プロジェクト名:　統計・情報技術を駆使したゲノム多型と表現型多様性の連関解析システムの開発

プロジェクトディレクター：　城石 俊彦

１．研究目標

複雑系としての生命システムの理解を深めるために、遺伝的分化を遂げた生物個体の多様性の比較解析が有力な方法論として浮かび上がってきている。そのためには、生物多様性を客観的に記載する数値計測技術と得られた計測値の違いをゲノム多型に結びつけるための統計データ解析手法の融合が必要である。また、研究材料となる生物遺伝資源としては、長い進化的時代を経て多数の遺伝子に塩基置換が蓄積し、多様な表現形質を示す生物系統の利用が有効である。国立遺伝学研究所は、古くから世界各地の自然集団から採集し育成してきた遺伝的多様性に富んだマウスやイネ等の多数のモデル生物系統を保有している。本提案プロジェクトでは、これらの生物系統の持つ表現形質の多様性を可能な限り客観的に数値計測化し、それらのデータを対象にして統計数理研究所が培ってきた統計解析技術と国立情報学研究所の情報処理技術を活用することにより、数値計測化された表現型多様性とゲノム多型を関連づけて遺伝子（ゲノム）機能と遺伝子間ネットワークを解明することを目標とする。

２．研究概要

動植物の代表的なモデル生物であるマウスやイネが示す生物多様性を対象に、統計数理技術と情報処理技術を活用してゲノム多型と数値計測化された表現型を関連づけて遺伝子機能と遺伝子パスウェーを体系的に解析するシステムを開発する。また、大学等の外部機関との共同研究により個々のゲノム機能と生物多様性を生み出した進化メカニズムの解明を行う。

３．年次計画

テーマ	１６年度	１７年度	１８年度	１９年度	２０年度	２１年度
全体	←予備研究	プロジェクト初年度		中間評価		国際シンポジウム開催→
表現型多様性データの数値計測システムの開発	←予備研究					→
ゲノム多型と表現型多様性を関連づけるための統計データ解析システムの開発			←			→

平成１６年度（予備研究）

「モデル生物のゲノム・表現型多様性の統制的データ解析システムの開発」という研究課題で予備研究をスタートした。統計数理研究所と国立遺伝学研究所の二研究所間で以下の打合せを行い、融合研究の実施内容について協議した。

全体会議：2004 年 12 月 27 日於国立遺伝学研究所（国立遺伝学研究所２名、統計数理研究所６名参加）

融合研究全体の基本方針を協議した。マウス・イネの表現型、遺伝子間相互作用の数理解析法について具体的な研究方針の検討を行った。

個別会議：2005 年 1 月 26 日於統計数理研究所（統計数理研究所６名、国立遺伝学研究所２名参加）

イネのマイクロアレイによる遺伝子発現解析と生殖的隔離障壁の研究について協議した。

平成１７年度（プロジェクト開始）

プロジェクト全体の研究目標を参加者が共有し、研究方法などを協議するためのワークショップを開催して情報交換を行った。ゲノム多型と表現型多様性の関連づけによるゲノム機能と遺伝子間相互作用の解明のために、主に以下の二つの項目について研究を実施した。(1)表現型多様性データの数値計測システムの開発、(2)ゲノム多型と表現型多様性を関連づけるための統計データ解析システムの開発。各項目についての主な成果を概略する。(1) 形態多様性解析として、マウス下顎骨の画像データについてP型フーリエ記述子を用いた形態数値化とそれを用いた主成分分析が系統間の多様性解析に有効であることを示した。X線CT値による分析において、マウス内蔵脂肪と皮下脂肪を自動的に判別して各脂肪量を定量化するためのソフトウエアの開発に着手した。マウス行動パターンの内、社会行動と自発活動の日周期変動について、客観的な数値計測化とモデルによるシミュレーションを行った。（2）遺伝的距離のある生物系統において、一方の系統のゲノムDNAのプローブセットを用いたマイクロアレイの統計解析についての検討を行い、SNP由来の見せかけのシグナル強度を判定するための方法論を検討した。イネ生殖隔離障壁を引き起こす遺伝子座間相互作用検出のため、検定統計量の相関構造の特定とそれに基づいた多重性の調整の方法について検討した。新しいQTL解析の開発とその有効性の検証を行うためのマウスF2交配による実験データの生産を行った。

平成１８年度

マウスやイネの表現型多様性に関する実験データの収集を精力的に行い、その生データを元にして情報技術を活用した表現型多様性の数値計測システムを完成させる。平行して、数値データを解析するための統計手法の確立を一段と推進する。

平成１９年度（中間評価）

生物多様性解明のため、遺伝子座間交互作用（エピスタシス）の検出システムにより、マウスのエネルギー代謝や種・亜種分化に伴うイネやマウスの生殖隔離、マウス行動に関係したエピスタシス解析を実施する。さらに、改良されたQTL解析系を使ったイネ・マウスの量的形質の統計遺伝解析を行う。また、マイクロアレイデータに基づいた遺伝子発現量の制御に係わる遺伝子群のe-QTLを実施する。プロジェクトの中間時点でのとりまとめを行い、関連分野の研究者に呼びかけたワークショップやシンポジウムを企画する。

平成２０年度

生物多様性解明を目指して、生物形態、エネルギー代謝及びマウス行動関連表現型解析を一段と進めてデータベース化する。QTL、e-QTL解析に基づいたゲノム多型と表現型多様性の連関解析によって、表現型を制御する責任遺伝子（群）のマッピングとそれに基づいた同定を行う。同定された遺伝子の自然集団中での多型性を解析し、生物多様性の実体を明らかにする。また、類縁生物種での関連遺伝子の多型解析により、生物多様化と進化の様相について解明する。

平成２１年度

生物多様性についてプロジェクトで得られた表現型データを統合したデータベースを構築する。表現型を制御する遺伝子の同定をさらに推進する。さらにエピスタシスに関与する複数の遺伝子群を同定する。自然集団での生物多様性データベースを完成させる。５年間の成果を基にして国内外の研究者に広く呼びかけて国際シンポジウムを企画する。

４．平成１７年度研究実施体制

本プロジェクトでは、参加者全員が一つの目標、すなわち「生物多様性」から遺伝子・ゲノム機能と遺伝子ネットワークの解明という共通のテーマを目指して有機的に連携しながら研究を推進するところに特徴がある。平成１７年度は、この点に鑑み、実質的にサブテーマに細分化しない研究体制をとって研究を行った。

具体的には、二つの大項目の下に複数の個別研究テーマを設定し、そのテーマ毎に以下のようなチーム体制を敷いて、研究を推進した。

（１）表現型多様性データの数値計測システムの開発

1) ３Ｄ画像データからのデータマイニング（マウス脂肪組織）

[国立情報学研究所] 北本 朝展、佐藤 真一、藤山 秋佐夫

[国立遺伝学研究所] 城石 俊彦、高田 豊行、前野 哲輝

2)フーリエ記述子を用いたマウス形態（下顎骨等）多様性データの数値計測システム

[統計数理研究所] 田村 義保、鄭 澤宇（総研大学生）

[国立遺伝学研究所] 城石 俊彦、細谷 正樹（総研大学生）

3)マウス行動（行動周期性、社会行動パターン）の自動数値計測システム

[統計数理研究所] 川崎 能典、種村 正美、土谷 隆

[国立遺伝学研究所] 小出 剛、西 明紀（総研大学生）

4)マイクロアレイによるイネ・マウス遺伝子発現データ解析システムの最適化

[国立遺伝学研究所] 倉田 のり、春島 嘉章、堀内 陽子、高田 豊行、城石 俊彦

[統計数理研究所] 江口 真透、藤澤 洋徳、川喜田 雅則（総研大学生）

（2）ゲノム多型と表現型多様性を関連づけるための統計データ解析システムの開発

1)生殖的隔離障壁遺伝子座間相互作用の検出

[統計数理研究所] 栗木　哲、藤澤 洋徳

[国立遺伝学研究所] 倉田 のり、春島 嘉章

５．平成１７年度の研究進捗

各テーマの進捗状況を以下に記述する。

5.1　表現型多様性データの数値計測システムの開発

5.1.1　画像認識の自動化による脂肪組織の数値計測法の開発

【参加研究者】

[国立情報学研究所] 北本 朝展、佐藤 真一、藤山 秋佐夫

[国立遺伝学研究所] 高田 豊行、前野 哲輝、城石 俊彦；[㈱インシリコバイオロジー]大山 彰

【研究の目的】

近年、X 線 CT(Computed Tomography:コンピュータ断層撮影) 装置の高速化に伴い、多数のスライス画像を短時間で取得することが可能となった。得られた画像データから対象臓器を正確に抽出し、客観的数値データに基づく測定をおこなう手法の開発は非常に重要であり、この解析手法の構築は本研究が目的としている CT 画像を使用したマウス表現型の収集を高速に行うこと以外にも、医療分野における診断の高速化、定量化などを目的とした診断支援に応用できると考えられ、非常に有効かつ汎用性がある。我々はマウス全身を対象とした CT 画像から脂肪組織（皮下・内臓脂肪）を自動測定し数値化するシステムを構築するため、複数個体を用いて撮影した CT 画像を用いて各組織の CT 値の抽出と画像毎の CT 値のばらつきを想定し、ノイズに強いアルゴリズムを実装した脂肪組織の自動分離計測ソフトウェアの開発を進めている。

【平成 17 年度の成果】

(1)脂肪組織の自動分離計測ソフトウェアの開発のための入力画像の取得、ならびに付属ソフトウエアを用いた脂肪組織計測における問題点の抽出。

我々は、脂肪組織抽出に適した高解像度の 3D 画像の構築のために、23 個体の CT 画像の収集をおこなった。具体的に CT 撮影に使用するためのマウスには、体長ならびに皮下・内臓脂肪の形状に多様性が期待できる汎用実験用マウスと日本産野生由来マウスの交配から得られた F2 世代を使用し、撮影については尾部を除く全身についておこなった。平成１７年度に導入した CT 撮影装置では最小距離であるスライス間隔 1㎜で撮影をおこない、得られた画像については、CT 撮影装置に付属しているソフトウエアで解析を行い、問題点を検討した。

