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 情報生命科学専攻

＜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を判断すること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、類似性の精度が低いと、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、遺伝子型の測定も高精度であることが必要となる。そこで超高速 DNA 解読装置 (Solexa, SOLiD) を活用して、転写開始点の網羅的収集、トランスクリプトーム全体の絶対定量化、全長 cDNA 配列の廉価で高速な決定法、ヌクレオソーム構造の描出等について取り組む。

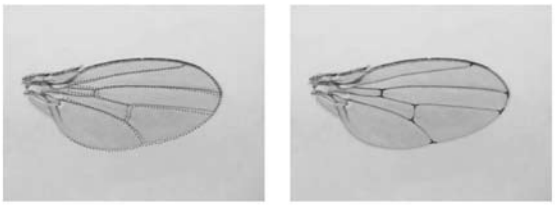
＜ 2008 年度の研究の当初計画＞

(1) 表現型の解析： 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研究を継続する (相垣班員との共同研究)。2005-7 年度の研究で構築した画像処理ソフトウエアは、野生型からの微小な変化を正確に定量化するのには問題がなく、90% 前後の翅脈の判定は既に可能になっている。しかし翅脈の喪失、ノッチの認識等の顕著な表現型の変化は、野生型の表現型に比べて大きすぎるため定量化は困難である。微小な変化の同定を目指した研究ではあったが、完全自動化を達成するための画像処理機能を改善してゆきたい。

(2) 遺伝子型の解析： 超高速 DNA 解読装置 (Solexa, SOLiD) は 25 ～ 50 塩基程度の短い配列を 1 週間程度の短期間に数千万配列も出力することが可能であり、1 日当たりの塩基産出量は 3-5 億塩基にも及ぶ。この利点はしばしば強調されるものの一方で、塩基読取エラー率は高く、しかも配列長が短いため、正確な結果を導くには工夫が必要である。たとえばエラーを除くために、配列の精度の高いゲノム上に短い配列をアラインメントして補正することが頻繁に行われる。この際、塩基読取エラーが後半に片寄る性質を考慮して塩基長とミスマッチの許容範囲パラメータを実験ごとに微調整する必要がある。また配列長が長くないと解きにくいゲノムアセンブリ等の問題は避け、大量の短い配列を活用できる応用例を慎重に選ぶ必要もある。そこで我々は、転写開始点の網羅的収集、遺伝子発現量の絶対定量 (橋本班員と共同)、全長 cDNA 配列の廉価で高速な決定法 (菅野・鈴木班員)、ヌクレオソーム構造の描出 (武田班員)、DNA メチル化 (伊藤班員) に取り組む計画である。

＜ 2008 年度の成果＞

(1) 表現型の解析： 微小な変化を正確に定量化する場合には問題は少なくなくなっており、90% 前後の変異体の翅脈を判定できる。非公開ではあるが翅脈の画像解析データベースを構築し、遺伝子機能解析に活用している。一方難問として残った、翅脈の喪失、ノッチの認識等の顕著な変化を示す表現型を認識できるようにするため、新しい画像解析方法を考案した。詳しくは野生型翅脈モデルを変異体翅脈へ変形する際に、Hausdorff 距離を最適化するアファイン変換を、幾何学的分岐限定方法を用いて高速に計算できるようにした (global matching, 下図左)。さらに翅脈の交差点を微調整して正確に位置合わせするための技法 (local matching, 下図右) も合わせて考案した。その結果、ソフトウエアのロバストネスは各段に向上し、ほぼ完全自動化を達成することができた（論文投稿準備中)。



