表現型を始めとする機能情報の解析技術

●森下 真一
東京大学大学院新領域創成科学研究科

＜研究の目的と進め方＞

ゲノムや遺伝子を操作した結果得られる変異体の表現型が、野生型の表現型と比べてどのように変化しているかを判断する作業は、研究者の主観に委ねられることが多い。微小な変化を同定することは難しく、たとえ同定できたとしても、どれだけ他の変化と類似しているか否かについては、主観的になりがちであり、研究者の経験に左右される。さらに進んで変化の類似性に基づいて表現型をグループ分けしたとしても、類似性の精度が不正確であるならば、そのグループがどのような機能を表現しているかについて正しい結論を導くことは困難になる。したがって、表現型を精密に定量化する視点が大切である。このような動機から、本研究では、遺伝子操作によって引き起こされる変異体の表現型を定量化することを研究する。さらに定量化した表現型と遺伝子型の相関をより正確に描出するには、遺伝子型の測定精度も高くなければならない。そこで近い将来に新型シークセンサーが産出する大量のゲノムおよび遺伝子の配列断片情報を高速処理する技術についても取組む。さらに siRNA/dsRNA 配列、ゲノミック PCR プライマの設計精度を上げる技術についても研究する。

＜ 2007 年度の研究の当初計画＞

(1) 数万点に及ぶショウジョウバエの翅の画像から、細胞の大きさがブロックごとにどのように変化するかを追跡することにより細胞増殖に関与する遺伝子、たとえばインスリンシグナル伝達経路に関与する様々な遺伝子の機能を推定する計画である。2005 および 2006 年度の研究で、画像処理ソフトウエアの精度は十分に向上したため、野生型と比較して遺伝子を強制発現もしくは RNAi により阻害した変異体の変化を統計的に検定し、異常値の分布に従ってグループ化し、さらにクラス分類アルゴリズムによる機能推定を行う（相垣班員との共同研究）。(2) 表現型の画像解析については、ジョウジョウバエだけでなく出芽酵母の遺伝子破壊株の画像を解析して機能を推定するフェノーム研究を 2001 年以降取り組んできた。2006 年度までにすべての非必須遺伝子の破壊株の画像解析と遺伝子機能解析が終わり、必須遺伝子を破壊したときの表現型を計測する基盤ができあがった。2007 年度はこれらの画像データを解析し、必須遺伝子のフェノーム解析を行う計画である。(3) 出芽酵母、ショウジョウバエ、メダカ、ヒトを対象に遺伝子転写開始点周辺のゲノム領域を、エピジェネティクス的解析を継続する。新型シークエンサが産出する大量の配列断片を利用して、転写開始点、ゲノム上のヌクレオソーム出現確率、ゲノムのメチル化、ヒストンのアセチル化、ゲノムの塩基変異率、遺伝子発現量を精度良く推定する効率的ソフトウエアを研究開発し、これらの情報の相関関係を解析する。

＜ 2007 年度の成果＞

ショウジョウバエおよび出芽酵母の変異体画像解析については、結果的には順調に計画どおり進んだが、その途中ではいくつもの技術的問題点が発生した。ジョウジョウバエの翅の細胞数の計測で最も解決に時間を要したのが、透明な翅の裏側に生えている細毛が表から透けて見えてしまう問題であった。人間の専門家が見ても判定が困難な場合も多いのではないかという疑問もわいてきた。そこで翅の 16 個の異なるブロック（細胞数は 100 から 1000 個程度）で 4 人の専門家による計測を実施したところ、専門家の間でも最大 20% 程度の食い違いが生じることがわかった。一方我々が研究開発したソフトウエアは、どの例でも専門家の計測した幅の範囲内で細胞数を勘定することができた。したがってソフトウエアの精度はほぼ満足できるレベルに達したと考えている。

このように細胞数計測のソフトウエアは完成したが、その結果、翅の画像を再度撮影し直すという軌道修正が必要となった。プロジェクトを開始した当時は画像を格納する二次記憶装置が高価であった為に、画像を jpeg 形式で圧縮して保存することとした。そのため認識する対象も翅脈のブロックの大きさ等のマクロな情報に限定されていた。ところが研究を進めてきた 2 年半の間に二次記憶装置の価格が急落し、研究予算の範囲で大容量のデータ収集が可能になってきた。そのため TIFF による高解像度のデータ（従来に比べて約 100 倍のサイズ）を蓄積することが可能になった。この結果、細胞数の計測というミクロな情報まで観測できるようになり研究範囲を広げることが可能になった。しかしながら従来のデータは jpeg でしか撮影していなかったため、すべての翅画像を TIFF で撮影しなおすこととした。さらに強制発現系の画像だけでなく、上田龍班員が作成してきている RNAi 変異体も多数蓄積されてきたため、こちらの画像も情報処理することとした。

