

機関番号：63904

研究種目：基盤研究(B)

研究期間：2008 ～ 2010

課題番号：20310125

研究課題名(和文) 微生物ゲノム比較基盤の確立に向けたゲノムコア構造解析法の開発

研究課題名(英文) Development of a method for core genome identification as a basis for microbial comparative genomics

研究代表者

内山 郁夫 (UCHIYAMA IKUO)

基礎生物学研究所・ゲノム情報研究室・助教

研究者番号：90243089

研究成果の概要(和文)：

水平移動を含む複雑な進化過程を持つ微生物ゲノムの比較基盤として、類縁ゲノム間で主として水平的に伝搬されたと考えられる「ゲノムコア構造」を抽出することを目的として、遺伝子の並び順が保存された領域を抽出するコアゲノムアライメント法を開発した。この手法をいくつかのゲノムデータに適用してその有効性を検証するとともに、公開ゲノム配列に適用した解析結果をデータベースとして蓄積する体制を作った。また、本手法のアライメント結果の可視化を含めた一連のプロトコルを実装した解析ツールを作成し、いくつかの近縁微生物ゲノムの比較解析に適用して、その有用性を検証した。

研究成果の概要(英文)：

As a fundamental basis for comparative analysis of microbial genomes that have undergone complex evolutionary processes including horizontal gene transfers, we have investigated a method to identify a “genomic core structure” among related genomes that is mainly inherited through vertical transfers. Our method is based on extracting the regions that have conserved gene order among the related genomes. We applied our method to various genomic data to test its validity and established a system to accumulate the resulting core structures of the published data into a database. We also developed a tool that implements the protocol for core genome analysis including visualization of core structures, and applied it to actual comparative studies of closely related microbial genomes.

交付決定額

(金額単位：円)

	直接経費	間接経費	合計
2008年度	5,100,000	1,530,000	6,630,000
2009年度	4,600,000	1,380,000	5,980,000
2010年度	4,900,000	1,470,000	6,370,000
年度			
年度			
総計	14,600,000	4,380,000	18,980,000

研究分野：微生物ゲノム学

科研費の分科・細目：ゲノム科学、ゲノム情報科学

キーワード：ゲノム、進化、微生物

## 1. 研究開始当初の背景

ゲノム情報の爆発的な蓄積に伴い、それらのデータを整理し、体系化していくことが重要な課題となっている。特に微生物のゲノムは、研究開発当初ですでに数百種類のゲノムが蓄積しており、さらにメタゲノム解析技術の進展により、これらのゲノム情報を活用して、自然界における微生物多様性の実態の解明が進むことが期待されていた。しかし、微生物ゲノムの進化過程には、通常の垂直的な遺伝プロセスに加えて、水平的な遺伝子の移動が多数生じているため、そのような複雑な進化過程を持つゲノムをどのように体系づけて整理するかという問題があった。

## 2. 研究の目的

本研究では、ゲノムを体系づける方策として、ゲノムの中で主として垂直的に伝搬する部分を抽出して、これを、各系統群を特徴付ける「コアゲノム構造」として集積していくアプローチを考える。コアゲノムを抽出する方法としては、従来は「すべてのゲノムで保存されている遺伝子（ユニバーサル遺伝子）」を考えるのが通例であったが、「垂直的に伝播する」という条件を考える場合は、そのような保存性の厳しい条件を考えるのは適当ではない。そこで本研究では、類縁ゲノム間で遺伝子の並び順が保存された構造をとることによって、保存性の条件を緩めつつ、主に垂直的に伝搬された構造をよりの確に抽出する方法を開発することを目指す。このため、以下のような観点からの研究を行う。(1) 遺伝子の並び順を解析して、このようなコア構造を効率よく抽出する手法を開発する。(2) その手法の有効性を明らかにするため、様々な類縁グループを対象として適用し、抽出されたコア構造の特徴を解析する。(3) 本手法を利用可能なデータにシステムティックに適用して、コア構造を抽出した結果をデータベース化する方策について検討する。(4) この手法を実装して、データベースと連携して効果的な解析を行えるシステムの構築を行う。(5) これを具的的な近縁ゲノム比較解析に適用し、比較ゲノム解析のプロトコルを確立する。

