統計数理(2012)
第 60 巻 第 2 号 263–278

特集「多様性と進化の統計解析」
[研究詳解]

α 多様性の測定と確率文字列の理論

小谷野 仁[†]

(受付 2012 年 1 月 4 日；改訂 5 月 29 日；採択 5 月 31 日)

要　　旨

ある領域に生息する生物の群集がどれぐらい多様な種や個体からなっているかを α 多様性と言う．本稿では，これまでに提案されてきた様々な α 多様性の測定方法を分類し，重要なものを概観した上で，文字列の集合上に Levenshtein 距離を定義して距離空間とし，そこで確率論を展開することによって，配列レベルで α 多様性を測定するための方法を提案した著者らの最近の研究を紹介する．この研究においては，位置の尺度として，平均の代わりにコンセンサス配列が，散らばりの尺度として，分散の代わりにコンセンサス配列からの Levenshtein 距離の平均が採用され，これらに対して展開された漸近理論が測定方法の数理的基礎になっている．最後に，動植物と比較して α 多様性の測定が困難であった微生物群集に対する著者らの方法の応用について述べる．

キーワード：α 多様性，16S リボソーム RNA 遺伝子配列，確率文字列，コンセンサス配列，Levenshtein 距離，階層分散．

1. はじめに

ある領域に，いやもっと分かりやすく，地球上に存在する種の数を数える問題を考えよう．対象はひとまず真核生物に限ることにする．この問題は地球や太陽の年齢の推定問題とどこか似た趣を持っているが，それと異なり，本質的にサンプリングの問題となっている．これは生態学において非常に基礎的な問題であるように見えるが，May (1988)によると，Darwin の時代，つまり 19 世紀から現代に至るまであまり組織的には取り組まれてこなかった．18 世紀中葉に von Linné による分類学が始まってから現在までのおよそ 250 年間に，陸上生物 100 万種と海洋生物 25 万種の合わせて 125 万種がデータベースに登録されており，これらの他に，まだ登録はされていないが知られているものが 70 万種存在する．しかし，上の問題に対する方法論は乏しく，専門家によってその推定値は 500 万種以下から 5000 万種以上まで大きく異なっている．最も新しい推定結果は，Mora et al. (2011)における 870 万種であり，内訳は動物 777 万種，植物 298000 種，菌類 611000 種，原生生物 36400 種，クロミスタ 27500 種となっている．

一般に，地球上の生物量の大部分が細菌と古細菌によって占められていると考えられている．このことは，地球上の生物多様性を考える時，微生物の多様性が大きな重要性を持つことを意味している．そこで，微生物について上と同種の問題を考えてみよう．微生物に対しては種が明確に定義されていないので，種を単位とする訳にはいかない．ひとまず，細胞数を見ると，Whitman et al. (1998)は，地球上の原核生物の数を 4–6 $\times 10^{30}$ と推定しており，途方もない

[†] 京都大学化学研究所 バイオインフォマティクスセンター：〒611–0011 京都府宇治市五ヶ庄

数となっている．生息領域での内訳は，海洋中，土壌中，海底面下，及び地底にそれぞれ 1.2 $\times 10^{29}$，2.6 $\times 10^{29}$，3.5 $\times 10^{30}$，及び 0.25–2.5 $\times 10^{30}$ である（これらの区分の詳細は Whitman et al., 1998 を参照）．存在数から多様性に戻り，群集がどれぐらい色々な種からなっているかではなく，どれぐらい色々な DNA を持つ個体からなっているかという観点で，微生物群集の多様性を測定することを考えてみる．DNA はヌクレオチドが結合してできたものであり，ヌクレオチドの一部をなすアデニン（A），グアニン（G），チミン（T），及びシトシン（C）という 4 種類の塩基に注目して，A，G，T，及び C の 4 つの文字からなる文字列として表された．そこで，微生物の多様性の測定問題を，標本抽出を行った微生物の母集団が持っている全ての DNA 塩基配列の分散の推定問題として定式化することが考えられる．

本稿は Koyano and Kishino (2010) の詳解であり，確率文字列やその分布を定義し，確率文字列に対して通常の確率変数の平均や分散の対応物を導入して，上の問題をできるだけ厳密に取り扱うことを目指して筆者らが提案した方法とその数理的基礎の概要及び応用例について述べる．DNA 全体を見るのは，微生物とは言え非常に大変であるから，データとしては，保存性が高く，関係の遠い生物同士でも比較が可能であるという特性を持っているために，系統学などで伝統的に用いられてきた 16S リボソーム RNA 遺伝子配列を用いることを念頭においている．また，地球全体でどれぐらいかというよりは，様々な環境間の比較，例えば，人間が手を入れたら溶けてしまうほど強い酸性やアルカリ性の環境にも微生物は存在するが，これらの 2 つの極限環境では，配列のレベルで見てどちらに多様な微生物が存在するのかとか，このような極限環境と一般的な環境ではどうかといった問題への応用が，最初の動機となっている．

2. α 多様性の測定方法の概観

生態学には，Whittaker (1960, 1972) によって導入された，α 多様性，β 多様性，及び γ 多様性 という 3 つの伝統的な多様性の概念があった．本稿の主題は，これらのうち α 多様性である．そこで，本節で α 多様性の測定方法の歴史的な流れを概観しておこう．α 多様性は，当初は種を単位として測定されていた．しかし，種を単位とする測定では，種の間で違いの程度が異なることが考慮されていないという問題が指摘されるようになり，種の間や個体の間のダイバージェンスを考慮して，α 多様性を測定する方法が提案されている．また，種を単位とする α 多様性の測定方法は，種数のみを考慮するものと，種の均等度も考慮に入れるものに分けられた．そこで，以下では，α 多様性の測定方法を，種を単位とするか，個体間のダイバージェンスを考慮するか，また均等度を考慮に入れるか，入れないかで分類して見ていくことにする．

各個体が k 個の種のうちの 1 つに属している集団を考える．各 $i=1,\ldots,k$ に対して，π_i を第 i 種に属する個体を観測する確率とし，$\lambda = 1-\sum_{i=1}^{k}\pi_i^2$ とおく．そうすると，$\pi_1=\cdots=\pi_k=1/k$ の時，$\lambda = 1-1/k^2$ となり，$i^*\in\{1,\ldots,k\}$ が存在して $\pi_{i^*}=1$ が成り立つ時，$\lambda=0$ となって，$0\leqq\lambda\leqq 1-1/k^2$ が成り立つ．今，この集団から大きさ n の標本を無作為抽出したとし，各 $i=1,\ldots,k$ に対して，n_i を第 i 種に属している標本の中の個体の数とする．λ の推定量 l を $l=1-\sum_{i=1}^{k}n_i^2/n^2$ と定義し，これを **Simpson 指数**と言う (Simpson, 1949)．Simpson の論文では，多様性ではなく，集中度 $\lambda'=\sum_{i=1}^{k}\pi_i^2$ の推定問題が考察されており，$l'=\sum_{i=1}^{k}n_i(n_i-1)/n(n-1)$ という形の不偏推定量が提案されているが，現在では，上の推定量を Simpson 指数と言うのが一般的である．また，各 $i=1,\ldots,k$ に対して $p_i=n_i/n$ とおき，$H'=-\sum_{i=1}^{k}p_i\log p_i$ を **Shannon 指数**と言う (Shannon, 1948a, 1948b)．これらの指数は現在でもよく使われているが，対象の間のダイバージェンスを考慮していないという批判がある．

次に，種を単位として，均等度を考慮に入れない尺度，すなわち種数の推定量を見ていこう．パラメトリックな方法では Preston (1948) や Hurlbert (1971) の方法があるが，ノンパラメトリック

