計画研究：2005 ～ 2009 年度

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

●久原 哲[1)]　◇内山 郁夫[2)]　◇黒川 顕[3)]　◇平川 英樹[4)]
1) 九州大学大学院農学研究院　2) 自然科学研究機構基礎生物学研究所　3）東京工業大学生命情報専攻　4）かずさ DNA 研究所

＜研究の目的と進め方＞

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。

＜研究開始時の研究計画＞

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。

MBGD のオーソログ解析機能を、メタゲノム解析に適用する方策について検討する。

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。今後はさらに解析の対象を拡げ、他の科、属に対しても予測確度を推定する。また、相違度データベースからの情報を取り入れることで、より詳細な分類推定を可能にする。

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。土壌から菌を分離し、その 16S rDNA 遺伝子を PCR で増幅し、決定された配列と文献データを加えた 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。プローブ配列の設計は、既知の 16S rDNA 配列を元にして、クロスハイブリが出来る限り少なくなるようにする。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。また、プローブ配列に 16S rDNA における他の可変領域、もしくは、DNA ポリメラーゼなどのハウスキーピング遺伝子を用いることを検討する。

＜研究期間の成果＞

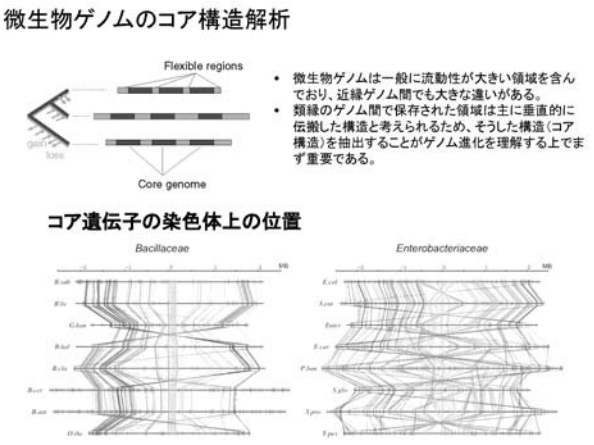
1）配列データの整備

従来の MBGD では公開されたゲノム配列のみが登録されているため、新規ゲノム配列の解析には直接つかえないという問題があった。そこで､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、次項で述べる 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｡CGAT は、アライメントの計算やデータの管理を行うサーバと、ドットプロットとアライメント表示とを組み合わせたビューアとからなっており、アライメントに種々の特徴セグメントの位置を重ねて表示することができる。特にゲノム上の繰り返し構造とゲノム多型部位とを重ねることにより、それらの関連を詳細に調べるのに適している。

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。



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。また、GOLD のデータを利用して、表現型に基づいて生物種セットを選択して解析に用いることができるようにした。さらに、菌類や原生生物などの真核微生物のデータの拡充を行うとともに、ヒトをはじめとするいくつかの高等生物も参照配列として加えた。

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。

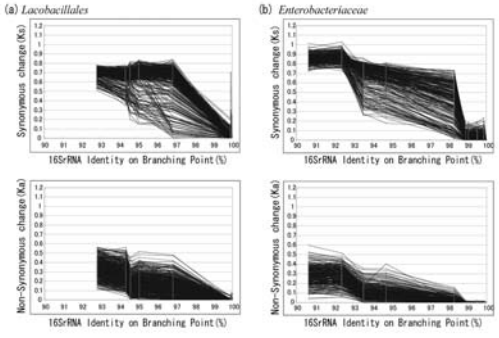
また、すべての遺伝子の進化的履歴を明らかにするにあたり、種の代表的な進化速度の指標となり得る必須遺伝子を、文献により報告されている各種において抽出し、それら必須遺伝子のオルソログ判定をおこなった。同時に、ＭＢＧＤとは独立して、ゲノム配列よりパラログ遺伝子群を抽出しクラスター化するソフトウェア ParalogCluster を開発し公開している。本ソフトウェアでは全遺伝子の相同性検索の結果をもちいて、シングルリンケージおよびドメインを考慮したクラスタリングによるパラログ遺伝子群の抽出が可能となっている。さらに、自動的に抽出したクラスターを研究者が独自に編集することが可能な GUI も併せて開発した。

