様式 C-19

科学研究費補助金研究成果報告書

平成２２年　４月　１日現在

研究種目：特定領域研究
研究期間：　２００５～２００９
課題番号：　１７０１８０２６
研究課題名（和文）　共生、相互作用など複雑なゲノム構成系を解析するための情報基盤研究

研究課題名（英文）　Computational Biology for complicated genome system

研究代表者
久原　哲　(KUHARA　Satoru)
九州大学・大学院農学研究院・教授
研究者番号：00153320

研究成果の概要（和文）：生物の特徴である普遍性と多様性のメカニズムを解明するために、配列データを整備し、ゲノム比較を行うシステムを開発した。メタゲノム情報から属分類解析を行うツールを開発し、腸内細菌叢を解析し、離乳後で細菌叢が劇的に変化することを明らかにした。また、細菌叢解析用の 16S r DNA 情報を用いたマイクロアレイを開発した。最後に、遺伝子発現データに基づく疾患等の責任遺伝子群の識別能力の高いカーネル部分空間法を開発した。

研究成果の概要（英文）：
Our intensive research using comparative genomics and computational analysis was responsible for the following successes:
1) We improved the MBGD database allowing the identification of core structures amongst moderately related microbial genomes based on multiple genome alignments. We constructed a comparative genome alignment analysis tool for the visualization of complex evolutionary changes between closely related genomes.
2) We created systematic tools for meta-genome analysis and performed a large-scale comparative meta-genomic analysis of fecal samples. Our data clearly demonstrated a difference in overall composition and gene repertoire between adult- and infant-type gut microbiomes.
3) We developed and applied a new filter based approach to gene subset selection for kernel-based classifiers. We derived kernel forms from several well-known class separability criteria and applied gene subset selection based on the kernelized criteria to microarray cancer classification. The results have demonstrated that our proposed strategy performs better than gene ranking and the conventional filter approach.

交付決定額　（金額単位：円）

	直接経費	間接経費	合　計
２００５年度	18,000,000	0	18,000,000
２００６年度	15,900,000	0	15,900,000
２００７年度	15,900,000	0	15,900,000
２００８年度	16,000,000	0	16,000,000
２００９年度	16,000,000	0	16,000,000
総　計	81,800,000	0	81,800,000

研究分野：バイオインフォマティクス
科研費の分科：ゲノム科学　細目：ゲノム情報科学
キーワード：モデル化、遺伝子、ゲノム、進化、発現解析

１．研究開始当初の背景

生命・生物の特徴である普遍性と多様性のメカニズムを解明するために、ゲノム配列の比較と遺伝子の使われ方あるいは相互作用・ネットワークの違いを解析する情報学的基盤システムの構築については、基盤データは蓄積されていたが、その解析のシステム化は道半ばであり、その完成が望まれていた。配列データに種々の生物学情報、遺伝子発現データ等を統合した配列比較情報解析システムへの拡張する必要性が望まれていた。また、今後の微生物研究の基盤となる難培養性微生物を含む微生物集団（土壌中、腸管内等）の解析は開始されたばかりであり、その解析ツール等の整備は急務であった。

２．研究の目的

生命・生物の特徴である普遍性と多様性のメカニズムを解明するために、ゲノム配列の比較と遺伝子の使われ方あるいは相互作用・ネットワークの違いを解析する情報学的基盤システムの構築を行う。基盤システムとしては、モデル生物として微生物間のオルソログ、パラログ遺伝子あるいは特異的な遺伝子の自動作成とデータベース化を行い、このデータに種々の生物学情報、遺伝子発現データ等を統合した配列比較情報解析システムに拡張する。また、最近、特に注目されている土壌や海洋あるいは腸内といった環境中における微生物の群集をまるごとゲノム解析するメタゲノム解析にも上記で開発した手法を適用する。このメタゲノム解析の困難さの原因とされている、遺伝子データベースのデータおよび解析ツールの不十分さに注目する。本研究では、これらの原因を解消するため、基盤となるデータベースの整備、ツールの開発も同時に行う。同時に今後の微生物研究の基盤となる難培養性微生物を含む微生物集団（土壌中、腸管内等）でのポピュレーションを計測する新チップの設計・製作も行う。