な方法の方がよく用いられる．古典的なものとして，**Chao 1** と呼ばれる推定量 $S_{\rm obs}+F_1^2/2F_2$ がある(Chao, 1984)．ここで，$S_{\rm obs}$ は観測された種の数，$i=1,2,\cdots$ に対して F_i は i 個体のみ観測された種の数である．更に，Chao and Lee (1992)は sampling coverage という概念に基づく新しい推定量を開発した．この推定量は，特に標本が小さい時，種数を過大推定することが指摘され(例えば Colwell and Coddington, 1994 を参照)，Chao et al. (1993)によって $S_{\rm comm}+S_{\rm rare}/C+F_1\gamma/C$ と改良された．ここで，$S_{\rm comm}$ は普通種(例えば 11 個体以上観測された種)の数，$S_{\rm rare}(=S_{\rm obs}-S_{\rm comm})$ は希少種の数である．また，$N_{\rm rare}=\sum_{i=1}^{10} iF_i$ と $M=\sum_{i=1}^{10} i(i-1)F_i$ に対して，$C=1-F_1/N_{\rm rare}$ であって，$\gamma=\max\{S_{\rm rare}M/\{CN_{\rm rare}(N_{\rm rare}-1)\},0\}$ である．この推定量は，Abundance-based Coverage Estimator の頭文字を取って **ACE** と呼ばれ，種数の推定量として現在よく使われている．

次に，個体間のダイバージェンスを考慮したものを見ていく．1982 年に Rao は 2 次エントロピーと呼ばれる量を導入した(Rao, 1982a, 1982b)．$\mathcal{P}$ を可測空間 $(\mathfrak{X},\mathfrak{A})$ 上の確率測度の凸集合とする時，関数 $Q:\mathcal{P}\to\mathbb{R}$ を $Q(P)=\int_{\mathfrak{X}}\int_{\mathfrak{X}} K(x,y)P(\mathrm{d}x)P(\mathrm{d}y)$ によって定義し，2 次エントロピーと言う．ここで，$K(x,y)$ は $\mathfrak{X}\times\mathfrak{X}$ 上の可測関数で，任意の $x_1,\ldots,x_n\in\mathfrak{X}$ と $\sum_{i=1}^n a_i=0$ となる任意の $a_1,\ldots,a_n\in\mathbb{R}$ に対して $\sum_{i=1}^n\sum_{j=1}^n K(x_i,x_j)a_ia_j\leqq 0$ を満たすものとする．P が，各 $i=1,\ldots,k$ に対して結果 i の起こる確率が p_i である多項分布である時には，$Q(P)=\sum_{i=1}^k\sum_{j=1}^k K(i,j)p_ip_j$ となる．生態学では，これは **Rao の多様性係数**と呼ばれ，多様性の尺度としてしばしば用いられる．$K(i,j)$ には，分析の目的に応じて様々な距離を使うことができる．Rao の多様性係数と類似の指数として，DNA 多型を研究するために集団遺伝学や分子進化学で古くから用いられてきた塩基多様度がある(Kimura, 1968; Nei and Tajima, 1981; Tajima, 1983)．k を集団の中の DNA の塩基配列の種類の数，各 $i=1,\ldots,k$ に対して p_i を種類 i に属する配列の頻度，d_{ij} を種類 i と j の配列の間で異なる塩基の数とする時，$\pi=\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} p_ip_jd_{ij}$ を**塩基多様度**と言う．塩基多様度はダイバージェンスとして **Hamming 距離**を用いた場合の総遺伝的変異で，Rao の多様性係数を 1/2 倍したものである．Hamming 距離は等しい長さを持つ 2 つの文字列に対して定義される距離で，2 つの文字列の対応する位置にある文字のうちのいくつが異なっているかを表す．生態学における応用としては，ダイバージェンスとして分類学的距離を使って海洋の環境汚染の影響を分析した Warwick and Clark (1995)や，ダイバージェンスとして Hamming 距離を使って微生物群集の多様性を分析した Martin (2002)などがある．

最後に，ダイバージェンスを考慮し，均等度を考慮に入れない α 多様性の尺度として，1992 年に Faith が導入した系統的多様性(Faith, 1992)を見よう．分岐図の中で，2 つの結接点の間にあって，そこに他の結接点がない部分を枝と言う．但し，分岐図の端点と内部の分岐点も結接点と見なす．T を種の全体とし，$S\subset T$ とする．また，B を T に対する分岐図の中の枝の全体とする．この時，$P\subset B$ が S に対する最短全長経路であるとは，S の中の任意の 2 つの種を P の中の適当な枝で結べるが，どの $P'\subset P$ もそれができない時に言う．そうして，S の**系統的多様性**を S に対する最短全長経路の中の枝の長さの和のことと定める．また，1998 年に Weitzman は，費用を最も有効に利用して生物多様性を保全する包括的な方法論を，経済学の学術誌である Econometrica に発表した(Weitzman, 1998)．彼はそれをノアの箱舟問題と名付けている．ノアの箱舟問題では，生物多様性の評価に系統的多様性が利用される．種の集合が与えられ，費用をかけるとそれぞれの種の生存確率が上昇する時，系統的多様性の将来の期待値を最大化するように，限られた資金を種の保全に配分しようとするのである．現在の所，この枠組みで最適解を見付けるのは難しく，例えば種の保全費用が全て等しいなど，制限付きのシナリオに対して貪欲アルゴリズムが提案されている段階であるが(Steel, 2005; Hartmann and Steel, 2006)，今後，もう少し一般的な条件の下で最適解が得られるようになると，この方法論

は生物多様性の保全のための学際的で体系的な方法論として広く応用されると予想される．

3. 確率文字列の理論

本節では，Koyano and Kishino (2010) において組織的に展開された確率文字列の理論のうち，α 多様性の推定量の一致性を保証するために必要となる，以下の定理 1 と 2 を記述するのに必要な部分を述べる．また，その前に第 2 段落と第 3 段落において，理論の背後にある考え方を述べる．しかし，なぜ，生物群集の α 多様性を測定するために，このような理論を展開する必要があるのだろうか．その理由は，ある生物群集の α 多様性を測定するためにそこからとられたデータはあくまで標本であり，α 多様性の測定は，実際には，標本に基づく母集団の α 多様性の推定であるということにある．前節で，これまでに提案されてきた α 多様性の指標のうち，歴史的に重要なものを概観した．それらのうち，Simpson 指数と塩基多様度についてはそれぞれ，初等的な，またはある程度のサンプリングの理論がある．しかし，群集生態学における多様性の測定において，このサンプリングの視点は現在ほぼ忘れられた状態にある．数からなるベクトルではない，配列データという生物学に特有のデータを用いて母集団の α 多様性を推定する問題を厳密に取り扱おうとする時，以下でその一部を紹介する理論が必要になったのである．

実数の全体と p 個の実数の組の全体をそれぞれ $\mathbb{R}$ と $\mathbb{R}^p$ によって表す．普通，統計学では，$\boldsymbol{x}_1=(x_1^{(1)},\ldots,x_1^{(p)}),\ldots,\boldsymbol{x}_n=(x_n^{(1)},\ldots,x_n^{(p)})\in\mathbb{R}^p$ を観測し，これらを足したり，引いたり，スカラー倍したり，(1 次元なら) 掛けたり，割ったり，べき乗したり，べき根をとったりして，母集団に関して何らかの推測をする．今，$A=\{a,b,c,\ldots,x,y,z\}$ とおき，A をアルファベットと呼ぶことにする．各 $i=1,\ldots,n$ と $j=1,\ldots,p$ に対して，$\boldsymbol{x}_i^{(j)}$ を $\mathbb{R}$ の要素ではなく，A の要素とすると，$\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ は p 次元実ベクトルではなく，長さ p の英単語となる．従って，例えば，DNA の塩基配列やタンパク質のアミノ酸配列の解析のために，$A=\{a,g,c,t\}$ または $A=\{a,r,n,d,c,q,e,g,h,i,l,k,m,f,p,s,t,w,y,v\}$ とおくことが考えられる．統計学では，観測値 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ をある確率空間上で定義されて $(\mathbb{R}^p,\mathfrak{B})$ ($\mathfrak{B}$ は $\mathbb{R}^p$ 上の Borel 集合族) に値をとる確率変数の実現値として扱う．A の p 個の要素の組の全体を A^p とし，$(A^p,2^{A^p})$ に値をとる確率変数を考えることによって，ランダムに生成された単語を扱うことができる．しかし，収集した標本の中の塩基配列やアミノ酸配列は，長さが揃っていないことが普通である．そこで，A の要素を有限個並べたものの後に何も文字がないことを表す特別な文字 e を無限個並べたものとして**文字列**を定義し，A の要素と e から作られる文字列の全体を A^* によって表すことにする．そうして，離散時間の確率過程で，A^* に値をとるものとして**確率文字列**を定義してみる．

