様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　　５　日現在

機関番号：　３２６６５
研究種目：　挑戦的萌芽研究
研究期間：　2015～2017
課題番号：　１５Ｋ１４４２３

研究課題名（和文）連続塩基出現頻度データベースの微生物ゲノム・遺伝子解析への応用

研究課題名（英文）Application of database on frequencies of nucleotide n-gram profiles to genomic analyses.

研究代表者

桑田　文幸（KUWATA, Fumiyuki）

日本大学・歯学部・特任教授

研究者番号：６０１２０４４０

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：　ゲノム塩基配列の報告されている真正細菌について４～６連続塩基の出現頻度を集計した。その頻度に基づき距離を計算し、16SrRNA遺伝子配列では系統分類が難しい属について系統解析が可能になった。菌叢構成種解析については、口腔内細菌の人工的な環境細菌叢試料を作り、５塩基配列の出現頻度による構成種比率を求めたところ精度の高い計算結果が得られた。水平伝播遺伝子解析については、同データベースを用いて１クラス・サポートベクターマシンを使って抽出された領域には遺伝子転位に係る遺伝子や転写・翻訳、移動性因子に関係する遺伝子が際立って多かった。機能がまだ報告されていない遺伝子がおよそ半数に達した。

研究成果の概要（英文）：In this study, we demonstrates construction of phylogenetic trees based on genome-wide comparisons based on n-gram profiles. n-gram frequency analysis was used to separate species that are difficult to distinguish based on 16S rRNA gene sequences. Next, the relative abundances of artificial bacterial species mixture were estimated. Frequencies of five continuous nucleotides were calculated from the obtained sequences and their proportional compositions were estimated by the combinations of frequencies from genomic data and samples. Origins of the fragments were determined by BLAST and were compared with the abundances based on nucleotide frequencies. The both results were agreed closely and furthermore the similar result obtained from 10 sequences of five continuous nucleotides with high entropy values. In addition, horisontal-transfered genes were estimated by one-class support vector machine based on continuous nucleotide frequencies within each genome.

研究分野：　口腔生化学

キーワード：　菌叢解析　サポートベクターマシン

１．研究開始当初の背景

生物種の塩基配列の並び方の組み合わせには、それぞれ偏りがある。これまで、連続配列の出現頻度を利用した解析では、自己組織化マップと組み合わせたメタゲノム解析などが報告されてきた［T. Abe, et al. Poler Res. 20: 103 (2006)］が、自己組織化マップ以外のパターン認識や機械学習等の手法と組み合わせた報告がほとんどない。
本研究では以下の３点に着目することにした。

(1) 細菌の系統解析：系統解析には主に 16S rRNA 遺伝子配列が利用されているが、近縁種の識別が難しく、水平伝播によって由来の異なる配列が混在する可能性などの問題がある。他の遺伝子を指標とする場合は、指標遺伝子の選択と評価が容易ではない。ゲノム全体の連続塩基配列出現頻度による比較ならば、どの種であっても等しく同じ指標で比較することが可能となる。

(2) 細菌叢の構成分析：環境由来の試料中の菌種の割合・分布を分析するためにも、16S rRNA 遺伝子配列による解析が行われている。16S rRNA 断片の種類と数の集計には、一つの菌が複数の 16S rRNA 遺伝子を保有していること、その数が種によって異なること、水平伝播の可能性などの問題がつきまとう。これを、連続配列出現頻度によって推定する方法が確立すれば、短時間に多数の試料の菌叢解析が可能になる。

(3) 細菌種間の水平伝播の検出：種ごとに固有の連続配列出現頻度を持つならば、水平伝播に由来するゲノム領域は、全体とは異なる頻度を示すはずである。遺伝子の機能に依存せず、網羅的に異種由来の領域を抽出できるはずである。そのためには、１クラス・サポートベクターマシンが適していると考えた。

２．研究の目的

研究の目的も大きく３つある。上記の細菌系統解析、細菌叢の種構成分析、そして、細菌ゲノム間の水平伝播の検出である。何れもゲノム中の４～６塩基の連続配列出現頻度の偏りを利用することで、これまでの解析方法とは異なる手法を開発するのが目的である。特に最初の２つは現在、16S rRNA 配列によって解析がなされているが、それに依存しない方法を開発できることになる。

