計画研究：2005 ～ 2009 年度

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、同定できたとしても、他の変化とどれだけ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。

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 (Illumina GA, ABI 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も取り組む。

＜研究開始時の研究計画＞

ゲノム配列が解読されたモデル生物におけるポストゲノム研究のアプローチのひとつとして、遺伝子コード領域の破壊／ RNAi による mRNA の働きの阻害／遺伝子を組織特異的に強制発現させるベクターのゲノムへの挿入等により遺伝子発現を操作し、その結果生じる表現型の変化を正確に計測し、類似した変化を生む遺伝子群を同定し、遺伝子の機能の解明に結びつけることが重要であると我々は考えている。このようなアプローチを実現し加速するには、さまざまな要素技術が生物系とバイオインフォマティクス系の共同研究から生み出されることが望ましい。そこで本プロジェクトでは、RNAi 用配列を設計評価し、遺伝子発現の絶対定量化し、表現型変化の絶対定量化し、取得された情報から遺伝子機能を予測するソフトウエアの研究を行う計画である。

表現型変化の絶対定量化に関しては、JST BIRD プロジェクトのサポートにより平成 13 年度より研究を開始した。平成 16 年 12 月現在、出芽酵母の非必須遺伝子破壊 4780 株すべてについて各々 200 個以上の細胞の計測を完了し、約 500 個の安定した形態パラメータを取得し、遺伝子機能の推定に役立てている。この成果を礎に、平成 16 年度より相垣研究室で収集しているショウジョウバエ変異体（遺伝子強制発現による）の翅脈画像を処理するソフトウエアの研究に取組んでいる。135個の形態パラメータのデータを取得して、野生株と変異体の間で有意な形態変化が起こっているか調査したところ、変異体の 2/3 以上において統計学的に極めて有意な変化が認められた（有意水準 0.002% の両側検定)。形態が極端に変化しない非必須遺伝子の破壊や強制発現によりえられた変異体を観測する限りにおいては、形態パラメータの再現性は高く、安定して取得することができる生物学的パラメータになりうる。そこで様々な形状をした変異体の翅脈画像から、ロバストに翅脈を画像認識するソフトウエアの構築に取組む計画である。

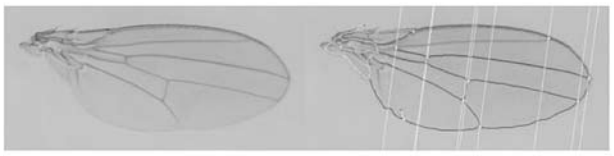
一方 遺伝子型の研究については、ヒト／マウスの遺伝子を効率的に阻害するための siRNA 配列設計法の研究を平成 15 年度より開始した。siRNA 配列は短いのでオフターゲット効果を十分考慮に入れ設計すべきであるが、遺伝子コード領域全域に対してオフターゲット効果の探索をおこなうため、現実に実行するには高速アルゴリズムが不可欠である。従来最も高速とされたのは平成 15 年に提案された Sung らの方法であるが、われわれのアルゴリズムのスピードは、Sung らに比べて一桁高速である。阻害率が高くオフターゲットの遺伝子配列に対する阻害効果が殆どない設計をおこなう高速なアルゴリズムを研究開発した。このような経験をもとに、DNA 解読速度の進展にともなって解決可能になる様々な問題に対するアルゴリズムを研究開発する計画である。