３．研究の方法

１）配列データの整備

網羅的なオーソログ解析に基づく微生物比較ゲノム解析データベースMBGDについて、データの拡充を進めるとともに、新規ゲノム解析に対するアノテーションづけに有用なデータベースに向けての開発を行う。このため、オーソロググループに対する機能アノテーションを充実させるとともに、利用者が持つゲノムを登録して比較解析に加える機能を追加する。

ごく近縁ゲノム間のゲノムアライメントに基づいて、挿入、欠失、逆位などのゲノムの構造変化の様子を詳細に解析するためのツールを開発する。

中程度の類縁度のゲノム間で遺伝子の並び順がよく保存性された構造（コア構造）を構築する手法の開発を行い、コア構造に含まれる遺伝子がどのような特徴を持つかについて調べる。

２）生物学データの整備

細菌のゲノムに存在するすべての遺伝子について、個々に系統関係を分子進化学的解析により明らかにし、オーソログ、パラログ、シングルトンに分類することで、種への分化途上で起きたイベントを詳細に記述する。メタゲノム解析により得られた遺伝子の断片配列を既存のデータベースから、種、属、科、目などの分類群と、遺伝子相同性との関連性を明らかにするツールを開発し、腸内細菌科等に適用する。

３）情報学的ツールの整備

細菌のゲノムに存在する遺伝子について、オルソログ遺伝子に注目し、各オルソログ遺伝子間の類似度を計算することで、オルソログマトリックスを構築し、そのマトリックスに細菌の形質などの生物学的情報を付加したゲノムマトリックス等を構築する。実際のデータとしては、黄色ブドウ球菌の臨床分離株における遺伝子（オルソログ）と形質（病状、部位、感染場所など）間でマトリックスを作成する。臨床分離株における遺伝子の有無は、アレイCGHデータにより調べる。その後、様々な病状に関連性が深い遺伝子群をクラス識別解析により明らかにする。クラス識別法には、幾つかの統計的な手法が提案されているため、それらの手法を適用することで、識別の精度を上げることを試みる。

４）新規微生物分野の情報収集のための基盤整備

難培養性微生物を含む混合菌叢の菌学的性状を比較的簡便に比較解析できるチップを作成する。16S rDNA遺伝子から、種、属、株間の比較ができるプローブ配列を設計する。さらに、実際に土壌や海洋といった環境中から16S rDNA配列を収集し、各微生物叢に生息する微生物群についてのプローブ設計をすることで、微生物叢に生育する微生物群を高い精度で検出することができるマイクロアレイを作製する。

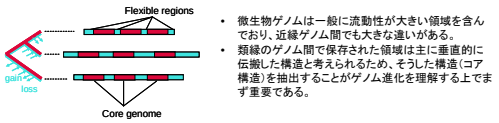
４．研究成果

１）配列データの整備

従来のMBGDから、利用者の持つゲノム配列をサーバ上に登録して、公表済みゲノムと組み合わせて、解析を行う機能、MyMBGDを作成した。また、近縁ゲノム比較向けの機能として、CGATインターフェイスをMBGDからも利用可能とした。挿入、欠失、逆位、重複などのゲノム構造変化の解析を主目的とした近縁ゲノム比較ツールCGATを開発、公開した。

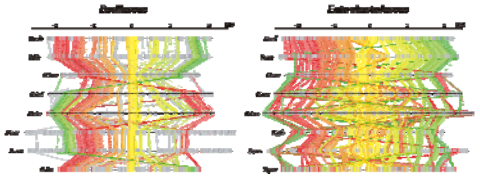
MBGDによるオーソログ分類結果を利用して、オーソログ間での遺伝子の並び順がよく保存された構造を抽出し、保存された順序で並べ替えた上でアライメントとして出力するプログラムCoreAlignerを開発した。この手法を用いてバチルス科と腸内細菌科のゲノムデータからコア構造を構築し、機能カテゴリ、必須遺伝子、GC含量、系統樹等の観点から、コア遺伝子の特徴付けを行った。その結果、抽出されたコア遺伝子群は必須遺伝子をはじめとする重要な遺伝子の多くを含んでおり、また非コア遺伝子と比べると垂直的に伝搬してきた割合が大きいことを示唆する結果を得た。

