Vol.2021-MPS-133 No.24
Vol.2021-BIO-66 No.24
2021/6/30

ディリクレ過程混合ガウスモデルを用いたメタゲノム配列クラスタリング法の開発

矢﨑堅太郎[1]　阿部貴志[1]

概要：次世代シークエンサーの登場により，環境中の微生物全体像を把握するためにメタゲノム解析が行われている．メタゲノムは複数の生物種が混在した状態で取得され新規性の高い配列を多く含んでいるため，既知のゲノム配列に依存する手法ではなく配列の類似性を用いて分離する手法が求められる．本研究では，新規微生物由来のクラスタを発見するために，教師なし学習であるディリクレ過程混合ガウスモデル (DPGMM) を使用して連続塩基頻度に基づくメタゲノム配列のクラスタリングを行った．DPGMM を用いることで同一生物が占有している割合の高いクラスタを多く作成することができた．一方で，異なる生物種が混在したクラスタも存在したため，クラスタのメンバ数とクラスタ内分散値を利用した閾値を設けることで生物種が混在したクラスタの除去を検討した．1,300 塩基長の生物種 10 種類からなるデータセットに対して検証した結果，同一生物の占有率が平均で 86%以上のクラスタのみを自動で抽出できた．そのため，メタゲノム配列から新規微生物由来のクラスタを検出する手法になり得ると言える．

キーワード：メタゲノム，クラスタリング，ディリクレ過程混合ガウスモデル

Development of metagenomic sequence clustering method using Dirichlet Process Gaussian Mixture Model

KENTARO YAZAKI[1]　TAKASHI ABE[1]

Abstract: With the advent of next-generation sequencers, metagenomic analysis is being carried out in order to grasp the overall picture of microorganisms among various environments. Since metagenomics are acquired in a mixed population of uncultured microorganisms and contain many sequences derived from novel species, a method of separating using sequence similarity is required instead of a method that relies on known genome sequences. In this study, in order to discover clusters derived from novel microorganisms, we performed clustering of metagenomic sequences based on oligonucleotde frequency using the Dirichlet Process Gaussian Mixture Model (DPGMM), which is unsupervised clustering method. By using DPGMM, we were able to create many clusters with a high proportion of the same organism. On the other hand, since there were clusters in which different species were mixed, we examined the removal of clusters in which species were mixed by setting a threshold using the number of cluster members and the dispersion value within the cluster. As a evaluation of examining a data set consisting of 10 species with a length of 1,300 bases, clusters with an occupancy rate of 86% on average of the same organism could be automatically extracted. Therefore, it can be a method for detecting clusters derived from novel microorganisms from metagenomic sequences.

Keywords: Metagenome, Clustering, Dirichlet Process Gaussian Mixture Model

1. はじめに

次世代シークエンサーによって環境中の微生物叢を網羅的に解析するメタゲノム解析が行われるようになった．メタゲノム解析は環境中に存在する新規微生物種の発見や遺伝子の探索に期待が持たれている[1]．メタゲノムは，生物種が混在した状態でゲノム配列が解読されるため，微生物叢を解明するには同一微生物群に分離する必要がある．また，環境中に存在する微生物の 99%は未解明であるため[2]，既知ゲノム配列に依存する方法ではなく，メタゲノム配列内の類似情報のみで分離する手法が望まれる．これまで，多くのメタゲノム配列に対するクラスタリング手法が提案されている[3]．我々は連続塩基組成を用いた一括学習型自己組織化マップ（BLSOM）を新規微生物の検出に使用している[4]．しかし，BLSOM ではクラスタの領域を定義することが難しく，新規微生物と判断された配列が同一の微生物であるかを推定できない．

本研究では，メタゲノム内で優占種となっている微生物由来のクラスタを発見するためにディリクレ過程混合ガウスモデル（Dirichlet Process Gaussian Mixture Model, DPGMM）を使用したメタゲノム配列クラスタリング法の開発を行う．DPGMM によるメタゲノム配列クラスタリングの有効性を確認した後，同一微生物群で構成されているクラスタを抽出できるか検証する．