＜研究期間の成果＞

表現型の解析：ショウジョウバエの変異体画像解析ではいくつもの予想外の技術的問題が発生し、相垣研究室と共同で、その都度解決してきた。以下順を追って成果を報告する。

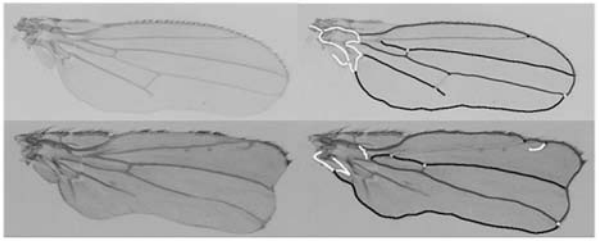
ショウジョウバエ翅脈の認識

これは、下図左のような画像から、翅脈が交わる点を認識し、翅脈を追跡して長さを計測し、そして翅脈により区画化された各ブロックの面積を定量化する問題である。



左図の翅脈を右図のように認識し翅脈長と面積を計測

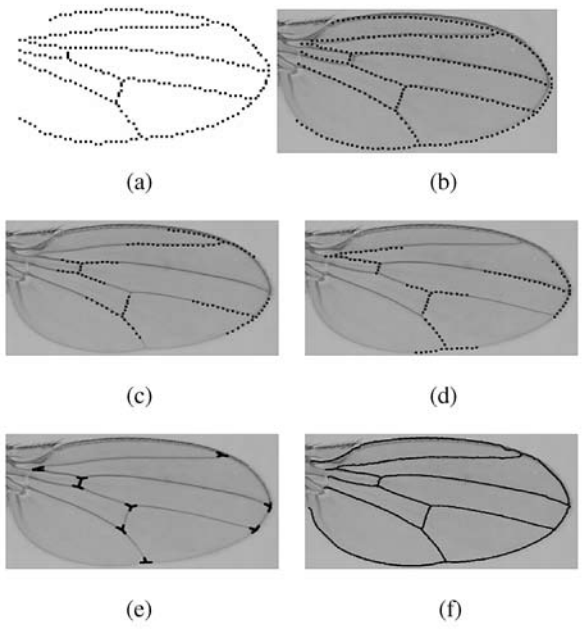
上には認識が比較的容易な例を示したが、変異体の翅脈は多様であり下図のような難しい例もある。



異常な形状を示す変異体の翅脈

異常な形状の翅脈はさまざまであり、すべてを定量化することの困難さは上の例からも明らかである。しかし、認識できる翅脈をできるだけ広げられることを目指し、異常形態にロバストな画像認識アルゴリズムの構築に腐心した。アルゴリズムは、(1) 翅の外周の認識 , (2) 翅脈と交差点の feature の抽出 (3) 翅脈の標準的なモデルをあてはめる操作 model matching の３つのステップに分かれる。詳しい説明は登録番号 0910131130 に報告したので、ここでは重要なステップ (3) に絞って説明する。

下図で (a) は標準的な翅脈のパターンを示している。このモデルと与えられた画像間のハウスドルフ距離が最小になるようにモデルをアフィン変換した結果が (b) である。距離が最小になるアフィン変換を見つけることは単純ではない。最適な変換を求めるために計算機科学で愛用されている分岐限定法を利用している。幾何学的な変換を求めるので geometric branch-and-bound (GBB) と呼ぶ。このあてはめ操作は画像全体に対して大域的に最適化しているために、個々の翅脈の交差点にモデルを完全にフィットさせることは難しい (図 c)。そこで交差点の周辺に限定して、再度 GBB を施したのが (d) であり、さらに局所的に GBB を実行した結果が (e) である。きめ細やかなチューニングにより翅脈の交差点は高い精度で確定できるようになった。交差点間を線で接続することで翅脈を (f) のように確定する。



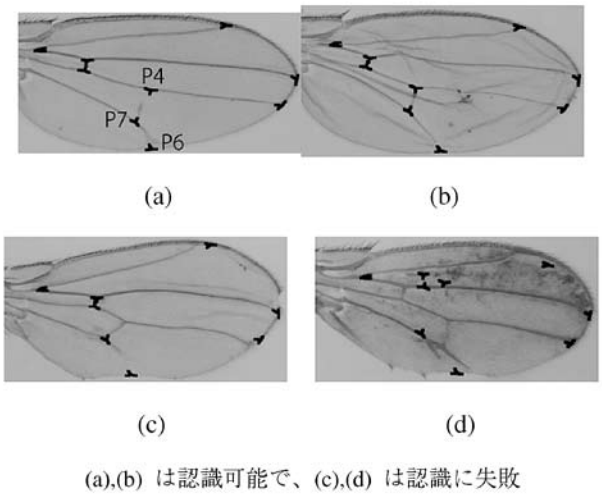
翅脈の標準的なモデルをあてはめる操作 model matching
点間を線で接続することで翅脈を (f) のように確定する。