同属同種のゲノムが多数報告されており、各種間が進化的に近縁である Enterobacteriaceae および Lactobacillales に焦点を絞り、全遺伝子の進化的分類を行っている。すべての遺伝子の進化的履歴を明らかにするにあたり、種の代表的な進化速度の指標となり得る参照遺伝子として、種分化解析で頻繁に用いられている16S rRNA 遺伝子をもちいた。参照遺伝子である 16S rRNA 遺伝子と、他の遺伝子との進化速度を比較し、種分化と遺伝子進化との関係を明らかにしようとしている。具体的には、参照遺伝子により得られた系統樹上での各分岐点でのオーソログ遺伝子を内山の MBGD にて抽出し、グループの祖先分岐点からすべての葉に至るまでの経路における遺伝子の進化履歴を明らかにした。下図は、横軸に各分岐点における 16S rRNA の相同性、縦軸にオーソログ遺伝子間での同義置換数（Ks）、非同義置換数（Ka）をとり、オーソログ遺伝子の進化速度が、進化の過程でどのように変化してきたのかを表したものである。この図から、Enterobacteriaceae および Lactobacillales のどちらにおいても、16S rRNA 遺伝子の相同性が増加すれば、置換数が減少するという比例関係が得られた。Konstantinidis et al. (2005) により、16S rRNA とその他の遺伝子の進化速度は、比較対象間での 16S rRNA の相同性が 60% 以上である時、良く比例することが知られている。しかし、本研究の結果からは、これらのグループにおいて過去に少なくとも２回（16S rRNA の相同性が 98%、94%）、進化速度が急激に変化する期間が存在したことを示唆している。この２回はそれぞれ１億年前、４～５億年前に相当し、哺乳類の誕生および陸上への生命の進出時期と重なることは非常に興味深く、今後は本解析を、ゲノムが明らかとなっているすべてのバクテリアで行うことで、ゲノム全遺伝子の進化的履歴の同定が可能となった。

微生物ゲノムのSNPs解析

1995年に始まるバクテリアのゲノム解析は、基本的にランダムショットガン法によりおこなわれてきた。本方法では冗長度が約８になる程度まで、ゲノム断片をシークエンスしアッセンブルすることでゲノム配列を決定する。ゲノム配列を確定する際、シークエンスクォリティと冗長度により、最終的には４塩基のいずれかに決定することになる（不確定の場合は、再シークエンスにより決定する）。サンプルとするゲノムは、バクテリア集団より採取するので、これらゲノム断片には個体特異的な点突然変異（SNPs）が含まれる。したがって、ショットガンリードのアッセンブル結果を注意深く観察することで、いわゆるhigh quality discrepancyと呼ばれるSNPsサイトを検出することが可能となる。そこで我々は、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。具体的には、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サイトを検出する予定にしている。はじめに、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

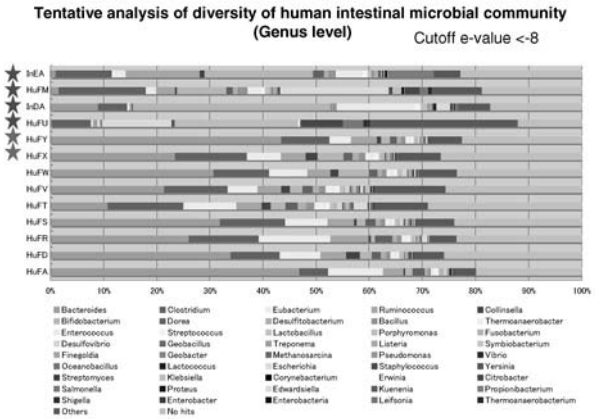
Lactobacillales、Enterobacteriaceae における オーソログ遺伝子の進化履歴