2. 実験内容

2.1 使用データ

メタゲノム配列クラスタリングのテストデータとして，BorkGroup（http://www.bork.embl.de/~mende/simulated_data）[5]が公開している生物属 10 種類からなる 660b（base, 塩基長）のデータセットを使用した．また，660b のデータセットから同一配列由来の Forward，Reverse を繋ぎ合わせて作

1 新潟大学大学院
Niigata University

Vol.2021-MPS-133 No.24
Vol.2021-BIO-66 No.24
2021/6/30

成した約 1300b のデータセットも使用した．データセットの生物属名及びデータ件数については表 1 に示す．

表 1．使用データの生物属名と件数

生物属名	データ件数	
	660b	1300b
Bacillus	30086	13538
Cyanothece	52682	23706
Escherichia	31358	14110
Lawsonia	10999	4949
Listeria	18451	8302
Methanococcus	13731	6178
Neisseria	16603	7472
Rhodopseudomonas	33625	15131
Staphylococcus	24237	10907
Thiobacillus	18228	8202
合計	250000	112495

DPGMM の入力には縮退 4 連続塩基頻度を用いた．連続塩基頻度とは，塩基配列の先頭から連続塩基の頻度を算出してベクトル化したものであり，生物種ごとに異なることが知られている[6]. 4 連続塩基頻度の場合, 塩基配列の先頭から 4 連続塩基を 1 文字ずつずらして頻度を算出することになる．また，縮退とは相補的な連続塩基を同一のものとみなすことで特徴量の次元削減を行ったものである(ex. AAAA = TTTT)[7].

2.2 ディリクレ過程混合ガウスモデル(DPGMM)[8]

メタゲノム配列は新規微生物の配列を多く含んでいるためクラスタ数を事前に知ることはできない．そのため，データからクラスタ数を自動で決定できる DPGMM を使用する．DPGMM は，混合ガウスモデルの混合比率の事前分布にディリクレ過程, 混合ガウス分布の事前分布にガウス・ガンマ分布を導入し，事後分布の近似分布を解くために EM アルゴリズムで最適化を行うクラスタリング法である.

混合比率を作成するために棒折り過程(Stick-Breaking Process, SBP)と呼ばれるディリクレ過程を導入する．SBP は長さ 1 の棒を$v_1:(1-v_1)$の比で折り，残りの棒から$v_2:(1-v_2)$の比で折ることを繰り返すことで，無限次元の混合比を得ることができる．ここでv_iは，ベータ分布から得られるため,

$$v_i \sim Beta(1,\alpha) \tag{1}$$

とする．SBP により無限個の混合比率を作成できるが，実際には無限次元の計算はできないため，十分大きな数で打ち切っている．

DPGMM では観測データXから潜在変数Zの事後分布$p(Z|X)$を計算することが目的であるが，解析的に求められない．そこで，変分ベイズ法を導入することで事後分布の近似解を求める[8][9][10]．変分ベイズでは対数周辺尤度$\ln p(X)$を目的関数としており,

$$\ln p(X) = \mathcal{L}(q) + \mathrm{KL}(q||p) \tag{2}$$

とする．$\mathrm{KL}(q||p)$は事後分布と近似分布の近さを表しており，0 に近いほど 2 つの分布が似ていることになる．$\mathcal{L}(q)$は変分下界と呼ばれる汎関数を表している．対数周辺尤度$\ln p(X)$は定数であるため，$\mathcal{L}(q)$が増加することで$\mathrm{KL}(q||p)$は減少し，事後分布$p(Z|X)$と近似分布qが近づく．変分下界$\mathcal{L}(q)$を最大化するためには，近似分布qの最適解を求める必要がある．