３．研究の方法

(1) 細菌系統解析
種ごとの連続塩基配列出現頻度をゲノム配列が既知の種において集計する。すなわち、例えば５塩基連続配列であれば、AAAAA、AAAAC、AAAAG……TTTTT までの 1024 の組合せとなる。ここで遺伝子の方向による配列の偏りを排除するために、相補配列を統合して、例えば AAAAA と TTTTT を同一配列として集計すれば、512 種の組合せとなる。これで、一つの種に対して 512 項のベクトルができる。得られたベクトルの距離、あるいは角度を、それぞれの種の間の距離として計算する。その計算は Euclidean 距離と Ward アルゴリズムで求めた。
比較する 16S rRNA 配列に基づく系統樹は当該遺伝子全長を用いて、Neighbor-joining method にて解析した。

(2) 細菌叢の構成分析
前項のベクトルを並べると、種の数が n、連続配列の種類が m であるとき、サンプル中のメタゲノム配列中に含まれる m 番目の連続塩基配列の出現数 b は、上記のように各細菌種 $\boldsymbol{x_n}$ の存在比を a とするならば、n 番目の種が全体に占める割合は下記のようになる。

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + a_{m3}x_3 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

各細菌ゲノム中の連続塩基配列出現頻度行列を $\boldsymbol{A}$ として、各種の構成比を並べたベクトルを $\boldsymbol{x}$ とするならば、メタゲノム解析結果全体の連続塩基配列出現数ベクトルを $\boldsymbol{b}$ とすると、

$$\boldsymbol{Ax = b}$$

という関係が成り立つ。このとき、

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

となっており、この式は、$\boldsymbol{A}$ の一般逆行列を $\boldsymbol{A^-}$とすれば、

$$x = A^{-}b$$

という形で解くことができる。この x は細菌種構成比の最適（近似）解となる。

512 種の連続塩基配列はベクトルの次元が多過ぎるので、ここで、連続塩基配列がゲノム集団全体を通して出現する頻度を F_i とすると、連続塩基配列 i がゲノム j に出現するという事象 E_j の確率は f_{ij}/F_i となるので、このときのエントロピー H は次で与えられる。

$$H = -\sum_{j=1}^{n} \frac{f_{ij}}{F_i} \log f_{ij} F_i$$

本実験で使用した細菌は、以前報告した論文に記載したとおりの方法で培養し、DNA を抽出した［Takeshita, T., Nakano, Y. & Yamashita, Y. Oral Microbiol. Immunol. 22, 419–428 (2007)］。精製した DNA の混合液を解析業者（北海道システムサイエンス）に委託して MiSeq 法 (Paired-end) で 100 塩基断片の配列解析を行った。通常の塩基配列の相同性による構成比解析では 100 塩基では足りないが、本法では十分であることを確認した。各サンプルから 2300 万リードの配列を得た。低精度の配列を除去した。the Fastx-toolkit (http://hannonlab.cshl.edu/fastxtoolkit/) を用いて、80%以上の配列において quality score >20 を示していないものを除去した。プライマー配列は、cutadap を用いて取り除いた。

(3) 細菌種間の水平伝播の検出

NCBI の Microbiol Genome Resources から古細菌を除いた真正細菌のゲノム配列 2604 を得た。全塩基配列から 2k 塩基ごとの断片を 1k 刻みで 5'末端から作成し、5 塩基配列出現頻度を集計した。R 3.1.0 のライブラリ kernlab を用いて、One-Class SVM を行い、平均的 5 塩基配列出現頻度から離れた 2kb 断片を抽出した。

得られた遺伝子のアミノ酸配列のうち、昨日の不明な遺伝子を cd-hit を用いて 70%以上の相同性のあるクラスターにまとめた。100 以上の配列を含むクラスターを抜きだしたところ 95 クラスターが得られた。得られたそれぞれの代表配列を、Pfam データベースに対してモチーフ検索を行った。