出芽酵母の変異体画像解析では、画像解析ソフトウエアの修正が必要になった。従来は、1 倍体を対象に画像ソフトを研究開発してきた。一方、必須遺伝子の破壊株を作成する際には、1 倍体の必須遺伝子を破壊すると致死的であるため、2 倍体を利用し、ハプロタイプの一方の必須遺伝子だけを破壊して表現型を観測するというアプローチを試みている。しかし 2 倍体の場合、1 倍体と違って出芽後の跡が細胞壁上に凹みとして残り、この凹みが誤って小さな芽として認識される問題が発生した。この問題の解決のため、画像処理ソフトウエアを作り直す作業を行った。

以上のように変異体画像処理ソフトウエアの研究開発と表現型の解析ではいくつかの問題点があったものの、研究計画の 4 年目に当たる来年度には雑誌へ研究成果を投稿できるような当初の計画通りに進んでいる。

メダカをモデル生物として発展させるためのゲノム解読および解析は 2006 年度に終了し、2007 年に関連論文を *Nature, Genome Research, Nucleic Acids Research* に報告した。過去 6 億年にわたる脊椎動物ゲノム進化を分析し、染色体再編成の様子を初めて再現したという評価を幸いにも受けた。メダカゲノム解読を礎に、新型シークエンサによる 5’ end タグ収集と、転写開始点解析、塩基レベルの進化解析と表現型への影響を考察する段階へと進むことができた。 哺乳類から線虫、イネ、ハエに至る幅広い生物種のための遺伝子阻害配列設計のサーバー siDirect(siRNA 用) および dsCheck(dsRNA 用) は幅広いユーザに利用されている。

さらに 2007 年度は、新型シークセンサー (Solexa/Illumina, ABI/SOLiD) の登場が研究を大きく進展させた。橋本真一博士と

共同研究している 5' SAGE 法は新型シークエンサの利用により大きく進展し、5' end の短いタグ (28-36nt) を大規模に収集することが可能になった（実験あたり数千万タグ)。1 週間あたりの収集量は数百倍になった。そのため 2007 年半ばより、ヒト、カイコ、ショウジョウバエ、メダカからタグを収集している。さらにメチル化等のゲノム修飾が転写開始点や転写量に与える影響を分析し、同様の解析を受精後の初期胚のステージで行いつつある。また新型シークセンサーを利用することでヌクレオソーム構造の分析が免疫沈降法との組み合わせで可能になりつつある。ヌクレオソーム構造が転写開始点周辺の遺伝子の機能および進化に与える影響を、メダカをモデル生物として解析した。その結果「転写開始点下流では遺伝子変異率が約 200bp 毎に周期的に変化しヌクレオソームのリンカー位置と同期している現象」を見出した。

＜国内外での成果の位置づけ＞

本研究が提案する網羅的遺伝子破壊／阻害による表現型変化をイメージ処理技術による定量化する試みは、研究代表者が世界に先駆けて取組んでいる研究テーマであり、類のない独創的なバイオインフォマティクス研究である。既に出芽酵母を対象にした研究では、出芽酵母遺伝子破壊株の形態パラメータから破壊した遺伝子機能を推定する生物学的知見と規則があたらしく得られている。本アプローチを他のモデル生物に展開し、新しい研究の方法論を定着させたい。

生体内の様々な部品が組合わさり全体として有機的なシステムとして振舞うという視座から生命現象を理解しようという生命システム情報の研究アプローチの中では、生命現象におこる変化を定量的に観測し、そこからシステムのモデルを立てて検証することが鍵となる。したがって生命システム情報の研究推進にも重要なアプローチとなっている。特に単細胞の真核生物である出芽酵母について、5000 個の遺伝子破壊株の顕微鏡画像処理技術を研究開発し、出芽酵母の表現型を定量化するためのパラメータを決め、遺伝子型との対応関係を推定することにある程度成功している。本研究ではこれらの成果を土台に、より複雑な多細胞のモデル生物としてメダカ・ショウジョウバエ・分裂酵母の表現型を解析し、さらにはヒトの臓器や腫瘍の表現型を定量化することに取組むことを目標としている。そのためには、領域に参加する研究者との行動研究が不可欠である。ショウジョウバエに関しては相垣班員、上田龍班員、メダカは武田班員、絶対定量 PCR プライマは伊藤隆司班員、5' end タグ解析は橋本班員、全長 cDNA ショットガンシークエンシングは菅野班員、鈴木班員と共同研究している。