変分ベイズ法では近似分布をパラメータごとに分解するため,

$$q(Z,\pi,\mu,\Lambda) = q(Z)q(v)q(\mu,\tau) \tag{3}$$

とする. ここで$q(Z)$は潜在変数Z，$q(v)$は混合比率π，$q(\mu,\tau)$は混合ガウス分布のパラメータについての近似分布を表す. 変分ベイズ法では$q(Z)$と$q(v), q(\mu,\tau)$の最適解を EM アルゴリズムによって繰り返し解くことで変分下界が増加し，結果として対数尤度が増加するようになる．

混合比率の近似分布$q(v)$は式(4)で与えられる．

$$q(v_k) = \mathbb{E}_q[\log V_i] + \sum_{i=1}^{k-1} \mathbb{E}_q\left[\log(1-V_i)\right] \tag{4}$$

この時,

$$\mathbb{E}_q[\log V_i] = \Psi(\gamma_{i1}) - \Psi(\gamma_{i1}+\gamma_{i2}) \tag{5}$$

$$\mathbb{E}_q[\log(1-V_i)] = \Psi(\gamma_{i2}) - \Psi(\gamma_{i1}+\gamma_{i2}) \tag{6}$$

とする．$\Psi(\cdot)$はディガンマ関数を表し，γ_iはディリクレ過程のパラメータ,kは混合分布数を表す．また，γ_iは式(7),(8)で与えられる．式(8)のαは，ディリクレ過程のハイパーパラメータである．

$$\gamma_{i1} = 1 + \sum_{n=1}^{N} r_{nk} \tag{7}$$

$$\gamma_{i2} = \alpha + \sum_{n=1}^{N}\sum_{j=k+1}^{K} r_{nj} \tag{8}$$

混合ガウス分布の近似分布$q(\mu_k,\tau_k)$は, ガウス・ガンマ分布を共役事前分布に取ることで計算している（式(9)）. μ_kはガウス分布の平均値，τ_kは分散を表すため，ガウス分布の共分散行列には対角行列を仮定している．

Vol.2021-MPS-133 No.24
Vol.2021-BIO-66 No.24
2021/6/30

$$q(\mu_k, \tau_k) = \mathcal{N}(\mu_k|\mu_0, {\lambda_k}^{-1})Gam(\tau_k|a_k, b_k) \quad (9)$$

この時,

$$\lambda_k = \lambda_0 + \tau_k N_k \quad (10)$$

$$\mu_k = \frac{N_k \tau_k \bar{x}_k + \lambda_0 \mu_0}{\tau_k N_k + \lambda_0} \quad (11)$$

$$a_k = a_0 + \frac{N_k}{2} \quad (12)$$

$$b_k = b_0 + \frac{1}{2}\sum_{n=1}^{N} r_{nk}(x_n - \mu_k)^2 \quad (13)$$

で計算される．潜在変数の近似分布$q(Z)$を求めるために，負担率$r_{n,k}$を式(14),(15)より計算する．

$$\log\phi_{nk} = \mathbb{E}_q[\log V_i] + \sum_{i=1}^{k-1}\mathbb{E}_q[\log(1 - V_i)] + \mathbb{E}_q[\log\tau_k] - \frac{1}{2}\log 2\pi - \frac{\tau_k}{2}(x_n - \mu_k)(x_n - \mu_k)^{\mathrm{T}} \quad (14)$$

$$r_{nk} = \frac{\phi_{nk}}{\sum_{j=1}^{K}\phi_{nj}} \quad (15)$$

$$q(Z) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk} \quad (16)$$

この時，2 つの統計量を計算しておく．

$$N_k = \sum_{n=1}^{N} r_{nk} \quad (17)$$

$$\bar{x}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk} x_n \quad (18)$$