マウスの全身撮影をおこなった結果、胸部皮下脂肪を測定する場合には周辺にほぼ同じ CT 値を持つ肺組織が存在しており、この部位は CT 値のみから判別する手法には限界があり、脂肪組織をうまく抽出できないことが判明した。また腹部においては皮下・内臓脂肪を分離するための腹筋線が不明瞭になる部位では、皮下・内臓脂肪の抽出がうまくおこなわれず、誤った抽出が頻繁におこなわれことが判明した。CT 撮影装置付属ソフトウエアによる解析の一例を図１に示したが、A) 全体像と解析に使用したスライス画像の位置（緑線)、 B) 解析処理前のスライス画像、C) 解析後、皮下脂肪を黄色、内臓脂肪を赤で表示する。皮下脂肪領域はうまく抽出できているが、内臓脂肪領域の抽出には失敗している（青楕円)。D) 手動更正後の画像であり、付属ソフトウエアにはこれらの「誤り」を手動で校正する機能が存在するが、マウス１個体は約７０枚の画像で構成されるため､すべてを校正するためには多大な時間を費やす必要があることが判明した。

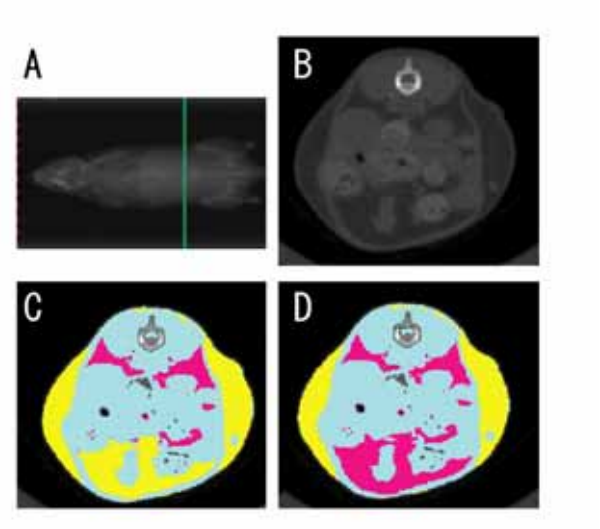


図１　CT 撮影装置附属ソフトウエアによる解析例

(2)画像認識の自動化のための脂肪組織計測ソフトウェアの開発と問題点

自動脂肪組織計測ソフトウェアの開発のための前段階として、CT 撮影装置から得られた画像データを独自に表示することが可能なソフトウエアを開発した。

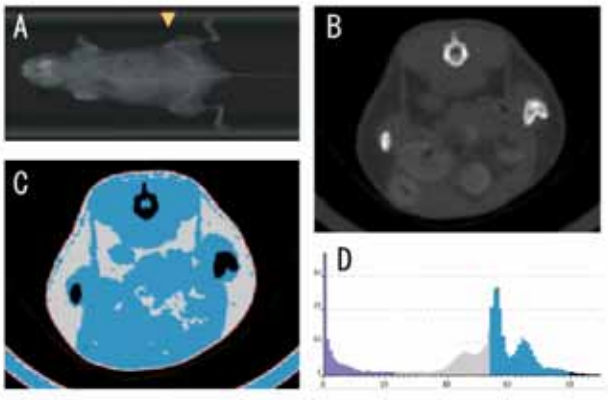


図２．開発中の脂肪組織計測ソフトウエアによる解析例

図２に開発中の脂肪組織計測ソフトウェアによる解析例を示した。A) 全体像と解析に使用したスライス画像の位置（矢印)、　B) 解析処理前のスライス画像、C) 解析後、脂肪を灰色、皮膚を赤、その他の組織を青で表示する。D) C で示した各部位の面積を自動計測するために表示したヒストグラム。横軸は CT 値、縦軸はピクセルの数、ヒストグラムの各色が占める面積は定義した組織の面積を表す。

このソフトウエアを利用し、CT 撮影装置より入手した画像の検証をおこない、撮影装置により得られた CT 値が各画像にどのように分布しているのかを観測したところ、各データセットごとにばらつきが存在することが判明した。すなわち、ある撮影に対するデータセットから脂肪ならびにその他の組織を分離するための最適な CT 値を設定しても、異なるデータセットではその最適な値にずれが生じ、脂肪を正確に抽出することが困難となり、データセットごとに最適な CT 値の設定を変更する必要があることが判明した。

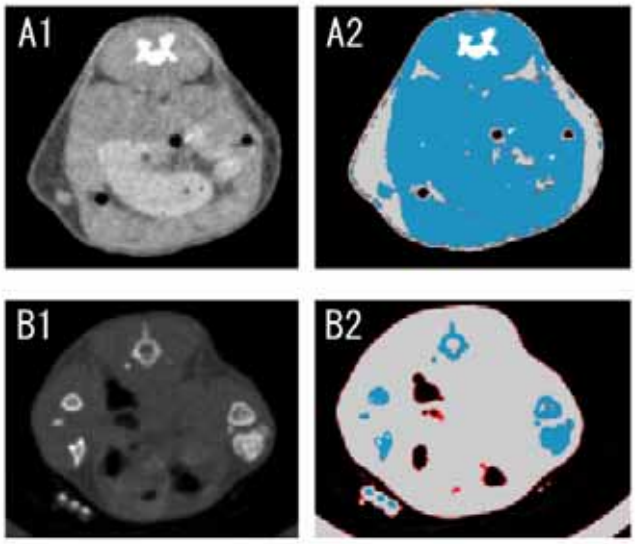


図３．CT 値の差異から脂肪量の自動抽出の例

開発中の脂肪組織計測ソフトウェアにより測定した異なるデータセットの CT 値の差が脂肪の自動抽出に与える影響を図３に示す。A1）全体像、A2）A1 画像における脂肪およびその他の組織を明確に分類するために設定した CT 値により認識された脂肪（灰色）とその他の組織（青）および皮膚（赤）。B1）異なるデータセットの全体像、B2）A1 で設定した脂肪およびその他の組織を分類する CT 値を適用した例。脂肪およびその他の組織の分類に失敗しており、現在、最適 CT 値の自動補正について検討をおこなっている。

最後に、皮下および内臓の脂肪を分離するための腹筋線抽出アルゴリズムの開発については、Deformable Models（変形モデル）を利用したアルゴリズムの開発が必要であり、パラメトリックな形状モデルとエネルギー関数（あるいは確率モデル）を定義し、制約つきエネルギー最小化（もしくは尤度最大化等）するという最適化問題を解くことによって、最適なパラメータを求めるという全体的な枠組みを構築している。具体的には、active contour（輪郭線抽出）、active net（面の抽出）、active tube（管の抽出）など、最適パラメータの意味付けを、画像からの情報抽出モデルとして使用するモデルの構築、とくに皮下と内臓の脂肪を抽出するために必要な腹筋線を指定する際の、「典型的な腹筋線」の同定、実際の画像からの腹筋線の候補となる画素の抽出、典型的なモデルからみた実際の画素の位置を変位とし、その歪みのエネルギーを最小化する、あるいは変位を確率変数と定義し、確率密度関数に基づいて尤度を最大化する、などの方法を検討中である。来年度はこれらアルゴリズムをソフトウェアに導入したい。

5.1.2 形態多様性の数値計測技術の開発

【参加研究者】

[統計数理研究所] 田村 義保、鄭 澤宇（総研大学生）

[国立遺伝学研究所] 城石 俊彦、細谷 正樹（総研大学生）

【研究目標】

生物の形態は多様であり、その形態差を比較することで種（亜種）を特定・分類することが古くから行われてきた。げっ歯目などの小型哺乳類では下顎骨の形態が指標として広く用いられており、特徴点間の距離を計測するのが一般的である。また最近では楕円フーリエ記述子を適用した輪郭全体の定量的評価による形態解析も行われている。このように形を数値計測することで生物形態の特徴を抽出することは可能であるが、計測の客観性や遺伝解析に結びつけていく際の取り扱いの簡便性などについては、未だに問題が残っている。本研究では、形態差を規定している遺伝的多型性を明らかにすることを最終目標として、より実用的な形態の数値計測技術を開発することを当面の目標としている。

【平成１７年度の成果】

遺伝的に異なるマウス近交系統である MSM/Ms 系統（ MSM 系統、*Mus musculus mollosinus* ）および C57BL/6J 系統（ B6 系統、*Mus musculus domesticus* ）（図１参照）、これら二系統間で F2 世代を作成し、その下顎骨の形態差を指標とした Quantitative Trait Loci (QTL) 解析を試みた。F2 世代の下顎骨は、MSM、B6 両系統の特徴を併せ持つ多様性に富んだ形態を示しており、これらを定量的に評価するには従来の計測手法では不十分な可能性がある。

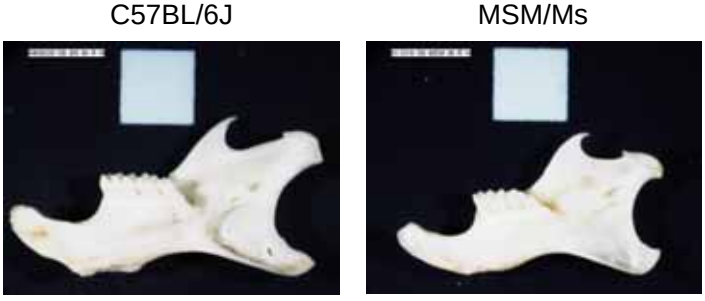


図１．MSM/Ms 系統と C57BL/6J 系統の下顎形態

そこで、形態的特徴の差をより詳細に比較するための解析手法として P 型フーリエ記述子の適用を試みた。P型フーリエ記述子とは、曲線を長さの等しい線分からなる折れ線として近似し、隣接する線分間の角度変化（偏角）を求め、その累積で定義される全曲率関数を基にしている。さらに、図形の回転、拡大、縮小に対して不変であるという性質を持つ。この手法では輪郭の部分形状（開曲線）を定量的に評価することが可能であることから、下顎骨を筋突起、下顎突起な