$\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\in\mathbb{R}^p$ を観測した時，これらの位置の尺度としては，平均ベクトル $\bar{\boldsymbol{x}}=1/n(\boldsymbol{x}_1+\cdots+\boldsymbol{x}_n)$ が最もよく用いられる．$\bar{\boldsymbol{x}}$ は加法とスカラー乗法を用いて定義されており，$\mathbb{R}^p$ のベクトル空間としての構造を使っている．A^* は連接によって非可換な半群になるが，ベクトル空間にはならず，$s_1,\ldots,s_n\in A^*$ に対して，これらの平均は定義されない．それでは，A^* 上の自然な位置の尺度は何だろうか．ここでは，位置の尺度としてコンセンサス配列 ($\bar{s}$ によって表す) を使うことにする．次に，$x_1,\ldots,x_n\in\mathbb{R}$ を観測した時，これらの散らばりの尺度としては，分散 $s^2=1/n\{|x_1-\bar{x}|^2+\cdots+|x_n-\bar{x}|^2\}$ が最もよく用いられる．s^2 は $\mathbb{R}$ 上の距離 (絶対値) を用いて定義されており，$\mathbb{R}$ の距離空間としての構造を使っている．A^* 上の自然な距離は何だろうか．ここでは，生物配列への応用を念頭に置いているため，距離として **Levenshtein 距離** (d_L によって表す) を使うことにする．2 つの文字列の間の Levenshtein 距離とは，一方の文字列を他方の文字列に移すのに必要な挿入，削除，及び置換という 3 つの操作の最少回数のことである．そうして，$s_1,\ldots,s_n\in A^*$ に対して，これらの散らばりの尺度を $v=1/n\{d_L(s_1,\bar{s})+\cdots+d_L(s_n,\bar{s})\}$

によって定義することにする．2 乗して定義してもよいが，それによって扱いやすくなることはない．ある正則条件の下で，$n\to\infty$ の時，$\bar{\boldsymbol{x}}$ と s^2 は母集団の平均と分散に収束することが知られている．$\bar{\boldsymbol{x}}$ と s^2 の A^* 上の対応物であって欲しい $\bar{s}$ と v はどのような漸近的な振る舞いをするだろうか．距離半群 A^* 上で $\bar{s}$ と v の漸近理論を作れないか考えてみよう．

最終的には，以下の定理 1 と定理 2 を述べたいが，それらを述べるための必要最小限の準備をする．証明を含め，詳細は Koyano and Kishino (2010) の EPAPS document を参照して欲しい．まず，唯 1 つの要素からなる文字列である文字を考える．$m-1$ 個の文字の集合 $A=\{a_1,\ldots,a_{m-1}\}$ をアルファベットと言うことにする．何も文字がないことを表す特別な文字 e を導入し，$\bar{A}=A\cup\{e\}$ とおく．例えば塩基配列を扱う場合，空文字 e も含めて $m=5$ である．確率空間 $(\Omega,\mathfrak{F},P)$ 上で定義されて $\bar{A}$ に値をとる確率変数を**確率文字**と言い，その全体を $\mathcal{M}(\Omega,\bar{A})$ によって表す．また，最大の頻度を持つ文字が一意に定まる $(x_1,\ldots,x_n)\in\bar{A}^n$ の集合を $[\bar{A}^n]$ によって表す．写像 $m:[\bar{A}^n]\to\bar{A}$ を

$$m(x_1,\ldots,x_n)=x_1,\ldots,x_n \text{ のうち最大の頻度を持つ文字}$$

によって定義し，$[\bar{A}^n]$ **上のコンセンサス文字**と言う．また，$x\in\bar{A}$ が存在して，任意の $y\in\bar{A}-\{x\}$ に対して，$q(x)>q(y)$ が成り立つ $\alpha\in\mathcal{M}(\Omega,\bar{A})$ の集合を $[\mathcal{M}(\Omega,\bar{A})]$ によって表す．ここで，q は α の分布の確率関数である．写像 $m':[\mathcal{M}(\Omega,\bar{A})]\to\bar{A}$ を

$$m'(\alpha)=\text{任意の } y\in\bar{A}-\{x\} \text{ に対して } q(x)>q(y) \text{ となる } x\in\bar{A}$$

によって定義し，$[\mathcal{M}(\Omega,\bar{A})]$ **上のコンセンサス文字**と言う．任意の $\omega\in\Omega$ に対して $\alpha_1(\omega),\ldots,\alpha_n(\omega)$ のコンセンサス文字が一意に定まる $(\alpha_1,\ldots,\alpha_n)\in\mathcal{M}(\Omega,\bar{A})^n$ の集合を $[\mathcal{M}(\Omega,\bar{A})^n]$ によって表す．写像 $\mu:[\mathcal{M}(\Omega,\bar{A})^n]\to\mathcal{M}(\Omega,\bar{A})$ を

$$\mu(\alpha_1,\ldots,\alpha_n)(\omega)=m(\alpha_1(\omega),\ldots,\alpha_n(\omega))$$

によって定義し，$[\mathcal{M}(\Omega,\bar{A})^n]$ **上のコンセンサス文字**と言う．

次に，一般の文字列を考える．$\bar{A}$ の要素の列 $s=\{x_j\in\bar{A}:j\in\mathbb{Z}^+\}$ ($\mathbb{Z}^+$ は正の整数の集合) が A **上の文字列**であるとは，それが 2 つの条件

$$\text{(i) } k\in\mathbb{Z}^+ \text{ が存在して } x_k=e,\quad \text{(ii) } l\in\mathbb{Z}^+ \text{ に対して } x_l=e \text{ ならば, } x_{l+1}=e$$

を満たす時に言う．A 上の文字列の全体を A^* によって表す．確率文字の列 $\sigma=\{\alpha_j\in\mathcal{M}(\Omega,\bar{A}):j\in\mathbb{Z}^+\}$ が**確率文字列**であるとは，それが 2 つの条件

(i) 任意の $\omega\in\Omega$ に対して，$k\in\mathbb{Z}^+$ が存在して，$\alpha_k(\omega)=e$,
(ii) $\omega\in\Omega$ と $l\in\mathbb{Z}^+$ に対して，$\alpha_l(\omega)=e$ ならば，$\alpha_{l+1}(\omega)=e$

を満たす時に言う．確率文字列の全体を $\mathcal{M}(\Omega,A^*)$ によって表す．確率文字列の有限次元分布と独立性の定義は，通常の離散時間の確率過程のそれらと同じである．例えば，$\sigma\in\{\alpha_j:j\in\mathbb{Z}^+\}\in\mathcal{M}(\Omega,A^*)$ とする時，$j_1<\cdots<j_k$ となる各 $k\in\mathbb{Z}^+$ と $j_1,\ldots,j_k\in\mathbb{Z}^+$ に対して，集合関数 $\boldsymbol{Q}_{j_1,\ldots,j_k}:2^{\bar{A}^k}\to[0,1]$ を

$$\boldsymbol{Q}_{j_1,\ldots,j_k}(B)=P(\{\omega\in\Omega:(\alpha_{j_1}(\omega),\ldots,\alpha_{j_k}(\omega))\in B\})$$

によって定義すると，$\boldsymbol{Q}_{j_1,\ldots,j_k}$ は $2^{\bar{A}^k}$ 上の確率測度になるから，これをサイト $j_1,\ldots,j_k$ における σ の有限次元分布と言う．そうして，関数 $\boldsymbol{q}_{j_1,\ldots,j_k}:\bar{A}^k\to[0,1]$ を

$$\boldsymbol{q}_{j_1,\ldots,j_k}(x_1,\ldots,x_k)=\boldsymbol{Q}_{j_1,\ldots,j_k}(\{x_1,\ldots,x_k\})$$