4．研究成果

(1) 細菌系統解析

16S rRNA 配列での系統解析が難しい ***E. coli*/*Shigella*** の解析を本方法にしたがって行ったところ、図 1 に示すように、***E. coli*** と ***Shigalla*** 属細菌を分けることができた。

Yersinia 属細菌も 16S rRNA 配列による解析が困難な種として知られている。この属には 3 つの種が含まれているが、それが本方法によりはっきり系統樹上で分けられた。***Y. pestis*** と ***Y. pseudotuber-***

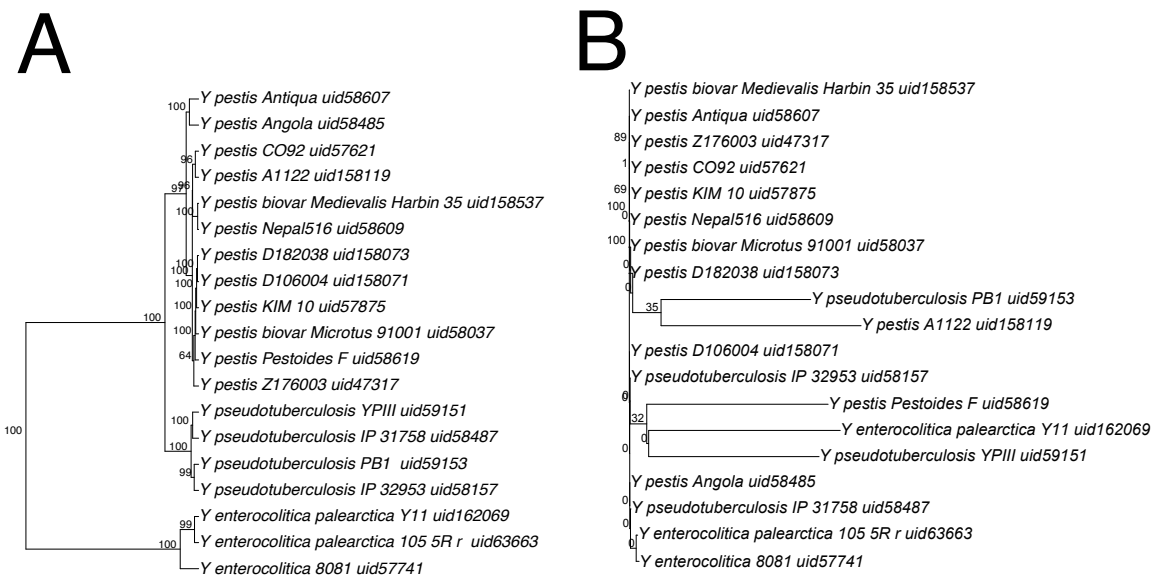

図 1． ***Yersinia*** 属細菌の系統樹解析。A: 5 塩基連続配列出現頻度に基づく系統樹 B: 16S rRNA 配列に基づく系統樹

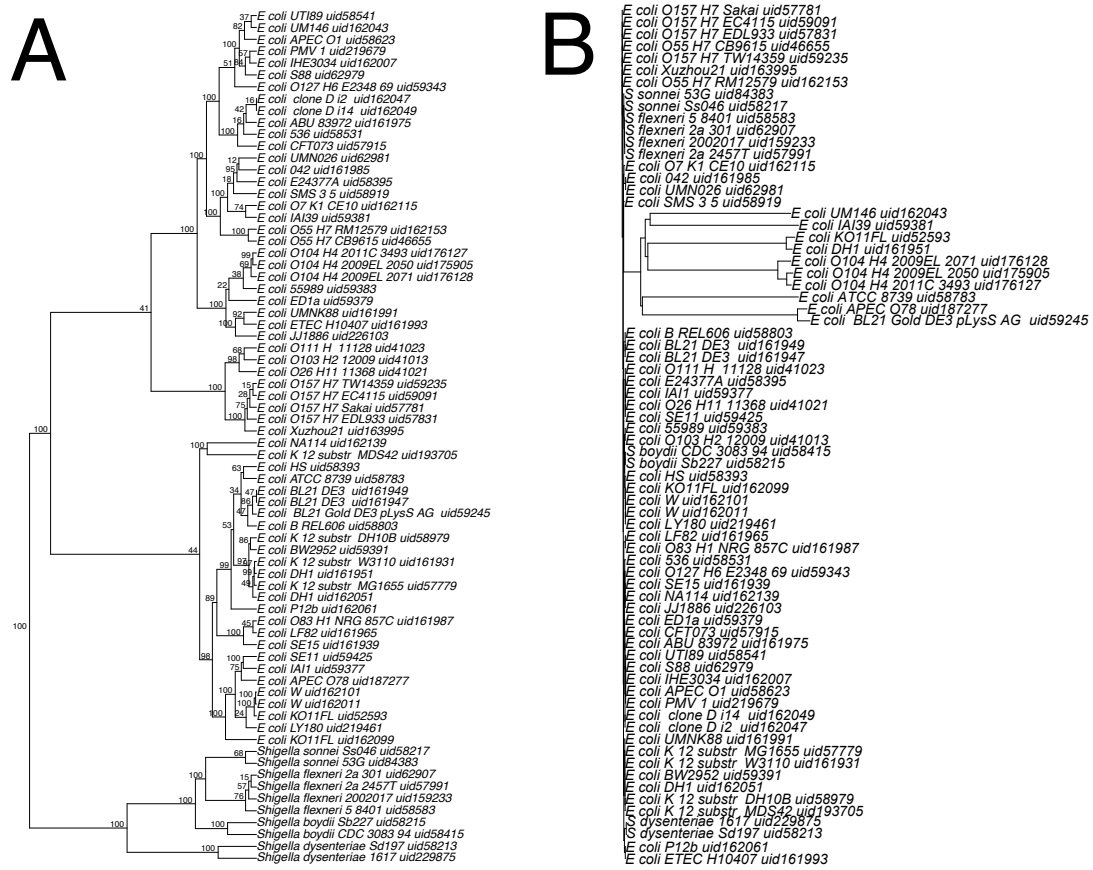

図２．***E. coli*** と ***Shigella*** 属細菌の系統樹解析。A: ５塩基連続配列出現頻度に基づく系統樹 B: 16S rRNA 配列に基づく系統樹

culosis の 16S rRNA 塩基配列はまったく同一である。したがって、16S rRNA の配列に基づいた系統解析で両種を解析することはできない。