## 3. 研究の方法

### (1) コア構造抽出手法の開発

近縁ゲノム間で遺伝子の並び順が保存された領域を抽出する手法を開発する。その際、これまでに作成したゲノム間のオーソログ解析手法(DomClust)を活かして、あらかじめ作成されたオーソログ対応データを入力とし、遺伝子の並び順が最も保持されるようなオーソロググループの並び順をとるというアプローチをとる。この手法を実際のデータ

に応用して、その効果を検証する。その際、特に従来「コアゲノム」の定義としてよく用いられている、遺伝子の（あるなしの）保存性のみに基づく「保存遺伝子」、あるいはすべてのゲノムで保存されている「ユニバーサル遺伝子」との違いに着目する。

### (2) コア構造の特徴づけ

遺伝子の並び順が保存された構造として抽出した「コア構造」の特徴付けを行う。特に「コア構造」は主に垂直伝搬された構造であるという観点に着目し、系統樹や GC 含量などの水平移動を検出する手法を適用して、コア遺伝子が主として垂直的に伝搬されているということを検証する。

### (3) コアゲノムデータベースの構築

本手法の結果を活用するため、微生物比較ゲノムデータベース MBGD に登録された科レベルの系統群の中で、一定の多様性を有するものを抜き出し、これに本手法を適用して系統群固有のコア構造を抽出してデータベース化する手続きを作成する。

### (4) コア構造解析を促進するツールの開発

本手法を、様々な比較ゲノム解析に適用するため、研究代表者らによって開発した比較ゲノムワークベンチ RECOG にコア構造解析機能を組み込んで、一連の解析を行って、その結果を詳細に検討するための体制を整える。

### (5) 近縁ゲノム比較解析への応用

本手法を具体的なゲノムプロジェクトに適用して、比較ゲノム解析プロトコルの確立を図る。具体的には、外部の共同研究者とともに以下の解析を行った。① 黄色ブドウ球菌を含む *Staphylococcus* 属のゲノム進化の解明を目指した、*Staphylococcus* 属の類縁種である *Macroccoccus caseolyticus* と他の *Stapylococcus* 属ゲノムとの比較解析、②ヘリコバクターピロリ菌の東アジア株の特徴付けを目指したピロリ菌ゲノム比較解析。

## 4. 研究成果

### (1) コア構造抽出手法の開発

CoreAligner プログラムを開発した。このプログラムは、ゲノム間の網羅的なオーソログ関係を入力として、オーソロググループを頂点、近傍関係を辺とする近傍グラフを作成し、各ゲノムで遺伝子の並び順が最もよく保たれるようなオーソロググループの配置をとることにより、ゲノム間で保存された構造の抽出を行う (図1)。

方法は以下のステップにより行う。①各オ

一ソログ遺伝子対のうち、ゲノム上の近傍関

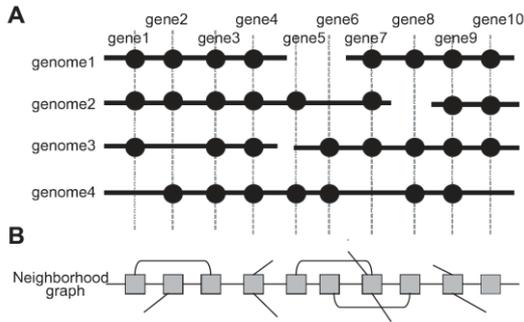


図1 コア構造アライメント

係が保存されている対（デフォルトでは半数以上のゲノムで保存されている対）を抽出し、近傍グラフを作成する。②ゲノム上の向きに基づいてグラフの辺の向き付けを行う。③グラフの辺を頂点と見なすことによって、3遺伝子の並びを単位とするグラフに変換する。④グラフからループを除去する。⑤動的計画法により最良のパスを選択する。⑥元のグラフに変換して遺伝子の並びを確定する。