と定義し，$\boldsymbol{Q}_{j_1,\ldots,j_k}$ の確率関数と言う．$(s_1,\ldots,s_n)\in(A^*)^n$ とし，各 $i=1,\ldots,n$ に対して $s_i=\{x_{ij}:j\in\mathbb{Z}^+\}$ とする．任意の $j\in\mathbb{Z}^+$ に対して $x_{1j},\ldots,x_{nj}$ のコンセンサス文字が一意に定まる $s_1,\ldots,s_n$ の集合を $[(A^*)^n]$ によって表す．写像 $\boldsymbol{m}:[(A^*)^n]\to A^*$ を

$$\boldsymbol{m}(s_1,\ldots,s_n)=\{m(x_{1j},\ldots,x_{nj}):j\in\mathbb{Z}^+\},\quad s_i=\{x_{ij}:j\in\mathbb{Z}^+\},\quad i=1,\ldots,n$$

によって定義し，$[(A^*)^n]$ 上のコンセンサス配列と言う．また，関数 $\boldsymbol{v}:[(A^*)^n]\to[0,\infty)$ を

$$\boldsymbol{v}(s_1,\ldots,s_n)=\frac{1}{n}\sum_{i=1}^{n}d_L(s_i,\boldsymbol{m}(s_1,\ldots,s_n))$$

によって定義し，$[(A^*)^n]$ 上の分散と言う．任意の $j\in\mathbb{Z}^+$ に対して α_j のコンセンサス文字が一意に定まる $\sigma=\{\alpha_j:j\in\mathbb{Z}^+\}\in\mathcal{M}(\Omega,A^*)$ の集合を $[\mathcal{M}(\Omega,A^*)]$ によって表す．写像 $\boldsymbol{m}':[\mathcal{M}(\Omega,A^*)]\to A^*$ を

$$\boldsymbol{m}'(\sigma)=\{m'(\alpha_j):j\in\mathbb{Z}^+\},\quad \sigma=\{\alpha_j:j\in\mathbb{Z}^+\}$$

によって定義し，$[\mathcal{M}(\Omega,A^*)]$ 上のコンセンサス配列と言う．また，関数 $\boldsymbol{v}':[\mathcal{M}(\Omega,A^*)]\to[0,\infty)$ を

$$\boldsymbol{v}'(\sigma)=\sum_{s\in A^*}d_L(s,\boldsymbol{m}'(\sigma))\boldsymbol{q}_{1,\ldots,|s|+1}(x_1,\ldots,x_{|s|},e),\quad s=(x_1,\ldots,x_{|s|},e,\ldots)$$

によって定義し，$[\mathcal{M}(\Omega,A^*)]$ 上の分散と言う．ここで，$|s|$ は文字列 s の長さである．$(\sigma_1,\ldots,\sigma_n)\in\mathcal{M}(\Omega,A^*)^n$ とし，各 $i=1,\ldots,n$ に対して $\sigma_i=\{\alpha_{ij}:j\in\mathbb{Z}^+\}$ とする．任意の $j\in\mathbb{Z}^+$ と $\omega\in\Omega$ に対して $\alpha_{1j}(\omega),\ldots,\alpha_{nj}(\omega)$ のコンセンサス文字が一意に定まる $(\sigma_1,\ldots,\sigma_n)$ の全体を $[\mathcal{M}(\Omega,A^*)^n]$ によって表す．また，$\mathcal{M}(\Omega,[0,\infty))$ は Ω 上で定義されて $[0,\infty)$ に値をとる確率変数の全体である．写像 $\boldsymbol{\mu}:[\mathcal{M}(\Omega,A^*)^n]\to\mathcal{M}(\Omega,A^*)$ を

$$\boldsymbol{\mu}(\sigma_1,\ldots,\sigma_n)(\omega)=\{\mu(\alpha_{1j},\ldots,\alpha_{nj})(\omega):j\in\mathbb{Z}^+\}$$

によって定義し，$[\mathcal{M}(\Omega,A^*)^n]$ 上のコンセンサス配列と言う．また，写像 $\boldsymbol{\kappa}:[\mathcal{M}(\Omega,A^*)^n]\to\mathcal{M}(\Omega,[0,\infty))$ を

$$\boldsymbol{\kappa}(\sigma_1,\ldots,\sigma_n)(\omega)=\frac{1}{n}\sum_{i=1}^{n}d_L(\sigma_i(\omega),\boldsymbol{\mu}(\sigma_1,\ldots,\sigma_n)(\omega))$$

によって定義し，$[\mathcal{M}(\Omega,A^*)^n]$ 上の分散と言う．

定理 1. $\{\sigma_n=\{\alpha_{nj}:j\in\mathbb{Z}^+\}:n\in\mathbb{Z}^+\}\subset[\mathcal{M}(\Omega,A^*)]$ であって，各 $j\in\mathbb{Z}^+$ に対して $\{\alpha_{nj}:n\in\mathbb{Z}^+\}$ が独立で同一の分布に従い，各 $n\in\mathbb{Z}^+$ に対して $(\sigma_1,\ldots,\sigma_n)\in[\mathcal{M}(\Omega,A^*)^n]$ が成り立つならば，$n_0\in\mathbb{Z}^+$ が存在して，任意の $n\geqq n_0$ に対して

$$\boldsymbol{\mu}(\sigma_1,\ldots,\sigma_n)=\boldsymbol{m}'(\sigma_1)\ \text{ a.s.}$$

が成り立つ．

定理 2. $\{\sigma_n:n\in\mathbb{Z}^+\}\subset[\mathcal{M}(\Omega,A^*)]$ が独立で同一の有限次元分布を持ち，各 $n\in\mathbb{Z}^+$ に対して $(\sigma_1,\ldots,\sigma_n)\in[\mathcal{M}(\Omega,A^*)^n]$ であるならば，

$$\boldsymbol{\kappa}(\sigma_1,\ldots,\sigma_n)\to\boldsymbol{v}'(\sigma_1)\ \ (n\to\infty)\quad\text{a.s.}$$

が成り立つ．

4. 提案された α 多様性の測定方法

図 1 はエジプトの強アルカリ，高塩湖(Hamra 湖，Um Risha 湖，及び Fazda 湖)において収集された微生物の 16S リボソーム RNA 遺伝子配列の環境標本に多次元尺度法を適用した結果とそれから作られたヒストグラムであり，図 2 は無根系統樹である(データの詳細については Mesbah et al., 2007 を参照)．配列の間の距離としては，Levenshtein 距離を用いている．これらの図から，標本中の 16S リボソーム RNA 遺伝子配列の分布が多峰性を持っていることと，配列がいくつかのグループに分かれていることが分かる．このような分布の多峰性と部分群集形成性は，他の環境下の微生物群集から抽出された 16S リボソーム RNA 遺伝子配列の環境標

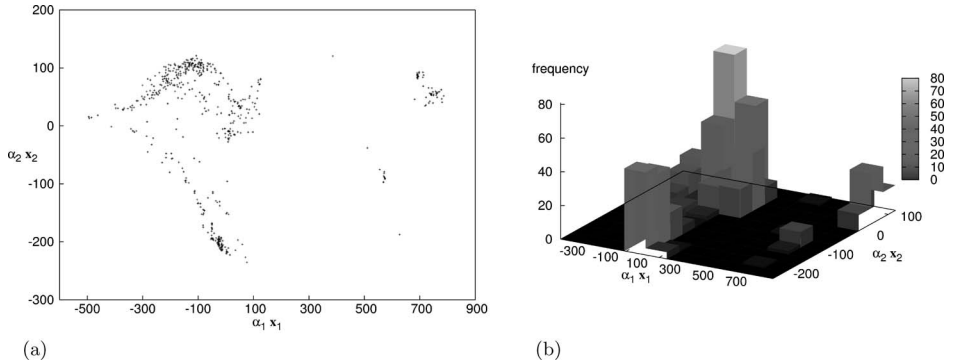


図 1. エジプトの強アルカリ，高塩湖で収集された微生物の 16S リボソーム RNA 遺伝子配列の環境標本に多次元尺度法を適用した結果 (a) とそれから作られたヒストグラム (b)．α_1 と α_2 は，それぞれ行列 $A=(a_{ij})$ の最大の固有値と 2 番目に大きな固有値であって，$\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ は，それぞれ α_1 と α_2 に対する固有ベクトルである．ここで，環境標本の中の配列 $s_1,\ldots,s_{558}$ に対して $a_{ij}=-(d_L(s_i,s_j)^2-\sum_{j=1}^{558}d_L(s_i,s_j)^2/558-\sum_{i=1}^{558}d_L(s_i,s_j)^2/558+\sum_{i=1}^{558}\sum_{j=1}^{558}d_L(s_i,s_j)^2/558^2)/2$ である (標本の大きさ $=558$)．

