Vol.2014-BIO-39 No.4
2014/9/19

リードの由来階級の既知・未知予測に基づくメタゲノム配列の系統分類手法

吉田　拓真　竹中　要一　松田　秀雄

概要：メタゲノム解析では，ある環境に存在する微生物叢から採取した大量のリード（ゲノム配列断片）を解析することで，微生物群全体の機能や有用微生物の調査を行う．メタゲノム分類はメタゲノム解析の 1 分野で，リードが由来する微生物群に存在する微生物種を同定することを目的としている．しかし，世界には数多くの未知の微生物種が存在するため，種の階級だけで分類することはできない．そこで，種より上位の階級を含め，階級に柔軟に分類単位を推定することで未知の微生物種を推定することが行われている．単純ベイズ分類器はメタゲノム分類において使用される分類器の一つで，高い精度・感度をもつ．しかし，この手法にはリードを特定の階級にしか分類できない問題があった．
本研究では，各階級でリードが由来している分類単位が既知か未知かを推定し，これをもとにリードを分類する生物階級を決定する手法を提案する．この手法により単純ベイズ分類器を拡張し，シミュレーションデータに基づく実験を行った．

キーワード：メタゲノム，単純ベイズ分類器，生物階級，未知微生物種

1. はじめに

海中，土壌，火口など，地球上の様々な環境に微生物は生息している．その種数は 10^6 から 10^8，総菌数は 10^{30} のオーダーと言われている [1]．多種多様な微生物の中には，食品の発酵や土壌の浄化など，有用な機能をもつものが多く存在する．微生物の多様性を明らかにし，その中から有用な機能を扱えるようになることが微生物研究の目的である．

微生物の研究には，メタゲノム解析という手法が用いられている．メタゲノム解析では，環境から取得した微生物を分離・培養することなく，すべて混ざり合ったままでゲノム配列断片を大量に取得する．この大量のゲノム配列断片の集合を，ある生物種のゲノムでなく環境中に存在するすべての多種多様なゲノムの総体の意味でメタゲノムという．このようにして，難培養微生物を含む，環境中に存在する多種多様な微生物のゲノム情報を取得することができる．ゲノム配列断片の塩基配列を決定したものをリードという．リードを取得する際，リードが由来する微生物種やゲノム上の位置といった情報は失われてしまう．そのため，リードが由来する微生物種を推定する必要がある．リードが由来する微生物種を推定することをメタゲノム分類という．メタゲノム分類により，メタゲノム中の微生物の多様性を明らかにすることができ，その後の微生物種ごとのゲノム解析も容易になる．

メタゲノム分類で用いられる分類手法は配列相同性に基づく手法 (Alignment-based method) と配列組成に基づく手法（Composition-based method）に分けることができる [2]．

配列相同性に基づく手法は，リードと既知の微生物ゲノムの塩基配列の類似性（相同性）によりリードを分類する．この手法を用いた代表的なソフトウェアには BLAST[3]，MEGAN[4]，MetaBin[5] がある．配列相同性に基づく手法は，ゲノム既知の微生物由来のリードは分類精度が良い一方で，未知の微生物由来のリードの分類精度が低い，計算量が大きい，実行時間が長いといった問題点がある．

配列組成に基づく手法は，リードと既知の微生物ゲノムの塩基配列の組成から特徴量を抽出し，特徴量の類似性によりリードを分類する．この手法を用いた代表的なソフトウェアには NBC[6] や Phymm[7]，TACOA[8] がある．配列組成に基づく手法は，実行時間が短く，計算量が小さい一方で，十分なリード長（リードに含まれる塩基対の個数）がなければ分類精度が良くならないという問題がある．

配列組成に基づく手法である NBC は単純ベイズ分類器を利用したソフトウェアである．少ない計算量で，高い感度と精度を達成する [9] が，単純ベイズ分類器にはリードを