翅脈認識の精度

翅脈認識アルゴリズムの精度は、モデルの形状が変異体画像群とどれだけ近いかどうかに依存する。モデルからかけ離れた形状の変異体にモデルをあてはめることは所詮難しい。それでは、今回利用したモデルを使って、一体どれぐらいの変異体の翅脈画像を正確に認識できたか？　遺伝子を過剰発現させた変異体 5000 個の翅脈画像（833 x 622 pixels）を認識させたところ 96.2% はモデルにあてはまる翅脈パターンをしていた。詳しくは、モデルと実画像の翅脈交差点の距離がすべて 10 pixel 以内のとき、正しく画像を認識したと判定した。この認識率は変異体の翅脈画像を分析した時に期待できる確度をおおよそ示している。翅脈認識アルゴリズムのロバスト性をより詳しく調べるために、100 個の野生型の翅脈画像、650 個のモデルに近い形状をした変異体の翅脈画像、250 個のモデルを逸脱した形状の変異体の翅脈画像に対してアルゴリズムを適用した結果を次の表に示す。

	画像数	認識数	認識率
野生型	100	100	100.0%
モデルに近い形状をした変異体	650	629	96.8%
モデルを逸脱した形状の変異体	250	157	62.8%

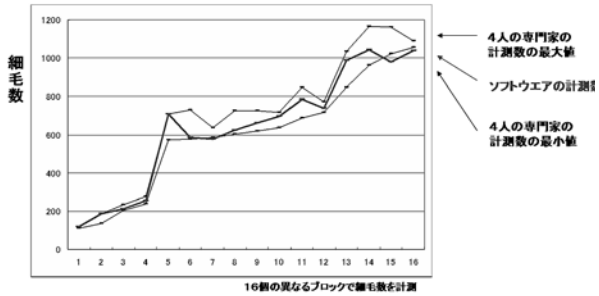
このようにモデルを逸脱した形状の変異体の認識率には改善の余地がある。しかし翅脈パターンが著しく変化した場合には、１つのモデルだけからマッチングをとるのが難しい例が数多くある。現在のモデルで認識可能な場合と、困難な場合の典型的な例を下図に示す。(c) や (d) のような翅脈パターンを認識するには、個々の例に特有のパターンをモデル化して用意すれば解決できる可能性がある。しかしそのようなアドホックなモデルは普遍的でない。様々な特殊なモデルを際限なく作ることにもなりかねない問題を考慮すると、スマートなアプローチとは言えない。認識率のさらなる改善は残された今後の課題である。



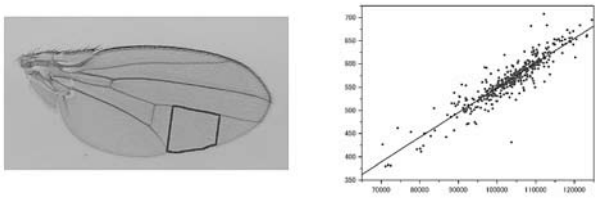
(a),(b) は認識可能で、(c),(d) は認識に失敗

ジョウジョウバエの翅の細胞数の計測と精度

細胞には細毛と呼ばれる毛が存在する。そこで細毛をカウントすることで､細胞数を間接的に計測できる。しかし単純ではなく、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）を 4 人の専門家が計測したところ、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。

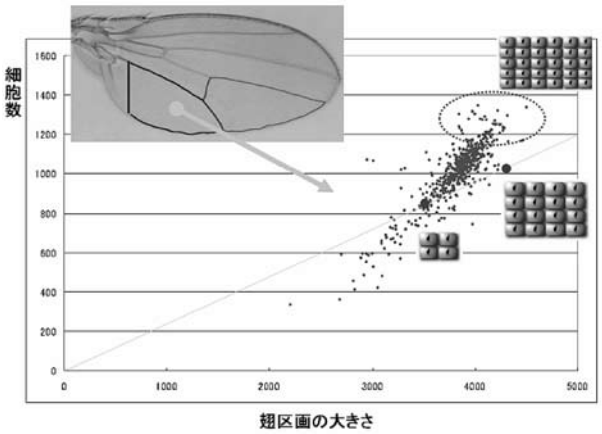


さらに、翅のブロック別に、ブロックの大きさと細毛数の関係を調べた結果、大部分の変異体では、ブロックの大きさと細毛数は線形的に比例すること、すなわち細胞数の変化が翅ブロックサイズに支配的であることもわかった。下図では左の翅のブロック領域について、各変異体のブロックの大きさをx軸、細毛数をy軸にとったのが右のグラフであり、線形相関がある。



