腸内細菌叢-ヒト属性関連性の統合的抽出方式のヒト性別属性分析への応用

引地　志織† 　佐々木 史織‡ 　清木 康†

†慶應義塾大学環境情報学部 〒252-0812 神奈川県藤沢市遠藤 5322
‡慶應義塾大学政策・メディア研究科 〒252-0812 神奈川県藤沢市遠藤 5322
E-mail: †{t13751sh, kiyoki}@sfc.keio.ac.jp, ‡sashiori@sfc.keio.ac.jp

あらまし 本稿では，腸内細菌叢データと国籍・性別・年齢などのヒト属性との関連性について，文脈解釈機能と意味的分析機構を用い統合的に抽出する方式を提案する．本方式では，細菌学の先行研究からヒューリスティクスとしてヒト属性に関する特定の腸内細菌種の組み合わせを文脈とし，網羅的に複数のクラスタリングアルゴリズム (k-means クラスタリング，階層的クラスタリング)を実行・結果比較を行うことにより，有意な腸内細菌叢関連性をヒトの具体的属性情報と腸内細菌種の数値的傾向として抽出する.本方式の特徴は,クラスタリングアルゴリズムの特徴，および，対象データに合わせた最適アルゴリズムを使用した分析を行うことにより，有意な腸内細菌叢-ヒト属性関連性を得ることを可能とする点にある．本稿では，被験者 60 人，61 種類の門レベルの細菌データを対象として，性別に関する文脈を使用し実験を行った結果，最適アルゴリズムとして階層的クラスタリングが選択され，女性に関連するクラスタ情報(*Actinobacteria* と *Bacteroidetes* の検出割合)が腸内細菌叢-ヒト属性関連性として抽出されることを示す.

キーワード 腸内細菌，データマイニング，個人化医療，次世代シーケンサー，特徴抽出

1. はじめに

ヒト腸内細菌叢とは，体細胞数を遥かに凌ぐ数の細菌の集合体を示し，生体に様々な影響を与えていることが知られている[1]．近年，次世代シーケンシング（NGS）により菌叢の解析技術が進み，腸内細菌叢の異常と癌や糖尿病[2], 炎症性大腸炎[3]といった一部の疾患の関連が報告されており，“Microbiomarker”という一種の生体指標として注目されている．

腸内細菌叢とヒト属性の関係性を効率的に抽出することによって，個人化医療への貢献が期待されているが，遺伝子などのゲノム配列情報を扱うゲノミクスや，mRNA の発現状況を扱うトランスクリプトミクスなどの他のオミックス解析に比べると解析手法が十分でなく，現在は主に主座標分析により全体データの傾向から国籍や年齢などのヒトの属性情報と腸内細菌種の大まかな関係を捉えることにとどまっている．他のオミックス解析では，付帯する情報も含めたデータ分析手法や個々の研究成果を蓄積した巨大なデータベースがあり，データの共有や再利用を行った様々な統合的な解析が可能となっているため，腸内細菌叢解析では統合的な解析手法の実現が大きな課題となっている．

そこで，本稿では腸内細菌叢に関するデータベースを対象とした文脈解釈機能[4]および意味的分析機構[5,6]による腸内細菌叢-ヒト属性関連性の統合的抽出方式を提案し，実例を用いて本方式の有効性及び実現可能性を検証する．本方式により，細菌学の先行研究からヒューリスティクスとしてヒト属性に関する特定の腸内細菌種の組み合わせを文脈として設定し，網羅的に複数のクラスタリングアルゴリズム（k-means クラスタリング，階層的クラスタリング）を実行・結果比較を行うことで，最も有意な腸内細菌叢関係性をヒトの具体的属性情報と腸内細菌種の数値的傾向として抽出する．本方式を利用することにより，腸内細菌叢とヒト属性の関係について包括的な理解が進み，将来的には細菌学的観点から新たな治療法の確立に繋がると期待される．

2. 関連研究

現在，腸内細菌叢解析において，UniFrac 距離[7]を使用した主座標分析[8]が主流となっており，全体データの傾向から国籍や年齢などのヒトの属性情報と腸内細菌種の関係[9,10,11]を捉えることが可能となっている.膨大な腸内細菌叢データを抽出出来る装置として，NGS が広く使用されているが, OMIM [12]や GAD[13], Gene Expression Omnibus[14]，ArrayExpress[15]，Stanford Microarray Database[16]，GO[17]などの他の遺伝子データベースに比べると，蓄積されているデータ数が少なく，既存の腸内細菌叢解析ツールである QIIME[18]でも解析システムの開発が大きな課題となっている．

NGS のデータ解析の第一段階として，定量的な細菌データとの関連づけを行う為の次元縮小として，クラ

スタリングが一般的に使用されており，データベース分野でもクラスタリングを用いた多様なデータ分析方式が提案されている[19,20].