式(17),(18)で計算される統計量は混合分布パラメータの更新に必要となる．表 2 に DPGMM のアルゴリズムを示す．

表 2．DPGMM のアルゴリズム

1. [初期化]
   各パラメータ$\gamma_{k1,k2}, \mu_k, \lambda_k, a_k, b_k$とハイパーパラメータ$\alpha, \mu_0, \lambda_0, a_0, b_0$を初期化する．
2. [E ステップ]
   式(14),(15)より負担率r_{nk}を計算し，式(17),(18)の統計量を求め，$q(Z)$を更新する．
3. [M ステップ]
   $q(v), q(\mu, \tau)$を最適化するため式(7),(8),(10~13)を計算し，パラメータを更新する．
4. [収束確認]
   変分下界の増加量を確認する．増加量が閾値よりも大きければ 2.3.を繰り返し，小さければ終了する．

2.3 クラスタの評価値

クラスタリングにより作成された複数の生物属が混在したクラスタを取り除くために評価値を用いる．評価値には，クラスタのメンバ数とクラスタ内分散値を使用し，各評価値で閾値を設定することでクラスタの抽出を行う．

2.3.1 クラスタ内分散値

クラスタ内分散値は，各クラスタに対して中心から各データ点との距離を測るため，クラスタの凝集性を評価できる．クラスタ内分散値CV_iは式(19)で表せる．ここで，n_{c_i}はi番目のクラスタc_iのメンバ数，$x_{c_{ij}}$はクラスタc_iのj番目のデータ，μ_{c_i}はクラスタc_iの平均値を表す．

$$CV_i = \frac{1}{n_{c_i}}\sum_{j}^{n_{c_i}}\left(x_{c_{ij}} - \mu_{c_i}\right)\left(x_{c_{ij}} - \mu_{c_i}\right)^{\mathrm{T}} \quad (19)$$

2.3.2 クラスタ評価値の閾値の設定

メンバ数とクラスタ内分散値を用いた閾値は以下のように設定した．

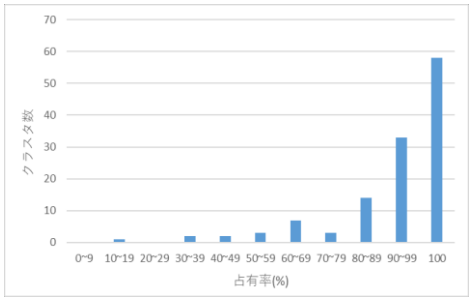
1. 全クラスタのメンバ数の平均値以下を取り除く．
2. 1.により抽出された全クラスタのクラスタ内分散値の平均値以上を取り除く．

以上のステップにより，クラスタ候補を抽出する．

3. 実験結果

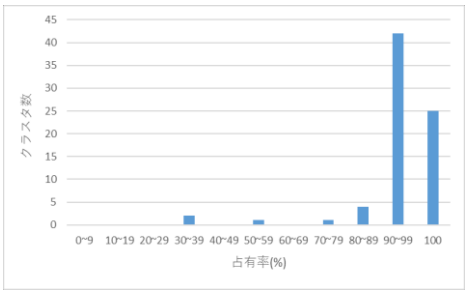
3.1 既知ゲノム配列のクラスタリング結果

異なる塩基長で本手法が有効であるかを評価するために 660b，1300b の生物属 10 種に対してクラスタリングを行った．図 1 に作成されたクラスタの占有率ごとのクラスタ数を示す．ここで，占有率とはクラスタ内に存在する生物属の内，最も多く存在している生物がそのクラスタ内で占める割合のことであり，同一生物で構成されているかの指標となる．図 1 より，660b，1300b 共に占有率の高いクラスタが多く作成される一方で，占有率の低いクラスタも存在していることがわかる．また，1300b の方が占有率の低いクラスタが少ないため，660b に比べて生物属ごとの分離ができている．



(a). 660b

Vol.2021-MPS-133 No.24
Vol.2021-BIO-66 No.24
2021/6/30



(b). 1300b

図 1. 生物属 10 種類に対して行ったクラスタリングの占有率ごとのクラスタ数のヒストグラム. (a)は 660b，(b)は 1300b についての結果を示す.

