2）樹の空間の幾何学，3）balanced minimum evolution（BME）法の幾何学的視点，である．BME 法は距離に基づく方法の 1 つで，距離に基づく方法のうち最もよく使われるものの 1 つである近隣結合（Neighbor-Joining）法は BME の貪欲法である．

2.1 系統学的不変量

系統学的不変量は代数生物学においてよく研究されている分野である（Allman and Rhodes, 2007 とその引用文献）．

T を n の葉ノードをもつ有根樹とし，$\mathcal{V}(T)$ を T のノードの集合とする．それぞれのノード $v\in\mathcal{V}(T)$ について X_v を k 通りの異なる値をとる離散確率変数とする．X_v が状態 i をとる確率 $P(X_v=i)$ を考えよう．π を根 r の確率変数 X_r の分布とする．それぞれのノード $v\in\mathcal{V}\backslash\{r\}$ について，$a(v)$ を v の一意に定まる親とする．$a(v)$ から v への推移を $k\times k$ 行列 $A^{(v)}$ とするとき，それぞれのノードの確率分布は以下の規則で再帰的に計算できる．

$$P(X_v=j)=\sum_{i=1}^{k}A^{(v)}_{ij}\cdot P(X_{a(v)}=i). \tag{2.1}$$

この規則はすべての確率変数 X_v の同時分布を与える．T の葉ノードを $1,2,\ldots,n$, と名付け，葉ノードの変数の周辺分布を次のように略記する．

$$p_{i_1i_2\ldots i_n}=P(X_1=i_1,X_2=i_2,\ldots,X_n=i_n). \tag{2.2}$$

モデルの系統学的不変量とは，葉ノードの確率 $p_{i_1i_2\ldots i_n}$ の多項式で，パラメタの選び方によらず 0 になるものをいう．これらの多項式の集合は不定元 $p_{i_1i_2\ldots i_n}$ 上の多項式環の素イデアルをなす（Allman and Rhodes, 2003; Sturmfels and Sullivant, 2005）．さらに，次の定理が成り立つ．

定理 1.（Sturmfels and Sullivant, 2005）系統樹 T 上の群に基づく任意のモデルにおいて，系統学的不変量の素イデアルは，T のそれぞれの内部ノードの周りの部分モデルの不変量と，T の任意の枝における分離が条件付き独立であることを保証する 2 次多項式により生成される．

ここで，群に基づくモデルとは，置換率行列が定める離散状態をとる連続時間マルコフモデルにおいて，パラメータがある種の対称性を持つものを指す（Pachter and Sturmfels, 2005）．

Casanellas et al.（2005）は，様々な異なるモデルの下で，系統学的不変量のアイディアを用いて小さい樹のカタログについて議論した．ここで，小さい樹とは 5 個以下の葉ノードをもつ樹を指す．その章と付随するウェブサイトは，小さい樹の上の異なる系統学的モデルの様々な代数的特徴を統一した形式で扱えることを目指している．

2.2 系統樹の幾何学

Billera et al.（2001）は，固定した葉ノードをもつすべての系統樹の集合を幾何学的にモデル化する連続空間について述べた．多面体幾何に関する定義の詳細は Grünbaum（2003）, Sturmfels（1996）, Ziegler（1995）を参照．

T を有限集合 $X:=\{1,2,\ldots,n\}$ によりラベルされた葉ノードをもつ重み（長さ）つき有根樹とする．これは写像 $d\colon\binom{X}{2}\to\mathbb{R}$ を定義する．ここで $\binom{X}{2}$ はすべての葉ノードのラベルの順序を区別しない対の集合を指す．d を葉ノードの対の距離とする（すなわち，非類似度写像 $d\colon X\times X\to\mathbb{R}$, ここで $d(x,x)=0$，$d_{x,y}=d_{y,x}$ である）．写像 $d\colon\binom{X}{2}\to\mathbb{R}$ がすべての $i,j,k\in X$ について

$$d_{ij}\le\max\{d_{ik},d_{jk}\}$$

を満たすとき，その写像を超計量（ultrametric）という．d は $\{d_{ij},d_{ik},d_{jk}\}$ の最大が少なくとも 2 度達成されるとき超計量であるといってもよい．超計量を考察するのは，分子時計（置換率が