一方，データマイニングの分野では，全体のおおまかな傾向を捉えるよりも，文脈解釈機能による間接的な分析対象データの指定を行い，意味的分析機構による部分的に特出した傾向抽出を行うことが多く，有効性が示されている[4,5,6]．これまでの研究では医療分野のドキュメント群に関する意味的分析機構の適用が行われていた[21,22]が，腸内細菌叢に関するデータベースへの適応はクラスタリング手法に依存的であり，複数のクラスタリング手法を用いた統合的な解析システムの実現はされていなかった[23].

本稿で提案する腸内細菌叢-ヒト属性関連性の統合的抽出方式では，散布図や系統樹を用いて全体の傾向を可視化した上で，最も有意な腸内細菌叢-ヒト属性の関連性を腸内細菌種の数値的傾向とヒトの具体的属性情報として抽出する．本方式では，一般的に使用されているクラスタリングアルゴリズムである k-means クラスタリング及び階層的クラスタリングによる分析を行うことで，文脈により指定されたデータの分布に関わらず，クラスタリングアルゴリズムの特徴や対象データに合わせた腸内細菌叢-ヒト属性関連性を得ることが可能となる．具体的なクラスタリングアルゴリズムの特徴の一つとして，クラスタ数の指定後，k-means クラスタリングではクラスタ内のデータ数が近似値となるのに対し，階層的クラスタリングでは距離行列に基づきクラスタの振り分けを行うため，クラスタ内のデータ数が近似値とは限らないというクラスタ内に含まれるデータ数の差が挙げられる．k-means クラスタリングの特異的な問題点として 6 割程度の低い精度が挙げられるが，本方式では精度保証のため，数十回（25 回程度）のクラスタリングを実行した上での結果を反映する．関連研究により得られた知見をヒューリスティクスとして利用することにより，絞り込まれたデータへの具体的なアプローチが可能となり，本方式の個人化医療への貢献が期待される.

3. 腸内細菌叢-ヒト属性関連性の統合的抽出方式の概要

3.1. データ構造

本方式では，関連研究からヒトの性別との関連性が示唆されている 2 種の細菌種の組み合わせを性別文脈とし，全パラメーターの中から特定のパラメーターを抽出し，分析対象とする部分空間を選択する．実現例として，対象データを公開されている 3 か国 60 人の腸内細菌の検出割合データ [11]とし，全 3 種のデータベースを対象とする．ER 図と各データベースに含まれているデータ例を図 1，図 2 を示す.

- **ヒト属性データベース（Human Attribute）**：関連研究から抽出した 60 人分の国籍・性別・年齢などの 11 種のヒトの属性が格納されている[11].
- **細菌データベース（Bacteria）**：60 人のヒトの糞便から検出された N 種（N=1254）の腸内細菌の検出割合を示す[11]．定量的な比較対象として，全リード数（次世代シークエンサーから得られた遺伝子断片の本数）に対するリード数の検出割合（%）が格納されている.
- **文脈データベース（Context）**：関連研究から抽出した 2 種の細菌種の組み合わせとして生成した，国籍，緯度，性別に関する 3 種の文脈が格納されている[11,24,25]．文脈により {N*(N-1)}/2 (=61*60/2 =1830) 通りの組み合わせの中から 1 通りの組み合わせを抽出する.

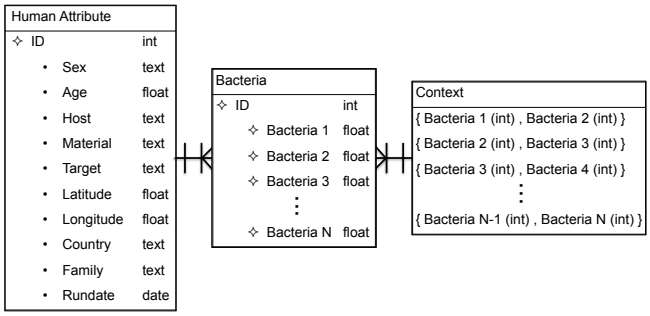


図 1　腸内細菌叢-ヒト属性関連性を示した ER 図

ID	Sex	Age	Host	Material	Target	Latitude	Longitude	Country	Family	Rundate
4489001	M	10	Human	Feces	V4	38.64699	-90.225	USA	Daughter	7/25/2011
4489002	F	49	Human	Feces	V4	-15.38	35.3	Malawi	Mother	8/1/2011
4489060	M	78	Human	Feces	V4	5.410833	-67.609	Venezuela	Father	7/25/2011

(a)

ID	Bacteria 1 (*Actinobacteria*)	Bacteria 2 (*Bacteroidetes*)	Bacteria 3 (*Firmicutes*)	⋯	Bacteria N-1	Bacteria N
4489357	5.0989323	19.1008194	58.7901226	⋯	1.8676578	6.5054534
4489360	4.0412602	21.6349149	48.992333	⋯	1.8537779	0.00838
4489363	1.6734524	9.0795053	68.1555154	⋯	1.1551756	0.059875

(b)

ID	Bacteria 1 (*Actinobacteria*), Bacteria 2 (*Bacteroidetes*)	Bacteria 2 (*Bacteroidetes*), Bacteria 3 (*Firmicutes*)	Bacteria 3 (*Firmicutes*), Bacteria 4 (*Proteobacteria*)	⋯	Bacteria N-2, Bacteria N-1	Bacteria N-1, Bacteria N
country	0	0	0	⋯	0	1
latitude	0	0	0	⋯	1	0
sex	1	0	0	⋯	0	0

(c)

図 2　本方式による腸内細菌叢-ヒト性別属性関連性抽出実験における使用データ例: (a) ヒト属性データ，(b) 細菌データ，(c) 文脈データ.