特定の階級においてしか分類できないという問題がある．例えば，微生物種を推定するために，種の階級において分類することを考える．このとき，ゲノム未知の微生物種由来のリードが存在した場合，種の階級で分類するとリードの分類先が不明であるため，誤った分類が行われてしまう．これを避けるため，階級に柔軟な分類が必要となる．もし未知の微生物種由来のリードが存在しても，属や科といった種より上位の階級では，既知の分類単位に分類できるかもしれない．このように階級に柔軟な分類を行うことで，未知の微生物種由来のリードに対しても誤った分類を避けて未知の微生物の様相を知ることができる．

本研究では，単純ベイズ分類器の特定の階級でしか分類できないという問題を改善し，階級に柔軟な分類をできるようになる手法を提案する．この手法では，単純ベイズ分類器ですべての階級においてリードのスコア付けを行い，このスコアをもとにどの階級に分類すべきか推定する．また，シミュレーションによる実験で，提案手法の性能を確認する．

2. 単純ベイズ分類器によるメタゲノム分類

2.1 単純ベイズ分類器

単純ベイズ分類器は，ベイズの定理に基づく教師あり学習の分類器である．

ベイズの定理は式 (1) で表される．

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{1}$$

式 (1) において，$P(A)$ は事象 A が起こる確率，$P(B)$ は事象 B が起こる確率，$P(B|A)$ は事象 A が起こった時に事象 B が起こる確率，$P(A|B)$ は事象 B が起こった時に事象 A が起こる確率である．ベイズの定理は，事象 A を原因，事象 B を観測結果と考えると，事象 A が事象 B を起こすという因果関係（式（1）左辺）を観測結果により測定できる（式（1）右辺）としている．

単純ベイズ分類器への入力として分類先となるクラス $C_1, C_2, \cdots, C_p$，あるデータから抽出した K 個の特徴量からなる特徴量ベクトル $\boldsymbol{d} = (d_1, d_2, \cdots, d_K)$ を考える．このとき，あるデータがクラス C_i に属する確率 $P(C_i|\boldsymbol{d})$ は式（1）より次のように計算できる．

$$P(C_i|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|C_i)P(C_i)}{P(\boldsymbol{d})} \tag{2}$$

さらに，各特徴量の出現確率が互いに独立であることを仮定する．これは一見すると非現実的な仮定だが，実際には高い精度で分類を行う [10]．仮定により，次の式が成り立つ．

$$P(C_i|\boldsymbol{d}) = \frac{\prod_{k=1}^{K} P(d_k|C_i)P(C_i)}{P(\boldsymbol{d})} \tag{3}$$

式 (3) において，$P(\boldsymbol{d})$ はどのクラスにおいても一定である．そのため，式（3）を最大にするようなクラス C_i を求める式は次のように表せる．

$$\hat{C} = \arg\max_i \prod_{k=1}^{K} P(d_k|C_i)P(C_i) \tag{4}$$

式（4）において，最大化を行う部分 $\prod_{k=1}^{K} P(d_k|C_i)P(C_i)$ をスコアとよぶ．

2.2 単純ベイズ分類器のメタゲノム分類への適用

単純ベイズ分類器をメタゲノム分類へ適用する方法について述べる．

メタゲノム分類において，分類先となるクラスは既知の同一階級である分類単位，分類対象となるデータはリードである．由来している可能性が最も高い分類単位にリードを分類することが目的である．

リードから抽出する特徴量としてモチーフプロファイルを用いる．モチーフとはある固定の長さの塩基配列のことである．モチーフの長さが n の塩基配列は塩基が 4 種類あることから 4^n 通り存在する．モチーフプロファイルにはリードに含まれるモチーフがすべて記録されている．もし同じモチーフが複数存在するならば，その回数だけ同じモチーフがプロファイル中に含まれる．

それぞれの既知の分類単位からはモチーフ頻度プロファイルを作成する．モチーフ頻度プロファイルとは，ゲノム配列に対してモチーフの数え上げを行い，すべてのモチーフの出現頻度を記録したものである．モチーフ頻度プロファイルは，まず，その分類単位に含まれる既知ゲノムすべてからモチーフ頻度プロファイルを作成し，次に，同一モチーフの出現頻度を合算することで作成する．