どの部位ごとに切り分けて解析することが可能となる。

実際に MSM、B6 両系統の筋突起について P 型フーリエ記述子を適用し、主成分分析を行った結果を図２に示す。解析結果をみると第一主成分において二系統間の違いが明確に表れている

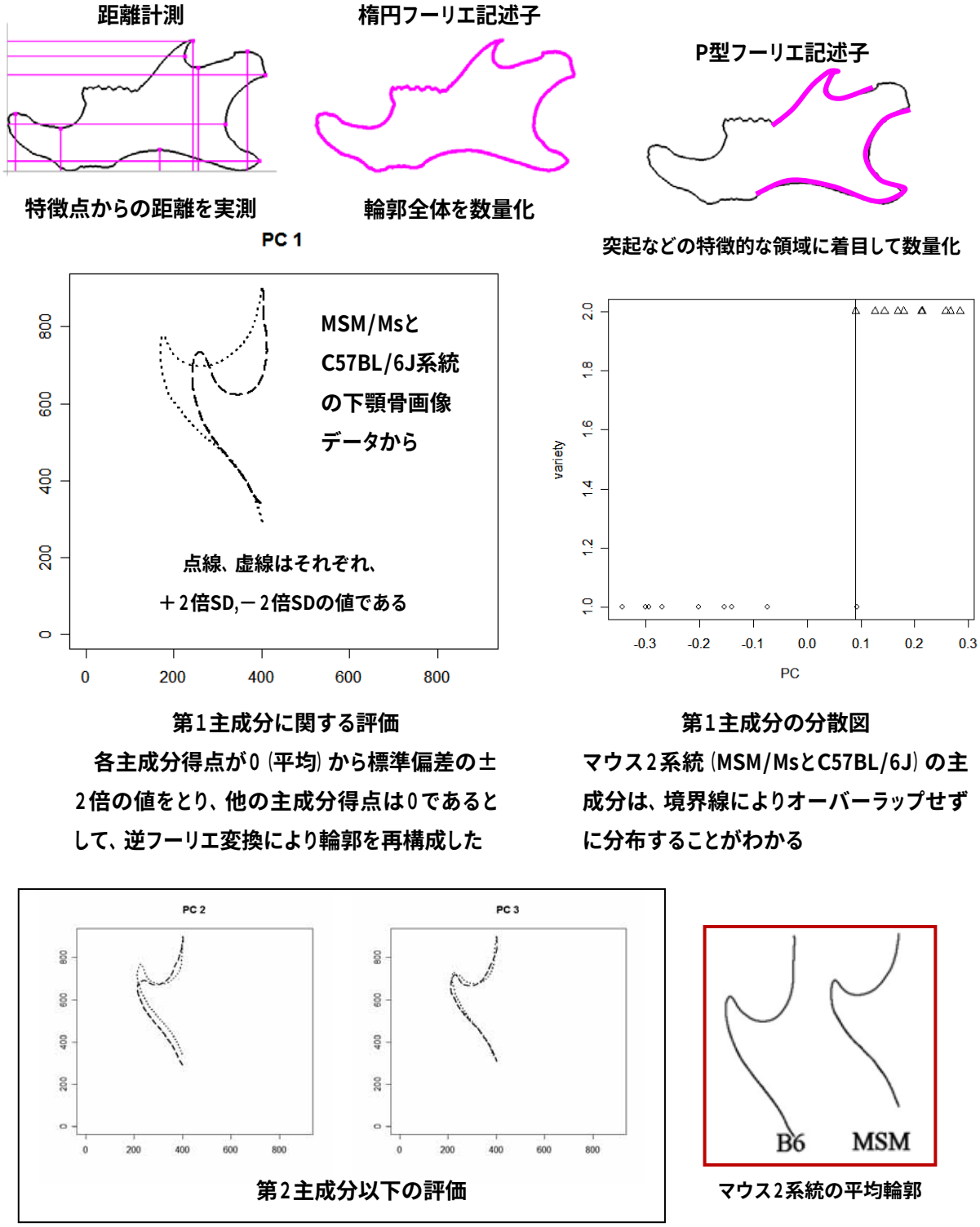


図２．P 型フーリエ記述子を用いた形態計測と解析結果

今後の方針としては、F2 世代を用いた QTL 解析に P 型フーリエ記述子を適用することで、従来の計測手法と比べてP型フーリエ記述子が優れているかどうかについて検証していくとともに、これ以外の計測手法の適用や、計測後の多変量データの取り扱い方法についても順次検討を行っていく予定である。

5.1.3 マウス行動（行動周期性、社会行動パターン）の自動数値計測システムの開発

【参加研究者】

[統計数理研究所] 種村 正美、土谷 隆、川崎 能典

[国立遺伝学研究所] 小出 剛、西 明紀（総研大学生）

【研究目的】

遺伝研において開発維持している多様な野生由来マウス系統は、高次機能の多様性を解析するための有用なリソースとなる。その高次機能の中で、行動表現型は形としては残らないため、その計測記録システムおよび解析には困難が伴う。また、行動は時間軸に沿った発現をするため、その定量的な評価は難しい。本研究では、マウス行動を定量的に計測し、得られたデータを数理的な手法を用いて解析して、将来の研究において発展性のある行動成分を抽出し、その形質の背後にある特性をモデル化により明らかにすることを目的とする。そのため、統計数理研究所と共同で、マウスの自発活動性や社会性などの行動を時系列に沿って効果的に解析するシステムの確立を進めた。

【平成１７年度の成果】

（1）マウス自発活動における時系列解析　（川崎 能典・小出 剛・西 明紀）

自発活動性は、動物にとってテリトリーの確保や餌の探索といった生存に深く関わる重要な行動形質の一つである。その活動には周期性が見られ、概日周期についてはすでによく知られている。しかし、動物は活動期においても一様に活動しているわけではなく短い時間で活動と休息を繰り返しているように見える。このような超日周期については、まだほとんど研究が進んでいない。その原因は、超日周期を解析するための有効な解析手法がこれまで確立されていなかったためである。そこで、遺伝的に異なったマウス系統である MSM と C57BL/6 (B6)を用いて、その活動性について解析系の確立を目指した。

解析に用いたマウスは、照明を 8:00 から 20:00 までを明期、20:00 から 8:00 までを暗期としてコントロールし、室温 22±2°C、湿度 50±10%に保った飼育室内で維持した。解析には MSM と C57BL/6 の生後 8～15 週齢の雌雄を解析に用いた。各マウスはテストの前に個別に飼育ケージに移し 1 日間飼育した。その後、ACTIVITY SENSOR（O’hara,Co., Ltd）の測定ケージにそれぞれ移し、4 日間連続でその活動量を自動計測した。最初の 1 日間は馴化期間として解析からは除外した。各個体の自発活動量は 1 分毎に表す時系列データとして記録した。

MSM と B6 の雄それぞれ 1 個体のデータを参考までに以下に示す。マウスは暗期に主に活動す

るため、20：00 から 8：00 までの活動量を示している。図１のように、夜間での活動に 2 系統間で違いが見られるようであるが、個体によるばらつきもあり、その判断は難しい。

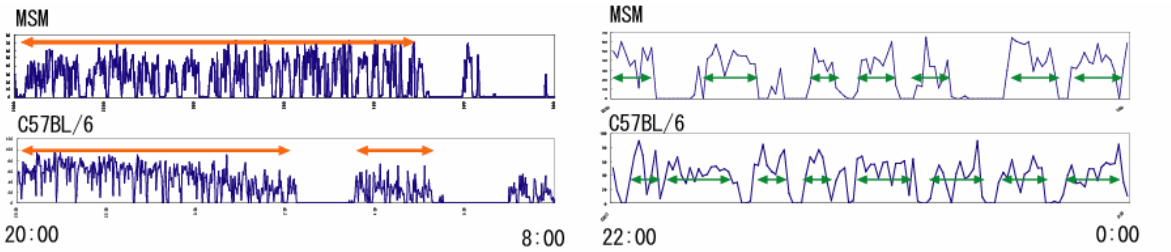


図１．MSM と B6 の活動性の例

そこで、数理的解析手法の導入を進めた。この解析では、点過程で活動をモデル化し、24時間でのセンサークロス頻度を強度関数により推定する手法を用いた。その後、個体ごとに推定した強度関数を主成分分析することで系統に共通の日周期を抽出することを目指した。

点過程とは、時間軸上で不等間隔に発生する事象列のことであり、ここではマウスのセンサークロスがそれに相当する。実際には1分あたりの交差回数しか計測されていないが、観測された回数で60秒を割り、各観測に便宜的な発生時間を与えることで、あたかも点過程のように扱うことが可能である。ここでは、横軸を細かく微小時間に区切り、各微小時間ごとにある高さの矩形（くけい）を考え、各微小時間区間内に事象が起こる確率が対応する矩形の高さにおよそ比例するという確率モデルを考える。矩形の高さは単位時間あたりの事象の起きやすさ＝強度であり、24時間を一周期とするフーリエ級数で表現する。その結果、活動強度が個体ごとに、時間の関数として推定される。推定されたフーリエ級数（連続的曲線）を24時間区間内で等間隔に離散化すれば、個体数×離散時点数の多変量データのように扱い、主成分分析を適用することができる。解析結果を観察すると、第一主成分はほぼ共通の日周期に対応する。それをB6とMSMで比較したのが図２である。