ブロックの大きさと細毛数の関係

一方で、ブロックによっては区画の大きさと細胞数は必ずしも線形に比例するわけではなく、区画が大きくなるに従って、単位面積当たりの細胞数は多くなるような現象も観測された（下図）。興味深い結果であるが、翅脈に囲まれた領域が、翅脈の外圧に逆らって大きくなるために、内部の単位当たり細胞数を増やして内部圧を上昇させていると考えられる。



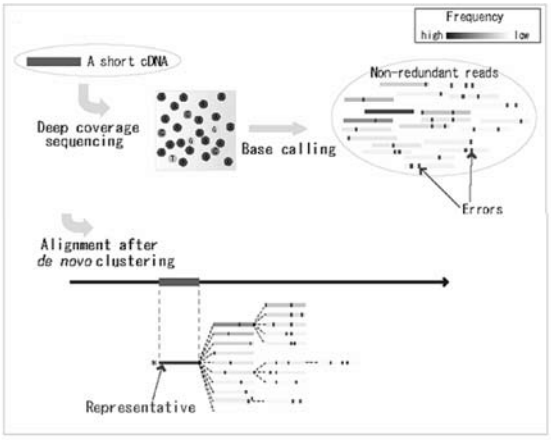
画像データベースの高解像度化

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。

遺伝子型の解析：2007年度から国内にも普及した超高速DNA解読装置(Solexa, SOLiD)は25～100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約20億塩基にも及ぶ(2009年)。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)に取り組んだ。

超高速DNA解読装置の塩基エラー修正

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 FreClu を考案した（下図にアルゴリズムの動作を示す）。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。頻度情報を利用してクラスタリングするという考え方はアルゴリズムとしても興味深く、さらにアルゴリズムは問題のサイズに対して線形時間で動作するように工夫しており高速である。(登録番号 0909041119)。

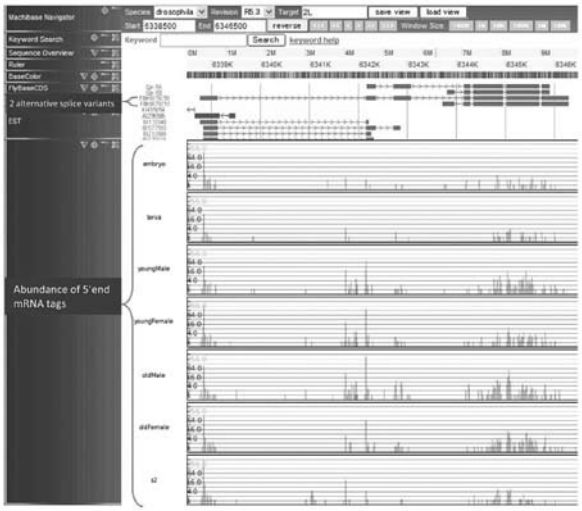


FreClu: 短いタグをゲノムへアラインメントする前に頻度情報を利用してクラスタリングし、その後にアラインメントする

超高速DNA解読装置を利用した転写開始点の網羅的収集

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 Illumina GA 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（登録番号　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（登録番号　0901131436)。下図にはショウジョウバエの組織から収集した転写開始点情報を示す。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（登録番号　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（登録番号 0901131410)。橋本班員との共同研究。



MachiBase: ショウジョウバエの７つの組織（初期胚、幼虫、オス、メス等）から収集した転写開始点と転写量を表示したデータベース

超高速DNA解読装置を使って全長cDNA配列を解読する手法

再構築の精度は 99% 以上、コストは cDNA あたり約３ドル (Illumina GA 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。

クロマチン構造と進化

DNA 配列の多様性は、生殖細胞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 Illumina GA 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（右上図参照)。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（登録番号　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。