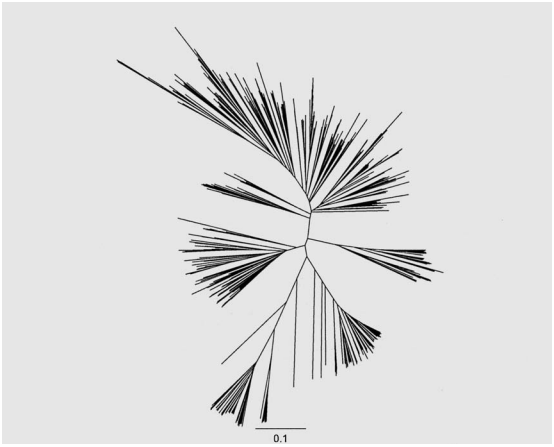


図 2. エジプトの強アルカリ，高塩湖で収集された微生物の 16S リボソーム RNA 遺伝子配列の環境標本から作られた無根系統樹．尺度バーは座位当たり 0.1 の置換数を表す．

本においても観察されることが知られている．

第 1 節で述べたように，本稿では，1 つの環境下の微生物群集の α 多様性の測定問題を，その環境下の微生物の 16S リボソーム RNA 遺伝子配列の全体の散らばりの測定問題として考察したい．しかし，1 つの環境下の全ての微生物の 16S リボソーム RNA 遺伝子配列を収集することは，到底不可能である．そこで，16S リボソーム RNA 遺伝子配列の母集団の多様性は標本から推定することとし，この推定問題の定式化と推定量の性質の考察に，第 3 節で紹介した確率文字列の理論を応用しよう．上で述べた分布の多峰性と部分群集形成性から，環境標本中の全ての 16S リボソーム RNA 遺伝子配列の同一分布性を仮定するのは難しいため，部分群集の中でのみ同一分布性を仮定する．また，推定対象である 16S リボソーム RNA 遺伝子配列の母集団の多様性を，分散分析のアナロジーとして，それを構成する各部分群集の中の多様性と異なる部分群集の間の多様性を反映する階層的な量として定義する．これらの 2 点が，本稿におけるこの推定問題の定式化の特徴となっている．本節では，以後，16S リボソーム RNA 遺伝子配列を単に配列と言う．

1 つの環境下で n 個の配列を観測する．各配列は k 個のグループのうちの 1 つに属しており，各グループに属する全ての配列は同一の分布から生成されたと仮定する．今，各グループの分布，グループの数 k，及び観測された各配列がどのグループに属しているかが未知である．従って，多様性の推定の前に配列の分類が必要になるが，これについては Koyano and Kishino (2010) を参照して欲しい．k-平均法などと比較した時，この論文で提案されているアルゴリズムは，事前にグループの数を指定しなくとも実行できるという特長を持つために，微生物配列の分類において比較的妥当に見える結果を返した．但し，配列に限らず，一般にデータを分類するという問題は，統計学における大きな問題のうちの 1 つであり，現在は，微生物 16S リボソーム RNA 遺伝子配列のための 1 つの手法を考案して実践してみたという段階にある．本稿では，配列の分類の問題には立ち入らないこととし，以下では，配列の分類後の多様性の推定に関して述べる．

各 $i=1,\ldots,k$ に対して，第 i グループから抽出された配列の数を n_i，各 $j=1,\ldots,n_i$ に対して，第 i グループから抽出された第 j 配列を s_{ij} とし，

$$C_1=\{s_{11},\ldots,s_{1n_1}\},\ldots,C_k=\{s_{k1},\ldots,s_{kn_k}\},\ \sum_{i=1}^{k}n_i=n$$

とおく．C_i に属する配列のコンセンサス配列と分散をそれぞれ ($\boldsymbol{m}(s_{i1},\ldots,s_{in_i})$ などと書かずに) $\boldsymbol{m}(C_i)$ と $\boldsymbol{v}(C_i)$ と略記する．$s_{i1},\ldots,s_{in_i}$ を，適当な確率空間上で定義されて，$(A^*,2^{A^*})$ に値をとる，独立で同一の有限次元分布を持つ確率文字列 $\sigma_{i1},\ldots,\sigma_{in_i}$ の実現値として扱う．σ_{i1} のコンセンサス配列と分散をそれぞれ (σ_{i1} を省略して) $\boldsymbol{m}'_i$ と $\boldsymbol{v}'_i$ と書く．

部分群集の間の散らばりと各部分群集の中の散らばりを組み合わせた量

$$\boldsymbol{v}(\boldsymbol{m}'_1,\ldots,\boldsymbol{m}'_k)+\frac{1}{k}\sum_{i=1}^{k}\boldsymbol{v}'_i \tag{4.1}$$

を導入し，16S リボソーム RNA 遺伝子配列を用いた多様性の推定問題を (4.1) の推定問題として定式化する．(4.1) の $\boldsymbol{v}$ と $\boldsymbol{v}'_i$ の定義に含まれる Levenshtein 距離をサイト当たりの Levenshtein 距離 (Levenshtein 距離を 2 つの配列のうち長い方の長さで割ったもの) に置き換えることによって，(4.1) の正規化を定義する．環境間の多様性の比較には，この正規化された多様性が便利である．2 階層分散 $\boldsymbol{v}_2(s_1,\ldots,s_n)$ を

$$\boldsymbol{v}_2(s_1,\ldots,s_n)=\frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i}\left\{d_L\big(\boldsymbol{m}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_k)),\boldsymbol{m}(C_i)\big)+d_L(\boldsymbol{m}(C_i),s_{ij})\right\} \tag{4.2}$$

$$= \boldsymbol{v}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_k)) + \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{v}(C_i)$$

によって定義する．また，

$$\boldsymbol{w}(s_1,\ldots,s_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{d_L(s_i,\boldsymbol{m}(s_1,\ldots,s_n))}{\max\{|s_i|,|\boldsymbol{m}(s_1,\ldots,s_n)|\}}$$

に対して，

$$\boldsymbol{w}_2(s_1,\ldots,s_n) = \boldsymbol{w}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_k)) + \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{w}(C_i) \tag{4.3}$$

を正規化 2 階層分散と呼ぶ．$0 \le \boldsymbol{w}_2(s_1,\ldots,s_n) \le 2$ である．多様性 (4.1) は未知パラメーター $\boldsymbol{m}_i'$ と $\boldsymbol{v}_i'(1 \le i \le k)$ の既知の関数である．これらをそれぞれ $\boldsymbol{m}(C_i)$ と $\boldsymbol{v}(C_i)(1 \le i \le k)$ によって，あるいは同じことであるが，(4.1) を 2 階層分散 (4.2) によって推定しよう．前節の定理 1 と定理 2 から，ある正則条件の下で，各 $i = 1,\ldots,k$ に対して，$n_i \to \infty$ の時，$\boldsymbol{m}(C_i)$ と $\boldsymbol{v}(C_i)$ (これらは n_i に依存する) はそれぞれ，$\boldsymbol{m}_i'$ と $\boldsymbol{v}_i'$ に収束する．従って，(4.1) は (4.2) によって一致推定される．