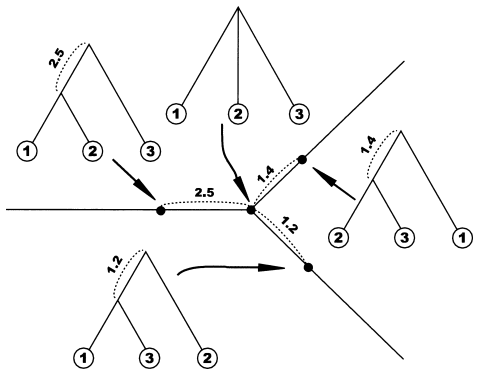

図 1. $n=3$ の樹の空間．それぞれの錐について，頂点から錐の点までの距離は対応する樹の内枝の長さを表す．それぞれの半直線が錐を表す．

一定であるがゆえに枝の長さで時間を測れること）が生物学的に重要な仮説だからである．分子時計を仮定すると，現在観察されるすべての生物種は共通祖先から同じ時間を経過しているため，系統樹は超計量をもつ．これはトロピカル超曲面を定義し（Pachter and Sturmfels, 2004b），この空間は $\mathbb{R}^{\binom{n}{2}}$ の部分集合として

$$T_n = \{d \in \mathbb{R}^{\binom{n}{2}} : d \text{ は超計量 }\}$$

のように与えられる．これは線形な等・不等式の解の共通部分を持たない集合の和集合（共通部分を持たない錐）である．T_n は $\mathbb{R}_+ \times \mathbb{R}^{n-2}_{\geq 0}$ に同型の錐の和集合である．すなわち，超計量の樹は，根までの距離を固定すれば内枝の長さのみで決まる（例えば，$n=3$ では樹のトポロジーはただ 1 つのパラメタである内枝の長さで定まる）．それゆえ錐 $\mathbb{R}^{n-2}_{\geq 0}$ は内枝の重みをパラメトライズする．例えば，T_3 は図 1 の 3 つの錐の交わりであり，トロピカル幾何では直線である．

同様に無根系統樹の幾何も定義できるが，その前に非類似度写像を定義しなくてはならない．

定義 1. $\{1,2,\ldots,n\}$ 上の非類似度写像 $\{d_{ij}\}_{i,j=1}^n$ は，$n \times n$ 対称行列で，対角成分は 0，その他のすべての成分は正である．

樹の計量を 4 点条件で定義できる．非類似度写像 d は三角不等式

$$d(i,j) \leq d(i,k) + d(k,j), \qquad i,j,k \in X := \{1,2,\ldots,n\} \tag{2.3}$$

を満たすとき計量とよばれるが，計量 d が，X でラベルされる葉を持ち，各枝が非負の長さを持ち，全ての葉 $x, y \in X$ について葉 x から葉 y の一意の経路の長さが $d(x,y)$ に等しい樹が存在するとき，樹の計量とよばれる．

定理 2.（4 点条件, Buneman, 1971） d を非類似度写像とする．d は，すべての可能な異なる葉ノード i,j,k,l について，$\{d_{ij}+d_{kl}, d_{ik}+d_{jl}, d_{il}+d_{jk}\}$ の最大が少なくとも 2 度達成されるとき，その時に限り，d は樹の計量である．

定理 2 より，$\{1,2,\ldots,n\}$ 上の樹の計量は $\mathbb{R}_+^{\binom{n}{2}}$ 中の錐の和集合として実現される（Pachter and Sturmfels, 2005）．

Owen and Provan (2011) は，樹の空間における 2 つの樹の測地線距離の計算（Billera et al., 2001）は多項式時間で可能であることを示した．また，Chakerian and Holmes (2010) は，樹の

測地線距離の系統樹と階層的クラスタリングの樹の評価への応用を示した.