次に， 660b，1300b ごとに生物属 10 種の中から任意に 5 種，7 種を選択してデータセットを作成し，10 回ずつクラスタリングを行う．これは生物属数を変更することでクラスタリングの分離精度が変わるかを検証するためと実験回数を多くすることでクラスタの占有率のばらつきを見るために行う．各クラスタリングで得られたクラスタの占有率の平均値を図 2 に示す．図 2 より，クラスタの占有率の平均値は 660b，1300b 共に 90%以上となっている．そのため，DPGMM をメタゲノム配列に適用することにより同一微生物で構成されているクラスタを多く作成できると言える．

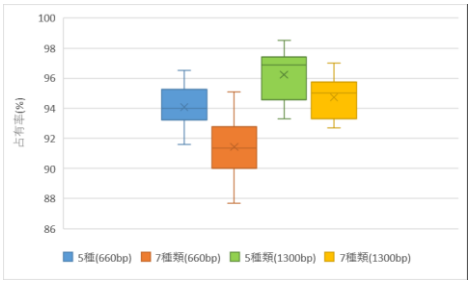


図 2. 10 回ずつクラスタリングを行った時の各クラスタリングで求めた占有率の平均値の箱ひげ図.

また，クラスタの占有率の最低値を図 3 に示す．図 3 より，塩基の長さや生物属数に関わらず，ほとんどの試行で占有率の低いクラスタが作成されている．そこで，クラスタのメンバ数とクラスタ内分散値を用いて占有率の低いクラスタの除去を行う．

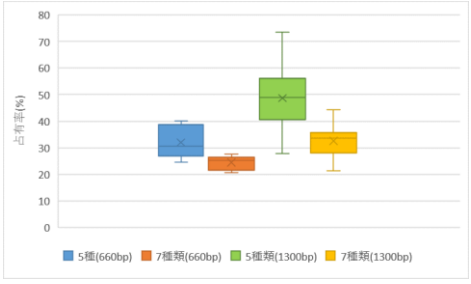
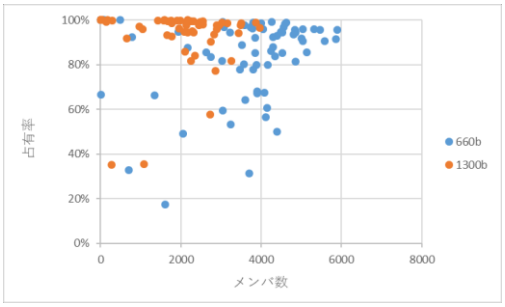


図 3. 10 回ずつクラスタリングを行った時の各クラスタリングで求めた占有率の最低値の箱ひげ図.