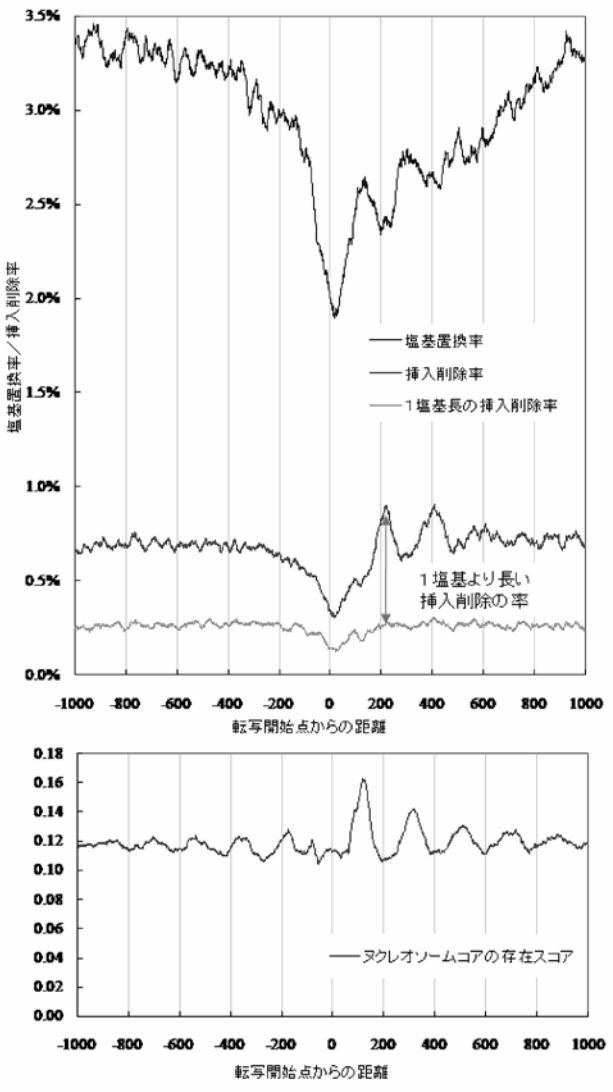
＜国内外での成果の位置づけ＞

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。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（登録番号　0612221541)。変異体の画像を解析して微小な形態変化から遺伝子機能を予測するフェノーム研究での国際的な評価は高い。例えば以下の記事で紹介されている。

Patrick Goymer: Shaping up the genome. ***Nature Review***, Vol.7, Feb. 2006; 79

一方、DNA 解読や超高速 DNA 解読装置のデータ分析についても国際的な評価を得ている。たとえば研究代表者が corresponding author として報告した下記の 2 報の論文が Faculty of 1000 Biology に収められている。

- Kasahara M *et al*. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447, 714-719 (2007)
- Sasaki S *et al*. Chromatin-Associated Periodicity in Genetic Variation Downstream of Transcriptional Start Sites. *Science* 323(5912), 401-404 (2009)



後者の論文が掲載された Science 誌の同一号では、以下の記事が研究の意義を解説している。

Semple and Taylor. MOLECULAR BIOLOGY: The Structure of Change ***Science*** 16 January 2009: 347-348.

国内では、以下に示すように、超高速 DNA 解読装置の解説特集号や解説記事を執筆し、分子生物学会年会でもシンポジウムおよびワークショップを開催し、アウトリーチ活動を行っている。

- 第 32 回日本分子生物学会　ワークショップ「バイオイメージからの情報抽出:定量化とその先にあるもの」(オーガナイザー　大浪修一　森下真一) 2009 年
- 第 31 回日本分子生物学会年会・第 81 回日本生化学会　シンポジウム「超高速シーケンサーとバイオインフォマティクス」(オーガナイザー　森下真一　鈴木穣) 2008 年
- 蛋白質 核酸 酵素　特集　次世代高速シークエンサーの応用と情報解析　「次世代高速シークエンサーの特性と情報処理」(森下真一) 2009 年 8 月
- 実験医学増刊号「生命研究への応用と開発が進むバイオデータベースとソフトウェア最前線」（森下　阿久津 編） 2008 年 4 月
- 細胞工学別冊 比較ゲノム学から読み解く生命システム「比較ゲノムを支える情報学 : 脊椎動物ゲノム進化を推定するロジック」（森下真一, 中谷洋一郎） 2007 年