2.3 近隣結合法の最適性

Gascuel and Steel (2006)は，最もよく使われる無根系統樹再構築アルゴリズムの 1 つである近隣結合法(Saitou and Nei, 1987)が非類似度写像により定まる BME 樹を見つけるための貪欲法であることを示した．Eickmeyer et al. (2008)は $n \leq 8$ の場合について多面体分割を比較した.

$d = \{d_{ij}\}_{i,j=1}^{n}$ を非類似度写像とする(対角成分が 0 の非負，実の要素をもつ $n \times n$ 行列である)．BME 問題は次を最小化する樹 T を見つけることである.

$$\frac{1}{|o(T)|} \sum_{(x_1,\ldots,x_n)\in o(T)} \left[\frac{1}{2}\sum_{i=1}^{n} d_{x_i x_{i+1}}\right], \qquad x_{n+1} = x_1. \tag{2.4}$$

ここで $o(T)$ は T の平面への埋め込みにおいて現われる葉ノードのすべての巡回置換の集合である．樹 T の葉ノード i と j を結ぶ経路にある内部ノードの集合を p_{ij}^T とかくと，(2.3)は次を最小化することに等しい.

$$\sum_{ij} \lambda_{ij}^T d_{ij}. \tag{2.5}$$

ここで，$i \neq j$ のとき $\lambda_{ij}^T = \prod_{v\in p_{ij}^T} (\deg(v)-1)^{-1}$，$\lambda_{ii}^T = 0$ である．これは NP 困難な線形計画問題(Day, 1987)であるが，その重要性は次の定理からわかる.

定義 2. T を n の葉ノードを持つ樹とし，$l : E(T) \to \mathbb{R}_+$ で各枝に長さを与えるとする．l による T の長さは次で定義される.

$$l(T) = \sum_{e\in E(T)} l(e).$$

定理 3. (Desper and Gascuel, 2004) T を樹長が $l : E(T) \to \mathbb{R}_+$ で与えられ，非類似度行列 $d = \{d_{ij}\}_{i,j=1}^{n}$ をもつ 2 分木とする．d_{ij} の分散が $2^{|p_{ij}^T|}$ に比例する(ある定数 c について $\mathrm{var}(d_{ij}) = c2^{|p_{ij}^T|}$)と仮定すれば，(2.4)を最小化したものは T の最小分散樹長推定量である．さらにそれは重み付き最小 2 乗樹長推定量に等しい.

この結果は(2.4)の方程式の最小化に重み付き最小 2 乗法を使う根拠になり，BME ポリトープを理解することの重要性を強調する.

定義 3. BME ポリトープとは次のベクトルの凸包である.

$$\{[\lambda_{12}^T, \lambda_{13}^T, \ldots, \lambda_{ij}^T, \ldots, \lambda_{n-1,n}^T] : T \text{ は葉ノードを } n \text{ 持つ樹 }\}.$$

ラベルづけられた樹のトポロジーは，各々の葉ノードを n もつ樹(T とする)にベクトル λ^T を与える．ここではそれを BME ベクトルとよぶ．任意の非類似度写像について，(2.4)を最小化する樹は BME ポリトープの頂点であり，常に 2 分木であるが，そのような樹では $\lambda_{ij}^T = 2^{-|p_{ij}^T|}$ である(Pauplin の公式という．Semple and Steel, 2004)．BME polytope は $\mathbb{R}^{\binom{n}{2}}$ にあり，次元は $\binom{n}{2} - 2$ である．BME ポリトープの正規扇(normal fan)は，非類似度写像 $\mathbb{R}_+^{\binom{n}{2}}$ の空間の多面体分割として BME 錐を導き，それらは，各々の樹 T について，T が(2.4)を最小化するような非類似写像の集合を表す.