しかし、今回の研究成果から、これらの種を含む３つの種を５塩基連続配列の出現頻度に基づく明確に分けることができた。***Y. pestis*** は 1,500–20,000 年前に ***Y. pseudotuberculosis*** から分岐したと考えられている。このことから、今回の結果は連続塩基配列の解析によってこれら２つの種を明確に識別できたことが示されているといえよう。

同様に ***E. coli*** と ***Shigella*** 属細菌の４つの種は、16S rRNA 遺伝子の塩基配列のみでははっきり分けることができない。ここでも５塩基連続配列の出現頻度に基づく解析で、高ブートストラップ値ではっきりと識別することができた。図に見られる３クラスター内で ***Shigella*** 属細菌は 35,000–270,000 年前に分かれたと考えられている。したがって、大腸菌から ***Shigella*** 属細菌を分けることは、1,500–20,000 年前に分かれた ***Y. pestis*** と ***Y. pseudotuberculosis*** を識別できた５塩基配列出現頻度に基づく解析で十分分離できるのは意外なことではない。

５塩基連続配列の組合わせは 512 種類であるが、もっと計算の負担を軽減するために、もっとも出現頻度の変化が大きくサンプルの特徴を強く反映する５塩基配列を、方法欄に記したエントロピーの値が最も高い 10 配列を選び、上記と同様の解析