モチーフプロファイルとモチーフ頻度プロファイルを用いて分類を行う．分類の際には，式（4）と等価な式（5）を用いる．

$$\hat{C} = \arg\max_i \sum_{k=1}^{K} \log P(d_k|C_i)P(C_i) \tag{5}$$

ここで，式（5）中の $\hat{C}$ はリードを分類する分類単位，C_i は各分類単位，K はリードに出現するモチーフの個数，d_k は k 個目のモチーフを表す．

式（5）右辺を計算するためには，$P(C_i)$ と $P(d_k|C_i)$ を計算する必要がある．$P(C_i)$ は分類単位に関わらず等しいある値だと仮定する．メタゲノム環境中にどの微生物種がどれだけ存在するかを予見するのは難しいからである．$P(d_k|C_i)$ は例えば次の式により計算する．

$$P(d_k|C_i) = \frac{f(d_k|C_i)}{|C_i|} \tag{6}$$

ここで，$f(d_k|C_i)$ は C_i におけるモチーフ k の出現頻度，$|C_i|$ は分類単位に含まれるすべてのゲノムの塩基長の和である．

表 1　階級ごとの既知・未知と分類先階級

	分類先階級						
	種	属	科	目	綱	門	未知
門	K	K	K	K	K	K	U
綱	K	K	K	K	K	U	U
目	K	K	K	K	U	U	U
科	K	K	K	U	U	U	U
属	K	K	U	U	U	U	U
種	K	U	U	U	U	U	U

2.3 問題点

未知の微生物種が多く含まれるメタゲノム環境においては，ある特定の階級だけでなく，様々な階級にリードを分類する方法が有効である．種が未知の微生物由来のリードであっても，属や科においては既知の分類単位に分類できるかもしれない．

しかし，単純ベイズ分類器によるメタゲノム分類では，階級別の分類ができない．単純ベイズ分類器でメタゲノム分類を行うことができる分類単位はすべて同じ階級である．異なる階級では分類単位に含まれるゲノムの数などが異なるために，スコアを単純に比較できないからである．

3. リードの由来階級の既知・未知予測に基づく系統分類

3.1 提案手法の概要

提案手法では，単純ベイズ分類器を階級に柔軟な分類をできるように拡張する．各階級でリードが既知か未知かを推定し，その結果をもとにリードを分類すべき階級を決定する．

リードの各階級における既知・未知と実際に想定される既知・未知のモデルを**表 1** に示す．ここで，リードがある階級で既知であるとは，リードが由来する微生物種が系統的に属する，既知ゲノムを含むある分類単位がその階級に存在するということである．表 1 において，K は既知（Known），U は未知（Unknown）を示す．このモデルでは，ある階級が分類先階級となる場合，その階級以上では既知，その階級未満では未知と考える．表には 6 階級を記載しているが，考慮する階級が増減しても同様に考える．

このモデルをもとにリードが各階級から由来している確率を求め，確率が最大となる階級にリードを分類する．

3.2 提案手法のアルゴリズム

分類先階級の決定は次の手順に従って行う．

(1) 単純ベイズ分類器による階級ごとのリード分類
(2) 階級ごとの既知・未知予測
(3) 分類先階級の決定

3.2.1 単純ベイズ分類器による階級ごとのリード分類

単純ベイズ分類器を用いて，分類先として考慮する階級すべてにおいてリードの分類をする．このとき，階級 $t(\in T)$ で推定した最適な分類単位 $\hat{c}_t(\in C_t)$ と，その分類単位を与える最大スコア s_t を記録しておく．ここで，T は分類先として考慮する階級の集合，C_t は階級 t に属する分類単位の集合である．

3.2.2 階級ごとの既知・未知予測

表 1 のモデルを適用するために，それぞれの階級において，リードが既知か未知かを予測する．式 (7) のように，各階級 t において最大スコア s_t と未知リード判定閾値 $threshold_t$ を比較することで，既知か未知か予測することができる．

$$ku_t = \begin{cases} known & (s_t \geq threshold_t) \\ unknown & (s_t < threshold_t) \end{cases} \tag{7}$$

単純ベイズ分類器は他の分類器に比べてこのような方法による未知リード判定に優れている [11]．未知リード判定閾値は予備実験によって求める．予備実験の委細は 4.2 を参照されたい．