近隣結合法は(2.4)の近似解を得るための貪欲法である．ここではアルゴリズムの詳細(Gascuel and Steel, 2006 を参照)は省くが，葉ノードの対をとり結合することを繰り返すものである．葉ノードの対をとるにあたっての選択基準は非類似度写像について線形(Bryant, 2005)という著しい性質がある．それゆえ，近隣結合法は，ノードの対ごとの距離が線形不等式を満たすとき，

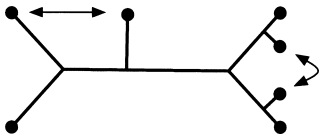

図 2. 7 次の BME ポリトープにおいて辺で結ばれない頂点の対．2 つの樹は，それらの樹が 3 つのサクランボをもち，図に示す 2 つの葉ノードの交換について異なるとき，またそのときに限り，辺で結ばれない．

またそのときに限り，特定の樹を出力する．ここで，線形不等式の解の集合は $\mathbb{R}^{\binom{n}{2}}$ の多面体錐を形成するが，これを近隣結合錐とよぶ．それゆえ，近隣結合法は，距離データが近隣結合錐の和集合にあるとき，またそのときに限り，特定の樹 T を出力する．Eickmeyer and Yoshida (2008) は，近隣結合錐は，$\mathbb{R}^{\binom{n}{2}}$ を分割するが，扇を形成しないことを示した．このことは近隣結合法の挙動において重要な意味があり，Eickmeyer and Yoshida (2008) はそれを $n \leq 7$ について調べた．

Eickmeyer et al. (2008) は葉ノードの数が 8 以下の無根系統樹について BME ポリトープを計算論的に調べ，さらに次の一般的補題を示した．

補題 1. (Eickmeyer et al., 2008 の補題 3.1) 任意の葉ノードの数 n について，BME ポリトープの頂点は，n の葉ノードを持つすべての無根 2 分木系統樹の BME ベクトルに対応する．星状系統樹 (すべての葉ノードとただ一つの内部ノードの間にのみ辺をもつ樹) の BME ベクトルは BME ポリトープの内部にあり，他のすべての BME ベクトルは BME ポリトープの境界にある．

Eickmeyer et al. (2008) は BME ポリトープの辺について計算論的に調べ，$n(\geq 6)$ では BME ポリトープの辺がなすグラフは n のノードをもつ完全グラフであることを示した．このことは，葉ノードの数が等しい樹の対について，常に 2 つの樹が (唯一の) 共に最適合な BME 樹になるような非類似写像があることを意味する．しかし，$n = 7$ のときは，BME ポリトープは辺で結ばれない組み合わせ論的な頂点の対をもつ．7 つの葉ノードと 3 つのサクランボ (2 つの葉ノードと 1 つの隣接する内点からなる部分樹) をもつ 2 つの 2 分木は，図 2 に示す 2 つの葉ノードの交換で関係づけられるとき，またそのときに限り，辺にならない．

Haws et al. (2011) は，任意の 2 分木から他の 2 分木への SPR (subtree-prune-regraft) 移動は BME ポリトープの辺に対応することを証明した．これはどのような 2 分木から別の 2 分木への NNI (nearest-neighbor-interchange) 移動も BME ポリトープの辺に対応することを示し，結果として，ソフトウェア `FastME` に実装されている BME 近似法は，NNI 移動により BME ポリトープの辺を移動するものである．さらに，彼らは，接続されていない系統群 (clade) によりパラメトライズされる BME ポリトープの面のすべてを定義し，記述した．

3. さらに知るために

モデルの識別可能性の証明は推測において重要である．進化モデル，混合モデルについて，代数学からの方法が使われてきた．最近の例として，Allman et al. (2008, 2011), Chai and Housworth (2011), Rhodes and Sullivant (2011), Sullivant (2011) がある．

幾何の視点から測地線的距離を調べた仕事がいくつかある (例えば Ardila et al., 2012 を参照)．

BME ポリトープの構造の辺と面による理解は，最適な BME 樹を見つけるための新しい最

適化戦略の開発に役立つかもしれない．例えば，BME 法は BME ポリトープ上の線形計画問題だから，BME ポリトープの辺の移動を用いるアプローチも可能だろう．また，Haws et al. (2011) は，BME ポリトープの未解決問題を列挙している．例えば，任意の 2 分木から 2 分木への TBR (tree-bisection-regrafting) 移動も BME ポリトープに対応する，すなわち，もし 2 つの 2 分木 T_1 と T_2 が TBR 移動により隣接であれば，λ^{T_1} と λ^{T_2} は BME ポリトープの辺をなす，という予想がある．