この手法を実際のデータに適用して評価した。入力の一ソログデータは、これまでに運用している微生物比較ゲノム解析データベース MBGD のデータを用いた。バチルス科と腸内細菌科のそれぞれ8種（近縁種をまとめるとそれぞれ6生物種グループ）を含むデータセットに適用したところ、それぞれおよそ1400、2100個の遺伝子を含むコア構造が抽出された。このコア遺伝子セットの中に、枯草菌と大腸菌で同定されている必須遺伝子がどの程度含まれているかを見たところ、いずれも90%以上の必須遺伝子を含んでいた。

次に、パラメータを変更したときの影響を、コア遺伝子数の変化を用いて調べた。保存性のパラメータを変更した場合は、最大で2倍程度と大きく変化したが、コア遺伝子に含まれる必須遺伝子の数で評価したところ、保存性50%でとったときが最大になる傾向を示したのでこれをデフォルト値として採用した。一方、「近傍」の条件を指定するパラメータを変更した場合は、コア遺伝子数に大きな変化は見られなかった。

また、対象とするゲノム数を変化させたときのコア遺伝子数の変化を調べるため、ゲノムのサブセットをとって、対象ゲノム数を増

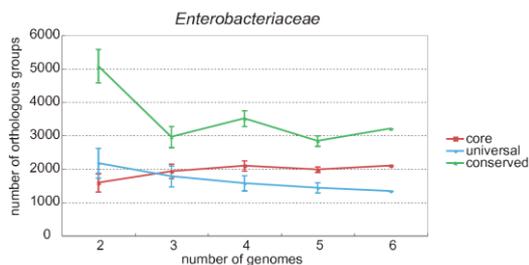


図2 ゲノム数によるコア遺伝子数の変化

やしながらコア構造抽出を繰り返すテストを行ったところ、対象が3-4ゲノム程度からコア遺伝子数はほぼ安定な値をとることが分かった。これに対して、「全部の生物種に含まれる」というユニバーサル遺伝子の数は、生物種を増やすにつれて単調に減少するため、安定な値とはならなかった。特に、腸内細菌科には共生細菌である *Sodalis glossinidius* が含まれていることもあって、我々の手法で定義されたコア遺伝子数は2000前後で安定であったのに対して、ユニバーサル遺伝子数は1300程度と大きく減少しただけでなく、サブセットの取り方による変動も大きかった（図2）。こうした安定性は、我々の方法によるコアゲノムの定義法が、ユニバーサル遺伝子をとるのと比べて、系統群の特徴付けを行う上で有利な特徴のひとつとなっている。

### (2) コア構造の特徴づけ

バチルス科の各遺伝子のGC含量の分布を調べた結果を図3に示す。ここでは、遺伝子の並び順の保存性に基づく「コア遺伝子」に加えて、あるなしで50%以上保存された「保存遺伝子」、もしくは100%保存された「ユニバーサル遺伝子」という条件を組み合わせ、遺伝子を6種類のクラスに分類してそれぞれの分布を示している。生物種によって程度に違いはあるものの、一般に「コア遺伝子」は「非コア遺伝子」と比べてGC含量の分散が小さいという結果が得られた。

また、系統樹については、コア遺伝子の連結配列に基づく系統解析によってリファレンスとなる種系統樹を作成し、それとトポロジーが一致しない系統樹を水平移動の可能性のある系統樹と評価した。「コア遺伝子」と「非コア遺伝子」それぞれで系統樹が一致しない遺伝子の割合を調べたところ、非コア遺伝子では20%程度一致しないケースが存在していたのに対して、コア遺伝子では5%前後