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ってきた。我が国では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が進行中である。本研究では、ショットガンリード合計数が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、現在各種バイオインフォマティクス手法を駆使して解析を実行中である。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実施した（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、つまりドミナントが存在している傾向にあることがわかった。



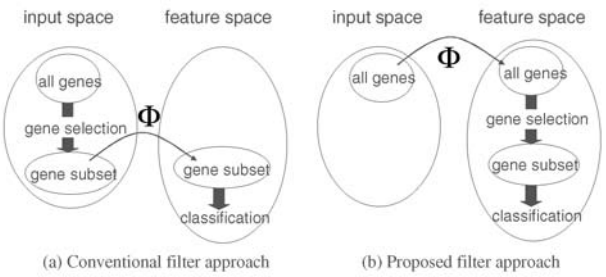
また、新型シークエンサーを用いた16S rRNAによる群集構造解析に向けたソフトウェア開発ならびにデータベース整備を行った。ヒト常在細菌の中で、Gut（大腸）、Skin（皮膚）、Lung（肺）、Oral（口腔）、Vagina（膣）の5部位に生息する細菌の16S rRNA遺伝子配列、計53,750本をGenBankから抽出し、菌種組成や配列相同性により比較し、各フローラにおける細菌群集の特徴を抽出した。ヒト常在細菌のうち未同定の新規の細菌と思われる細菌種については、16S rRNA遺伝子配列の相同性により他の環境中に生息する細菌と比較し、新規の細菌の由来と環境中における分布についても明らかにした。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。カーネル部分空間法は高い識別性能を有するだけでなく、マイクロアレイデータに特徴的な高次元でかつサンプル数が少ない場合において計算が効率的であり、また多クラスの取り扱いが容易であるという利点を有する。

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。遺伝子選択によって、まず識別能力の高い遺伝子を同定し、それらを識別に用いることにより、識別精度を向上させることができ、また計算コストを削減できる。さらに、同定した少数の遺伝子を標的にすることにより、効率的な実験設計が実現できるだけでなく、そのようにして同定した遺伝子は、研究対象とする生物現象のメカニズムを解明する上で、重要な知見をもたらすものと考えられる。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。これらの基準はクラス分離度を特徴空間において測ることができるため、カーネルパラメータの選択基準や遺伝子群の選択基準として応用することができる。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。

数多く提案されている遺伝子選択の手法の中でも、SVMに基づく多変量のアプローチは極めて有効であることが報告されている。しかし、SVMの判別関数の構成原理から考えると、ノイズやはずれ値に敏感であり、特にノイズが多くサンプル数が少ないマイクロアレイデータにおいては必ずしも有効であるとは考えられない。ソフトマージンパラメータを導入し適切に設定することでノイズの影響を軽減し、さらにオーバーフィッティングを回避できるが、その性能評価は限定的なものにとどまっている。本研究では、原理的にノイズやはずれ値に影響を受けにくい判別関数に基づく多変量の遺伝子選択の手法としてMMC-RFE (Maximum Margin Criterion based Recursive Feature Elimination) を新規に提案し、SVM-RFEとの性能比較を行った。9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、2クラスのデータセットについては、MMC-RFEはハードマージン SVM-RFEとパラメータを適切に設定したソフトマージンSVM-RFEとの中間的な性能を示す傾向にあり、多クラスのデータセットについてはSVM-RFEよりも有効であった。提案したMMC-RFEに基づく遺伝子選択のアルゴリズムは計算が安定で効率的であり、またパラメータの設定を必要としない。さらにSVM-RFEと異なり、多クラスへの拡張が容易であるという利点を有する。したがって、マイクロアレイにとどまらず、ハイスループット技術によって生成される膨大かつノイズを多く含むデータからバイオマーカーを探索するための一般的な手法として実用性が期待される。



gene selection をカーネル識別法と併用する場合，あらかじめ input space において遺伝子選択を行ってから，feature space において識別を行うというやり方(a)が考えられるが，提案手法(b)では gene selection も feature space において行う．