global matching（左） と local matching （右）

(2) 遺伝子型の解析（超高速 DNA 解読装置を利用した成果)：

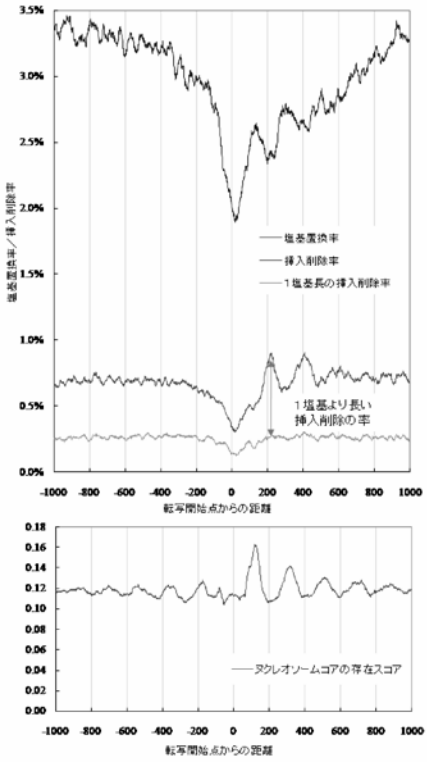
(2-a) 塩基エラー率が高い場合には、ミスマッチを許して配列をゲノムにアラインメントすると、位置を誤認識する率が高くなる。そこでアラインメント前に短いタグをクラスタリングし、各クラスターから塩基エラーがない（もしくは少ない）高頻度の代表配列を選択し、その代表配列をゲノム上にアラインメントする方法を考案した。転写開始点タグや small RNA タグ中の読み取りミスを補完し、全タグの 5% 程度のタグを正確な位置へとアラインメントできるようになった（論文投稿中)。

(2-b) 橋本班員が構築した mRNA の 5’ 端 27 塩基を収集する方法では 1 つのサンプルから Solexa の 1 レーンを使って約 500 万タグを約 3 日間で収集できる能力がある。しかも、同じサンプルを独立に観測しても、同じ配列が観測される回数の再現性は著しく高く (相関係数 0.99 以上)、ダイナミックレンジは 3-4 桁程度もある（論文番号 0901111023)。ハエおよびメダカの異なるサンプル (胚等) 間の違いを調べたが、転写開始点周辺では 1 塩基レベルで分布が正確に保存される傾向が顕著であった（論文番号 0901131436)。さらに、後に述べるメダカ初期胚を使った転写開始点下流における DNA クロマチン構造とゲノム変異の周期的相関を発見するためにも活用され（論文番号 0901131406)、解読したカイコゲノムの遺伝子アノテーションにも利用された（論文番号 0901131410)。橋本班員との共同研究。

(2-c) 超高速 DNA 解読装置を使って全長 cDNA 配列を解読する手法を研究開発した。再構築の精度は 99% 以上、コストは cDNA あたり約 1 ドル (Solexa の試料代) となり、実用に供するようになった（論文投稿中)。菅野・鈴木班員との共同研究。

(2-d) DNA 配列の多様性は、生殖細胞系列における遺伝子の働きやクロマチン構造を反映しているのであろうか？ という疑問に答えるため、2 系統のメダカ（*Oryzias latipes* の Hd-rR と HNI 系統）のゲノム配列を比較し多様性を描出した。さらに Hd-rR 系統の胞胚から Solexa により得た約 3730 万個のヌクレオソームコアのゲノム上の位置を同定し、6 段階の胚形成期における代表的な転写開始点 11,654 箇所周辺で分析した。その結果、転写開始

点下流において、DNA変異率が約200塩基対（bp）の周期で変化することを観察した（下図参照）。具体的には、1bpよりも長い挿入削除率は、転写開始点からの距離がおよそ+200bp、+400bp、+600bpの位置で最大となる一方で、点突然変異率はこれらの位置で最小になっていた。この約200bpの周期性はクロマチン構造と相関しており、ヌクレオソームコアが存在している率は0bp、+200bp、+400bp、および+600bpの位置で最も低くなっていた。これらのデータは、進化過程において、遺伝子の働き（転写）やクロマチン構造が、DNA配列の形成に寄与する可能性があることを例示している(論文番号　0901131406)。Andrew Fire博士、武田洋幸班員、橋本班員、菅野・鈴木班員、小原班員との共同研究。



＜国内外での成果の位置づけ＞

ショウジョウバエ翅脈画像解析ソフトウエアの研究開発は世界的に類似研究がなくユニークである。一方、超高速ＤＮＡ解読装置を活用した研究は世界的に発展が目覚ましく競争の激しい分野である。きびしい状況であるが、(2-a) ～ (2-d) の研究項目はどれも国際的にリードしている。

＜達成できなかったこと、予想外の困難、その理由＞

超高速ＤＮＡ解読装置周辺の研究競争は激しい。着想したアイデアを素早く検証し、論文として発表するまでの時間をできるだけ短くすることが望ましいが、(2-a) ～ (2-c) の研究テーマは、今から考えると研究過程で試行錯誤が多く、予想外の問題を解決するために時間を費やした。