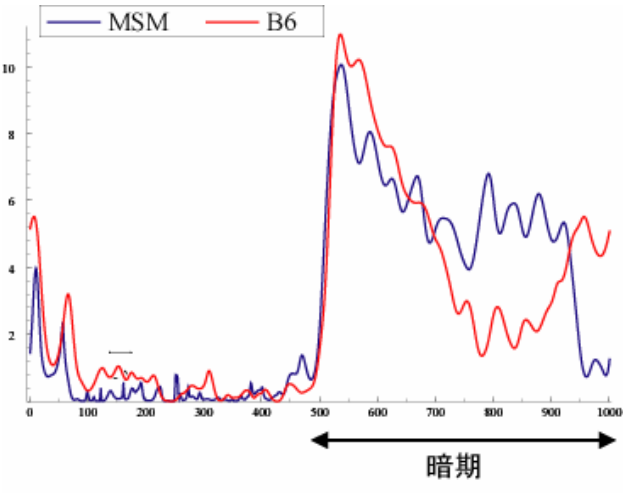


図２．第１主成分の系統間比較

今後の方針としては、データは本来非負整数値（計数時系列）なので、それを直接表現するようなモデルを与え、統計的モデル選択の議論の中で、例えばB6とMSMとで活動周期が同じか否か等の客観的比較を行う。分布関数としてはポアソン分布が適切だが、1分ごとの観測区間の背後に、時間固有のポアソン母数を想定する。（点過程との対比でいえば、強度関数を1分区間ごとに積分した量がポアソン母数である。）ここではポアソン母数のダイナミクスを、離散時間時系列モデルで表現する。

具体的には、ポアソン母数（の対数値、非負制約を課すため）を、集団に共通の周期成分と、個体特有の変動成分とで表現する。実際に観測された計数時系列から、それと整合的であるような時変ポアソン母数を求めること、ひいてはポアソン母数のダイナミクスを構成する周期関数や個体固有成分の時系列変動モデルに含まれる未知母数を推定することが、ここでの統計的推定問題になる。

（２）マウスにおける社会行動解析　（種村 正美・土谷 隆・小出 剛・西 明紀）

社会的動物であるマウスにとって、同種の他個体を認識し、コミュニケーションを行うことは非常に重要である。この社会性には個体差があり、遺伝的寄与が示唆されているが、その遺伝的基盤については解明が進んでいない。その原因は、動物が示す社会性を定量化する効果的な手法が確立されていないことが大きな原因の一つであると考えられる。そこで、本研究では、マウス 2 個体が示す社会行動を定量化する手法の確立を目指した研究を行った。

テストに用いる 2 個体は同リターの同姓の個体であり、9～10 週齢時に個別飼育し、テストは個別飼育から

図３．オープンフィールドテスト

約 10 日後に行った。テストでは、2 個体のマウスをオープンフィールド（60×60×40 ㎝）に入れ、10 分間自由に探索させ、2 個体の行動をビデオで記録した（図３参照)。その後、コンピュータ上で Image SI（O’hara,Co., Ltd）を用いて画像解析を行い、2 個体の接触時間、接触回数、3 count／1 sec での各個体の位置情報を算出した。

これまでに、MSM と B6 系統、更に B6 系統の１本の染色体を MSM 由来のものに置換したコンソミック系統を用いた社会行動の解析を進めた。すでに MSM と B6 系統、更に半数以上のコンソミック系統の解析が終了した。親系統である B6 と MSM は社会相互行
動に顕著な違いを示し、MSM は雌雄共に接触時間が長く、接触回数も多かった。一方、コンソミック系統においては、B6 と差を示した系統は 17 系統中 2 系統のみであった（図４参照)。

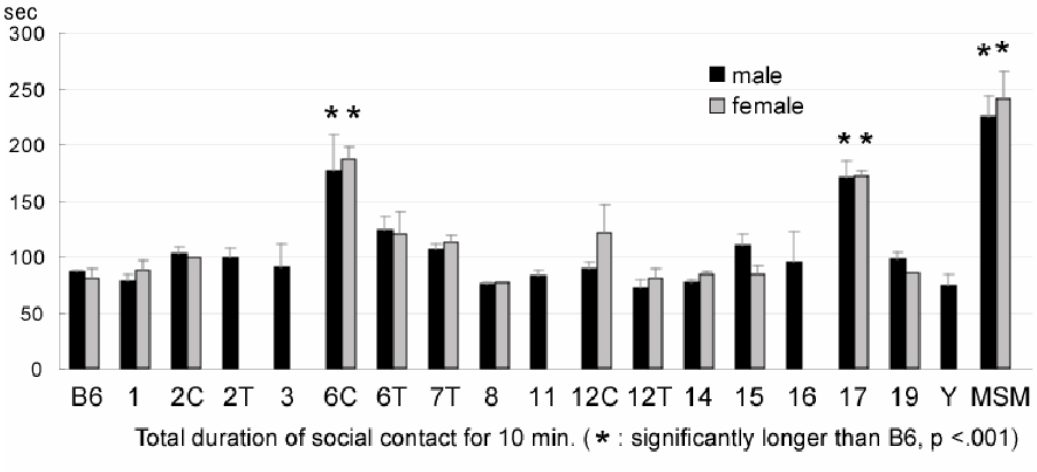


図４．コンソミック系統の社会行動特性

画像解析ソフトによる解析からは、顕著な差が検出されなかったにもかかわらず、実際の映像を見るといくつかの系統に特徴が見受けられた。例えば、B6-6TMSM は MSM に特徴的な追随行動が多く出現する系統であった。このような、画像解析ソフトからは検出されなかった差を検出するために、動き情報を抽出し、数値的に解析する手法の開発が不可欠であることが判明した。本年度は、以下の二つのアプローチによりマウス社会行動の解析手法確立を試みた。

１）　２次元 unit vector chain を用いた解析

$$f_\xi = -\xi \, S_i \cdot S_j \qquad\qquad S_i \cdot S_j = \cos(\phi_{12})$$

この解析では、単位ベクトルの系列の特徴をあるパラメータで数値的に表現できる統計的手法（Tanemura, 1994）を用いる。例えば、隣接するベクトルがほぼ同じ向きにそろったベクトルの系列の場合、ベクトル間に誘因相互作用が働いていると考え、逆に、隣接するベクトルが逆向き

になる傾向のあるベクトル系列の場合、反発相互作用が働くと見なす。隣接するベクトル間の相互作用エネルギーを示す数式がスライドの最初に与えてある。この相互作用の系列に関する和についての統計分布（ギブス分布）から、パラメータの尤度推定をするのが上記のわれわれの開発した統計的手法である。誘因相互作用の傾向のあるとき、パラメータは正の値をとり、反発相互作用の場合は負の値をとる。その絶対値は相互作用の強さに対応する。この解析法をマウスの社会行動に応用するのが本研究の主旨である。（エフグザイの意味は相互作用エネルギーを表し、隣接する２つのベクトル s_i と s_j の間のエネルギーを表す。マイナスがついているのは、慣習的な事情からで、統計物理学においてイジングモデル（ising model）などに使われる。マイナスグザイと s_i ・ s_j とは掛け算であり、エフグザイのグザイはエフの添え字であり、グザイがパラメータであることを強調する記法である。ファイ１２は個体１と個体２が示すベクトルの角度を表し、二つのベクトルの内積はコサインで表される。最終的な尤度推定法で求められた推定値としてのグザイはハットをつけて表される。）

そこで、まずマウス２個体の動きの時系列を方向ベクトルの系列（２次元の単位ベクトル鎖）に次のやり方で変換する。すなわち、最初の単位ベクトルは一定の向きに設定しておく。そして、その後の単位ベクトルを２個体の行動パターンに対応した角度を設定することによって､次々とベクトルの鎖（unit vector chain）構成する。例えば、ある時刻から次の時刻の間に２個体が接近行動をとるとき、0 度に近い角度を割り当て、逆に回避行動をとるとき 180 度に割り当てる。s_i 、 s_j が同じ向きだと角度はゼロになり、COSは１になる。0 度から 180 度の間はさまざまな割り当て方が可能であるが、今年度は一つの割り当て方に基づいた解析の中間結果を出し､マウスの系統ごとにパラメータ推定値が得られ､系統ごとの社会性の特徴が数量的に表現できる可能性が示唆されている(表参照)。

Data	$\hat{\xi}$	Mean dist.
B6 female	1.2306	57.60
B6 male	1.1684	55.37
Chr2C female	2.0657	50.09
Chr2C male	2.1387	60.61
Chr6T female	2.4680	60.72
Chr6T male	2.5811	71.98
MSM female	3.5891	47.48
MSM male	2.9604	46.44

現在、どのような角度の割り当て法が良いかを検討中で、来年度は本格的にデータ解析に取り組む予定である。

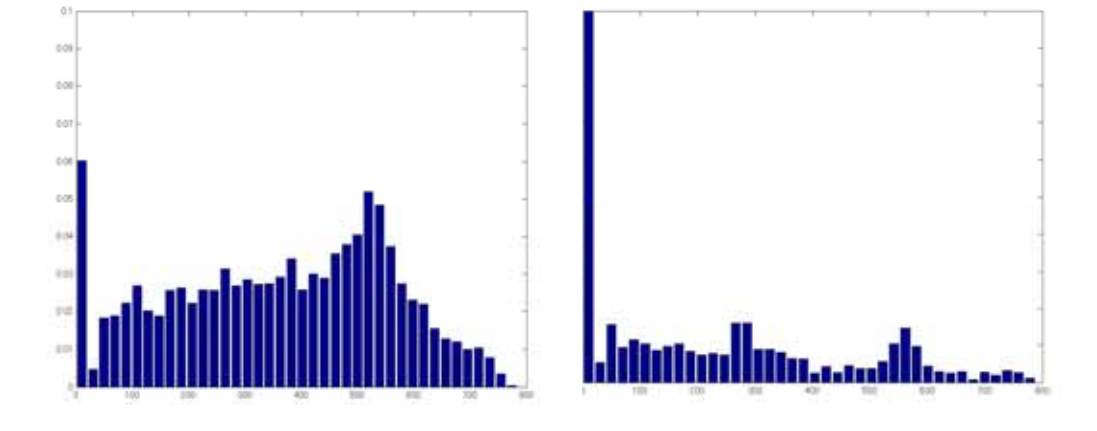
２） 2個体の距離情報を用いた解析

この解析では「B６とMSMの２個体の距離分布を精密に求め、両者の混合分布としてコンソミック系統の距離分布を推定し､混合パラメータの値で社会性を特徴付ける」ことを目標とする。(研究の展開によっては「距離」以外の特徴量の分布を用いることも考えられる。）

２個体の距離は、0ｃｍからケージの対角線の長さ84cm の間で変化する。そこで、２個体の距離の分布（確率密度関数）を、0から84の間で（非負の値をとる）多項式で表す統計モデルを