＜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、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、解読装置が出力する 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。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、展望＞

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。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。今後５年間の研究の展開を、ゲノム解析周辺に絞って考えてみると、１分子実時間で塩基解読を可能にする Pacific Biosciences 社のシーケンサーの市場化（2010 年下半期の予定）が大きな反響をもつだろうと現在のところ予測している。いくつもの革新的特長がこの技術には備わっており、われわれも１分子実時間計測に立脚した新しい観測法、新しい情報学的分析技術、新しい個人ゲノム解析等に取組み始めている。このように進展が楽しみな未来像を描けるのも、本ゲノム特定領域が育んできた国内ゲノム研究者のネットワークのおかげである。その中で機会を与えていただいたことに深く感謝いたします。

＜研究期間の全成果公表リスト＞

1）論文／プロシーディング

1. 0601311151
   Naito Y, Yamada T, Matsumiya T, Ui-Tei K, Saigo K, Morishita S. dsCheck: highly sensitive off-target search software for dsRNA-mediated RNA interference. *Nucleic Acids Research* 2005 33(Web Server issue):W753-W757
2. 0601311156
   Yamada T, Morishita S. Accelerated off-target search algorithm for siRNA. *Bioinformatics*, 21(8):1316-1324, 2005
3. 0601311201
   Kasai Y, Hashimoto S, Yamada T, Sese J, Sugano S, Matsushima K, Morishita S. 5'SAGE: 5'-end Serial Analysis of Gene Expression database. *Nucleic Acids Research* Database Issue 33: D550-D552, 2005.
4. 0601311209
   Saito TL, Sese J, Nakatani Y, Sano F, Yukawa M, Ohya Y, Morishita S. Data Mining Tools for the Saccharomyces cerevisiae Morphological Database. *Nucleic Acids Research* 2005 33(Web Server issue):W589-W591
5. 0612221541
   Ohya Y, Sese J, Yukawa M, Sano F, Nakatani Y, Saito TL, Saka A, Fukuda T, Ishihara S, Oka S, Suzuki G, Watanabe M, Hirata A, Ohtani M, Sawai H, Fraysse N, Latgé JP, François JM, Aebi M, Tanaka S, Muramatsu S, Araki H, Sonoike K, Nogami S, Morishita S. High- dimensional and large-scale phenotyping of yeast mutants. *Proc Natl Acad Sci U S A*. 102(52):19015-20 (2005).
6. 0606191356
   T.Yamada, H.Souma, S.Morishita. PrimerStation: a highly specific multiplex genomic PCR primer design server for the human genome. *Nucleic Acids Research* 34: W665-W669; (2006).
7. 0612231206
   Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, and Ito T. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci U S A*. 103(47):17846-51 (2006)
8. 0612221507
   Suzuki G, Sawai H, Ohtani M, Nogami S, Sano-Kumagai F, Saka A, Yukawa M, Saito TL, Sese J, Hirata D, Morishita S, and Ohya Y. Evaluation of image processing programs for accurate measurement of budding and fission yeast morphology. *Curr Genet*. 6; 1-11 (2006)
9. 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
10. 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)

11. 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 Research* 36(Database issue): D747-52 (2008)

12. 0901161248
Miyagawa T, Kawashima M, Nishida N, Ohashi J, Kimura R, Fujimoto A, Shimada M, Morishita S, Shigeta T, Lin L, Hong SC, Faraco J, Shin YK, Jeong JH, Okazaki Y, Tsuji S, Honda M, Honda Y, Mignot E, Tokunaga K. Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. *Nature Genetics*, 40(11):1324-8, (2008)

13. 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*, 38(12), 1036-1045 (2008)

14. 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)

15. 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)

16. 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* 323(5912), 401-404 (2009)

17. 0909041119
Qu W, Hashimoto S, Morishita S. Efficient frequency-based de novo short read clustering for error trimming in next-generation sequencing. *Genome Research* 19(7): 1309-1315 (2009)

