マルチモーダル統合バイオＤＢ

東京大学大学院新領域創成科学研究科
森下　真一

Multimodal BIODB

Shinichi Morishita
Graduate School of Frontier Sciences, The University of Tokyo

A variety of multimodal omics data in biology, such as genome, nucleosome, methylome, transcriptome, proteome, and phenome, have been available to date. In particular, collecting massive genome/nucleosome/methylome/transcriptome data has been becoming increasingly feasible because of the wide-spread availability of next-generation sequencing technology. Since the volume of data has been growing exponentially for the last several years, the computational analysis is likely to dominate the overall task, demanding an efficient way of handling a flood of data. The “Multimodal BIODB” project aims at developing a uniform computational framework for integrating massive multimodal biological data so as to facilitate the process of scientific research and discovery efficiently. We will illustrate a couple of new findings using the multimodal BIODB to demonstrate its usefulness.

1. はじめに

近年、DNA 配列、組織別遺伝子発現量、バイオイメージ、パスウエイ、文献などの異なるタイプの生物データが急速に蓄積されてきている。膨大なデータから目的の情報を抽出するには BLAST、クラスタリング、PubMed などのソフトウェアが欠かせない。これらのソフトウェアは配列・文献など単一のデータタイプを処理するために特化している。しかし現実には単独のデータタイプだけが利用されることは少ない。むしろ異なるデータタイプを組合せながら慎重に遺伝子情報を絞り込み、失敗の少ない実験を計画できるように利用することがおおい。また異なるデータタイプの組み合わせから新しい科学的知見のヒントを得ることもある。

マルチモーダル統合バイオ DB とは、異なるデータタイプに対する問合せを実現したシステムを意味している。さまざまなデータタイプが蓄積されている出芽酵母では、このような研究目標をもった国際的研究グループ YSBN が既に存在し、我々もそのなかでマルチモーダル問合せの萌芽的研究を試みてきた[1]。この取り組みを本プロジェクトが発足した 2006 年以来、出芽酵母に加えて、発生遺伝学のモデル生物であるメダカ[2][5]、ショウジョウバエ[9]、害虫研究のモデル生物であるカイコ[3]そして、ヒトへと広げてゆき、当初の研

究計画に沿って順調に研究開発をすすめた。

2. 研究開発の成果

多様な生物種のマルチモーダル DB サーバーの構築を可能にしたのは UTGB Toolkit の研究開発である。本ツールキットは本プロジェクトが開始された 2006 年度より継続して研究開発されているが、以下のようにシステムの中身は充実した。

- ゲノムブラウザーUTGB の公開　（以下の特長をそなえる。図 1 参照）
  - 初心者でも個人用にゲノムブラウザー構築できる機能
  - データベースの移植性の向上　SQLite JDBC ライブラリによる研究開発
  - シェル機能の開発 (UTGB Shell) によるデータベース管理の効率化
- 次世代シークエンサーデータ処理のための高速ソフトウェアの研究開発と公開

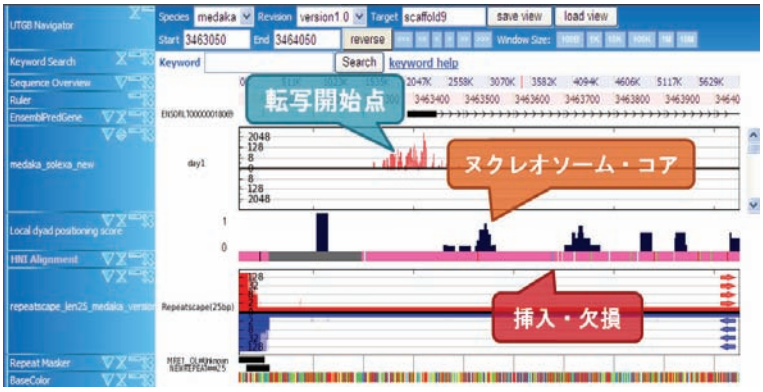

図 1　UTGB メダカゲノムブラウザーでの表示例。転写開始点下流のヌクレオソーム構造の位置情報と、メダカゲノムの挿入・欠損の位置を並べて比較している

上記にもリストしたように、当初の研究計画では予想されていなかった革新的シークエンシング技術が 2007 年より普及した。そのため、研究計画を前向きに軌道修正し、マルチモーダル統合バイオ DB を利用できるように腐心した。超高速シークエンサー(Illumina/HiSeq) は、解読可能配列長は　75～150 塩基と短いものの、単位時間当たりの塩基解読量は 2011 年 2 月には約 200 億塩基／日に達し、従来のサンガー法に比べ約 10,000 倍の能力がある。われわれはこのデータ分析にマルチモーダル統合バイオ DB を利用できないか模索した。なぜなら超高速シークエンサーを利用すると以下に示すような未曾有のマルチモーダルデータを高速に収集可能だからである。