微生物ゲノムのコア構造解析



- 微生物ゲノムは一般に流動性が大きい領域を含んでおり、近縁ゲノム間でも大きな違いがある。
- 類縁のゲノム間で保存された領域は主に垂直的に伝搬した構造と考えられるため、そうした構造(コア構造)を抽出することがゲノム進化を理解する上でまず重要である。

コア遺伝子の染色体上の位置



MBGDのオーソロググループごとのアノテーションを充実させるため、グループに属する個々の遺伝子のアノテーションやクロスリファレンスに基づいてグループ全体のアノテーションを付与する手続きを実装した。

ホモロジー検索結果から、ベストヒットに基づく簡単なルールによって既存のオーソロググループに遺伝子を割り当てる手法を作成し、メタゲノムデータに対応したMyMBGD機能の一環として加えた。より高度な手法として、メタゲノムのデータをMBGDのオーソログクラスタリングアルゴリズムDomClustによる階層的クラスタリングの枠組みに取り込んで解析する手法についても実装を行った。

2）生物学データの整備

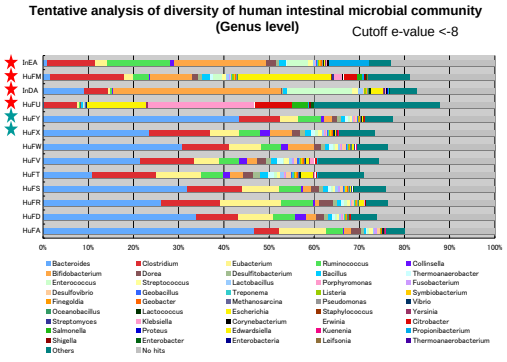
同属同種のゲノムが多数報告されている Streptococcus 属および Bacillus 属、さらには腸内細菌群に焦点を絞り、遺伝子の進化的分類を行った。また、ＭＢＧＤとは独立して、ゲノム配列よりパラログ遺伝子群を抽出しクラスター化するソフトウェア ParalogCluster を開発し公開している。本ソフトウェアでは全遺伝子の相同性検索の結果をもちいて、シングルリンケージおよびドメインを考慮したクラスタリングによるパラログ遺伝子群の抽出が可能となっている。さらに、自動的に抽出したクラスターを研究者が独自に編集することが可能な GUI も併せて開発した。

微生物ゲノムの SNPs 解析

国内で終了または解析中のゲノムプロジェクトで解読されたショットガンリードから、バクテリアの SNPs サイトを抽出することを試みた。東京大学の服部教授および海洋科学技術センターの高見博士の協力の元、約 50 種におよぶバクテリアの SNPs サイトを検出し、高見博士より提供頂いた、Bacillus halodurans, Oceanobacillus, iheyensi, eobacillus kaustophilus の３種における SNPs サイトの検出を試み、本方法により検出できた SNPs サイトと、Bacillus 属および Streptococcus 属における比較ゲノム解析から得られた遺伝子ごとの分子進化解析を組み合わせる事で、ゲノムレベルでの遺伝子進化を体系的に説明した。

ヒト腸内メタゲノム解析

環境中に存在する微生物を群集まるごとゲノム解析するメタゲノム解析が世界的に盛んに行われるようになり、我が国でも、2 家族、乳児、子供から大人に至る 13 人の腸内細菌叢のメタゲノム解析を行った。その結果、総塩基数 727Mb、予測した遺伝子合計数が 462、223 遺伝子と極めて大量の情報が得られ、得られた遺伝子を既存のデータベースに対して相同性検索を行い、腸内細菌叢の属分類解析を行った（下図）。