この節の最後に，2 階層分散の計算の手順を簡単な具体例に沿って述べておきたい．次のような大きさ 9 の標本を得たとする．

$$s_1 = \mathrm{AAGCT},\ s_2 = \mathrm{AAGCTT},\ s_3 = \mathrm{ACGCTT},$$
$$s_4 = \mathrm{GCGCAT},\ s_5 = \mathrm{GCGCAC},\ s_6 = \mathrm{GCGCTTG},$$
$$s_7 = \mathrm{ACGCAA},\ s_8 = \mathrm{ACGCA},\ s_9 = \mathrm{ACCCA}.$$

ステップ 1. 標本を $C_1 = \{s_1, s_2, s_3\}$，$C_2 = \{s_4, s_5, s_6\}$ 及び $C_3 = \{s_7, s_8, s_9\}$ という 3 つのグループに分ける．1 つの分類のアルゴリズムが Koyano and Kishino (2010) において提案されているが，他の方法によって分類してももちろん良い．各グループの中では，それに属する配列がそれらのコンセンサス配列を中心として分布するようにすることが，1 つの分類の基準と考えられる．

ステップ 2. 各グループのコンセンサス配列を構成する．

$$\boldsymbol{m}(C_1) = \mathrm{AAGCTT},\ \boldsymbol{m}(C_2) = \mathrm{GCGCAT},\ \boldsymbol{m}(C_3) = \mathrm{ACGCA}.$$

ステップ 3. 各グループにおいて，それに属する配列とそれらのコンセンサス配列の Levenshtein 距離を計算する．

$$d_L(s_1,\boldsymbol{m}(C_1)) = 1,\ d_L(s_2,\boldsymbol{m}(C_1)) = 0,\ d_L(s_3,\boldsymbol{m}(C_1)) = 1,$$
$$d_L(s_4,\boldsymbol{m}(C_2)) = 0,\ d_L(s_5,\boldsymbol{m}(C_2)) = 1,\ d_L(s_6,\boldsymbol{m}(C_2)) = 2,$$
$$d_L(s_7,\boldsymbol{m}(C_3)) = 1,\ d_L(s_8,\boldsymbol{m}(C_3)) = 0,\ d_L(s_9,\boldsymbol{m}(C_3)) = 1.$$

正規化 2 階層分散の計算に必要なサイト当たりの Levenshtein 距離は，2 つの配列のうち長い方の長さで割ったものであるから，例えば

$$\frac{d_L(s_1,\boldsymbol{m}(C_1))}{\max\{|s_1|,|\boldsymbol{m}(C_1)|\}} = \frac{1}{\max\{5,6\}} \approx 0.16667$$

である．現在では，Levenshtein 距離を計算するプログラムが，様々な言語で実装され公開されている．

ステップ **4.** 各グループの中で文字列版の分散を計算する.

$$\boldsymbol{v}(C_1)=\frac{1+0+1}{3}=\frac{2}{3},\ \boldsymbol{v}(C_2)=\frac{0+1+2}{3}=\frac{3}{3},\ \boldsymbol{v}(C_1)=\frac{1+0+1}{3}=\frac{2}{3}.$$

ステップ **5.** コンセンサス配列のコンセンサス配列を構成する.

$$\boldsymbol{m}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_3))=\mathrm{ACGCAT}.$$

ステップ **6.** コンセンサス配列とそれらのコンセンサス配列の Levenshtein 距離を計算する.

$$d_L\big(\boldsymbol{m}(C_1),\boldsymbol{m}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_3))\big)=2,\ d_L\big(\boldsymbol{m}(C_2),\boldsymbol{m}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_3))\big)=1,$$
$$d_L\big(\boldsymbol{m}(C_3),\boldsymbol{m}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_3))\big)=1.$$

ステップ **7.** コンセンサス配列に対して，文字列版の分散を計算する.

$$\boldsymbol{v}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_3))=\frac{2+1+1}{3}=\frac{4}{3}.$$

ステップ **8.** 2 階層分散を計算する.

$$\begin{aligned}\boldsymbol{v}_2(s_1,\ldots,s_9) &= \boldsymbol{v}(\boldsymbol{m}(C_1),\ldots,\boldsymbol{m}(C_3))+\frac{1}{3}\sum_{i=1}^{3}\boldsymbol{v}(C_i)\\ &= \frac{4}{3}+\frac{1}{3}\Big(\frac{2}{3}+\frac{3}{3}+\frac{2}{3}\Big)\ \approx\ 2.1111.\end{aligned}$$

ここまでの計算に用いた Levenshtein 距離をサイト当たりの Levenshtein 距離に置き換えると，正規化 2 階層分散が得られる.

5. 微生物生態学への応用

第 1 節で述べたように，現在，地球の生物量のかなりの部分が細菌と古細菌によって占められていると考えられており，微生物多様性を知ることは，生態学における基礎的な主題のうちの 1 つとなっている．しかし，微生物生態学においては，微生物群集の多様性を明確に定義して，それを測定しようとしたり，測定方法を考案しようとする研究は，近年になるまでほとんど行われてこなかった．1980 年代に入ってから，Shannon 指数や Simpson 指数を使って微生物多様性を測定しようとする研究が現れ出した．初期の研究としては，例えば Bianchi and Bianchi (1982), Torsvik et al. (1990), Kühn et al. (1991), Moyer et al. (1994) などがある．しかし，Shannon 指数や Simpson 指数では，種が明確に定義されていることや各個体をあいまいさなく同定できることが前提になっており，これらの指標の微生物群集への適用には問題があることが当初から指摘されていた (Staley, 1980; Torsvik et al., 1990)．また，図 1 と 2 において示されているように，微生物群集からの 16S リボソーム RNA 遺伝子配列のデータの 1 つの特徴として強いクラスター性があるが，種数に対応する異なる配列数を数える指標や，塩基多様度のような異なるサイト数に基づく指標では，このようなクラスター構造を反映できないという問題がある．代替的な方法の研究として，Watve and Gangal (1996), Hughes et al. (2001), Hong et al. (2006) などがあるが，理論的な基礎を持つ体系的な方法とは言えなかった．第 3 節と第 4 節において，文字列の集合上に確率論を展開し，それに基礎を置いた α 多様性の測定方法を紹介した．本節では，いくつかの環境の微生物群集から収集された 16S リボソーム RNA 遺伝子配列の環境標本にこの方法を適用して，これらの母集団の α 多様性を推定した結果を紹介する.

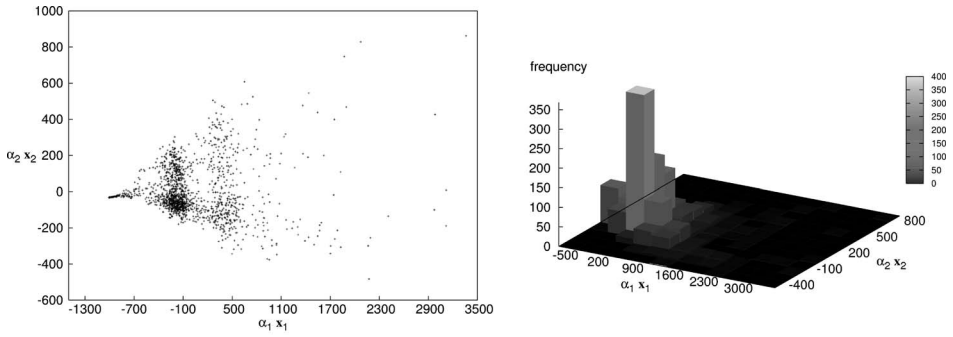
環境として，アメリカのイェローストーン国立公園の温泉 (高温，強酸性環境，HS)，エジプトの強アルカリ，高塩湖 (強アルカリ性，高塩環境，AL)，メキシコの塩田 (高塩環境，HL)，南

極(低温環境，AT)，マリアナ海溝などの深海海底(高圧，低酸素，低温環境，DT)，及び海底熱水孔域(高圧，低酸素，高温環境，DV)の 6 つの極限環境を選ぶことにする．また，比較のために，サルガッソー海(SS)で収集された全ゲノム・ショットガン配列の環境標本も用いる．データは全てデータベース Nucleotide に登録されているものを利用する．データに関する詳しい情報は Koyano and Kishino (2010)を参照して欲しい．推定量としては，0 以上 2 以下に基準化された 2 階層分散(4.3)を使う．そうして，(i)温度，(ii)水素イオン濃度，(iii)酸素分圧，(iv)塩濃度，及び(v)圧力の 5 つの環境パラメーターに注目し，これらの環境パラメーターと環境微生物の多様性の関係を調べてみよう．