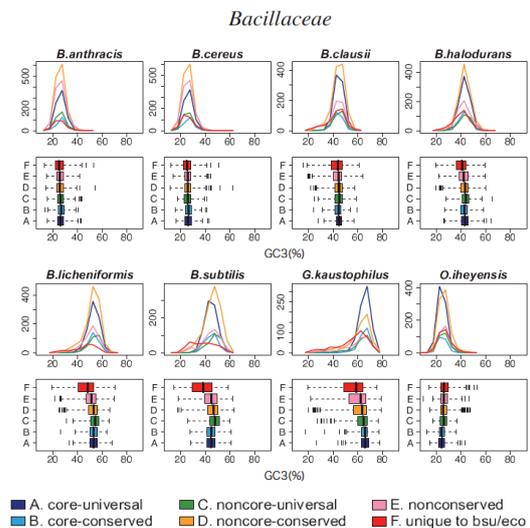


図3 クラスごとのGC含量の分布

であった。この傾向は遺伝子の保存性とは関係なく成立していた。

### (3) コアゲノムデータベースの構築

微生物ゲノムデータベース MBGD から、「科」レベルの系統群で6種以上のゲノムが存在する系統群（最新版では44科）に、本手法を機械的に適用してコア遺伝子を抽出した。ただし、これらの中にはマイコプラズマ科やサーモトガ科など、並び順の保存性が低くてほとんどコア構造が抽出できないものもあったので、並び順の保存性に基づくコア遺伝子のほかに、遺伝子の保存性のみに基づくコア遺伝子も加えた。それらのコア遺伝子のコンセンサスの並び順やゲノム上の配置に加え、各コア遺伝子のマルチプルアライメントの情報なども計算してデータベースに格納した。これにより、指定した遺伝子がコア構造に含まれるかどうかを検索したり、その周辺のコア構造の様子を調べたりできるようにした。

### (4) コア構造解析を行うシステムの開発

比較ゲノム解析ワークベンチ RECOG にコア構造解析機能を組み込んで、指定した系統群のゲノムについてオーソログ解析を行い、それに基づいてコア構造構築を行うまでの一連の作業をこのシステム上で行えるようにした。また、コア構造アライメント結果の可視化や、ゲノムマップ上の表示を行って詳細な解析を行えるようにした（図4）。

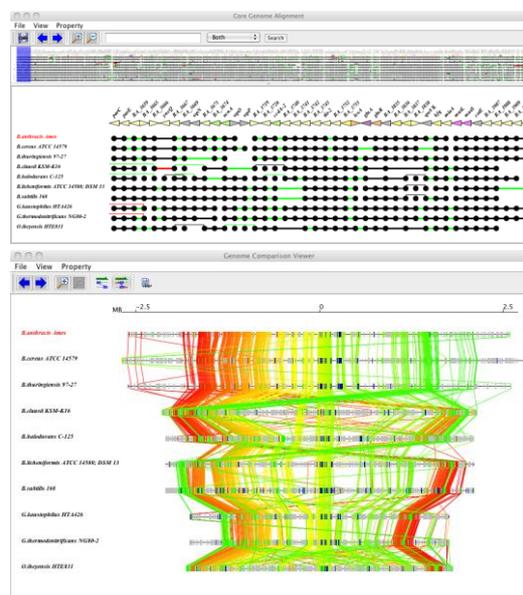


図4 コア構造の可視化

### (5) 近縁ゲノム比較解析への応用

本手法を①*Macrococcus caseolyticus* のゲノム解析、および、②*Helicobacter pylori* 東

アジア株のゲノム解析へ適用した。①ではスタフィロコッカス科およびバチルス科を交えた比較を行い、②ではピロリ菌の東アジア株とヨーロッパ、アフリカ株との比較を行った。いずれの研究においても、コア遺伝子を抽出した後、コア遺伝子の連結配列を用いて精密な種系統樹を作成し、さらに抽出されたコア遺伝子をゲノム上に配置して、コア構造に対する挿入、欠失、逆位などのゲノム変化のパターンを解析する、という手順で進めた。その結果、①ではマクロコッカスはスタフィロコッカス属の共通祖先の上流から生じ、かつスタフィロコッカス固有遺伝子を含む領域の多くを欠いたコンパクトなゲノムを持っていることから、スタフィロコッカスのゲノム進化についての推測を行うことが可能になった。また、②では東アジア株固有に生じた逆位のパターンの詳細な解析からゲノム進化の新規メカニズムの発見などにつながった。これらの解析を通じて、本研究におけるコアゲノム解析のアプローチは、複雑なゲノム進化における本質的な部分を抽出することによって、その解釈を促進するという効果を持ち、近縁ゲノム比較解析のプロトコルの一つとして有用であることを示すことができた。