Haws et al. (2011) に関連して，Bordewich et al. (2009) は SPR 移動による山登り法によって，どのような非類似度写像の入力についても BME 最適解を得ることができることを示した．このことは，BME 法と SPR 移動が無矛盾であることを示す．さらに，TBR 移動は SPR 移動の一般化だから，このことは BME 法が TBR 移動と無矛盾であることを示唆する．BME 法が NNI 移動と無矛盾であることを示すのも興味深い．

参 考 文 献

Allman, E. and Rhodes, J. (2003). Phylogenetic invariants for the general Markov model of sequence mutations, *Mathematical Biosciences*, **186**, 133–144.

Allman, E. and Rhodes, J. (2007). Phylogenetic invariants, *Reconstructing Evolution*: *New Mathematical and Computational Advances* (eds. O. Gascuel and M. Steel), Chapter 8, Oxford University Press, Oxford, U.K.

Allman, E., Ané, C. and Rhodes, J. (2008). Identifiability of a Markovian model of molecular evolution with Gamma-distributed rates, *Advances in Applied Probability*, **40**, 229–249.

Allman, E., Petrović, S., Rhodes, J. A. and Sullivant, S. (2011). Identifiability of two-tree mixtures for group-based models, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **8**, 710–722.

Ardila, F., Owen, M. and Sullivant, S. (2012). Geodesics in CAT(0) cubical complexes, *Advances in Applied Mathematics*, **48** (1), 142–163.

Barluenga, M., Stölting, K. N., Muschick, W. S. M. and Meyer, A. (2006). Sympatric speciation in Nicaraguan crater lake cichlid fish, *Nature*, **439**, 719–723.

Beerenwinkel, N. and Sullivant, S. (2009). Markov models for accumulating mutations, *Biometrika*, **96**, 645–661.

Beerenwinkel, N., Pachter, L. and Sturmfels, B. (2007). Epistasis and the shapes of fitness landscapes, *Statistica Sinica*, **17**, 1317–1342.

Billera, L. J., Holmes, S. P. and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees, *Advances in Applied Mathematics*, **27**, 733–767.

Bordewich, M., Gascuel, O., Huber, K. T. and Moulton, V. (2009). Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **6**, 110–117.

Bryant, D. (2005). On the uniqueness of the selection criterion in neighbor-joining, *Journal of Classification*, **22**, 3–15.

Buneman, P. (1971). The recovery of trees from measures of similarity, *Mathematics of the Archaeological and Historical Sciences* (eds. F. R. Hodson, D. G. Kendall and P. Tautu), 387–395, Edinburgh University Press, Edinburgh, U.K.

Buot, M. G. and Richards, D. P. (2006). Counting and locating the solutions of polynomial systems of maximum likelihood equations, *Journal of Symbolic Computation*, **41**, 234–244.

Casanellas, M. and Fernández-Sánchez, J. (2007). Performance of a new invariants method on homogeneous and non-homogeneous quartet trees, *Molecular Biology and Evolution*, **24**, 288–293.

Casanellas, M., Garcia, L. and Sullivant, S. (2005). Catalog of small trees, *Algebraic Statistics for Computational Biology* (eds. L. Pachter and B. Sturmfels), 298–311, Cambridge University Press, New York.

Catanese, F., Hoşten, S., Khetan, A. and Sturmfels, B. (2006). The maximum likelihood degree, *American Journal of Mathematics*, **128**, 671–697.

Chai, J. and Housworth, E. A. (2011). On Rogers' proof of identifiability for the GTR $+\Gamma+I$ model, *Systematic Biology*, **60**, 713–718.