18. 0909041126
Saito TL, Yoshimura J, Sasaki S, Ahsan B, Sasaki A, Kuroshu R, Morishita S. UTGB Toolkit for Personalized Genome Browsers. *Bioinformatics* 25(15):1856-1861 (2009)

19. 0910131130
Hatsuda H, Muramatsu K, Aigaki T, and Morishita S. Robust and Accurate Recognition of Veins in Fruit Fly Wings. *IEEE 6th International Symposium on Image and Signal Processing and Analysis*. pp. 146-151, Satzburg, Austria (2009)

2) 図書

1. 0612221534 M.Kasahara and S.Morishita. *Large-scale genome sequence processing*. Imperial College Press, 248pp. (2006).

3) データベース／ソフトウェア

1. 0612221516 http://yeast.utgenome.org/
UT Genome Browser (Yeast) 出芽酵母ゲノムにおける転写開始点を完全長 cDNA により再アノテーションしたデータベースを公開。11,575 個の転写開始点が 3,638 個の遺伝子と対応付けられており、出芽酵母においても複数の転写開始点が存在することが示されている。

2. 0507070222 http://medaka.utgenome.org/
UT Genome Browser (Medaka) メダカゲノムブラウザー

3. 0507070220 http://yeast.gi.k.u-tokyo.ac.jp/
SCMD (*Saccharomyces Cerevisiae* Morphological Database) 出芽酵母の遺伝子破壊株（EUROSCARF）の顕微鏡画像をイメージ処理し、形態に関する様々な定量的パラメータを抽出。２００１年末よりデータ収集を開始し、２００４年８月に約５０００個の非必須遺伝子破壊株、約１８０万細胞の形態パラメータの計測を完了。

4. 0507070216 http://design.rnai.jp/
siDirect（ヒト、マウス、ラット、ドッグ等の siRNA 配列設計サーバー）

5. 0606191356 http://ps.cb.k.u-tokyo.ac.jp/
PrimerStation マルチプレックスゲノミックＰＣＲプライマ設計サイト。SNP タイピング , 絶対定量プライマ設計、オリゴプローブアレー設計等に利用可能。

6. 0507070219 http://5sage.gi.k.u-tokyo.ac.jp/
5' SAGE（ヒト 5SAGE タグブラウザー）5'-end SAGE tag がヒトゲノム上でどの位置に出現し、出現回数が何回あり、周辺にどのような遺伝子がエンコードされているかを、わかりやすく表示する web server。

7. 0901131451 http://machibase.gi.k.u-tokyo.ac.jp/
MachiBase: a Drosophila melanogaster 5'-end mRNA transcription database. ショウジョウバエの７つの組織（胚、幼虫、老 / 若 x 雄 / 雌 , S2）から収集した各々 3-400 万個の転写開始点タグ情報の公開データベース。

8. 0911241516 http: //mlab.cb.k.u-tokyo.ac.jp/~quwei/DeNovoShortReadClustering
FreClu 5'SAGE および miRNA 等の同じローカスから由来する短いタグ配列の誤りを訂正するためのソフトウエア。次世代シーケンサーが出力する数億個のタグ配列を高速に処理できるような線形時間アルゴリズムである。

9. 0909041128 http://utgenome.org/
UTGB toolkit for personalized genome browsers. 各研究室レベルで容易にゲノム解析ができるように、Personalized Genome Browser を構築するツールキットを研究開発し、無料で公開した。実際数分でゲノムブラウザーを立ち上げ、データを secure な環境で分析することも、一般に公開することも可能である。

10. 0911241513 http://dscheck.rnai.jp/
dsCheck 線虫、ショウジョウバエ等のモデル生物において RNA 干渉を行う配列を設計する web サイト。cross reaction を防止できるように配慮してある。

11. 0911241510 http://musica.gi.k.u-tokyo.ac.jp/
MuSICA2 全長 cDNA 配列を求めるために、複数の cDNA クローンを短く断片化して、次世代シーケンサーで解読し、参照ゲノムへとアラインメントし、情報を補って cDNA 配列を解読するソフト。