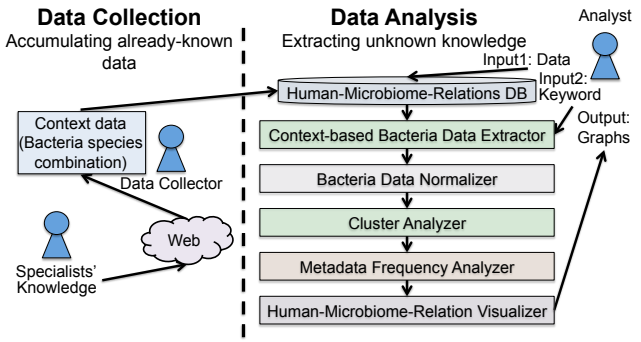
3.2. 文脈定義及び分析関数

本方式の文脈データベースにおいて，各文脈を以下のように定義する．

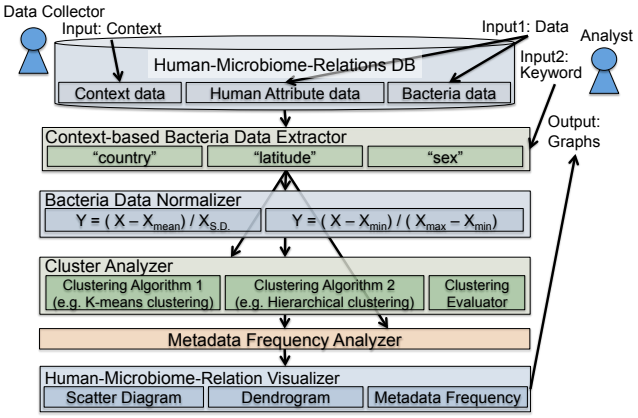
- **国籍文脈（country）**は，検索クエリを指定する（例：{country: Venezuela}）ことで，国籍と関連している 2 種の腸内細菌種（例：{Bacteria1: *Bacteroides*, Bacteria2: *Prevotella*}）を指定し，細菌データベースから分析データの限定を行う[11].
- **緯度文脈（latitude）**は，検索クエリを指定する（例：{latitude: 38.64699}）ことで，緯度と関連している 2 種の腸内細菌種（例：{Bacteria1: *Bacteroides*, Bacteria2: *Firmicutes*}）を指定し，細菌データベースから分析データの限定を行う[20].
- **性別文脈（sex）**は，検索クエリを指定する（例：{sex: M}）ことで，性別と関連している 2 種の腸内細菌種（例：{Bacteria1: *Actinobacteria*, Bacteria2: *Bacteroidetes*}）を指定し，細菌データベースから分析データの限定を行う[21].

図 3 は腸内細菌叢-ヒト属性関連性の統合的抽出方式の概略図を示す．文脈によるデータ分析を行う本方式は 5 つの分析関数により実現する．本方式の特徴は次の 3 点である．

1. キーワードにより文脈として指定されたヒト属性（e.g.国籍，性別，緯度）を対象としてクラスタを形成し,その文脈に対応する細菌データの特徴を抽出する．
2. 文脈により指定された 2 種の細菌の割合を散布図や系統樹を用いて可視化する．
3. ヒト属性メタデータの出現頻度を用いた意味分析により，未知の腸内細菌叢-ヒト属性関連性を抽出する．



(a)



(b)

図 3 本腸内細菌叢-ヒト性別属性関連性抽出方式の概略図: (a) システム構造，および，(b) データフロー．

(1) データ指定（Context-based Bacteria Data Extractor）: $f_{extraction}$

$f_{extraction}$ では，細菌学の先行研究からヒューリスティクスとしてヒト属性に関する特定の腸内細菌種の組み合わせを示す文脈を使用し，対象数値データの限定を行い，再成形された数値データを出力する．データ分析者の知識に応じてキーワードを入力すると，$f_{extraction}$ によりキーワードと一致する細菌データを抽出する．地理や性別に関係する細菌データは国籍や緯度，性別文脈により選択される．

文脈データベース（e.g. 国籍や緯度，性別）において，文脈により指定される細菌 2 種の組み合わせ A=$\{a_i, a_j\}$とキーワード $keyc$ を設定し，細菌データベースに含まれている一連の検出細菌割合データ B=$\{b_1, b_2, \ldots, b_x\}$を参照し，$f_{extraction}$ により式(1)のようにパターン認識を利用して細菌データを抽出する．

$$f_{extraction}\,(keyc, \mathrm{A}, \mathrm{B}) \rightarrow \{b_e \mid b_e = a_{keyc}\} \qquad (1)$$