この結果、離乳前後で細菌叢が劇的に変化すること、大人の属分布は個人間で比較的安定していること、反対に乳児では個人間での分散が大きいこと、さらには乳児では大人と比

較して属の多様性が低い、傾向にあることがわかった。

また、新型シークエンサーを用いた 16S rRNA による群集構造解析に向けたソフトウェア開発ならびにデータベース整備を行った。ヒト常在細菌の中で、Gut（大腸）、Skin（皮膚）、Lung（肺）、Oral（口腔）、Vagina（膣）の 5 部位に生息する細菌の 16S rRNA 遺伝子配列、計 53,750 本を GenBank から抽出し、菌種組成や配列相同性により比較し、各フローラにおける細菌群集の特徴を抽出した。また、取得したヒト常在細菌の 16S rRNA 遺伝子配列情報をデータベース化するとともに、群集構造を分子系統的に容易に理解するための可視化ソフトウェア、ヒト常在細菌の配列&メタ情報を検索できる web アプリケーションも作成した。また、454 シークエンサーによる大規模群集構造解析を効率良く推進するためのユニバーサルプライマー設計手法を開発し、これまで公開されている 724 の細菌ゲノム配列から各分類群をターゲットとした 454 シークエンサーに特化したプライマーセットを設計した。

３）情報学的ツールの整備

近年、遺伝子発現データに基づく癌等の疾患の識別問題において、多くの教師付き機械学習法が応用されている。特に Support Vector Machine（SVM）は、最も有効な手法の一つとして主流になっている。しかしながら、他のカーネル識別法の応用に関する研究報告はほとんどなされていない。そこで、本研究では、カーネル識別法の一つであるカーネル部分空間法を癌組織の多クラス識別問題に適用し、複数のタイプの多クラス SVM と識別性能を比較した。７つの癌マイクロアレイデータセットを用いた比較実験の結果、カーネル部分空間法は高次元データに対して、多クラス SVM に匹敵する高い識別性能を示すことを明らかにした。

識別器の設計とならび遺伝子選択は、 マイクロアレイデータに基づく識別問題において重要な役割を担っており、データ解析における中心的なテーマになっている。本研究ではまず、Fisher の基準をはじめとするクラス分離度を測る複数の基準がカーネル化できることを示した。実験では、Fisher の基準がカーネル部分空間法におけるカーネルパラメータの選択において有効であることおよび、遺伝子群の選択基準として用いた場合、遺伝子ランキングや従来の考え方（識別は特徴空間において行うが、遺伝子群の選択はもとの空間において行う手法）と比べ、より高い識別性能をもたらす遺伝子群を同定できることを示した。これらの基準は識別のためのカーネル設計に応用できる他、マイクロアレイデータに代表されるベクトルデータのみならず、配列,グラフ,木のようにヘテロなデータ構造を有する多様な生物データを解析する上で有用であると考えられる。

これらの識別問題を微生物の株間の性質の違いに応用した。複数の株における黄色ブドウ球菌の性質（病状、部位、感染場所などのクラス）と関連性が深い遺伝子群を明らかにするために、クラス識別の一つである遺伝子選択を行った。まずは、黄色ブドウ球菌の MW2 株をプローブとしたアレイ CGH（Comparative Genomic Hybridization）解析を行うことで、各株における遺伝子の有無を判定した。様々な病状のうち、「膿痂疹」と「SSSS」に着目し、病状に違いが生じる原因遺伝子を特定するため、「膿痂疹」を起こす 46 株と「SSSS」を起こす 27 株に対して遺伝子選択を行なったところ、関連の深い遺伝子として 140 個と 122 個の遺伝子がそれぞれ抽出された。

４）新規微生物分野の情報収集のための基盤整備

微生物叢に生息する微生物の種を特定するためのマイクロアレイの作製を行った。約 4,000 種の基準株の 16S rDNA 配列における可変領域（約 500bp）の配列間のホモロジー検索を行い、それぞれの配列に特異的な配列を探索し、その中から種に特有な 23mer の配列をプローブ配列とし最終的に、ミスマッチを考慮した 2,718 本のプローブ配列の設計を行った。