3.2.3 分類先階級の決定

表 1 のモデルを用いて分類先階級を決定する．階級 t における最大スコアが s_t であるリードが階級 t から由来している確率は式 (8) で与えられる．

$$P(r,t) = \prod_{t' \in T} predictedValue(ku_t, t, t', s_t) \tag{8}$$

$predictedValue$ は既知あるいは未知が ku と予測されたリードが，実際に既知ないし未知である確率を表す．具体的には式 (9) で与えられる．

$$predictedValue(ku, t, t', s) = \begin{cases} ppv(s,t') & (ku = known \wedge t \geq t') \\ 1 - ppv(s,t') & (ku = known \wedge t < t') \\ npv(s,t') & (ku = unknown \wedge t < t') \\ 1 - npv(s,t') & (ku = unknown \wedge t \geq t') \end{cases} \tag{9}$$

式 (9) において，$ppv(s,t')$ は階級 t' でスコア s であるリードが既知だと予測された時に実際に既知である確率，$npv(s,t')$ は階級 t' でスコア s であるリードが未知だと予測された時に実際に未知である確率である．階級 t と t' の間の等号・不等号は t と t' が同じ階級であるとき $t = t'$ が，t が t' より上位の階級であるとき $t > t'$ が成立する．

分類先の階級 $\hat{t}$ は式 (8) を最大化する t である．これは式 (10) により与えられる．

$$\hat{t} = \arg\max_{t \in T} P(r,t) \tag{10}$$

4. 実験と考察

4.1 実験概要

2 つの実験を行った．

第 1 に，予備実験として，提案手法で使用する未知リード判定閾値を決定をするための実験を行った．シミュレー

表 2　階級ごとの既知ゲノム配列と未知ゲノム配列の数

階級	既知	未知
門	1813	1
綱	1803	11
目	1798	16
科	1770	44
属	1716	98
種	1647	167

ションリードを作成し，NBC を用いてシミュレーションリードのスコアを計算し，スコアとリードの既知・未知から未知リード判定閾値を求めた．

第 2 に，従来手法と提案手法を比較するための実験を行った．従来手法と提案手法それぞれを用いて種・属・科・目・綱・門の 6 つの階級で分類を行った．感度，分類的中率，正解率の 3 つの尺度について結果を求めた．

どちらの実験でも単純ベイズ分類器の実装として NBC[6] を用い，モチーフ長は 15 とした．

4.2　予備実験：未知リード判定閾値の決定

4.2.1　実験データ・条件

4.2.1.1　実験データ

NCBI[12] から 1814 のバクテリアゲノム配列を取得した．そのうち 1579 ゲノムからモチーフ頻度プロファイルを作成し，他 235 ゲノムからは作成しなかった．NBC にとっては，モチーフ頻度プロファイルの作成に使用したゲノムは既知のゲノム，使用しなかったゲノムは未知のゲノムだと言える．各階級における既知ゲノム配列と未知ゲノム配列の数を**表 2** に示す．

すべてのゲノムからリード長が 25bp，100bp，500bp のシミュレーションリードをそれぞれ 100 本ずつ作成した．なお，bp とは塩基対（base pare）の意味である．リードの作成には wgsim[13] を用いた．このとき，エラー率 0，インデル率 0 とした．なお，エラーとは塩基配列の読み取り誤り，インデルとは塩基の挿入や削除のことである．エラーはシーケンサーの誤りであるのに対し，インデルは生物の塩基配列自体がデータベースの上のゲノムと異なることだといえる．

4.2.1.2　未知リード判定閾値の求め方

未知リード判定閾値は次のようにして求めた．

(1) シミュレーションリードを NBC によってスコア付けする．
(2) リードが既知か未知かという情報とスコアを対応付ける．
(3) 感度（sensitivity）と特異度（specificity）の和が最大となるスコアを未知リード判定閾値とする．
(4) 以上の操作を各階級で行う．