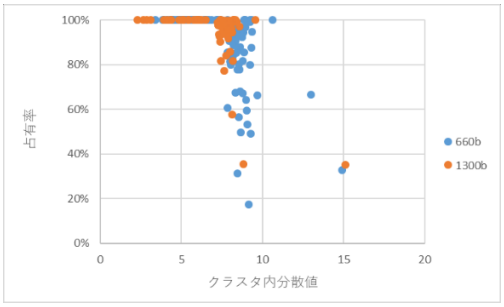
3.2 クラスタの評価値の検証

3.2.1 メンバ数とクラスタ内分散値の有効性

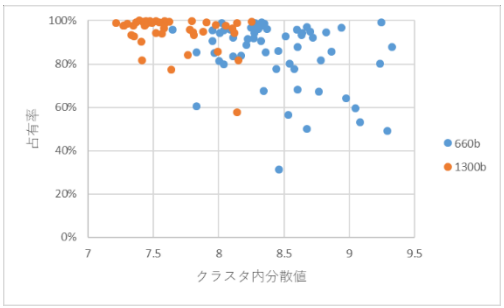
2.3.2 で設定した閾値により占有率の高いクラスタを抽出できるか確認する．生物属 10 種類に対して行ったクラスタリング結果について，各クラスタのメンバ数とクラスタ内分散値をそれぞれ図 4 に示す．図 4(a)よりメンバ数の小さいクラスタに占有率の低いクラスタが含まれている．図 4(a)において，メンバ数の平均値はそれぞれ 660b では 2032.5，1300b では 1499.7 であったため，平均値を閾値とすることでメンバ数が小さく占有率の低いクラスタを取り除くことができる．図 4(b)よりクラスタ内分散値の大きいクラスタは占有率が低くなる傾向がある．また，図 4(c)において，クラスタ内分散値の平均値は 660b では 8.4，1300 では 7.6 であった．そのため，メンバ数の平均値を閾値としてクラスタを抽出した後，クラスタ内分散値の平均値を閾値とすることで占有率の高いクラスタを抽出できることがわかった．



(a). メンバ数



(b). クラスタ内分散値



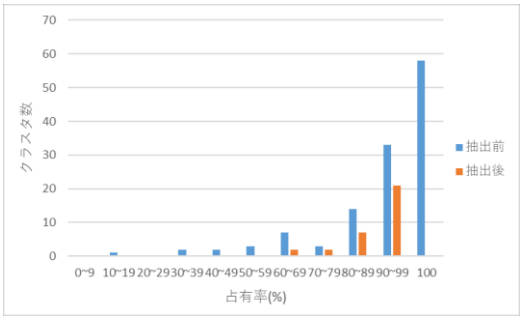
(c). メンバ数による抽出後のクラスタ内分散値

図 4. 生物属 10 種類に対して行ったクラスタリングの各クラスタについてそれぞれ占有率と(a)メンバ数，(b)クラスタ内分散値，(c)メンバ数による抽出後のクラスタ内分散値の散布図を示す.

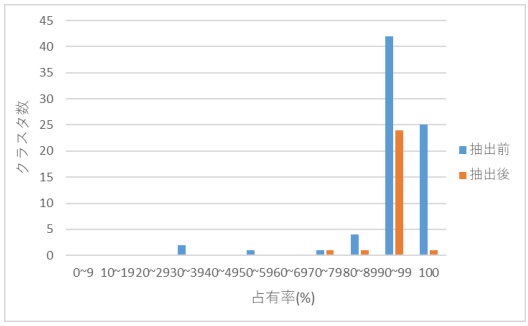
Vol.2021-MPS-133 No.24
Vol.2021-BIO-66 No.24
2021/6/30

3.2.2 メンバ数とクラスタ内分散値によるクラスタの抽出結果

660b, 1300b の生物属 10 種からなるデータセットに対してクラスタリングされたクラスタについて，2.3.2 で設定した閾値によるクラスタ抽出を行った結果を図 5 に示す．図 5 よりクラスタ評価値を用いることで占有率の低いクラスタが取り除かれたことがわかる．660b では占有率 60%以上，1300b では占有率 77%以上のクラスタを抽出できた．また，占有率の平均値は 660b では 90%, 1300b では 96%となっていた．



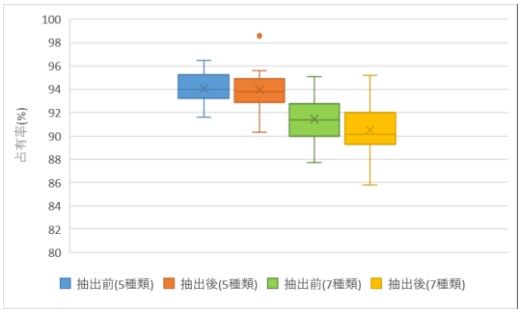
(a). 660b



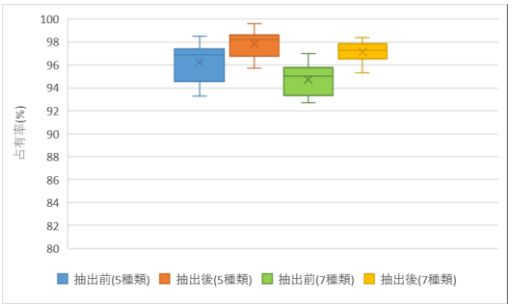
(b). 1300b

図 5. 生物属 10 種類のクラスタリング結果に対して 2.3.2 の閾値により抽出を行った時の占有率ごとのクラスタ数のヒストグラム．(a)は 660b，(b)は 1300b についての結果を示す．