### 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5件）

1. Kawai, M., Furuta, Y., Yahara, K., Tsuru, T., Oshima, K., Handa, N., Takahashi, N., Yoshida, M., Azuma, T., Hattori, M., Uchiyama, I., Kobayashi, I., BMC Microbiology, 11, 104 (2011) 査読有り

2. Furuta, Y., Kawai, M., Yahara, K., Takahashi, N., Handa, N., Tsuru, T., Oshima, K., Yoshida, M., Azuma, T., Hattori, M., Uchiyama, I., Kobayashi, I., Birth and death of genes linked to chromosomal inversion, Proc. Natl. Acad. Sci. USA, 108, 1501-1506 (2011) 査読有り

3. Uchiyama, I., Higuchi, T., Kawai, M., MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity, Nucleic Acids, Res., 38, D361-D365 (2010). 査読有り

4. Baba, T., Kuwahara-Arai, K., Uchiyama, I., Takeuchi, F., Ito, T., Hiramatsu, K., Complete Genome Sequence of *Macrococcus caseolyticus* Strain JSCS5402, Reflecting the Ancestral

Genome of the Human-Pathogenic Staphylococci, J. Bacteriol., 191, 1180-1190, (2009). 査読有り

5. Uchiyama, I., Multiple genome alignment for identifying the core structure among moderately related microbial genomes, BMC Genomics, 9, 515 (2008) 査読有り

〔学会発表〕 (計 10 件)

1. Uchiyama, I., RECOG: a framework for systematic approaches to microbial genome comparison, International Symposium on Omics Approach to Microbes, 2011 年 2 月 17-18 日, NEERI, Nagpur, India

2. Uchiyama, I., Systematic approach for large-scale microbial comparative genomics, The 4<sup>th</sup> Global COE International Symposium 2009 joint with the 19<sup>th</sup> Hot Spring Harbor Symposium “Molecular Evolution and Bioinformatics”, 2009 年 11 月 1-2 日, 九州大学, 福岡

3. 内山郁夫, 大規模比較ゲノム解析の推進に向けて: 比較ゲノムワークベンチ RECOG の開発と、ピロリ菌東アジア株ゲノム解析への応用, 第 11 回日本進化学会大会, 2009 年 9 月 2-4 日, 北海道大学, 札幌

4. Uchiyama, I., CoreAligner: Multiple genome alignment for identifying the genomic core among moderately related microbial genomes, 19th International Conference on Genome Informatics, 2008 年 12 月 1 日, Marriot Surfers Paradise, Gold Coast, Australia

5. 内山郁夫, ゲノム進化過程の解明に向けた近縁細菌ゲノムのインフォマティクス, 第10回日本進化学会大会, 2008年8月24日, 東京大学駒場キャンパス

〔図書〕 (計 2 件)

1. 内山郁夫, 比較ゲノムにおけるバイオインフォマティクス基盤, 服部正平監修「メタゲノム解析技術の最前線」, 54-65, CMC 出版, 2011

2. Uchiyama, I., Functional inference in microbial genomics based on large-scale comparative analysis, in “Omics approaches to protein function prediction” Daisuke Kihara ed.), Springer, 2011.

〔その他〕

ホームページ等

微生物比較ゲノムデータベース MBGD

<http://mbgd.genome.ad.jp>

CoreAligner プログラム

<http://mbgd.genome.ad.jp/CoreAligner>

6. 研究組織

(1) 研究代表者

内山 郁夫 (UCHIYAMA IKUO)

基礎生物学研究所・ゲノム情報研究室・助教

研究者番号 : 90243089