各標本の大きさとそれから推定された母集団の α 多様性の推定値を，それぞれ n と $\hat{\alpha}$ に上の環境の略記号を添え字として付けて表すことにすると，$n_{\mathrm{HS}}=1068$，$n_{\mathrm{AL}}=558$，$n_{\mathrm{HL}}=1655$，$n_{\mathrm{AT}}=1056$，$n_{\mathrm{DT}}=441$，$n_{\mathrm{DV}}=800$，及び $n_{\mathrm{SS}}=1500$ であって，$\hat{\alpha}_{\mathrm{HS}}=0.4332$，$\hat{\alpha}_{\mathrm{AL}}=0.4180$，$\hat{\alpha}_{\mathrm{HL}}=0.3705$，$\hat{\alpha}_{\mathrm{AT}}=0.2843$，$\hat{\alpha}_{\mathrm{DT}}=0.3490$，及び $\hat{\alpha}_{\mathrm{DV}}=0.4156$ を得る．また，サルガッソー海で収集された全ゲノム・ショットガン配列の環境標本に機械的に前節の方法を適用すると，0.5629 となる．この最後の値を α 多様性の推定値と解釈することはできないが，ここで重要なことは，2 種類の標本の性質から判断して，16S リボソーム RNA 遺伝子配列の環境標本に前節の方法を適用して得た α 多様性の推定値が，全ゲノム・ショットガン配列の環境標本に機械的に前節の方法を適用して得た値を超えることはなさそうであるということである．そうすると，上の結果は，推定誤差の存在を考慮しても，南極において多様性が著しく低いことを示している．また，ここで取り上げた 6 つの極限環境では，高温環境において相対的に高い推定値が，低温環境において相対的に低い推定値が得られている．これらから，上の 5 つの環境パラメーターのうち，特に温度が多様性に大きな影響を与えること，低温は多様性を大きく制限することが示唆される．陸上生物の種多様性の大域的なパターンのある程度一般的な法則として，北半球では緯度が高いほど種多様性が低いということ(例えば, Rohde, 1992; Willig et al., 2003)と，標高が高いほど種多様性が低いということ(例えば，Rahbek, 1995; Sanders, 2002)が知られている．上の結果は微生物群集に関するものであるが，陸上生物の配列レベルの α 多様性でも，これらの法則はある程度成り立つのかも知れない．推定値の頑健性や消化器官の細菌叢への応用については Koyano and Kishino (2010)を参照して欲しい．

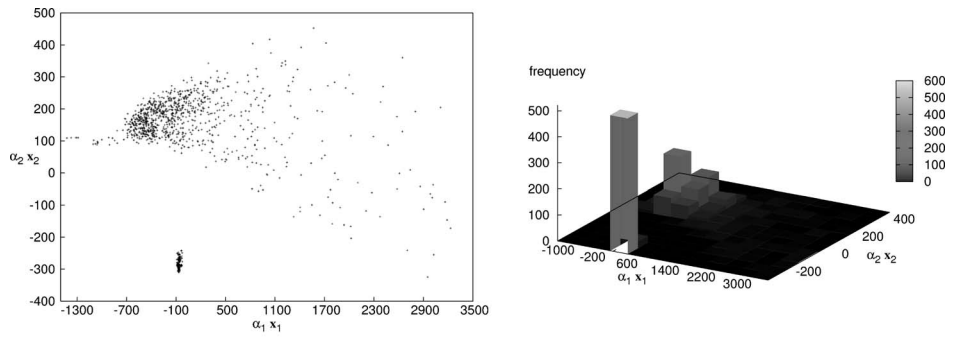
6. まとめ

2000 年代の半ばから，1 つの環境下で収集された，必ずしも相同でない配列の極めて大規模な標本であるメタゲノム・データ(全ゲノム・ショットガン配列の環境標本)が，様々な環境下の微生物群集から収集されるようになってきた．しかし，現在の所，α 多様性の測定などの最も基本的な解析を含め，メタゲノム・データの解析法で，厳密な理論的基礎を持つものを展開するには様々な問題がある．最も根本的な問題は，一般に相同でない全ゲノム・ショットガン配列の集合上にどのように距離を定義するか，あるいはそもそもそのようなことが可能なのかという問題である．また，大規模なメタゲノム学と言えども，全数調査ではないから，研究の目的によっては，標本から母集団に関して推測するという枠組みを使わざるを得ないが，上の問題から母集団のモデリングの問題も生じる．図 3 と図 4 は，それぞれサルガッソー海とヒトの腸の微生物群集から収集された全ゲノム・ショットガン配列の環境標本から機械的に Levenshtein 距離に関する距離行列を作って，多次元尺度法を適用した結果とそれから作られたヒストグラムであり，図 5 は無根系統樹である(データの詳細については Venter et al., 2004 と Gill et al., 2006 を参照)．他に候補が見当たらないために，距離として Levenshtein 距離を使ったが，全ゲノム・ショットガン配列の環境標本においては，配列の間の Levenshtein 距離



(a) (b)

図 3. サルガッソー海で収集された微生物の全ゲノム・ショットガン配列の環境標本に多次元尺度法を適用した結果(a)とそれから作られたヒストグラム(b)．α_1 と α_2 は，それぞれ行列 $A=(a_{ij})$ の最大の固有値と 2 番目に大きな固有値であって，$\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ は，それぞれ α_1 と α_2 に対する固有ベクトルである．ここで，環境標本の中の配列 $s_1,\ldots,s_{1500}$ に対して $a_{ij}=-(d_L(s_i,s_j)^2-\sum_{j=1}^{1500}d_L(s_i,s_j)^2/1500-\sum_{i=1}^{1500}d_L(s_i,s_j)^2/1500+\sum_{i=1}^{1500}\sum_{j=1}^{1500}d_L(s_i,s_j)^2/1500^2)/2$ である (標本の大きさ $=1500$)．



(a) (b)

図 4. ヒトの腸で収集された微生物の全ゲノム・ショットガン配列の環境標本に多次元尺度法を適用した結果 (a) とそれから作られたヒストグラム (b)．α_1 と α_2 は，それぞれ行列 $A=(a_{ij})$ の最大の固有値と 2 番目に大きな固有値であって，$\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ は，それぞれ α_1 と α_2 に対する固有ベクトルである．ここで，環境標本の中の配列 $s_1,\ldots,s_{1500}$ に対して $a_{ij}=-(d_L(s_i,s_j)^2-\sum_{j=1}^{1500}d_L(s_i,s_j)^2/1500-\sum_{i=1}^{1500}d_L(s_i,s_j)^2/1500+\sum_{i=1}^{1500}\sum_{j=1}^{1500}d_L(s_i,s_j)^2/1500^2)/2$ である (標本の大きさ $=1500$)．

に必ずしも意味がある訳ではないため，これらを図 1 及び図 2 と単純には比較できない．しかし，図 3 と図 4 においては，配列が単峰型に近い分布を持っており，16S リボソーム RNA 遺伝子配列の環境標本において観察されたような分布の多峰性が見られない(ヒトの腸における高頻度の配列の離れた集団について検討が必要であるが)．また，図 5 においては，配列のグループの形成の仕方が，16S リボソーム RNA 遺伝子配列の環境標本においてと全く異なっている．これらから，全ゲノム・ショットガン配列の環境標本は，16S リボソーム RNA 遺伝子配列の環境標本と大きく異なる構造を持っていると推測される．しかし，図 3 から図 5 の背後では Levenshtein 距離に関する距離行列が使われているため，このように構造の違いが示

(a) (b)

図 5. サルガッソー海 (a) とヒトの腸 (b) で収集された微生物の全ゲノム・ショットガン配列の環境標本から作られた無根系統樹．尺度バーは座位当たり 0.1 の置換数を表す．