次に，660b，1300b ごとに 5 種，7 種と生物属数を変更して 10 回ずつクラスタリングした結果に対して ，2.3.2 の閾値を用いて抽出を行った．まず，クラスタ抽出前後の占有率の平均値の変化について図 6 に示す．図 6 より 660b, 1300b 共に占有率の平均は高いままである．



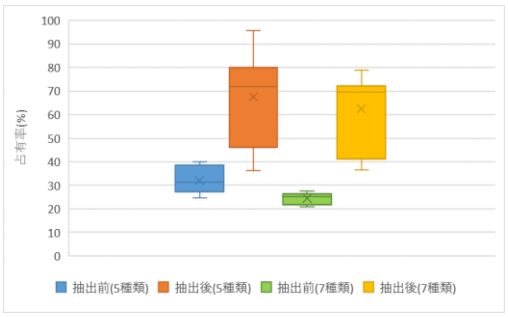
(a). 660b



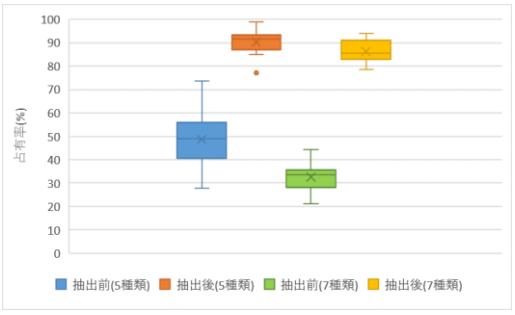
(b). 1300b

図 6. 10 回ずつ行ったクラスタリングに対して 2.3.2 の閾値により抽出したクラスタの占有率の平均値の箱ひげ図．(a)は 660b，(b)は 1300b についての結果を示す．

次に，占有率の最低値について抽出前後で比較したものを図 7 に示す．



(a). 660b



(b). 1300b

図 7. 10 回ずつ行ったクラスタリングに対して 2.3.2 の閾値により抽出したクラスタの占有率の最低値の箱ひげ図．(a)は 660b，(b)は 1300b についての結果を示す．

図 7 より，660b，1300b 共に占有率の最低値が上がっていることからクラスタ評価を行うことで，占有率の高いクラスタを抽出できていることがわかる．660b では最低占有率の平均値が 5 種で 68%，7 種で 62%となっており，1300b では 5 種で 90%，7 種で 86%となっていた．抽出前と比べると占有率が約 40%以上上がっていることになる．また，660b の 5 種，7 種について抽出後の最低占有率にばらつきが見られるが 10 回中 50%を下回ったのは 3 回であり，7 回は占有率が 66%以上となっていたため，多くの試行では占

Vol.2021-MPS-133 No.24
Vol.2021-BIO-66 No.24
2021/6/30

有率の高いクラスタのみを抽出できた．

また，クラスタの抽出を行うことで抽出できなかったクラスタに属するデータは取り出せない．全データ数に対して抽出できたクラスタに属するデータの割合を図 8 に示す．図 8 より，塩基長と生物属数に関わらず約 50%のデータが抽出できていることがわかる．