(2) 正規化（Normalizer）: $f_{normalization}$

$f_{normalization}$ では，分析対象数値データに対して，2 種の正規化手法（#1: $\mathrm{Y} = (\mathrm{X} - \mathrm{X}_{\mathrm{mean}}) / \mathrm{X}_{\mathrm{S.D.}}$, #2: $\mathrm{Y} = (\mathrm{X} - \mathrm{X}_{\mathrm{min}}) / (\mathrm{X}_{\mathrm{max}} - \mathrm{X}_{\mathrm{min}})$）を使用し，正規化後の数値データを出力する．データ分析者が入力した細菌データ B に応じて，正規化手法 # を m として指定することで，細菌データの正規化を行う．

$f_{normalization}$ により，式(2)のように指定した正規化手法を利用して，細菌データを抽出する．

$$f_{normalization}\,(m, b_e) \rightarrow \{b_n\} \qquad (2)$$

(3) クラスタリング（Cluster Evaluator）：$f_{clustering}$

入力された数値データは網羅的に複数のクラスタリングアルゴリズム（K-means クラスタリング，階層的クラスタリング）を $f_{clustering}$ により実行され，各数値データが属する一連のクラスタ番号として $C = \{c_1, c_2, \ldots, c_x\}$，図 g として散布図や系統樹を出力する．各アルゴリズムを用いた結果評価方法として，式(3)のように，全クラスタに属するデータ数における各クラスタのデータ数の割合（%）と，各性別に属するデータ数における性別及びクラスタごとに属するデータ数の割合（%）の積を利用することにより，最も高い値を有意な結果として抽出する．下記のような最適クラスタリング評価方法を用いることにより，クラスタリングアルゴリズムごとに生成されるクラスタ内データ数の差及び細菌データベース内に属する性別ごとのデータ数の差を考慮することが可能となる．非階層的クラスタリングでは k-means クラスタリング（k=3），非階層的クラスタリング（k=3）ではユークリッド距離，重み付き平均法（WPGMA）を用いてクラスタリングを実行される．

$f_{clustering}$ により，式(4)のようにクラスタ番号データや図を抽出する．

$$\text{最適クラスタリング評価方法} = \left\{\frac{\text{各クラスタに属するデータ数}}{\text{全クラスタに属するデータ数}} \times 100(\%)\right\} \times \left\{\frac{\text{性別およびクラスタごとに属するデータ数}}{\text{各性別に属するデータ数}} \times 100(\%)\right\} \quad (3)$$

$$f_{clustering}(b_e) \rightarrow \{C, g\} \; or \; f_{clustering}(b_n) \rightarrow \{C, g\} \quad (4)$$

(4) 意味的データマイニング（Metadata Frequency Analyzer）：f_{mining}

f_{mining} では，$f_{extraction}$ や $f_{normalization}$ により得られた細菌データと $f_{clustering}$ により得られたクラスタ番号データ，ヒト属性データ $D=\{d_1, d_2, \ldots, d_y\}$ を用いて分析を行い，各クラスタ内メタデータカウント数（メタデータの出現頻度）k または図 g を出力する．データ分析者の知識に応じてヒト属性データベース内に存在するキーワード $keyh$（e.g. country, sex and age）を入力することにより，キーワードにより指定された属性に注目した分析を行うことが可能となる（例： country の入力時には USA, Malawi, Venezuela を示す）．クラスタリング前のヒト属性データ確認時など，クラスタ番号データを必要としない場合も，f_{mining} によりクラスタ番号データを入力せずにメタデータカウントを行うことが可能である．

f_{mining} により，式(5)のようにメタデータカウント数や図，キーワード指定時のみヒト属性データを抽出する．

$$f_{mining}(keyh, B, C, D) \rightarrow \{k, d_{keyh}, g\}$$
$$or \; f_{mining}(keyh, B, D) \rightarrow \{k, g\} \quad (5)$$

(5) 可視化（Visualizer）：$f_{visualization}$

$f_{visualization}$ では，$f_{clustering}$，f_{mining} により得られた図を下記のように散布図や系統樹として表示する．データ分析者は $f_{visualization}$ により，式(6)のように，腸内細菌叢とヒト属性の関連性を視覚的に確認することが可能となる．

$$f_{visualization}(b_e, b_n, C, k, d_{keyh}, g) \rightarrow screen \quad (6)$$

3.3. 文脈定義及び分析関数

3.2 の方式により，プロトタイプシステムの実装を行った．数値計算や可視化には Numpy[26]，Scipy[27]，Matplotlib[28]を使用し，散布図や系統樹を作成した．各クラスタ内のメタデータカウント（メタデータ出現頻度）分析には sqlite3 を使用した．