カーネル部分空間法を遺伝子発現データのクラス識別問題に適用し、新規の識別器、及び遺伝子選択のシステムを開発した。識別器、遺伝子選択は、マイクロアレイ等の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。遺伝子選択によって、まず識別能力の高い遺伝子を同定し、それらを識別に用いることにより、識別精度を向上させることができ、また計算コストを削減できる。さらに、同定した少数の遺伝子を標的にすることにより、生物現象のメカニズム解明の足がかりとなる。成果としては、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ことを明らかにした。

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マイクロアレイを作製し、様々な病状から得られた 176 の株についてハイブリ実験した後、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」に着目した場合、水疱形成に関与する原因遺伝子は共に表皮剥脱毒素（ET）と考えられるが、前者は軽症であるのに対し、後者は重篤となる。このように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。各病状について、クラス分類を行った結果、各株の性質を特徴付ける遺伝子群を推定することができた。また、マルチクラス識別も試みたところ、2 クラス予測の精度の方が高かった。

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た．DNA チップの検出感度を調べるために、1bp および 2bp のミスマッチを含めた 154 本のテスト用プローブ配列を設計し、テストチップを作製した．テストチップに対するハイブリダイゼーションの結果、1bp のミスマッチまでハイブリダイズすることが分かった．あわせて、ハイブリする際の蛍光試薬やバッファーの検討をしたところ、検出感度を上げることができた．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。

作製した微生物叢解析用マイクロアレイに対してハイブリ実験を行ったところ、クロスハイブリが見られた。その原因としては、プローブ設計の際に、ミスマッチが正確に見積もられていないこと、プローブ配列の二次構造が考慮されていないことが考えられた。また、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ために、ハイブリダイゼーションの温度を上げ、特異性を高める実験条件の検討を行なった結果、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ことができるようになった。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いたが、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、微生物叢の違いを反映しているものと考えられた。作製したマイクロアレイは、クロスハイブリが生じるプローブが存在しているが、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膨大な量の DNA 断片を読み取ることで菌叢解析を行った。それ以来、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。本研究では、16S rDNA 配列を用いたオリゴチップを作製し、土壌菌叢を調べることで土壌の性質を比較するということを目的とした。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。

＜達成できなかったこと、予想外の困難、その理由＞

1）配列データの整備

当初は MBGD データベースに基づいて、もう少し具体的なゲノム解析を多方面に展開したいと考えていたが、思ったほどできていない。具体的な解析を行うには結局個々に検討すべきことがいろいろ出てくるため、ある程度のマンパワーがないと多方面に展開することは難しい。

2）生物学データの整備

メタゲノム解析では膨大なデータを扱うため、計算機資源の確保が重要な課題である．また、解析が進むにつれ過去の計算結果も併せて保存する必要があるため、ストレージ領域も予想外に大量に確保する必要があった.

GenBank から 16S rRNA 遺伝子データを抽出しデータベース化を行ったが、新規の大規模 16S rRNA シークエンスが猛烈な勢いで登録されており、既存の計算機設備を用いた配列アラインメントおよびクラスタリングなどの計算が困難であったため、GenBank 上のすべての 16S rRNA 遺伝子データを対象とする事ができなかった。しかしながら、メタ情報を伴った配列情報のみを慎重に抽出した上でデータベース化し、大規模シークエンスに関しては既データベースにマッピングすることで、すべてのデータを対象とした解析を実施することが可能となった。

3）情報学的ツールの整備

黄色ブドウ球菌の CGH 解析における遺伝子選択については、「膿痂疹」と「SSSS」に着目した場合、それらに関係すると思われる遺伝子群を推定することができた。現在、その結果を元にして、それぞれの病状に関与する可能性がある遺伝子を実験により確認している。このように、遺伝子選択については順調に研究を進めることができた。一方、それぞれの病状に特徴的な遺伝子群を用いてクラス識別を行ったところ、その精度はあまり高くなかった。その理由としては、CGH 解析に用いるアレイ実験の数が少ないことと病状の分類が難しく、その精度が十分ではないことが考えられた。