微生物叢解析用マイクロアレイの問題点であるクロスハイブリを解消するには、ハイブリの際に増幅断片に対するフラグメンテーションをすることが効果的であった。これらの検討により、クロスハイブリが少なくより高い精度で微生物の種を同定するプロトコルを完成した。さらに、微生物叢解析用マイクロアレイを幾つかの海水土壌に対して適用した結果、全体的な発光のパターンは類似してるものの、各サンプルにのみ発光が見られるプローブも存在しており、微生物叢を識別することができるため、様々な環境に対して適用することができることを明らかにした。

＜国内外での成果の位置づけ＞

１）配列データの整備

MBGD データベースを基盤とした、微生物ゲノムを系統的に比較する環境の整備は着々と進んでおり、国内外でユニークな位置づけの微生物ゲノム解析システムとして確立しつつある。MBGD およびその元になるオーソログ分類手法 DomClust については、2009 年 7

月に行われたオーソログ解析に関連する主要な研究者が集まって開かれた会議（Quest for Orthologs）において、国内から唯一の参加者として発表した。一方、オーソログ解析を基にしたゲノムコア構造アライメント CoreAligner は、共通祖先から主として垂直的に伝搬し保存されたオーソログ遺伝子をリストする手段としてユニークであり、今後近縁微生物ゲノムの大規模比較を行う際に、有力なアプローチになるものと考えている。こちらも複数の国際会議で口頭発表として採択されるなど、一定の評価を得ている。

２）生物学データの整備

メタゲノム解析

ヒト腸内細菌叢はヒトと密接に関わりながら複雑な共同体を形成している。それらヒト腸内細菌叢に共通なゲノムの特徴を特定するために、乳児から大人まで様々な年令の 13 人の健常人を対象として大規模な比較メタゲノム解析を実施した。その結果、幼児における腸内細菌叢は大人とは大きく異なり単純で、かつ個人間における変異が極めて大きかった。反対に幼児を含む大人における腸内細菌叢はより複雑であるものの、年齢や性別に関わらず機能的に一様であった。さらに、大人では 237、乳児では 136 の特異的に強化された遺伝子ファミリーを特定することができた。それら遺伝子ファミリーは、腸の環境に順応するために、腸内細菌の種類によって利用される様々な戦略を示していると考えられる。また、特異的な conjugative トランスポゾンが、ヒト腸内において爆発的に増幅したことを発見した。このことは、ヒト腸内が細菌間における遺伝子水平伝播のホットスポットであることを示唆していることを論文発表した。

この論文発表と同時に国内の新聞各紙に記事が掲載された。さらに科学的な側面だけでなく、ネット上の多くのウェブサイトにて取り上げられたことから、一般に対する興味の啓発にも大きく貢献したと考えられる。また Science 誌の主編集者からも取材を受けた。さらに、日米欧を中心として、国際的なコンソーシアムが発足した。この中でも本研究は大きく取り上げられたことからも、本研究が先導的な研究として国際的に認識されていると考えられる。

３）情報学的ツールの整備

クラス識別やその一つの手法である遺伝子選択は、癌患者に対するアレイ解析において適用され、癌の診断や予後予測に使われている。このクラス識別を黄色ブドウ球菌の病状と遺伝子群の関連解析に用いる試みは、これまでになされていないため、新たな解析手法として考えられる。また、黄色ブドウ球菌が引き起こす様々な病状の原因遺伝子や発症メカニズムは、殆ど明らかにされていないため、クラス識別の結果は重要な情報になると考える。

４）新規微生物分野の情報収集のための基盤整備

2004年にVenterらがサルガッソー海における菌叢解析以来、土壌菌叢、海底における鯨骨周辺の菌叢、マウスの腸内菌叢といった様々な環境化での菌叢解析が行われている。近年、国内外では菌叢を調べる研究が進められているため、膨大な量の16S rDNAの配列が決定されている。難培養性微生物を含む混合菌叢の菌学的性状を比較的簡便に解析できる独自のオリゴチップの作製を行うことができれば、様々な環境における菌叢を正確に把握することができると考えられる。