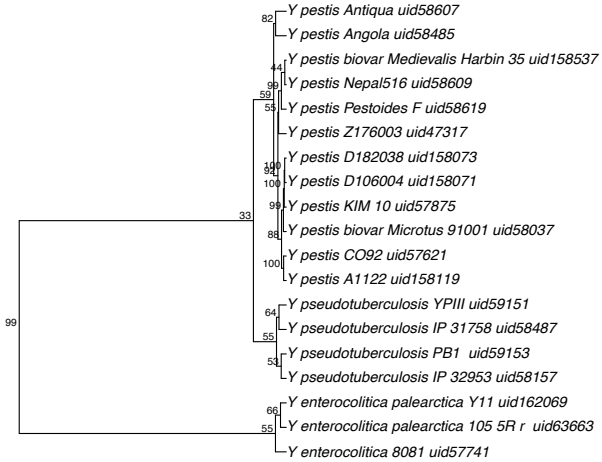

図３．図１と手法は同じだが、エントロピー値の高い 10 種の配列で解析した系統樹。

に使用した。その結果、図３に示したように、512 種を使ったときと遜色のない解析結果が得られた。

(2) 細菌叢の構成分析

方法欄に記した方法で５塩基連続配列の出現頻度に基づき、人工的に調製した混合菌液の菌叢解析を、構成する菌の割合を変えて行った。表１に示したように、Blast 解析の結果と極めてよく一致した結果が得られた。さらに菌種を増やした表２の場合でも、構成比率を変えて解析してみても、いずれの場合も良い結果が得られた。このことから、５塩基連続配列の出現頻度により菌叢を構成する種の種類と割合が推定できることが示唆された。16S rRNA 遺伝子の配列に依存しないので、増幅による偏り、rRNA 遺伝子数の違い等の影響を受けない。また、特定の菌の特定の領域にプライマーを設計する必要もない。さらに、高エントロピー値を示した連続配列のみで解析を行ったときでも、両表に示したように、512 種を使った結果と遜色のない値が得られた。現在、実際の口腔内サンプルを使って、菌叢構成をどれくらい正確に推定できるかを検証中である。

(3) 細菌種間の水平伝播の検出

方法の欄に記した手順に従って、抽出した 2 kb 断片を Blast 解析した結果、遺伝子の移動や伝播に関係する遺伝子、RNA 関連遺伝子、複製・修復に関係する遺伝子等が多く含まれていることが判った（図４及び５）。

このうちおよそ半数が機能未知の遺伝子だった。そこで、その遺伝子をまとめて cd-hit を用いて相同性に基づくグループを作ったところ、100 以上の配列を含むクラスターが 95 も得られた。得られたそれぞれの代表配列を、Pfam データベースでモチーフ検索した結果が図６である。98 配列のうち、46 配列が機能未知のモチーフとヒットした。残り 52 配列は、図６に示すように、DNA 結合領域や発現制御タンパク等のモチーフと一致するものが見られるものの、特徴的な特性は見出せなかった。

５塩基連続配列の出現頻度の偏りによって水平伝播がもたらした遺伝子を見出せる可能性が示唆されたが、その領域の解析は本研究の範囲を越えるものであり、今後の検討課題としたい。

表１．５連続配列の出現頻度に基づく菌叢構成種の解析（４種混合）

Bacterial species	5gram	5gram with entropy	Blast
A. actinomycetemcomitans	0.19	0.19	0.20 (739708)
F. nucleatum	0.29	0.27	0.29 (1091592)
P. gingivalis	0.20	0.19	0.22 (820849)
S. mutans	0.31	0.31	0.30 (1139395)
A. actinomycetemcomitans	0.32	0.32	0.32 (1814712)
F. nucleatum	0.14	0.14	0.13 (1210886)
P. gingivalis	0.042	0.050	0.046 (534209)
S. mutans	0.49	0.48	0.49 (187945)

表２．５連続配列の出現頻度に基づく菌叢構成種の解析（７種混合）

Bacterial species	5gram	5gram with entropy	Blast
A. actinomycetemcomitans	0.12	0.12	0.13 (452129)
A. naeslundii	0.050	0.057	0.34 (117545)
E. coli	0.16	0.16	0.17 (611121)
F. nucleatum	0.28	0.30	0.29 (1016541)
P. gingivalis	0.065	0.061	0.077 (269528)
P. intermedia	0.20	0.15	0.17 (606719)
S. mutans	0.13	0.16	0.12 (433699)
A. actinomycetemcomitans	0.072	0.071	0.078 (267213)
A. naeslundii	0.13	0.14	0.082 (279458)
E. coli	0.11	0.095	0.12 (418595)
F. nucleatum	0.24	0.19	0.25 (865380)
P. gingivalis	0.033	0.030	0.045 (154092)
P. intermedia	0.11	0.095	0.096 (329682)
S. mutans	0.31	0.35	0.32 (1109353)
A. actinomycetemcomitans	0.019	0.017	0.020 (70987)
A. naeslundii	0.050	0.053	0.031 (111009)
E. coli	0.20	0.19	0.21 (746480)
F. nucleatum	0.12	0.13	0.12 (439608)
P. gingivalis	0.015	0.021	0.021 (77319)
P. intermedia	0.039	0.038	0.029 (104222)
S. mutans	0.56	0.57	0.57 (2051907)