考える。新しい最適化手法である半正定値計画法を用いると、このモデルのデータへのあてはまりの一番良いパラメータ、およびそのあてはまりの良さの尺度であるＡＩＣ(赤池情報量規準)を効率よく厳密に計算することができる。いろいろな次数の多項式のモデルが考えられるが、AICを計算し、一番良い次数のモデルを選択する。

上に述べた手法により、まず、基本となる、Ｂ６の２個体の距離の確率密度$P_{B6}(x)$と、MSMの２個体の距離の確率密度$P_{MSM}(x)$を求める（Ｂ６の２個体の距離のヒストグラムとＭＳＭの２個体の距離分布のヒストグラムによる推定を下図に示す。推定されるP_{B6}やP_{MSM}の形はこれらのヒストグラムと似た格好となると思われる)。



Ｂ６（雄）　　ＭＳＭ（雄）

（MSMの一番左のビンは打ち切られており実際の値は約0.75.）

そして、各コンソミック系統の２個体の距離の分布（確率密度関数）を、上で推定した$P_{B6}(x)$と$P_{MSM}(x)$を元に、

$\alpha P_{B6}(x)+(1-\alpha)P_{MSM}(x)$　　x：2個体の距離

と表現し、コンソミック系統ごとに、観測データに一番良く合う最適なα（αは0と1の間の数）を最尤法によって求め、これを社会性の指標とすることを考えている（この部分の計算は容易である)。αが１に近ければ、そのコンソミック系統はB6に振る舞いが近く、０に近ければ、MSMに振る舞いが近いとみなす。MSMの方が振る舞いについては社会性が高いと考えられているので、αが１に近い方が社会性が高いと考えられる。

本年度は、ヒストグラムの作成等データの予備的な解析、そして、本研究で必要となる半正定値計画法のアルゴリズムの理論的解析を行った。来年度は、上に述べた方針で研究を進める予定である。

5.1.4 マイクロアレイを用いた発現遺伝子の質および量と形質多様性の相関研究

【参加研究者】

[統計数理研究所] 川喜田 雅則、藤沢 洋徳、江口 真透

[国立遺伝学研究所] 倉田 のり、春島 嘉章、堀内 陽子、城石 俊彦、高田 豊行

【全体の研究目的】

現在、様々な生物種において網羅的な遺伝子発現解析のために、マイクロアレイを用いた方法が主流となっている。本研究の目的は、その中でも世界的に汎用されている Affymtrix 社の GeneChip Genome Array を用いて、高頻度で単塩基多型（SNP）を示す遺伝的に多様性な生物系統について、精度良く遺伝子発現量を測定するためのシステムを開発し、それによって多様なマウスおよびイネ系統において、体系的な遺伝子発現解析を行うことである。

（1） マウス亜種系統を用いたゲノム多型に配慮したトランスクリプトーム解析

（高田 豊行、城石 俊彦、江口 真透、川喜田 雅則）

【研究の目的】

実験用マウス系統であるC57BL/6J(B6)と国立遺伝学研究所において樹立された日本産野生由来近交系統であるMSM/Ms(MSM)には、100万年におよぶ地理的隔離に起因した約1%のSNPが存在すると考えられる。これらマウス系統間に存在する多様な表現型の違いがこれら約1%のゲノム多型に起因すると考えられることから、これらゲノム多型が支配する表現型の多様性について、特に遺伝子発現差異を中心に検討するためマイクロアレイによる網羅的遺伝子発現解析をおこなう。さらに膨大なデータを適切に解析するための新しい統計学的手法を検討し、表現型の多様性に関連付けた網羅的遺伝子発現ネットワークの構築をおこなう。本年度は以下1）および2）について検討をおこなった。

【平成17年度の成果】

1) マイクロアレイによる網羅的遺伝子発現解析をおこなうための25mer x 11 perfect match プローブ情報の取得ならびに B6-MSMゲノム間で異なる塩基配列の検定

MSM-Whole Genome Shotgun データよりMSM系統のＳＮＰ情報の取得をおこない、Affymetrix Mouse Genome 430 2.0 Array に搭載されている B6 系統の塩基配列情報を基に設計されたプローブについてマッピングをおこない、422,162 のプローブの約 44%を MSM ゲノム中にマップすることに成功した。さらにこれらをプローブセットごとに集計したところ、所属する Probe 配列が 11 個共すべて MSM read により完全に覆われている 5,492 組の Probe セットを明らかにした。さらに、これらの Probe セットを構成する各プローブにおける SNP の詳細な情報を明らかにした。表 1 に Probe 配列が 11 個共すべて MSM read により完全に覆われている 5492 組のプローブセットに含まれる SNP の頻度を示した。

表 1　MSM readにより完全に覆われているプローブセットに含まれるSNPの頻度

No. of SNP-free probe pairs	Frequency in 5,492 probe sets
11	2474
10	1116
9	826
8	501
7	287
6	136
5	87

２）GeneChipを用いたB6およびMSM系統の肝臓における遺伝子発現量測定

B6-MSMマウス系統間における発現遺伝子の質および量と形質多様性を検討するため、DNA チップ（ GeneChip ）により、10週齢雄 の肝臓のサンプルの発現遺伝子を測定した。DNA チップは、 Mouse Genome 430 2.0 Array （ Affymetrix 社製）を使用した。Mouse Genome 430 2.0 Arrayには、約45,000遺伝子が搭載されており、内訳は 最新のアノテーション情報で定義された37693個のUnigene IDを持つProbe setとなっている。図1 に、GeneChip で測定したB6およびMSM系統の肝臓における遺伝子発現量の蛍光シグナル値のスキャタードプロットを示す。両系統の肝臓サンプルについてDNA チップで測定した約45,000遺伝子のうち、MSM系統で2 倍以上の発現増加が認められた遺伝子数は607 個、1/2 以下の発現減少が認められた遺伝子数は930個であった。

現在、 MSM系統の塩基配列差（ＳＮＰ）による遺伝子発現量の蛍光シグナル値の補正をおこなうため、表1で示したプローブセットを利用したMAS5およびRMAアルゴリズムによる補正結果の検討をおこなっている。

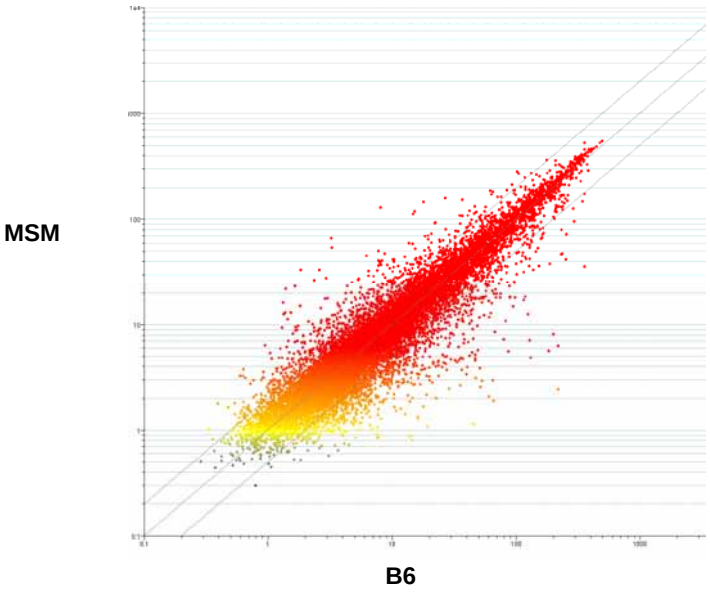


図１．GeneChip で測定したB6およびMSM系統の肝臓における遺伝子発現量スキャッタードプロット。使用個体、 B6およびMSM系統、 10週齢、肝臓、チップ： Mouse Genome 430 2.0 Array

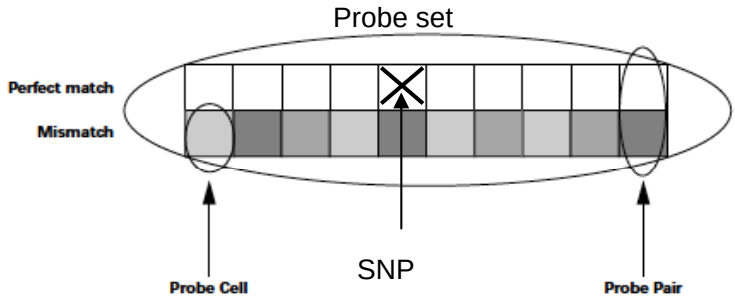
（２） イネ多様性研究におけるゲノム複合データの融合的解析

（倉田 のり、春島 嘉章、堀内 陽子、江口 真透、川喜田 雅則）

【研究目的】

物の形質や形態の多様性は発現する遺伝子の構造（遺伝子塩基配列）や遺伝子の発現量の差によってもたらされる。本研究の目的は、Affymtrix 社の GeneChip Rice Genome Array を用い、遺伝子構造の差と遺伝子発現量の差を別々に見積もり、多様な遺伝的背景のイネの形質や形態の差を解析することである。

Rice Genome Array は UniGene Build#52、Genbank mRNAs (July 13, 2004)と predicted genes from TIGR's osa1 version 2.0 をクラスタリングし、ジャポニカの 48,564 の転写物とインディカの 1,260 の転写物がターゲットとして設定されている。各転写物の 3 ' UT 側 32～600bp のターゲット配列中に 8～11 カ所の 25mer Perfect Match probe (PM)と、
そのプローブの中央にコンプリメント塩基のミスマッチを入れた Mismatch probe (MM)がペアで設定されている。これらプローブペアを1セットとしてターゲット遺伝子発現量をハイブリダイゼーションのシグナル強度で計測、推定するという設計である。