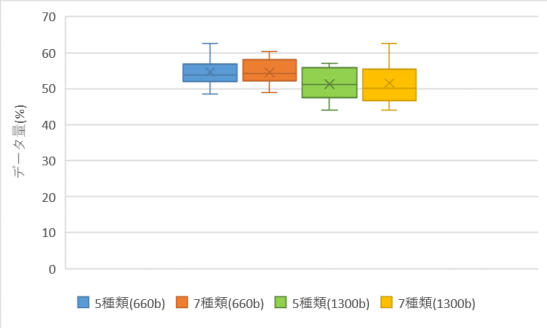


図 8．10 回ずつ行ったクラスタリングに対して 2.3.2 の閾値によるクラスタの抽出を行った結果，取り出せるデータ割合の箱ひげ図．

4. 結論

DPGMM を使用することで占有率の高いクラスタが多く作成されたことから，メタゲノム配列のクラスタリング法としての有用性が示唆された．しかし，占有率の低いクラスタも一定数存在していたため，それらを取り除くためにクラスタのメンバ数とクラスタ内分散値をクラスタの評価値として使用することで占有率の高いクラスタの抽出が可能となった．

660b の生物属 10 種類のデータをクラスタリングした後にクラスタ評価を行った結果，占有率 60%以上のクラスタを抽出できており，占有率の平均値は 90%となっていた．また，生物属 10 種の内から 5，7 種と変更して 10 回ずつ検証を行った結果，5 種類のデータセットでは平均で占有率 68%以上，最低で 36%以上のクラスタ，7 種類のデータセットでは平均で占有率 62%以上，最低で占有率 37%以上のクラスタのみを抽出できた．1300b の生物属 10 種類のデータに対して行った結果，占有率 77%以上のクラスタを抽出できており，占有率の平均値は 96%となっていた．また，生物属数を 5，7 種と変更して 10 回ずつ検証を行った結果，5 種類のデータセットでは平均で占有率 90%以上，最低で占有率 77%以上のクラスタ，7 種類のデータセットでは平均で占有率 86%以上，最低で占有率 79%以上のクラスタのみを抽出できることがわかった．

生物属 10 種類のデータに対する検証の結果，660b と比べて 1300b の方が生物属ごとの分離精度が良いことがわかった．また，生物属数を 5，7 種として 10 回ずつ検証を行った結果，660b，1300b 共に多くの試行で占有率の高いクラスタを抽出できたことから，メタゲノム内の生物属数に関わらず同一微生物ごとに分離可能であることがわかった．

そのため，本手法はメタゲノム配列クラスタリングに有効な手法であると言える．また，抽出されたデータの割合は全データ数の約 50%となっていることからメタゲノム内の優占種のクラスタは取得できると考えられる．そのため，新規微生物が優占種となっているメタゲノム配列に適用することで新規微生物由来のクラスタを発見できる期待が持てる．

参考文献

[1] Venter J.C. et al., Environmental genome shotgun sequencing of the sargasso sea, Science, vol 304:66-74(2004).
[2] Amann R, Ludwig W, Schleifer K.H, Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59: 143–169(1995).
[3] Sedlar K, Kupkova K, Provaznik I, Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomes, Computational and Structural Biotechnology Journal 15:48-55(2017).
[4] Abe T, Hamano Y, Ikemura T, Visualization of genome signatures of eukaryote genomes by batch-learning self-organizing map with a special emphasis on Drosophila genomes, Biomed Res(2004).
[5] Mende, D. R. et al. Assessment of metagenomic assembly using simulated next generation sequencing data. Plos One 7(2): e31386(2012).
[6] Karlin S, Burge C, Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11, 283–290(1995).
[7] Abe T, Kanaya S, and et, Informatics for unveiling hidden genome signatures, Genome Research, 13: 693-702(2003)
[8] Blei D.M, Jordan M.I, Variational inference for Dirichlet Process Mixtures, Bayesian Analysis 1, Number 1, pp.124-144(2006).
[9] 上田修功，ベイズ学習のアルゴリズム-高次元積分の近似手法-，人工知能学会, 19 巻 6 号 656-663(2004).
[10] 著:Bishop M.C，監訳:元田浩，栗田多喜夫，樋口知之，松本裕治，村田昇，パターン認識と機械学習 下，丸善出版株式会社(2012).