なお，感度とは実際は既知であるリードを既知と予測する割合，特異度とは実際は未知であるリードを未知と予測する割合のことである．

表 3　リード長と階級ごとの未知リード判定閾値

	25	100	500
門	-279.771	-1729.316	-9932.902
綱	-222.851	-1746.813	-9994.860
目	-223.586	-1728.889	-9834.971
科	-215.039	-1736.082	-9789.985
属	-211.574	-1661.998	-9535.597
種	-200.296	-1653.616	-9308.978

表 4　階級別適切な階級のシミュレーションリードの数

階級	リードの数
未知	1800
門	1600
綱	1600
目	8300
科	21900
属	29000
種	14000

4.2.2　結果と考察

各階級における未知リード判定閾値を**表 3** に示す．表から，上位の階級であるほどに未知リード判定閾値が上昇していることがわかる．上位の階級になるほど未知の分類単位が減るためだと考えられる．また，この結果はやはり単純ベイズ分類器で異なる階級間のスコアを単純に比較してはならないことを示唆している．

4.3　実験：シミュレーションデータの分類

4.3.1　実験データ・条件

4.3.1.1　実験データ

NCBI[12] から新たに 782 のバクテリアゲノム配列を取得した．これらからリード長が 25bp，100bp，500bp のシミュレーションリードをそれぞれ 100 本ずつ作成した．リードの作成には wgsim[13] を用いた．このとき，エラー率 0，インデル率 0 とした．

これらの新たなバクテリアゲノムとシミュレーションリードを予備実験で使用したデータセットに加え，新たなデータセットとした．結果，階級別の適切な分類先のシミュレーションリードの数は**表 4** のようになった．例えば，family(科) が適切な階級であるようなシミュレーションリードは 21900 本であることをす．また，既知ゲノムと全く系統が関係しない未知のリード (unknown) が 1800 本ある．

4.3.1.2　実験手法

NBC では，階級全体で分類を行う提案手法と比較するため，6 つの階級（種，属，科，目，綱，門）それぞれで分類を行った．

提案手法では，6 つの階級と，どの階級にも属さないリードが分類される未知の階級を設けて分類を行った．提案手

表 5　手法と階級別の感度

	従来手法			提案手法		
	25bp	100bp	500bp	25bp	100bp	500bp
未知	*	*	*	18.2%	93.4	96.3
門	11.0	15.3	16.3	7.0	0.1	0.1
綱	21.4	33.3	37.2	0.9	1.6	2.2
目	16.0	26.3	32.1	3.6	7.1	10.2
科	4.1	6.5	6.5	0.1	0.2	0.1
属	17.8	24.8	27.8	0.8	0.7	0.9
種	41.1	53.1	58.9	36.5	35.3	35.9

法の Predicted Value は NBC のスコアに依存せず固定の値（$PPV = NPV = 0.9$）とした．

4.3.1.3　指標

分類器の性能を表すための指標として，分類的中率，感度，正解率を用いた．

分類的中率とは，ある階級に分類したリードのうち，正しくその階級に分類できたリードの割合である．実際に分類を行ったとき，その結果がどれだけ信用できるかの尺度と言える．

感度とは，ある階級に分類するのが正解であるようなリードのうち，その階級で正しく分類されたリードの割合である．最適な階級に分類できているかどうかの尺度と言える．

正解率とは，ある階級について，その階級以下の階級において正しい系統に分類が行われたリードの割合である．例えば，種の分類単位 A に分類するのが最適なリードが，A が属する属の分類単位 B に分類されたとしても，正解率は上昇する．最適ではないが，正しい系統に分類できているかどうかの尺度と言える．

4.3.2　結果

4.3.2.1　感度

各手法と階級別の感度を**表 5** に示す．表において，各従来手法の結果として，各階級で分類した結果を 1 列にまとめて表記している．これは，従来手法が各階級においてしか分類しないためである．なお，従来手法の未知の階級に表記している「*」は，未知の階級にリードを分類しないことを表している．

提案手法は未知の階級にリードを分類しており，特に 100bp と 500bp においては非常に高い感度となった．一方で，その他の階級においては，従来手法に劣る結果となった．また，従来手法がリード長が長くなるに連れてどの階級でも感度が上昇しているのに対し，提案手法は 4 つの階級（未知，綱，目，属）では感度が上昇したものの，その他 3 つの階級（門，科，種）では下降した．