細菌データベース，ヒト属性データベース，文脈データベースは関連研究を基に作成した[11]．本システムでは細菌データ，ヒト属性データ，文脈データを一括入力し，散布図または系統樹などのグラフを出力する．

本実装では，性別文脈を設定し，性別と腸内細菌叢の関連性抽出を試みた．2 種のクラスタリング手法として，階層的クラスタリング（metric = “eculidean”, method = “weighted”），k-means クラスタリング（k = 3）を使用する．

図 4 は散布図や系統樹，メタデータ頻度として性別文脈を用いた実装結果例を示す．

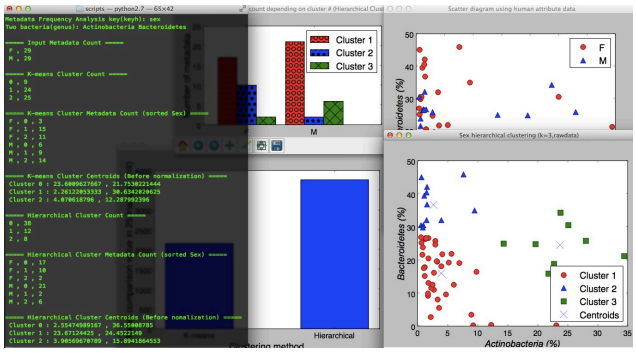


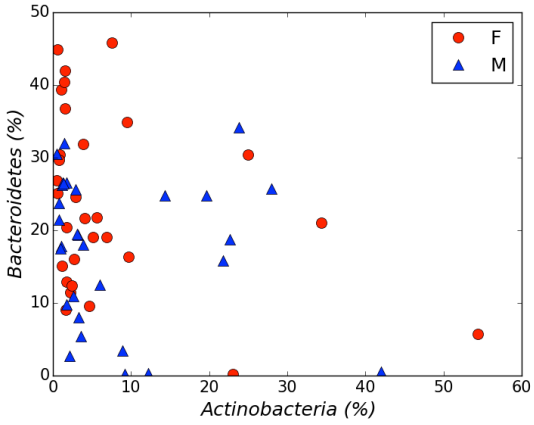
図 4　本方式により実装した実験システム出力例：性別文脈を用いた実験結果とその可視化．

4. 実験：性別文脈を用いた階層的クラスタリングによる腸内細菌叢-ヒト属性関連性抽出

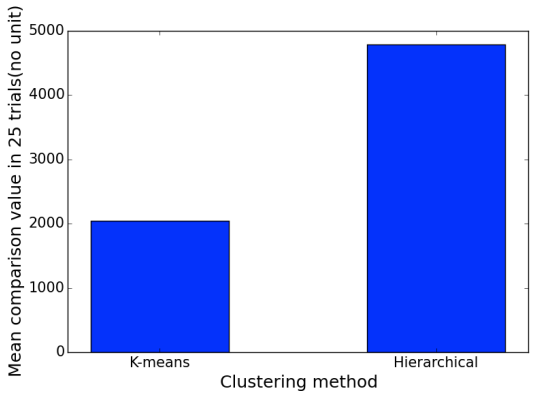
4.1. 実験 A：全 60 人分の腸内細菌叢データを使用した腸内細菌叢-性別属性関連性抽出

以前の実験において文脈解釈機能及び複数のクラスタリング手法を用い，国籍文脈により指定された一部の腸内細菌叢データにおいて，腸内細菌叢-ヒト属性の抽出を行った[23]．最適クラスタリングアルゴリズムとして階層的クラスタリングが選択され，USA に属するヒトのみが含まれるクラスタ情報（*Bacteroides* と *Prevotella* の検出割合）を発見的に抽出することが可能となった[29]．これに対し，今回は被験者 60 人，61 種類の門レベルの細菌データを対象として，性別に関する文脈を使用し実験を行った．

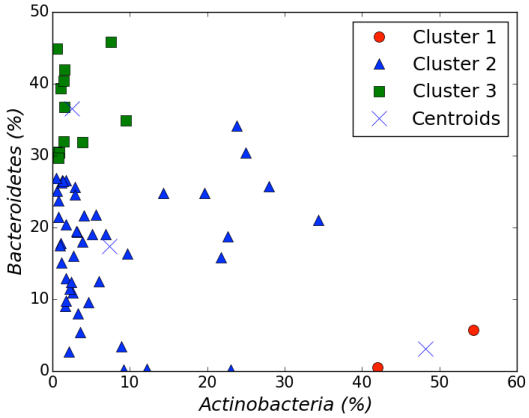
本実験の目的は，性別文脈を用いた腸内細菌叢-ヒト属性関連性を抽出することにより，ヒト属性と腸内細菌種の数値的傾向として定量的に抽出する本方式の実現可能性を示すことにある．結果を図 5 に示す．



(a)



(b)



(c)