Chakerian, J. and Holmes, S. (2010). Computational tools for evaluating phylogenetic and hierarchical clustering trees, arXiv: 1006.1015.

Craven, K. D., Hsiau, P. T. W., Leuchtmann, A., Hollin, W. and Schardl, C. L. (2001). Multigene phylogeny of *Epichloë* species, fungal symbionts of grasses, *Annals of the Missouri Botanical Garden*, **88**, 14–34.

Day, W. H. E. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices, *Bulletin of Mathematical Biology*, **49**, 461–467.

Desper, R. and Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting, *Molecular Biology and Evolution*, **21**, 587–598.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions, *Annals of Statistics* **26**, 363–397.

Drton, M. and Sullivant, S. (2007). Algebraic statistical models, *Statistica Sinica*, **17**, 1273–1297.

Edwards, S. V., Liu, L. and Pearl, D. K. (2007). High-resolution species trees without concatenation, *Proceedings of the National Academy of Sciences, USA*, **104**, 5936–5941.

Eickmeyer, K. and Yoshida, R. (2008). Geometry of neighbor-joining algorithm for small trees, *The Proceedings of Algebraic Biology, Springer LNC Series*, 82–96.

Eickmeyer, K., Huggins, P., Pachter L. and Yoshida, R. (2008). On the optimality of the neighbor-joining algorithm, *Algorithms for Molecular Biology*, **3**, http://www.almob.org/content/3/1/5.

Eriksson, N. (2005). Tree construction with singular value decomposition, *Algebraic Statistics for Computational Biology* (eds. L. Pachter and B. Sturmfels), 347–358, Cambridge University Press, New York.

Felsenstein, J. (2003). *Inferring Phylogenies*, Sinauer Associates, Inc., Sunderland, Massachusetts.

Fienberg, S. E. and Slavkovic, A. B. (2004). Making the release of confidential data from multi-way tables count, *Chance*, **17**, 5–10.

Friesen, T. L., Stukenbrock, E. H., Liu, Z., Meinhardt, S., Ling, H., Faris, J. D., Rasmussen, J. B., Solomon, P. S., McDonald, B. A. and Oliver, R. P. (2006). Emergence of a new disease as a result of interspecific virulence gene transfer, *Nature Genetics*, **38**, 953–956.

Gascuel, O. and Steel, M. (2006). Neighbor-joining revealed, *Molecular Biology and Evolution*, **23**, 1997–2000.

Grünbaum, B. (2003). *Convex Polytopes*, 2nd ed., Graduate Texts in Mathematics, **221**, Springer-Verlag, New York.

Haarmann, T., Machado, C., Lübbe, Y., Correia, T., Schardl, C. L., Panaccione, D. G. and Tudzynski, P. (2005). The ergot alkaloid gene cluster in Claviceps purpurea: Extension of the cluster sequence and intra species evolution, *Phytochemistry*, **66**, 1312–1320.

Haws, D., Hodge, T. and Yoshida, R. (2011). Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope, *Bulletin of Mathematical Biology*, **73**, 2627–2648.

Heitsch, C. and Condon, A. E. (2002). From RNA secondary structure to coding theory: A combinatorial approach, *8th International Meeting on DNA Based Computers, Sapporo, Japan*, Springer-Verlag, New York.

Hobolth, A. and Jensen, J. L. (2005). Statistical inference in evolutionary models of DNA sequences via the EM algorithm, *Statistical Applications in Genetics and Molecular Biology*, **4**, 18.

Jarrah, A. and Laubenbacher, R. and Stigler, B. and Stillman, M. (2007). Reverse-engineering of polynomial dynamical systems, *Advances in Applied Mathematics*, **39**, 477–489.

Laubenbacher, R. and Stigler, B. (2004). A computational algebra approach to the reverse engineering of gene regulatory networks, *Journal of Theoretical Biology*, **229**, 523–537.

Levy, D., Yoshida, R. and Pachter, L. (2006). Neighbor joining with phylogenetic diversity estimates. *Molecular Biology and Evolution*, **23**, 491–498.