4.3.2.2　分類的中率

各手法と階級別の分類的中率の結果を**表 6** に示す．感度と同様に，表において，各種法の結果として，各階級で分類した結果を 1 列にまとめて表記している．また，従来手法の未知の階級に表記している「*」は，未知の階級にリードを分類しないことを表している．

表 6　手法と階級別の分類的中率

	従来手法			提案手法		
	25bp	100bp	500bp	25bp	100bp	500bp
未知	*	*	*	2.9%	2.8	2.7
門	0.2	0.3	0.3	0.3	0.1	0.1
綱	0.4	0.7	0.8	0.6	1.4	1.9
目	1.7	2.8	3.4	4.8	21.4	33.1
科	1.2	1.8	2.1	1.4	2.5	2.6
属	6.6	9.2	10.3	5.0	29.7	47.5
種	7.4	9.5	10.5	34.9	54.2	59.7

表 7　リード長 25bp　手法と階級別の正解率

	従来手法						提案
	門	綱	目	科	属	種	手法
未知	20.3%	22.1	22.6	18.5	15.3	7.4	16.0
門	20.8	22.6	23.1	18.9	15.6	7.5	16.0
綱	*	23.1	23.6	19.3	16.0	7.7	9.7
目	*	*	24.1	19.8	16.3	7.8	9.3
科	*	*	*	22.3	18.4	8.9	8.9
属	*	*	*	*	27.8	13.4	12.9
種	*	*	*	*	*	41.1	36.5

提案手法は 6 つの階級（未知，綱，目，科，属，種）で従来手法より良い結果となった．また，その中でも未知と門を除いた 5 つの階級では，従来手法に比べてリード長が長くなるにつれて分類的中率が高くなった．

4.3.2.3　正解率

リード長ごとの各手法の正解率を**表 7**，**表 8**，**表 9** に示す．表において，未知の階級の正解率は，リード全体の分類における正解率を表す．正解率ではある階級以下で分類されたリードの数を考えるが，未知の階級以下の階級において分類されたリードとは，すべての階級において分類されたリードのことだからである．また，表中の「*」はその階級に分類されたリードがなかったことを表す．

提案手法は，全体の正解率で，リード長が 25bp のとき科より上の階級で分類した従来手法を下回り，100bp と 500bp では属より上の階級で分類した従来手法を下回った．どのリード長でも種の階級で分類した従来手法よりはよい結果を残した．また，提案手法の各階級の正解率と各階級で分類した従来手法の各階級での正解率を比較すると，どのリード長においても従来手法が提案手法を上回った．

4.4　考察

3 つの尺度全体を通してみると，提案手法は分類的中率では従来手法に優る結果を残したが，感度と正解率では従来手法に劣る結果となった．理想的な階級予測能力をもつ手法ならば，従来手法から，感度と正解率を落とさずに，分類的中率を上昇させることができる．しかし，提案手法

表 8　リード長 100bp　手法と階級別の正解率

	従来手法						提案
	門	綱	目	科	属	種	手法
未知	25.6%	30.6	30.6	25.4	19.9	9.5	12.8
門	26.2	31.3	31.3	26.0	20.4	9.7	10.9
綱	*	32.0	31.9	26.5	20.8	9.9	10.7
目	*	*	32.6	27.1	21.2	10.2	9.9
科	*	*	*	30.6	24.0	11.5	9.0
属	*	*	*	*	36.2	17.3	12.1
種	*	*	*	*	*	53.1	35.3

表 9　リード長 500bp　手法と階級別の正解率

	従来手法						提案
	門	綱	目	科	属	種	手法
未知	27.6%	34.2	34.9	28.4	21.6	10.5	14.0
門	28.3	35.0	35.7	29.1	22.1	10.8	12.0
綱	*	35.8	36.5	29.7	22.6	11.0	12.1
目	*	*	37.2	30.3	23.0	11.3	10.7
科	*	*	*	34.2	26.0	12.7	9.2
属	*	*	*	*	39.2	19.2	12.7
種	*	*	*	*	*	58.9	35.9