上図に１個のプローブセットのイメージ図を示した。通常の遺伝子発現量の推定は、PM と MM のシグナル強度差の平均値から推定している。この解析手法で、多様な遺伝的背景をもつ野生イネ等の遺伝子発現解析を行うと、プローブが設定されているターゲットの一部の塩基配列の差によってターゲット遺伝子全体の発現量の推定を誤ってしまう。そこで、我々は遺伝子構造の差と遺伝子発現量の差を別々に見積もるため、以下の２種類の異なる手法の開発を試みている。

i) 野生イネ等の塩基配列が分からない遺伝子の発現量を GeneChip で解析する前に、そのゲノム DNA を GeneChip にハイブリダイッゼーションすることにより、プローブのターゲット配列にある SNP 等の遺伝子構造の差を推定する。次に遺伝子構造に差の無いプローブのみを選んでセットを再構築し、各遺伝子の発現量のみのデータを拾い上げ、解析する。

ii) プローブ設計の基となった系統のイネを用い、様々な組織より mRNA を抽出し遺伝子の発現解析を行う。次に発現遺伝子のプローブセット内の PM 値または(PM-MM)値等の値が発現量に応じどのような相関になるかプロファイリングする。野生イネ等の塩基配列が分からない遺伝子の発現量を GeneChip で解析する時に、このプロファイリングからずれたプローブに SNP 等の遺伝子構造の差があると判定し、遺伝子構造に差が無いと判定されたプローブのみでセットを再構築し各遺伝子の発現量を推定する。

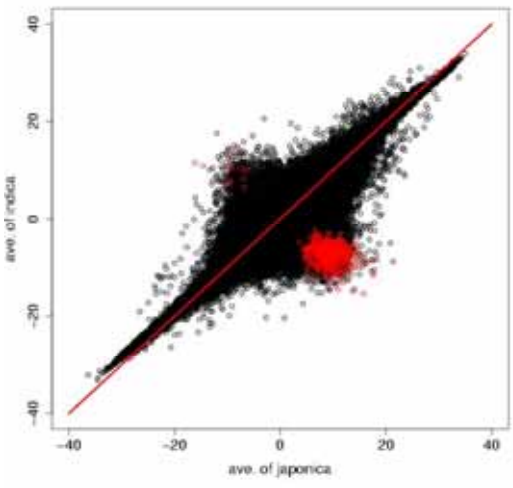
Rice Genome Array のプローブ設計の基となった塩基配列は、主に国際イネゲノム塩基配列解読プロジェクトで用いられたジャポニカ品種「日本晴」に由来するのものである。イネでゲノムの塩基配列が分かっているものとしてはインディカ品種の「93-11」があり、両者の塩基配列の差を右表に示す。これらゲノム塩基配列が既知の日本晴と 93-11 を用い構造と発現量の差を区別できるか本年度より試みている。

Category	Region	SNP/kb	Indel/kb	SNP/Indel
Gene regions	5′ UTR	4.72	1.14	4.15
	Coding region	3.00	0.22	13.71
	Nonsynonymous Ka	2.10		
	Synonymous Ks	5.93		
	Introns	6.07	1.28	4.76
	3′ UTR	4.50	1.01	4.46
Total genome	Genome-wide	15.13	2.89	5.25
	Copy number ≤ 10	13.74	2.66	5.17
	Transposons	27.64	4.61	6.00

【平成 17 年度の研究成果】

1）マイクロアレイ実験

今年度は、Rice Genome Array を用いて手法 i)の実験を行った。つまり日本晴及び 93-11 よりゲノム DNA を抽出し、それぞれ４回の繰り返し実験によりハイブリダイゼーションシグナルを得た。右の図は日本晴及び 93-11 のプローブ毎の PM-MM の平均値を使った SAM(significance analysis of microarray)解析で、対角線から大きく外れた右下の赤い点が有意に日本晴が 93-11 よりシグナル強度の高いプローブで、左上の赤い点が93-11の方が日本晴より有意に大きな値を示すプローブである。



これらの有為差を示したプローブについては、今後 2.2.2 および 2.2.3 で収集した塩基配列レベルの情報と比較検討することにより、どのように配列とシグナル強度が相関しているかの解析を進める。

また手法 ii)のために下表に示した日本晴の 10 の組織より mRNA を抽出し、ハイブリダイゼーションによる発現シグナルを得た。

Tissue	Number of repetitions
Callus on callus induce medium	3
Callus two days after transplanted on regenaration medium	3
Callus four days after transplanted on regenaration medium	2
Callus six days after transplanted on regenaration medium	3
Callus eight days after transplanted on regenaration medium	2
Shoot	2
Young leaf	3
Young panicle	2
Flowering panicle	2
Root	2

これらのデータについては、発現シグナルの妥当性がアレイごとにどのように判断できるか、など基本データについて検証の後、実際の比較解析に入る予定である。

２）プローブおよびターゲット塩基配列の検索

日本晴と 93-11 のそれぞれのゲノム塩基配列、Rice Genome Array の 57,381 個のターゲット塩基配列（コントロールを含む）、631,066 個の 25bp のプローブ塩基配列について以下の解析を行った。

i) BLAT 検索によるターゲット塩基配列の日本晴と 93-11 のそれぞれのゲノム上へのマッピング及びターゲット配列のエクソン・イントロン構造の特定。

ii) ローブ塩基配列を用い、ターゲット塩基配列に対する blastn による検索、及び 1 により再構築した93-11のターゲット配列に対しても検索を行った。(*in silico* mRNAs hybridization)

iii)プローブ塩基配列を用い、日本晴と 93-11 のそれぞれのゲノム塩基配列対する blastn による検索。(*in silico* genomic DNA hybridization)

BLAT 検索の結果 57,381 中 55,080 個のターゲット配列が日本晴ゲノム中にマップされ、54,019 個のターゲット配列が 93-11 ゲノム中にマップされた。

ゲノム塩基配列に対するプローブ塩基配列の blastn による検索では、ゲノム上に似た塩基配列配列がどのぐらいあるかも調べた（下表参照）。blastn の通常の scoring では 25bp の perfect match は score 25、Mismatch probe に相当する score は 21 だが、念のため score 18 まで調べた。

	Above score 18						Perfect match					
	Nipponbare			93-11			Nipponbare			93-11		
	Copies	Probes	(%)	Copies	Probes	(%)	Copies	Probes	(%)	Copies	Probes	(%)
Max	2,079	1	(0.0)	3,109	2	(0.0)	952	1	(0.0)	1415	1	(0.0)
Single	1	455,389	(72.2)	1	403,630	(64.0)	1	512,987	(81.3)	1	407,572	(64.6)
No	0	45,373	(7.2)	0	76,957	(12.2)	0	75,964	(12.0)	0	166,504	(26.4)

プローブ塩基配列がゲノム塩基配列に対してマッチしないものは、コントロール（イネ以外の配列）プローブ、あるいはイントロンをまたぐプローブ、ターゲット配列そのものがイネゲノムに位置づけられないものなどである。詳細は今後の解析となる。

３）データ解析のためのデータベース構築

塩基配列の解析結果と Rice Genome Array を用いた実験結果を統合化した in house のデータベースの構築を行い、GeneChip Rice Genome Array を用い、遺伝子構造の差と遺伝子発現量の差を別々に見積もり、多様な遺伝的背景のイネの形質や形態の差を解析可能か検討する。ゲノ

ムシークエンスを行わずにハイブリダイゼーションシグナルの差のみで、各プローブの構造差を見積もることができるようになると、副産物として、多様な系統間で構造変異の大きな遺伝子のみを抽出する方法への応用も可能となる。このことにより、多様性・進化研究への異なる切り口での展開へも視野に入れたい。

プローブ毎の情報として、以下の情報を取り込み中。

i)塩基配列の解析結果（Genome Blast による copy 数の結果、ゲノム上の位置、イントロンの有無）

ii)物理化学的特性（GC contents, 最近接法で推定した各 probe の結合力）

iii)Microarray 実験結果（シグナル強度、各種統計量）

プローブセット毎の情報として、以下の情報を取り込み準備中

iv)塩基配列の解析結果（ゲノム上の位置、ターゲット遺伝子情報）

v)構成プローブの情報（Affymetrix 社の設定、ゲノム上の single copy の probe 数）

vi)Microarray 実験結果（Affymetrix 社の設定による解析値、SNP を考慮した解析値）

４）イネアレイについての問題点

Affymetrix 社がプローブの設計に用いた推定遺伝子構造が、現時点で推定されるものと異なるものがあるため、幾つかのプローブがイントロンに設定されたり、現時点で１個の発現遺伝子に対して複数のプローブセットが設定されたりしている。対処の方法としては、現時点で最も信頼出来る推定遺伝子構造に基づき新たにプローブセットを再定義する必要がある。現在公開されているなかで、推定遺伝子構造が最も信頼出来るものを調査中。

【次年度以降の計画】

i)データベース化されたデータをもとに、プローブ毎のゲノム DNA を用いたハイブリダイゼーションのシグナル強度、塩基配列検索の結果、プローブの物理化学的特性との相関関係を明らかにし、遺伝子構造の差と遺伝子発現量の差を別々に見積もるための手法(1)の妥当性を検討する。

ii)塩基配列が知られていない野生イネのゲノム DNA を用いたハイブリダイゼーションを行い、プローブ設定領域の塩基配列の差を見積もる。

iii)日本晴の様々な組織から mRNA を抽出して遺伝子の発現を Rice Genome Array で解析し、プローブセットによる推定遺伝子構造の妥当性の検討と、セット内のプローブ間のシグナル強度の相関のプロファイリングを行う。

iv)上記(1)(3)の解析結果と、現時点で最も信頼出来るデータセット（TIGR または RAP）の推定遺伝子構造に基づき新たにプローブセットを再定義する。

5.2 ゲノム多型と表現型多様性を関連づけるための統計データ解析システムの開発

5.2.1 生殖的隔離障壁遺伝子座間相互作用の検出

【参加研究者】

[国立遺伝学研究所] 春島 嘉章，倉田 のり

[統計数理研究所] 藤澤 洋徳，栗木 哲

【研究の背景】

生物学的「種」は「互いに交配可能な自然集団で，他のその様な集団から生殖的に隔離されている」と定義され，生殖的隔離機構は「種」を分ける遺伝的しくみである．生殖的隔離機構の解明の重要性は古くから指摘されているが，生殖的隔離障壁となる遺伝子が同定され，その機構解明が進められている例は動植物を通じ少ない．研究を困難にしている要因は，主に原因遺伝子のマッピング手法がなかったことと，生殖的隔離が相互作用によってもたらされることである．Harushima ら（2001，Genetics，883–892）は，イネ高密度連鎖地図作製に関与し，連鎖地図上のマーカーの遺伝子型分離比が生殖的隔離障壁によって連続的に変化することに着想を得，回帰分析を用いて網羅的に生殖的隔離障壁をマップすることに成功した．つぎに，生殖的隔離をもたらす相互作用の検出を試みている．生殖的隔離をもたらす相互作用は以下 4 種類が考えられる．(1) 対立遺伝子座間の相互作用，(2) 配偶体又は接合体内の異なる遺伝子座間の相互作用，(3) 配偶体又は接合体の遺伝子座と親の遺伝子座との相互作用，(4) 配偶体又は接合体のゲノム遺伝子座と細胞質との相互作用．このうち本研究は (2) に関するものである．