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 12 件）

１）論文

1. Hideki Hirakawa, Hidenori Akita, Tamaki Fujiwara, Motoyuki Sugai and Satoru Kuhara, Structural insight into the binding mode between the targeting domain of ALE-1 (92AA) and pentaglycine of peptidoglycan. Protein Engineering Design and Selection 22, :385-391, (2009)
2. Uchiyama, I., Higuchi, T., Kawai, M.: MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity, Nucleic Acids Res. in press
3. Uchiyama, I.: Multiple genome alignment for identifying the core structure among moderately related microbial genomes, BMC Genomics, 9, 515 (2008)
4. Izutsu, K. et al., Comparative genomic analysis using microarray demonstrates a strong correlation between the presence of the 80-kilobase pathogenicity island and pathogenicity in Kanagawa phenomenon-positive *Vibrio parahaemolyticus* strains., *Infect. Immun.*, 76, 1016-1023, (2008)
5. Ogura, Y. et al., Extensive genomic diversity and selective conservation of virulence-determinants in enterohaemorrhagic *Escherichia coli* strains of O157 and non-O157 serotypes.,

Genome Biol., 8, R138 (2007).
6. Uchiyama, I.: MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups, Nucleic Acids Res. 35, D343-D346 (2007)
7. Uchiyama, I., Higuchi, T., Kobayashi, I.: CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes, BMC Bioinformatics, 7, 472 (2006)
8. Uchiyama, I., Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. Nucleic Acids Res.. 34、 647-658 (2006).
9. Niijima, S., Kuhara, S. Gene subset selection in kernel-induced feature space. Pattern Recognition Letters.27,1884-1892 (2006)
10. Niijima, S., Kuhara, S. Recursive gene selection based on maximum margin criterion: a comparison with SVM-RFE. BMC Bioinformatics 7:543 (2006)
11. Niijima, S., and Kuhara, S., Multiclass molecular cancer classification by Kernel subspace methods with effective Kernel parameter selection. J.B.C.B.,3(5)、1071-1088 (2005)
12. Kurokawa, K. et al., Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes., *DNA Res.*, 14, 169-181 (2007).

〔学会発表〕(計 57 件)

1. Naoya Kawamura, Hideki Hirakawa, Masato Oono, Kaori Ootani, Touru Shimizu, Satoru Kuhara (2009) The gene expression networks of Clostridium perfringens strain 13 from gene expression profiles, 第 32 回日本分子生物学会年会
2. Ikuo Uchiyama, MBGD and RECOG: integrated platform for comparative genomics based on large-scale ortholog grouping. “Quest for Orthologs”Wellcome Trust Conference Centre, July 3-5, 2009, Cambridge.

〔図書〕(計 0 件)

〔産業財産権〕
○出願状況 (計 0 件)

○取得状況 (計 0 件)

〔その他〕
ホームページ等

データベース/ソフトウェア

1. MBGD (Micribial Genome Database for Comparative Analysis):微生物比較ゲノムデータベース。 http://mbgd.genome.ad.jp/
   CGAT (Comparative Genome Analysis Tool):近縁ゲノム比較解析ツール。 http://mbgd.genome.ad.jp/CGAT/
   CoreAligner:類縁ゲノム間での遺伝子の並び順の保存性に基づいてゲノムのコア構造を構築するプログラム。 http://mbgd.genome.ad.jp/CoreAligner
2. Ortholog Group Management System: オルソログドメイン解析システム。 http://jamboree.grt.kyushu-u.ac.jp:9001/

６．研究組織
(1)研究代表者
久原 哲 (KUHARA SATORU)
九州大学・大学院農学研究院・教授
研究者番号：00153320

(2)研究分担者
なし

(3)連携研究者
黒川 顕 (KUROKAWA KEN)
東京工業大学・大学院生命理工学研究科・教授
研究者番号：20343246

内山 郁夫 (UCHIYAMA IKUO)
大学共同利用機関法人自然科学研究機構・助教
研究者番号：20343246

平川 英樹 (HIRAKAWA HIDEKI)
かずさ DNA 研究所・植物ゲノム研究部・研究員
研究者番号：80372746