4）新規微生物分野の情報収集のための基盤整備

アレイ実験については、ハイブリの条件を検討したところ、特異性を高めることができた。このことから、より高い精度で微生物叢における微生物の種を特定できることができるようになった。しかし、依然として幾つかのスポットにおいてクロスハイブリが見られた。微生物叢を対象としたマイクロアレイの設計では、対象となる微生物の配列が未知であるため、クロスハイブリが生じることは避けられない。

＜今後の課題、展望＞

1）配列データの整備

これまでの研究により、オーソログ解析に基づいて微生物ゲノムの大規模な比較解析を行うための基盤はかなり整備されたと考えている。ただし、近縁ゲノム比較のコア構造解析については、この手法を主要な系統グループのデータに網羅的に適用して、データベースとして整備するという課題が残っており、今後さらに検討を進めていく予定である。その過程で、大規模データを活用したゲノム進化プロセスに関する基礎的な研究も進めていきたい。また、メタゲノム解析は、近縁ゲノム比較と並んで、今後の比較ゲノムデータベースの重要なアプリケーションのひとつであると考えているが、解析の目的やデータの種類や質が多岐にわたるため、具体的なデータへの適用を通じた活用方法の検討については今後の課題である。

2）生物学データの整備

構築したヒト常在細菌の 16S rRNA 遺伝子データベースを公開する予定である。また、設計した 454 シークエンサーに特化したプライマーセットが、実際の細菌群集構造解析に対して有効であるかを判断するために、細菌群集を対象として PCR や DGGE 等による多様性解析実験を実施する。

3）情報学的ツールの整備

現在、クラス識別としては、サポートベクターマシンやニューラルネットワークなどを用いた幾つかの方法が提案されているため、それらを用いて識別の精度を上げる必要がある。今回提案した MMC-RFE に基づく遺伝子選択のアルゴリズムは計算が安定で効率的であり、またパラメータの設定を必要としない。さらに SVM-RFE と異なり、多クラスへの拡張が容易であるという利点を有する。したがって、マイクロアレイにとどまらず、ハイスループット技術によって生成される膨大かつノイズを多く含むデータからバイオマーカーを探索するための一般的な手法として実用性が期待される。

本研究の成果として、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へと拡張していくことが期待される。

4）新規微生物分野の情報収集のための基盤整備

クロスハイブリが生じている特定のスポットについてはプローブ配列の見直しを行う。また、微生物の検出に用いるプローブは16S rDNAでは十分でないと考えられているため、これまでに全ゲノム配列が決定された微生物における必須遺伝子をプローブとして検討する必要がある。

＜研究期間の全成果公表リスト＞

1）論文

1. 0912021035
   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. 0911301733
   Uchiyama, I., Higuchi, T., Kawai, M.: MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity, Nucleic Acids Res. in press
3. 0901122149
   Uchiyama, I.: Multiple genome alignment for identifying the core structure among moderately related microbial genomes, BMC Genomics, 9, 515 (2008)
4. 0801242034
   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. 0801231328
   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. 0702071708
   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. 0702071629
   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. 0602231719
   Uchiyama, I., Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. Nucleic Acids Res.. 34、647-658 (2006).
9. 0702101408
   Niijima, S., Kuhara, S. Gene subset selection in kernel-induced feature space. Pattern Recognition Letters.27,1884-1892 (2006)
10. 0702101416
    Niijima, S., Kuhara, S. Recursive gene selection based on maximum margin criterion: a comparison with SVM-RFE. BMC Bioinformatics 7:543 (2006)
11. 062231733 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. 0801220119
    Kurokawa, K. et al., Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes., *DNA Res.,* **14**, 169-181 (2007).

2）データベース / ソフトウェア

1. 0702081633
   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. 0911201336
   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/
   遺伝子の配列に基づいたクラスター分類によるオルソログ、ドメイン解析を行い、インターネットを用いた閲覧機能を備えたシステムにドメイン単位の分類とタンパク質単位の分類とを同時に表示し、相互の関係を自由に閲覧できる機能を付加したもの。