Mishra, B. and Mysore, V. (2007). Algorithmic algebraic model checking IV: Metabolic networks, *Algebraic Biology, Linz, Austria*, Springer-Verlag, New York.

Owen, M. and Provan, S. (2011). A fast algorithm for computing geodesic distances in tree space, *IEEE/ACM Transactions of Computational Biology and Bioinformatics*, **8**, 2–13.

Pachter, L. and Sturmfels, B. (2004a). Parametric inference for biological sequence analysis, *Proceedings of National Academy of Sciences, USA*, **101**, 16138–16143.

Pachter, L. and Sturmfels, B. (2004b). Tropical geometry of statistical models, *Proceedings of National Academy of Sciences, USA*, **101**, 16132–16137.

Pachter, L. and Sturmfels, B. (2005). *Algebraic Statistics for Computational Biology*, Cambridge University Press, New York.

Pedersen, A. G. (2006). The Phylogeny of HIV, http://www.cbs.dtu.dk/dtucourse/cookbooks/gorm/phd.malign.phylo/.

Pillay, D., Rambaut, A., Geretti, A. M. and Brown, A. J. L. (2007). HIV phylogenetics, *Bulletin of Mathematical Biology*, **335**, 460–461.

Pistone, G., Riccomagno, E. and Wynn, H. P. (2001). Computational commutative algebra in discrete statistics, *Algebraic Methods in Statistics and Probability, Notre Dame, IN, 2000*, Contemporaly Mathematics, Vol. 287, 267–282, American Mathematical Society, Providence, Rhode Island.

Rhodes, J. and Sullivant, S. (2011). Identifiability of large phylogenetic mixture models, *Bulletin of Mathematical Biology*, **74**, 212–231.

Rusakov, D. and Geiger, D. (2005). Asymptotic model selection for naive Bayesian networks, *Journal of Machine Learning Research*, **6**, 1–35.

Saitou, N. and Nei, M. (1987). The neighbor joining method: A new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, **4**, 406–425.

Semple, C. and Steel, M. (2004). Cyclic permutations and evolutionary trees, *Advances in Applied Mathematics*, **32**, 669–680.

Sitharam, M. and Agbandje-McKenna, M. (2006). Modeling virus self-assembly pathways: Avoiding dynamics using geometric constraint decomposition, *Journal of Computational Biology*, **13**, 1232–1265.

Sturmfels, B. (1996). Gröbner bases and convex polytopes, *University Lecture Series*, Vol. 8, AMS, Providence, Rhode Island.

Sturmfels, B. and Sullivant, S. (2005). Toric ideals of phylogenetic invariants, *Journal of Computational Biology*, **12**, 204–228.

Sullivant, S. (2011). The disentangling number for phylogenetic mixtures, arXiv: 1107.2880.

Watanabe, S. (2001). Algebraic analysis for non-identifiable learning machines, *Neural Computation*, **13**, 899–933.

Ziegler, G. (1995). *Lectures on Polytopes*, Springer-Verlag, Vienna.

288 Proceedings of the Institute of Statistical Mathematics Vol. 60, No. 2, 279–288 (2012)

Algebraic Methods for Molecular Phylogenetics

Ruriko Yoshida[1] (Shuhei Mano, trans. to Japanese)

[1]Department of Statistics, University of Kentucky, Lexington, KY 4050, U.S.A.

Recently there have been much work and collaborations between modern biology and higher mathematics. A number of important connections have been established betweeen computational biology and the emerging field of "algebraic statistics", which applies tools from combinatrics, computational algebra, and polyhedral geometry to statistical computational problems and statistical modeling. Phylogenetics has provided an abundunt source of applications for algebraic statistics, with research areas including phylogenetic invariants, the geometry of tree space, and analysis of phylogenetic reconstruction. The purpose of this review is to provide the reader with an introduction to this subject, a noncomprehensive guide to further reading, and a collection of more detailed case studies that provide examples of how algebraic methods have been used in the context of molecular phylogeny.

Key words: Algebraic statistics, phylogenetics.