図 5　性別文脈を用いた階層的クラスタリングによる腸内細菌叢-ヒト属性関連性抽出結果: (a) 性別属性ごとのオリジナルデータの散布図，(b) k-means クラスタリングと階層的クラスタリングを 25 回実行した結果の平均評価値比較，(c) オリジナルデータを対象として階層的クラスタリングを実行した結果のデータ散布図．

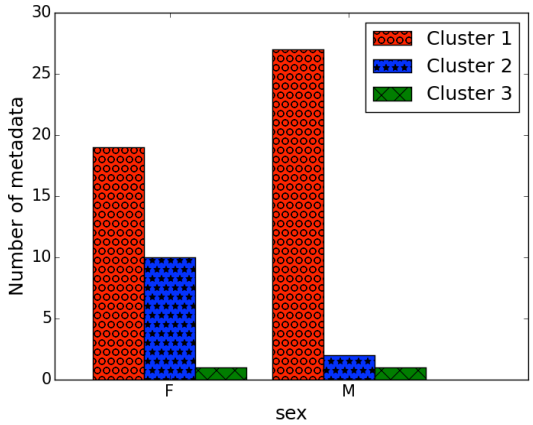
正規化手法による影響を考慮し，正規化前のオリジナルデータを使用した．性別文脈を使用した分析データの限定後，60 人の細菌データを対象として，性別属性ごとのオリジナルデータの散布図を作成した（図 5(a)）．クラスタリングアルゴリズムの評価方法として，全クラスタに属するデータ数における各クラスタのデータ数の割合（%）と，各性別に属するデータ数における性別及びクラスタごとに属するデータ数の割合（%）の積を設定した．各クラスタリングアルゴリズムにより得られた評価値を比較し（図 5(b)），数値的傾向を抽出するために，性別文脈により指定された 2 種の細菌種の組み合わせを用いた散布図を作成した（図 5(c)）．性別属性メタデータをカウントした結果（出現頻度）により 2 種の散布図（性別属性ごとのオリジナルデータの散布図とオリジナルデータを対象とした階層的クラスタリング結果のデータ散布図）を比較し，各クラスタの重心値（centroid）として数値的傾向を抽出する．

これらの結果により，12 人のヒトに特徴的な細菌データクラスタ（Cluster 2）の内，性別属性メタデータをカウントした結果，10 人の女性のメタデータが確認されたことから，Cluster 2 は女性が持つ腸内細菌叢の特徴的な傾向を示すと考えられる（図 5(d)）．先行研究では性別属性ごとの *Bacteroidetes* の差について示されており，本実験により検証することができた．

クラスタ情報である *Actinobacteria* と *Bacteroidetes* の検出割合の重心値を確認すると，Cluster 2 は *Actinobacteria* の検出割合が低く，*Bacteroidetes* の検出

割合が高いという他のクラスタとは異なる傾向を示す（図 5(e)）．この結果は腸内細菌叢とヒト属性（女性）との知られていない関係性を示していると考えられる．現在の一般的な治療においては，ヒトの属性は考慮されていないが，本実験の結果により，少なくとも *Actinobacteria* と *Bacteroidetes* については，個人ごとにより適した治療法として，性別ごとに異なる治療法を提案出来る可能性が見られた．

しかし，本実験では，女性が持つ腸内細菌叢の特徴を捉えることは可能となったが，男性については腸内細菌叢-ヒト属性の関連性を抽出することは出来なかった．このため，階層的クラスタリングを用いた腸内細菌層-ヒト属性関連性抽出方式は，女性と腸内細菌叢の関連性抽出には適しているが，同じ性別文脈により可視化することが出来る男性が持つ腸内細菌については一概に階層的クラスタリングが適しているとは言えない．データ非依存的な有意な関連性抽出を行うためには，使用する文脈のみではなく，着目する性別属性によってもクラスタリングアルゴリズムの検討が必要であると考えられる．



(d)

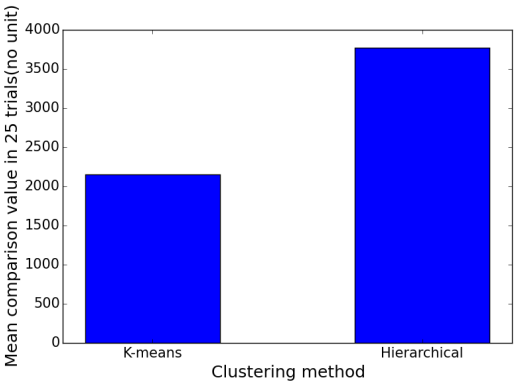
(%)	*Actinobacteria*	*Bacteroidetes*
Cluster 1	48.1818175	3.1057506
Cluster 2	2.55474989	36.55008785
Cluster 3	7.34318324	17.38253923

(e)