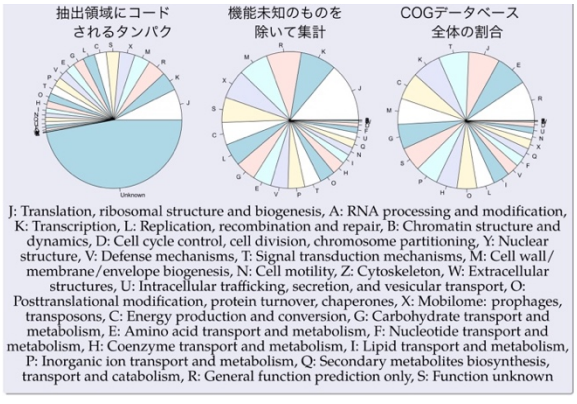

図４．OC-SVM によって抽出した領域に含まれていた遺伝子の Blast 検索結果による機能予測。右は対象のため、COG データベース全体の機能分布。

以上、短い連続配列の出現頻度を利用した、系統解析、細菌叢構成種の計算、水平伝播の予測について研究を行い、それぞれにおいて、本方法が有効であることが示さ

れた。

ATP_bind_2	Cluster314478	9.1E-147	P-loop_ATPase_protein_family
BMFP	Cluster719205	1.2E-28	Membrane_fusogenic_activity
CreA	Cluster609727	1.9E-57	CreA_protein
DUF1451	Cluster557116	1.7E-53	Zinc-ribbon_containing_domain
DUF150	Cluster509801	3.4E-28	RimP_N-terminal_domain
DUF150_C	Cluster509801	6E-22	RimP_C-terminal_SH3_domain
DUF3811	Cluster872585	8.6E-39	YjbD_family_(DUF3811)
DUF463	Cluster101113	3.8E-182	YcjX-like_family,_DUF463
EpmC	Cluster551242	2.8E-82	Elongation_factor_P_hydroxylase
Fe-S_assembly	Cluster10133542	3.E-32	Iron-sulphur_cluster_assembly
HNH	Cluster719680	7E-12	HNH_endonuclease
KGG	Cluster956554	1.5E-17	Stress-induced_bacterial_acidophilic_repeat_motif
Kinase-PPPase	Cluster302622	6.1E-74	Kinase/pyrophosphorylase
Lysine_decarbox	Cluster92819	1.5E-32	Possible_lysine_decarboxylase
MatP	Cluster573603	3.6E-42	MatP_N-terminal_domain
MatP_C	Cluster573603	1.7E-31	MatP_C-terminal_ribbon-helix-helix_domain
OsmC	Cluster715424	3.1E-20	OsmC-like_protein
Peptidase_M15_3	Cluster426358	7E-13	Peptidase_M15
Pmp3	Cluster874503	7.4E-19	Proteolipid_membrane_potential_modulator
Polyketide_cyc	Cluster580800	3.6E-33	Polyketide_cyclase_/_dehydrase_and_lipid_transport
Rick_17kDa_Anti	Cluster468105	3.1E-11	Glycine_zipper_2TM_domain
SNARE_assoc	Cluster422700	3.2E-25	SNARE_associated_Golgi_protein
Sdh5	Cluster888522	5.3E-24	Flavinator_of_succinate_dehydrogenase
Sel_put	Cluster831139	4.6E-27	Selenoprotein,_putative
Smr	Cluster508786	5E-18	Smr_domain
Sua5_yciO_yrdC	Cluster466975	2.1E-46	Telomere_recombination
TamB	Cluster6334	2.7E-69	TamB,_inner_membrane_protein_subunit_of_TAM_complex
ThrE_2	Cluster629659	1.6E-34	Threonine/Serine_exporter,_ThrE
Transcrip_reg	Cluster315387	1.6E-87	Transcriptional_regulator
Transcrip_reg	Cluster362603	9.8E-94	Transcriptional_regulator
Transcrip_reg	Cluster364712	4.7E-81	Transcriptional_regulator