唆されても，これらの図に基づいたモデリングに進めない．

しかし，メタゲノム・データは，16S リボソーム RNA 遺伝子配列の環境標本にはない多くの情報を持っていると考えられている．例えば，Venter et al. (2004) は，サルガッソー海で収集した微生物の DNA の約 1 Gbp のショットガン配列の配列決定を行い，それらが分類学的に新しい門に属する 148 種を含む 1800 種以上の微生物に由来することを明らかにしている．また，Tringe et al. (2005) は，土壌と海洋の微生物群集の比較メタゲノム分析を行って，環境特性と遺伝子分布の間に相関関係があることを報告している．第 1 節で述べたように，16S リボソーム RNA 遺伝子配列は高い保存性を持っているため，生物多様性分析においてこれまでよく利用されてきた．しかし，微生物生態学の動向を見ると，メタゲノム・データは今後ますます急速に蓄積，整備されていくと予想される．メタゲノム・データの解析法の開発が現在強く望まれている．

謝　辞

多くの大変に貴重なコメントを下さった匿名の査読者に感謝の意を表する．

参 考 文 献

Bianchi, M. A. G. and Bianchi, A. J. M. (1982). Statistical sampling of bacterial strains and its use in bacterial diversity measurement, *Microbial Ecology*, **8**, 61–69.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population, *Scandinavian Journal of Statistics*, **11**, 265–270.

Chao, A. and Lee, S. M. (1992). Estimating the number of classes via sample coverage, *Journal of the American Statistical Association*, **87**, 210–217.

Chao, A., Ma, M.-C. and Yang, M. C. K. (1993). Stopping rules and estimation for recapture debugging with unequal failure rates, *Biometrika*, **80**, 193–201.

Colwell, R. K. and Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation, *Philosophical Transactions of the Royal Society B*, **345**, 101–118.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity, *Biological Conservation*, **61**,

1–10.

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M. and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome, *Science*, **312**, 1355–1359.

Hartmann, K. and Steel, M. (2006). Maximizing phylogenetic diversity in biodiversity conservation: Greedy solutions to the Noah's Ark Problem, *Systematic Biology*, **55**, 644–651.

Hong, S. H., Bunge, J., Jeon, S. O. and Epstein, S. S. (2006). Predicting microbial species richness, *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 117–122.

Hughes, J. B., Hellmann, J. J., Ricketts, T. H. and Bohannan, B. J. M. (2001). Counting the uncountable: Statistical approaches to estimating microbial diversity, *Applied and Environmental Microbiology*, **67**, 4399–4406.

Hurlbert, S. H. (1971). The nonconcept of species diversity: A critique and alternative parameters, *Ecology*, **52**, 577–586.

Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles, *Genetical Research*, **11**, 247–270.

Koyano, H. and Kishino, H. (2010). Quantifying biodiversity and asymptotics for a sequence of random strings, *Physical Review E*, **81**, 061912.

Kühn, I., Allestam, G., Stenström, T. A. and Möllby, R. (1991). Biochemical fingerprinting of water coliform bacteria, a new method for measuring phenotypic diversity and for comparing different bacterial populations, *Applied and Environmental Microbiology*, **57**, 3171–3177.

Martin, A. P. (2002). Phylogenetic approaches for describing and comparing the diversity of microbial communities, *Applied and Environmental Microbiology*, **68**, 3673–3682.

May, R. M. (1988). How many species are there on earth?, *Science*, **241**, 1441–1449.

Mesbah, N. M., Abou-El-Ela, S. H. and Wiegel, J. (2007). Novel and unexpected prokaryotic diversity in water and sediments of the alkaline, hypersaline lakes of the Wadi An Natrun, Egypt, *Microbial Ecology*, **54**, 598–617.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. and Worm, B. (2011). How many species are there on Earth and in the ocean?, *PLoS Biology*, **9**, e1001127.

Moyer, C. L., Dobbs, F. C. and Karl, D. M. (1994). Estimation of diversity and community structure through restriction fragment length polymorphism distribution analysis of bacterial 16S rRNA genes from a microbial mat at an active, hydrothermal vent system, Loihi Seamount, Hawaii, *Applied and Environmental Microbiology*, **60**, 871–879.

Nei, M. and Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases, *Genetics*, **97**, 145–163.

Preston, F. W. (1948). The commonness, and rarity, of species, *Ecology*, **29**, 254–283.

Rahbek, C. (1995). The elevational gradient of species richness: A uniform pattern?, *Ecography*, **18**, 200–205.

Rao, C. R. (1982a). Diversity and dissimilarity coefficients: A unified approach, *Theoretical Population Biology*, **21**, 24–43.

Rao, C. R. (1982b). Diversity: Its measurement, decomposition, apportionment and analysis, *Sankhya A*, **44**, 1–22.

Rohde, K. (1992). Latitudinal gradients in species-diversity: The search for the primary cause, *Oikos*, **65**, 514–527.

Sanders, N. J. (2002). Elevational gradients in ant species richness: Area, geometry, and Rapoport's rule, *Ecography*, **25**, 25–32.

Shannon, C. E. (1948a). A mathematical theory of communication, *Bell System Technical Journal*,

27, 623–656.

Shannon, C. E. (1948b). A mathematical theory of communication, *Bell System Technical Journal*, **27**, 379–423.

Simpson, E. H. (1949). Measurement of diversity, *Nature*, **163**, p. 688.

Staley, J. T. (1980). Diversity of aquatic heterotropic bacteria, *Microbiology-1980* (ed. D. Schlessinger), 321–322, American Society for Microbiology, Washington, D. C.

Steel, M. (2005). Phylogenetic diversity and the greedy algorithm, *Systematic Biology*, **54**, 527–529.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations, *Genetics*, **105**, 437–460.

Torsvik, V., Goksøyr, J. and Daae, F. L. (1990). High diversity in DNA of soil bacteria, *Applied and Environmental Microbiology*, **56**, 782–787.

Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P. and Rubin, E. M. (2005). Comparative metagenomics of microbial communities, *Science*, **308**, 554–557.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H. and Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea, *Science*, **304**, 66–74.

Warwick, R. M. and Clarke, K. R. (1995). New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress, *Marine Ecology Progress Series*, **129**, 301–305.

Watve, M. G. and Gangal, R. M. (1996). Problems in measuring bacterial diversity and a possible solution, *Applied and Environmental Microbiology*, **62**, 4299–4301.

Weitzman, M. L. (1998). The Noah's Ark Problem, *Econometrica*, **66**, 1279–1298.

Whitman, W. B., Coleman, D. C. and Wiebe, W. J. (1998). Prokaryotes: The unseen majority, *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 6578–6583.

Whittaker, R. H. (1960). Vegetation of the Siskiyou mountains, Oregon and California, *Ecological Monographs*, **30**, 279–338.

Whittaker, R. H. (1972). Evolution and measurement of species diversity, *Taxon*, **21**, 213–251.

Willig, M. R., Kaufman, D. M. and Stevens, R. D. (2003). Latitudinal gradients of biodiversity: Pattern, process, scale, and synthesis, *Annual Review of Ecology, Evolution, and Systematics*, **34**, 273–309.

Measuring α Diversity and the Theory of Random Strings

Hitoshi Koyano

Bioinformatics Center, Institute for Chemical Research, Kyoto University

Diverse species or individuals belonging to a biological community in a certain area are termed α diversity. In this paper, we first classify various methods for measuring α diversity and historically review them. We then describe the author's and his coworker's recent study in which a method for measuring α diversity at the sequence level was proposed by developing the theory of probability on a set of strings with Levenshtein distance. We outline the theoretical basis for our proposed method, which was provided by taking a consensus sequence as a measure of location and a mean of the Levenshtein distances from a consensus sequence as a measure of dispersion instead of a usual mean and variance, respectively, and developing asymptotic theory for them. Lastly, we describe an application of our method to microbial communities, whose α diversity is more difficult to measure compared with those of animals and plants.

Key words: α diversity, 16S ribosomal RNA gene sequence, random string, consensus sequence, Levenshtein distance, hierarchical variance.