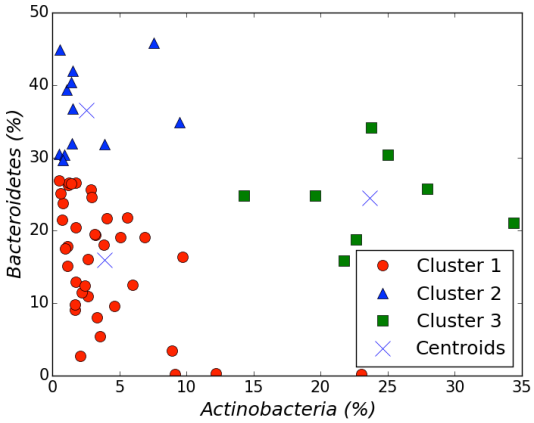
図 5　性別文脈を用いた階層的クラスタリングによる腸内細菌叢-ヒト属性関連性抽出結果: (d)各クラスタに含まれる性別属性メタデータをカウントした結果，(e)各クラスタの重心値（centroid）

4.2. 実験 B: 外れ値を除外後，58 人分の腸内細菌叢データを使用した腸内細菌叢-性別属性関連性抽出

本実験の目的は，外れ値を考慮した腸内細菌叢-ヒト属性関連性抽出により，腸内細菌叢の全データを使用時より有意な腸内細菌叢-ヒト属性関係性を抽出する本方式の実現可能性を示すことにある．結果を図 6 に示す．



(a)



(b)

図 6　性別文脈を用いた階層的クラスタリングによる腸内細菌叢-ヒト属性関連性抽出結果: (a) k-means クラスタリングと階層的クラスタリングを 25 回実行した結果の平均評価値比較，(b) オリジナルデータを対象として階層的クラスタリングを実行した結果のデータ散布図．

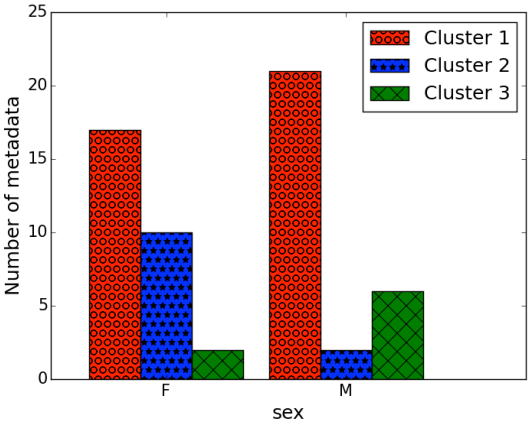
正規化手法による影響を考慮し，正規化前のオリジナルデータを使用した．性別文脈を使用した分析データの限定後，実験 A と同様に各クラスタリングアルゴリズムにより得られた評価値を比較し（図 6(a)），数値的傾向を抽出するために，性別文脈により指定された 2 種の細菌種の組み合わせを用いた散布図を作成した（図 6(b)）．各クラスタ内に含まれる性別属性メタデー

タをカウントし（図 6(c)），各クラスタの重心値（centroid）として数値的傾向を抽出した（図 6(d)）．

これらの結果により，実験 A と同様に 12 人のヒトに特徴的な細菌データクラスタ（Cluster 2）の内，性別属性メタデータをカウントした結果，10 人の女性のメタデータが確認されたことから，Cluster 2 は女性が持つ腸内細菌叢の特徴的な傾向を示すことが考えられる．さらに，8 人のヒトに特徴的な細菌データクラスタ（Cluster 3）の内，性別属性メタデータをカウントした結果，6 人の男性のメタデータが確認されたことから，Cluster 3 は男性が持つ腸内細菌叢の特徴的な傾向を示すという実験 A では得られなかった結果を得ることが出来た．先行研究では *Actinobacteria* については性別属性との有意な関連性が抽出されていなかったため，実験 B により腸内細菌叢-ヒト属性関連性を発見的に抽出できる可能性について示すことができた．

クラスタ情報である *Actinobacteria* と *Bacteroidetes* の検出割合の重心値を確認すると，Cluster 3 は *Actinobacteria* の検出割合が高いという他のクラスタとは異なる傾向を示す．この結果は腸内細菌叢とヒト属性（男性）との知られていない関係性を示していると考えられる．現在の一般的な治療においては，ヒトの属性は考慮されていないが，本実験の結果により，少なくとも *Actinobacteria* と *Bacteroidetes* については，個人ごとにより適した治療法として，性別ごとに異なる治療法を提案出来る可能性が見られた．

実験 B では，先行研究により性別属性と細菌データの関連性が示唆されている *Actinobacteria* と *Bacteroidetes* については，*Bacteroidetes* と女性の関連性という先行研究で得られた知見の検証，および，*Actinobacteria* と男性の関連性という先行研究では得られなかった知見を発見的に抽出することができたが，細菌データに含まれる他の細菌種については腸内細菌叢-ヒト属性の関連性を抽出することは出来なかった．このため，腸内細菌叢-ヒト属性関連性の統合的抽出方式により得られる腸内細菌叢-ヒト属性関連性が先行研究の知見に影響される可能性がないとは言えない．データ非依存的な有意な関連性抽出を行うためには，使用する文脈に含まれる細菌種の検討が必要であると考えられる．



(c)