【研究目標】

Harushima らは，イネの F_2 集団（186 個体）において，異なる遺伝子座間で，特定の遺伝子型組合せにより配偶体又は接合体で選抜を受けた可能性があると考え，マーカーの遺伝子型分離の独立性検定を行うことにより，遺伝子座間の相互作用の検出が可能ではないかと考えた．そこで，12 本の染色体上の異なる座位のマーカー間（1055×1055）で，遺伝子型分離の独立性についてカイ 2 乗検定を行った．その結果を図（後述の図 1）にプロットすると，第 9 染色体と第 12 染色体間にみられるカイ 2 乗値 33.6 を最大のピークとして，$p < 0.001$ を示すピークは 27 箇所検出された．しかしながら，この交雑組合せの F_1 稔性は高く，本当に相互作用によって特定の遺伝子型の組合せ個体が失われ，分離の独立性が失われたのかどうかは疑問である．そこで，同じ組合せで別の F_2 集団（約 300 個体）について前の集団でピークを示したマーカー間で遺伝子型の独立性検定の再現性を確かめた．すると，多くのピークの再現性は得られなかった．

本研究では，上記解析手法の統計学的妥当性を検討するとともに，生殖的隔離に関与する遺伝子座間の相互作用を検出する手法の検討を行い，実施可能な手法を開発することを目標とする．

【平成１７年度の成果】

（1）相互作用検出と多重性調整

1）実験データ

最初に Harushima らが本研究に先立って行った，相互作用検出のためのデータ解析について説明する．イネの代表的な品種である Nipponbare（Japonica 種）と Kasalath（Indica 種）を交配した F_1 を自殖して得られた F_2 集団の 186 個体について，マーカー座の遺伝子型がデータとして得られている．イネの染色体 12 本（全長 1650cM）に座数 1055 のマーカーが張られている．マーカーは共優性であるので，遺伝子型としては 3 通り 00, 01, 11 が観測される． マーカー座 i とマーカー座 j の組み合わせについて，観測される遺伝子型の頻度は次の 3×3 表にまとめられる．

表 1．3×3 表

$i \setminus j$	00	01	11
00	$n_{00,00}$	$n_{00,01}$	$n_{00,11}$
01	$n_{01,00}$	$n_{01,01}$	$n_{01,11}$
11	$n_{11,00}$	$n_{11,01}$	$n_{11,11}$

この 3×3 表の独立性のカイ 2 乗統計量（自由度 4）を T_{ij} とおく．T_{ij} の値が大きいとき，座間 i, j に相互作用があると判定することができる． T_{ij} の等高線図を図 1 に示す．

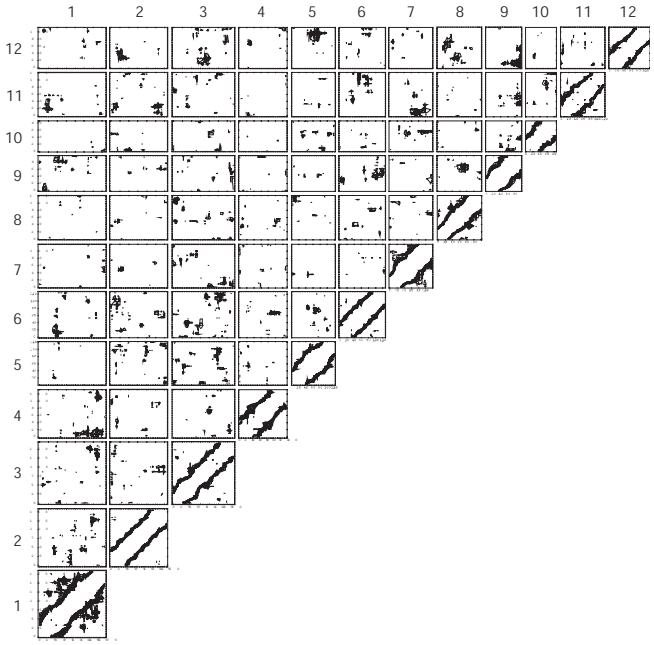


図 1．3×3 表のカイ 2 乗値

先に述べたように，図 1 における最大値は第 9 染色体と第 12 染色体間にみられるカイ 2 乗値 33.6（これは自由度 4 のカイ 2 乗分布で $p = 0.89\times10^{-7}$ に相当）であるが，再現性がなく見せかけのピークの可能性が高いと思われる．

2）統計的問題点 – 多重性調整の必要性

一般に検定を多数繰り返して行う多重検定においては， 見せかけの発見（偽陽性）が多いことが問題となる．例えば，独立な 5%検定を 1000 回行うと，帰無仮説が全て正しい場合であっても平均$1000\times0.05 = 50$ 回の棄却（偽陽性）が生じる．

遺伝子座間相互作用の検定では，マーカー座のペアの総数は $\binom{1055}{2} = \frac{1055\times1054}{2} = 55$万 ペアあり，非常に多数の検定を行うことになる．そのために見せかけの発見に対する配慮，すなわち多重性調整が必須のものとなる．

多重性調整のもっとも簡便な方法はボンフェロニ法である．これは，p 値に検定回数を乗じたものを多重性調整 p 値と定義するものである．例えばいまの場合は$0.89\times10^{-6}\times55$万 $= 0.49$ となる．しかしながらこの方法はしばしば非常に保守的な，実用的ではない結果を与える．より実際的な多重性調整法は，

$$\text{多重性調整 } p \text{ 値} = P(\max_{ij} T_{ij}(\text{確率変数}) > \max_{ij} T_{ij}(\text{実現値}))$$

と定義するものである．この多重性調整 p 値の計算のためには， 相互作用が存在しないという帰無仮説の下での T_{ij} の同時分布が必要である．マーカー座が連鎖するため，各検定統計量 T_{ij} は独立ではないことに注意する．次節でその相関構造を確定する．

3）統計量の相関構造

ある一対の染色体上の m 座のマーカーを考える． x_i , x'_i は 0 または 1 の値をとるものとし，ハプロタイプを

$$X = (x_1,\ldots,x_m) \in \{0,1\}^m, \qquad X' = (x'_1,\ldots,x'_m) \in \{0,1\}^m$$

とする．第 i 座の遺伝子型は $x_i x'_i$ である．

交差がポアソン点過程でおこるという過程のもとでは， X , X' の確率分布は以下の 1〜3 から定まる．

1. X と X' は独立に同じ分布に従う．

2. $P(x_i = 0) = P(x_i = 1) = \frac{1}{2}$

3. 事象 $\{x_i \neq x_{i+1}\}$ は各 i で独立．$P(x_i \neq x_{i+1}) = \frac{1}{2}(1-e^{-2d})$ ．ただし d は座 i と座 $i+1$ の間の遺伝的距離.

最後の性質は，Haldane の地図関数に他ならない．この X, X' の確率構造を出発点として，検定統計量 T_{ij} の相関構造を求めることができる．

表 1 の 3×3 表を 4 つの 2×2 表に分解する．

表 2．3×3 表の分解

$i \setminus j$	00	11
00	*	*
11	*	*

$i \setminus j$	00，11	01
00	*	*
11	*	*

$i \setminus j$	0/0	1/1
00，11	*	*
01	*	*

$i \setminus j$	00，11	01
00，11	*	*
01	*	*

4 つの表に対応するカイ 2 乗検定統計量（自由度 1）を $T_{ij}^{(k)}$ ($k = 1,2,3,4$ ）とおく．

命題 1．i, j を異なる染色体上のマーカー座とする．相互作用が存在しないという帰無仮説の下での T_{ij} の分布について，個体数 n が大きいときに以下がなりたつ.

1. (統計量の分解)

$$T_{ij} \approx T_{ij}^{(1)} + T_{ij}^{(2)} + T_{ij}^{(3)} + T_{ij}^{(4)}$$

この 4 つのコンポーネントは漸近的に独立に分布する.