超高速ＤＮＡ解読装置が市場に出回った当初の2006年夏ごろ、出力する25-36塩基の短い配列の品質は低く、QV値から推定されるエラー率に比べ現実のエラー率は遙かに大きかった。そのためアドホックにエラーを除去する作業に時間が取られ、エラー除去そのものも研究テーマとした（研究項目 2-a)。ところが研究過程で、超高速ＤＮＡ解読装置のエラー率は徐々に低減し、解読可能な塩基配列長も25塩基から50塩基近くまで伸び（2008年末頃)、エラー除去は楽になっていった。この結果、研究テーマ (2-b), (2-c) で作成したアドホックなエラー処理方法は整理することとなった。このように急速な技術革新と並行して研究する場合、技術改善状況に注意を払いながら、研究方針の舵をタイムリーに切り直すことが多くなる。論文発表までの時間をもう少し短縮できたのではないかと感じる。また、計算機資源が不足し計算が律速条件となった時期があった。東京大学情報基盤センターに導入された並列計算機システムを利用し、研究計画の遅延を補うことができた。

＜今後の課題＞

ショウジョウバエ翅脈の画像解析は平成２１年度中にはまとめデータベースを公開して研究を締めくくりたいと考えている。DNA多様性とクロマチン構造の関係は反響があった（論文番号　0901131406)。現在その周辺の問題、たとえば、(1) どのような配列モチーフがクロマチン構造の安定性に寄与しているのか？　(2) ＤＮＡメチル化等はクロマチン構造にどの程度影響し転写量を制御しているのか？　(3) 転写量が複数の遺伝子を含む領域で同時に増加もしくは減少する現象が観測されるがクロマチン構造の構造変化はどのていど関与しているか？　等をメダカ、線虫、ヒトを対象に考察している。

＜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）

0901131406 Sasaki S, Mello C, Shimada A, Nakatani Y, Hashimoto S, Ogawa M, Matsushima K, Gu S G, Kasahara M, Ahsan B, Sasaki A, Saito T, Suzuki Y, Sugano S, Kohara Y, Takeda H, Fire A, Morishita S*. Chromatin-Associated Periodicity in Genetic Variation Downstream of Transcriptional Start Sites. *Science*, (Published Online December 11, 2008)

0901131436 Ahsan B, Saito T, Hashimoto S, Muramatsu K, Tsuda M, Sasaki A, Matsushima K, Aigaki T, and Morishita S*. MachiBase: a Drosophila melanogaster 5'-end mRNA transcription database. *Nucleic Acids Research*, Vol. 37, Database issue D49-D53 (2009)

0901111023 Hashimoto S, Qu W, Ahsan B, Ogoshi K, Sasaki A, Nakatani Y, Lee Y, Ogawa M, Ametani A, Suzuki Y, Sugano S, Lee C C, Nutter R C, Morishita S, Matsushima K. High-resolution analysis of the 5'-end transcriptome using a next generation DNA sequencer. *PLoS One* 4(1):e4108. Epub (2009)

0901131410 The International Silkworm Genome Consortium (Morishita S is one of the corresponding authors). The genome of a lepidopteran model insect, the silkworm Bombyx mori. *Insect Biochemistry and Molecular Biology*, (in press)

0801270044 Ahsan B, (36 authors), Morishita S*. UTGB/medaka: genomic resource database for medaka biology. *Nucleic Acids Res.* 36 (Database issue): D747-52 (2008)

0806251650 Nakatani Y and Morishita S*. Vertebrate genome evolution examined by comparing the human and fish genomes. *Encyclopedia of Life Sciences*, John Wiley & Sons (2008)

2) データベース／ソフトウエア

0901131451 ショウジョウバエゲノムの転写開始点ブラウザー
http://machibase.gi.k.u-tokyo.ac.jp/

0507070222 メダカゲノムブラウザー
http://medaka.utgenome.org/

0612221516 出芽酵母ゲノムブラウザー
http://yeast.utgenome.org/

0507070216　siRNA設計　http://design.rnai.jp/

0507070220　出芽酵母表現型解析
http://scmd.gi.k.u-tokyo.ac.jp/

0606191356 Multiplex Genomic PCR design
http://ps.cb.k.u-tokyo.ac.jp/

論文投稿中のために非公開(未登録)のサーバー
ショウジョウバエ翅脈画像サーバー、UTGBゲノムブラウザー
(ショウジョウバエ, ヒト, マウス、カイコの5-endタグ情報)