では感度と正解率も低下してしまった．

この原因は，最適な階級は未知でないようなリードの多くを未知由来だと予測したことである．最適な階級は未知でないようなリードも未知由来だと予測したために，未知以外の階級が最適だと予測するリードが減って分類的中率は上昇し，未知以外の階級で最適な階級に分類されるリードが減って感度は低下した．

多くのリードが未知の階級由来だと判定されたのは，未知リード判定閾値の設定に問題があるためだと考えられる．今回，未知リード判定閾値は既知であるリードを既知と判定する割合と未知であるリードを未知と判定する割合がともに高くなるように設定した．しかし結果は，単純ベイズ分類器のスコアをもとに既知と未知を予測すると，既知と予測されたリードは実際に既知である確率は高くなり未知である確率は低くなるが，未知と予測されたリードは実際に既知である確率も高いということを示している．つまり，この閾値を用いて未知リード判定を行っても，未知と判定されたリードは実際に既知か未知かわからない．このような状況では，今回設定した NPV（未知と予測したリードが実際に未知である確率；$NPV = 0.9$）は実態にそぐわない高い見積りだったといえる．

5. おわりに

本研究では，階級ごとにリードが既知なのか未知なのか予測し，その予測をもとにリードを分類する手法を提案した．シミュレーションリードに基づく実験を行い，提案手法は階級別の分類を提案手法は従来手法に比べて良い分類的中率をもつが，階級ごとの感度と正解率では悪い結果であった．階級別の分類は可能になったものの，従来手法の分類性能を保ったまま階級予測を行うことはできなかった．

今後の課題は感度と正解率の向上である．そのためには，未知リード判定の方法を変えることが考えられる．例えば正しい分類単位に分類できているか否かの閾値を設定する．正しい分類単位に分類できているリードは既知リードとし，誤った分類単位に分類しているリードは未知リードだとする．ただしこの場合も，誤った分類単位に分類しているリードのうち実際にリードが未知である割合と，既知だが誤った分類単位に分類しているリードの割合に十分注意しなければならない．

参考文献

[1] W. B. Whitman, D. C. Coleman, and W. J. Wiebe.: Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. U.S.A., Vol. 95, No. 12, pp. 6578-6583, Jun 1998.
[2] Sharmila S.Mande,Mozoorul Haque Mohammed,Tarini Shankar Ghosh: Classification of metagenomic sequences: methods and challenges. Briefings in bioinformatics,Vol.13,Issue 6,pp.669-681,Jul 2012.
[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.: Basic local alignment search tool. J. Mol. Biol., Vol. 215, No. 3, pp. 403-410, Oct 1990.
[4] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster.: Megan analysis of metagenomic data. Genome Research, Vol. 17, No. 3, pp. 377-386, 2007.
[5] V. K. Sharma, N. Kumar, T. Prakash, and T. D. Taylor.: Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin. PLoS ONE, Vol. 7, No. 4, p.e34030, 2012.
[6] Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B: Metagenome fragment classification using N-mer frequence profiles. Adv. Bioinformatics 2008.
[7] A. Brady and S. L. Salzberg.: Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat. Methods, Vol. 6, No. 9, pp.673-676, Sep 2009.
[8] N. N. Diaz, L. Krause, A. Goesmann, K. Niehaus, and T. W. Nattkemper.: TACOA:taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics, Vol. 10, p. 56, 2009.
[9] A. L. Bazinet and M. P. Cummings.: A comparative evaluation of sequence classification programs. BMC Bioinformatics, Vol. 13, p. 92, 2012.
[10] I. Rish.: An empirical study of the naive bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence, Vol. 3, pp. 41-46, 2001.
[11] G. L. Rosen, R. Polikar, D. A. Caseiro, S. D. Essinger, and B. A. Sokhansanj.: Discovering the unknown: improving detection of novel species and genera from short reads. J. Biomed. Biotechnol., Vol. 2011, p. 495849, 2011.
[12] NCBI-Taxonomy.: 入手先 ⟨http://www.ncbi.nlm.nih.gov/taxonomy⟩ (2014-02-05).
[13] wgsim.: 入手先 ⟨https://github.com/lh3/wgsim⟩ (2014-02-05).