(%)	*Actinobacteria*	*Bacteroidetes*
Cluster 1	3.90569671	15.89418646
Cluster 2	2.55474989	36.55008785
Cluster 3	23.67124425	24.4522149

(d)

図 6　性別文脈を用いた階層的クラスタリングによる腸内細菌叢-ヒト属性関連性抽出結果: (c)各クラスタに含まれる性別属性メタデータをカウントした結果，(d)各クラスタの重心値（centroid）

5. 結論

本稿では，文脈解釈機能及び複数のクラスタリング手法を用い，性別文脈により指定された一部のデータにおいて，腸内細菌叢データとヒトの属性との関係性の統合的抽出方式を提案した．本方式により，腸内細菌叢とある特定のヒト属性との関連性として，最適アルゴリズムとして階層的クラスタリングが選択され，女性に関連するクラスタ情報 (*Actinobacteria* と *Bacteroidetes* の検出割合)が腸内細菌叢-ヒト属性関連性として抽出することが可能となった．

今後は，対象細菌データに応じた適切な正規化手法の検討，及び，本方式に各種文脈や異なる細菌データを適用する場合の有効性についても検証する実験を行う予定である．

参 考 文 献

[1] Bäckhed,F. et al,2005. Host-bacterial mutualism in the human intestine. *Science*,307,1915-1920.

[2] Rehman,A. et al,2015. Geographical patterns of the standing and active human gut microbiome in health and IBD. *Gut.*

[3] Yoshimoto,S. et al,2013. Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. *Nature*,499,97-+.

[4] Takano,K. et al,2005. A semantic Associative Search Method with Dynamic Context-Awareness Functions for Computing Causal Relations of Event Data Sets. *TOD*,46,SIG5(TOD25),40-55.

[5] Kiyoki,Y., Kitagawa,T. and Hayama,T.,1994. A metadatabase system for semantic image search by a mathematical model of meaning. *ACM SIGMOD Record*,23(4),34-41.

[6] Kiyoki,Y. and Kitagawa,T.,1995. A semantic associative search method for knowledge acquisition. *Information Modelling and Knowledge Bases (IOS Press)*,VI,121-130.

[7] Chen,J. et al,2012. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*,28,2106-2113.

[8] Lozupone,C. and Knight,R.,2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*,71,8228-8235.

[9] Clarke,S.F. et al,2014. Exercise and associated dietary extremes impact on gut microbial diversity. *Gut*,63,1913-1920.

[10] Walters,W.A. et al,2014. Meta-analyses of human gut microbes associated with obesity and IBD. *Febs Letters*, 588, 4223-4233.

[11] Yatsunenko,T. et al,2012. Human gut microbiome viewed across age and geography. *Nature*, 486, 222-+.

[12] McKusick,V.A.,2007. Mendelian inheritance in man and its online version, OMIM. *American Journal of Human Genetics*,80,588-604.

[13] Becker,K.G. et al,2004. The Genetic Association Database. *Nature Genetics*,36,431-432.

[14] GeneExpressionOmnibus:http://www.ncbi.nlm.nih.gov/geo/

[15] ArrayExpress:http://www.ebi.ac.uk/microarray-as/ae/

[16] StanfordMicroarrayDatabase:http://smd.stanford.edu/

[17] GO:http://www.geneontology.org/

[18] D'Argenio,V. et al,2014. Comparative Metagenomic Analysis of Human Gut Microbiome Composition Using Two Different Bioinformatic Pipelines. *Biomed Research International.*

[19] Han,J. and Kanber,M., 2000. Data mining: concepts and techniques. *Morgan Kaufmann Publishers*.

[20] Jain,A.K., Murty,M.N. and Flynn,P.J.,1999. Data clustering: a review. *ACM Computing Surveys*,31(3).

[21] Zushi,T. et al,2002. A semantic knowledge discovery method by recursively applying context dependent dynamic clustering to document data. *IPSJ Journal,*43,216-230.

[22] Kawamoto,M. et al,2003. An implementation method of semantic associative search spaces for medical documents.

[23] Hikichi,S. et al,2015. Human-microbiome-relations extraction and visualization system with context-dependent clustering and semantic analysis. *12th International Conference on Applied Computing 2015 (AC2015), Maynooth, Greater Dublin, Ireland,* accepted 8 pages, October 24-26.

[24] Suzuki,T.A. and Worobey,M.,2014. Geographical variation of human gut microbial composition. *Biology Letters*, 10.

[25] Dominianni,C. et al,2015. Sex, Body Mass Index, and Dietary Fiber Intake Influence the Human Gut Microbiome. *Plos One,*10(4).

[26] Numpy:http://www.numpy.org/

[27] Scipy:http://www.scipy.org/

[28] Matplotlib:http://matplotlib.org/

[29] Hikichi,S. et al,2015. Integrated Human Gut Microbiome Analysis System with Context-Awareness and Semantic-Analysis Functions, WebDB Forum 2015,Tokyo, Japan, November 24-25.