2. (相関構造)

$$T_{ij}^{(k)} \approx (z_{ij}^{(k)})^2$$

ここで $z_{ij}^{(k)}$ は平均 0，分散 1，相関構造

$$\mathrm{Cov}(z_{ij}^{(k)}, z_{i'j'}^{(k')}) = \exp\{-\rho_k \mid d_i - d_{i'} \mid -\lambda_k \mid \%d_j - \%d_{j'} \mid\} \delta_{kk'}$$

を持つ正規変量である．ただし d_i ， $\%d_j$ は遺伝子座 i, j の染色体上の位置（単位は M)，また $\delta_{kk'} = 1$ （$k = k'$），0 （$k \neq k'$），

$$(\rho_k, \lambda_k) = \begin{cases} (2,2) & (k=1) \\ (2,4) & (k=2) \\ (4,2) & (k=3) \\ (4,4) & (k=4) \end{cases}$$

異なる染色体上に位置する 2 つのマーカー座の遺伝的距離は無限大である．このことから，T_{ij} と $T_{i'j'}$ は，i と i' が異なる染色体上のマーカー座のとき独立である．（i と j'，j と i'，あるいは j と j' が異なる染色体上の場合も同様．）

4）多重性調整 p 値の計算

命題 1 の結果を用いて多重性調整 p 値を計算することができる．具体的には，以下の手順による：

1. $k=1,2,3,4$ の4つの場合について，係数

$$\alpha_i = e^{-\rho_k |d_i - d_{i-1}|}, \qquad \beta_j = e^{-\lambda_k |\tilde{d}_j - \tilde{d}_{j-1}|}$$

を用いて，ホワイトノイズ

$$\varepsilon_{ij} \sim N(0,1) \quad \text{i.i.d.}$$

から，以下の漸化式によって逐次的に乱数 z_{ij} を生成する．

$$\begin{aligned}
z_{11} &= \varepsilon_{11} \\
z_{i1} &= \alpha_i z_{i-1,1} + \sqrt{1-\alpha_i^2}\,\varepsilon_{i1} \\
z_{1j} &= \beta_j z_{1,j-1} + \sqrt{1-\beta_j^2}\,\varepsilon_{1j} \\
z_{ij} &= \alpha_i z_{i-1,j} + \beta_j z_{i,j-1} - \alpha_i \beta_j z_{i-1,j-1} + \sqrt{1-\alpha_i^2}\sqrt{1-\beta_j^2}\,\varepsilon_{ij}
\end{aligned}$$

2. 4 つの場合のそれぞれを $z_{ij}^{(k)}$ $(k=1,2,3,4)$ とおき，

$$W_{ij} = (z_{ij}^{(1)})^2 + (z_{ij}^{(2)})^2 + (z_{ij}^{(3)})^2 + (z_{ij}^{(4)})^2$$

とおく．

3. 上記の乱数生成過程を10000回程度繰り返し，経験分布（ヒストグラム）を描くことによって

$$\text{多重性調整 } p \text{ 値} = P(\max_{ij} W_{ij} > \text{最大カイ2乗値})$$

の近似値を得ることができる．

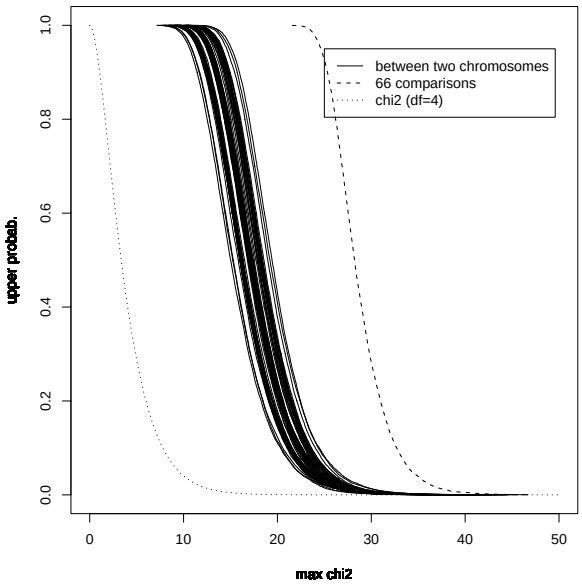


図 2． $\max W_{ij}$ の上側確率

図 2 において，点線は自由度 4 のカイ 2 乗分布の上側確率，66 本の実線は 12 本の染色体の $\binom{12}{2} = 66$ 組のペアのそれぞれについての $\max_{ij} W_{ij}$ の上側確率である．また破線は 66 個の $\max_{ij} W_{ij}$ の全体にわたる最大値の上側確率である．

最後の破線が多重性調整 p 値を与える．最大値 33.6 の調整 p 値は 0.068 であり，5%有意でもないことが分かる．これは，図 1 のピークの最大値が偽陽性であったことと整合している．

なお上の計算には，統計数理研のスーパーコンピュータ（SGI Altix3700）で 14 日 8 時間（延べ時間）必要とした．今後より解析的な計算方法や効率的なアルゴリズムの開発が必要である．

５） 別のアプローチ – コンポーネント毎の解析

カイ 2 乗統計量 T_{ij} を分解して得られる 4 つのコンポーネントは，QTL 解析における 4 つのエピスタシス項（加法効果× 加法効果，加法効果× 優性効果，優性効果× 加法効果，優性効果× 優性効果）と統計的にほとんど同じ構造をもつ．ここでは各コンポーネントを別々に取り扱うことを考える．

各コンポーネントは自由度 1 のカイ 2 乗分布となり，数学的な取り扱いがより容易になるとともに，そのそれぞれは表 1 の 3×3 表の独立性からの乖離の方向を表す情報を持つと考えられる．

とくにマーカーが等間隔という仮想的な状況では，非線形再生理論によって各コンポーネント毎の最大値の分布の近似式が得られる．

命題 2．染色体の一つは全長 L (M)でその上には m 個のマーカーが等間隔 Δ (M)で配置されているとする．もう一方は全長 $\tilde{L}$ (M)でマーカー $\tilde{m}$ 個が等間隔 $\tilde{\Delta}$ (M)で配置されているとする．このとき b が大きいとき，近似式

$$P(\max_{ij} T_{ij}^{(k)} > b^2) \approx 2L\tilde{L} \times \rho_k \lambda_k b^3 \phi(b) \nu(b\sqrt{2\rho_k \Delta}) \nu(b\sqrt{2\lambda_k \tilde{\Delta}})$$

がなりたつ．ここで $\phi(\cdot)$ は標準正規分布の密度関数，また $\Phi(\cdot)$ を標準正規分布の分布関数とするとき

$$\nu(x) = 2x^{-2} \exp\{-2\sum_{n=1}^{\infty} n^{-1}\Phi(-\frac{1}{2}x\sqrt{n})\} \qquad (x > 0)$$

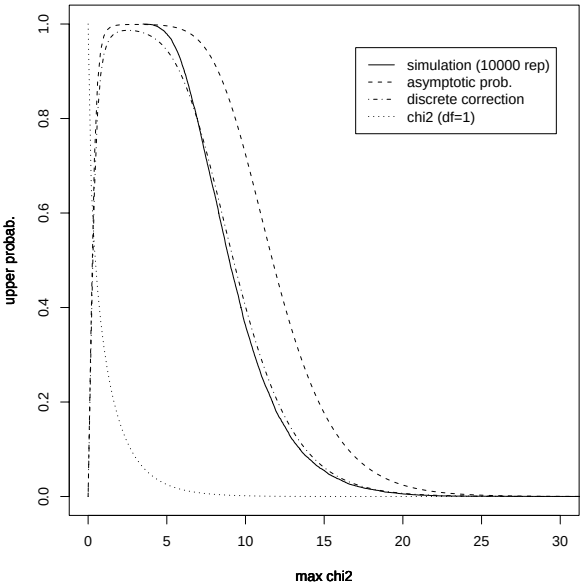


図 3．自由度 1 のコンポーネントの上側確率（マーカーが等間隔の場合)

図 3 は，染色体長が $L = \tilde{L} = 100$ cM，マーカー間隔は $\Delta = \tilde{\Delta} = 10$ cM，また $\rho_k = \lambda_k = 2$ の場合である．図において，一点鎖線が命題 2 による近似式，実線が乱数シミュレーションによる経験分布である．近似精度はかなり良いことが分かる．

現実にはマーカーが等間隔ということは考えられない. しかしながらマーカーが不等間隔の場合であっても，マーカー間隔の平均値を Δ あるいは $\tilde{\Delta}$ とおき命題 2 の近似式を用いると，よい近似式を与えることが別の数値実験によって確認されている．

【次年度以降の研究】

本年度研究においては，遺伝子座間検出のためのカイ 2 乗検定統計量の相関構造を確定し，多重性調整 p 値を計算する方法を与えた．

Harushima らがイネ F_2 集団 186 個体で計算した相互作用の検出のための独立性カイ 2 乗統計量の最大値 33.6 は，多重性調整 p 値では 5%有意ではなく，それが見せかけのものであることが裏付けられた．

一般に分割表の独立性検定は，その自由度が高くなればなるほど，モデルからの様々な乖離を検出できるが，一方で検出力が低くなることがよく知られている．今回，統計的方法によって有意なピークを検出できなかったのは，自由度が 4 のカイ 2 乗検定を用いたことによると思われる．次年度の研究方針としては，生殖的隔離障壁を具体的にモデル化しそのモデルの下で検出力の高い検定方法を用いることが考える．統計量 T_{ij} の 4 つのコンポーネント $T_{ij}^{(k)}$ のそれぞれは，生殖的隔離障壁のいくつかのモデルの下で高い検出力を与える統計量であることが想定されるが，それ以外にもいろいろな統計量の可能性がある．(1) 生殖的隔離障壁の適切なモデル化，(2) そのモデルの下で高い検出力を与える検定統計量の構成，ならびにその (3) 多重性調整の方法の開発，が次の課題である．

６．平成１７年度の研究成果

（１）知見・成果物・知的財産権等

知見は、進捗状況に記載した。成果物、知的財産権統については、平成１７年度は該当するものは無い。

（２）成果発表及び著書執筆等

平成１７年８月１９日、統計数理研究所において平成１７年度　新領域融合研究プロジェクト「生物多様性解析」第１回ワークショップを開催した。

平成１７年度　新領域融合研究プロジェクト
「生物多様性解析」第１回ワークショップ

日時：平成１７年８月１９日（金曜日）午後１時半～午後５時
会場：統計数理研究所 研修室（新館２Ｆ）

プログラム

第一部　分担課題担当者からの話題提供
13:30-14:00「生物多様性解析プロジェクトのめざすもの」
　　城石　俊彦
14:00-14:20「マウスにおける行動表現型とその関連遺伝子解析」
　　小出　剛
14:20-14:40「点過程モデルによるマウス行動周期性の解析」

川崎 能典
14:40-15:00「方向相互作用によるマウス社会行動の統計解析（中間報告)」
種村　正美
15:00-15:30　休憩
15:30-15:50「イネ多様性研究におけるゲノム複合データの融合的解析」
倉田　のり
15:50-16:10「遺伝子座間相互作用の検出における多重性調整」
栗木　哲
16:10-16:30「マイクロアレイデータと生物多様性」
舘野　義男
16:30-16:50「遺伝研のマイクロアレイデータの統計的な問題点」
江口 真透

第二部　　個別検討会
17:00-18:00