＜達成できなかったこと、予想外の困難、その理由＞

研究計画はほぼ達成され、あたらしい研究テーマにも取り組むことができたものの、いくつかの予想外の困難な問題もあり解決に時間がかかった。前にも述べたが、透明な翅の裏側に生えている細毛が表から透けて見えてしまう問題のため、ショウジョウバエの翅の細胞数の計測が予想外に難しかった。最終的に人間の専門家の意見のばらつきの範囲内にソフトウエアによる推定がおさまることができた。また翅画像の解像度が低いことも分かり、高解像度の TIFF 画像の収集に方針も変更した。新型シークエンサの情報処理のための計算機資源の整備と、高速なソフトウエアパイプラインの研究開発も容易ではなかった。

＜今後の課題＞

画像解析については、細胞数の計測精度は実用に耐えうるようになった。若干の性能の努力を行い、ソフトウエアと画像データベースの公開を進めたい。同一遺伝子の強制発現と RNAi を組み合わせた機能予測で成果を収めたい。

新型シークセンサーでは、5' end タグ解析に加えて、菅野班員・鈴木班員と全長 cDNA ショットガン法のソフトウエア開発、ヒトゲノムの疾患関連領域 (1-10Mb) を効率的に re-sequencing する方法を highly specific multiplex genomic PCR primers を併用して開発したい。情報支援班を通じて班員の方々に利用していただきたい。

新型シークエンサが出力するデータ量は従来のそれより 2 桁ぐらい上である。大きな二次記憶装置での格納が必要になる。さらにゲノムへのアラインメントはタグの QV が必ずしも高くないことを考慮して時間がかかっても丁寧に行うことが必要になる。クラスター型並列計算機に均等にジョブを上手に割り当て、大きなデータの転送がボトルネックにならないようなソフトウエアの研究が最も重要なテーマになると考えている。

ヌクレオソーム構造解析については免疫沈降法を利用したヌクレオソーム位置の確定方法でスタンフォード大学と共同研究を開始した。エピゲノムが表現型に与える影響も研究してゆきたい。その際には、ヌクレオソーム位置の推定のためには、どのぐらいタグを読めばよいのか？ 偽陽性データを効率的に除去するための方法は何か？　これらの問題をクリアすることを目標に研究を進めたい。

＜成果公表リスト＞

1) 論文／プロシーディング（査読付きのものに限る）

1. 0801270031　Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y, Jindo T, Kobayashi D, Shimada A, Toyoda A, Kuroki Y, Fujiyama A, Sasaki T, Shimizu A, Asakawa S, Shimizu N, Hashimoto S, Yang J, Lee Y, Matsushima K, Sugano S, Sakaizumi M, Narita T, Ohishi K, Haga S, Ohta F, Nomoto H, Nogata K, Morishita T, Endo T, Shin-I T, Takeda H*, Morishita S*, Kohara Y*. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447, 714-719 (2007) * corresponding
2. 0801270040　Nakatani Y, Takeda H, Kohara Y, Morishita S*. Reconstruction of the Vertebrate Ancestral Genome Reveals Dynamic Genome Reorganization in Early Vertebrates. *Genome Research* 17(9): 1254-1265 (2007)
3. 0801270044　Ahsan B, Kobayashi D, Yamada T, Kasahara M, Sasaki S, Saito TL, Nagayasu Y, Doi K, Nakatani Y, Qu W, Jindo T, Shimada A, Naruse K, Toyoda A, Kuroki Y, Fujiyama A, Sasaki T, Shimizu A, Asakawa S, Shimizu N, Hashimoto S, Yang J, Lee Y, Matsushima K, Sugano S, Sakaizumi M, Narita T, Ohishi K, Haga S, Ohta F, Nomoto H, Nogata K, Morishita T, Endo T, Shin-I T, Takeda H, Kohara Y, Morishita S*. UTGB/medaka: genomic resource database for medaka biology. *Nucleic Acids Res.* 36(Database issue): D747-52 (2008)

2) データベース／ソフトウエア

1. 0507070222 メダカゲノムブラウザー
   http://medaka.utgenome.org/
2. 0612221516 出芽酵母ゲノムブラウザー
   http://yeast.utgenome.org/
3. 0507070216　siRNA 設計
   http://design.rnai.jp/
4. 0507070220　出芽酵母表現型解析
   http://scmd.gi.k.u-tokyo.ac.jp/
5. 0606191356　Multiplex Genomic PCR design
   http://ps.cb.k.u-tokyo.ac.jp/