Transcrip_reg	Cluster413975	2.1E-76	Transcriptional_regulator
TusA	Cluster893998	2.4E-19	Sulfurtransferase_TusA
Ub-RnfH	Cluster775173	5.1E-34	RnfH_family_Ubiquitin
VEG	Cluster868841	1.4E-23	Biofilm_formation_stimulator_VEG
Vut_1	Cluster402719	1.7E-37	Putative_vitamin_uptake_transporter
YaiA	Cluster10267602	9.E-41	YaiA_protein
YbaB_DNA_bd	Cluster684063	2.8E-35	YbaB/EbfC_DNA-binding_family
YbaB_DNA_bd	Cluster794566	2.9E-26	YbaB/EbfC_DNA-binding_family
YbjQ_1	Cluster792077	1.1E-39	Putative_heavy-metal-binding
YcaO	Cluster44892	1E-81	YcaO_cyclodehydratase,_ATP-ad_Mg2+-binding
YccF	Cluster573053	5.9E-28	Inner_membrane_component_domain
YceI	Cluster506223	9.7E-42	YceI-like_domain
YcgL	Cluster761241	4.4E-32	YcgL_domain
YdfA_immunity	Cluster234027	1.7E-163	SigmaW_regulon_antibacterial
YejG	Cluster765317	1.9E-49	YejG-like_protein
YfbU	Cluster577121	2.8E-69	YfbU_domain
YfcL	Cluster891856	3.6E-37	YfcL_protein
YicC_N	Cluster262704	4E-44	YicC-like_family,_N-terminal_region
YjbR	Cluster790446	2.6E-18	YjbR
Zn_peptidase	Cluster310997	1.5E-139	Putative_neutral_zinc_metallopeptidase
tRNA_edit	Cluster632257	4.7E-24	Aminoacyl-tRNA_editing_domain

図６．Blast 検索で機能が見出せなかった遺伝子群を cd-hit でまとめた代表配列によるモチーフ検索

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

1. N. Suzuki, Y. Nakano, T, Watanabe, M. Yoneda, T. Hirofuji, T. Hanioka (2018) Two mechanisms of oral malodor inhibition by zinc ions. J. Appl. Oral Sci. e20170161 査読有
DOI: 10.1590/1678-7757-2017-0161

〔学会発表〕（計５件）

① 中野善夫、谷口奈央、桑田文幸　深層学習による口腔内菌叢解析に基づく口臭の判別、2017 年度生命科学系学会合同年次大会（2017 年）
② 中野善夫、谷口奈央、桑田文幸、埴岡隆：細菌叢解析に基づく機械学習による口臭の判別、第 59 回歯科基礎医学会（2017 年）
③ 中野善夫：口腔内細菌叢の 5 連続塩基出現頻度に基づく解析法、第 90 回日本細菌学会（2016 年）
④ 中野善夫、谷口奈央、桑田文幸：連続塩基出現頻度に基づいた菌叢構成種解析、第 40 回日本分子生物学会（2016 年）
⑤ 中野善夫、桑田文幸、谷口(鈴木)奈央：n-gram 塩基出現頻度に基づく微生物系統樹解析、第 38 回日本分子生物学会（2015 年）

〔図書〕（計０件）

〔産業財産権〕

○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
桑田 文幸（KUWATA, Fumiyuki）
日本大学・歯学部・特任教授
研究者番号：60120440

(2)研究分担者
谷口 奈央（TANIGUCHI, Nao）
福岡歯科大学・口腔歯学部・准教授
研究者番号：60372885

中野 善夫（NAKANO, Yoshio）
日本大学・歯学部・教授
研究者番号：80253459