- エピゲノム情報（DNA メチル化、ヌクレオソーム位置の分布、ヒストン修飾）
- SAGE 法を拡張した定量的転写開始点情報、および広いダイナミックレンジの遺伝子発現情報（2009 年 ***Genome Research*** 誌[8]および ***PLoS One*** 誌[10]に報告）
- 全長 cDNA 配列決定（cDNA クローンの配列決定[11]、および RNA-Seq 法）
- 遺伝子多型　（SNP だけでなく、1bp~100Kbp 長の挿入削除および逆位）

このような大量のデータを処理できるデータベースシステムを構築するため、相応の計算資源を用意しソフトウェア最適化を行う定石的戦略を実施した。しかしデータベースの公開だけでは研究テーマとしては陳腐に陥りがちでもある。われわれはもう一歩踏み込んで、あたらしい科学的結果を導くためにマルチモーダル統合バイオ DB の考え方が役立つ

ことを示したいと考えた。そこで上記のマルチモーダルデータを分析し「クロマチン構造が遺伝的多様性に影響を与えるか否か？」という問題分析に取り組んだ。さいわいスタンフォード大学医学部・Andrew Fire 教授および東京大学理学系研究科・武田洋幸教授のグループとの共同研究が実り、転写開始点下流において影響があるという結果を得ることができ（図２）、2009 年に ***Science*** 誌に報告することができた[6]。

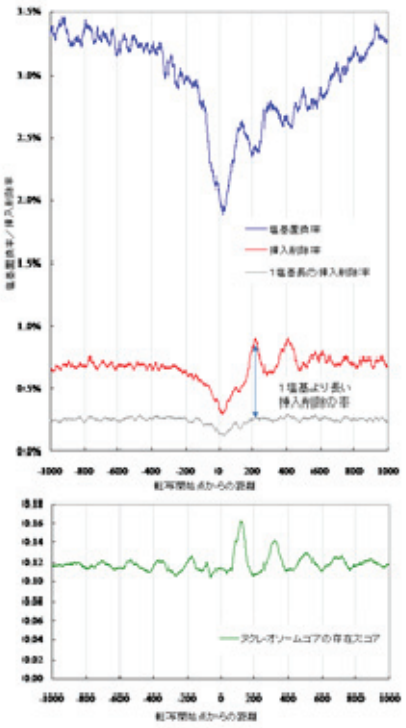

図２　転写開始点周辺での、塩基置換率、挿入削除率、ヌクレオソームコアの存在スコア

また SAGE 法を拡張した定量的転写開始点情報を処理する際に、塩基読み取りミスにより 5-10%のリードをゲノム上の正確な位置にアラインメントができないという問題があった。この問題を解決するために、各リードの頻度分布を利用したクラスタリングをするというあたらしいアルゴリズムを研究開発して評価してみたところ、大幅な性能向上がみられた(2009 年 ***Genome Research*** 誌に報告[8]、図３)。このソフトウェア FreClu を公開した。

コンピュータ科学的な観点からは、大量のデータから科学的分析を迅速に実施するには、マルチモーダルデータの表現形式を定式化し、データ変換の数学的な枠組みを考え、データ変換操作を代数的に組み合わせることができるような理論的な基盤をつくることがエレガントな解法につながると我々は考えた。この方針は 1970 年代に達成された関係データベースの理論構築に倣っているが、関係データベースと異なり、多様なデータ構造をもつマルチモーダルデータを扱った先行研究例は少なかった。研究の結果、満足のゆく理論と実装方法を構築することができた。成果は、計算機科学のデータベース分野で最も評価の高い ***ACM SIGMOD*** において 2008 年に論文発表することができた[4]。その後も研究を重ね、関係データベース上の SQL を利用することによるポータビリティの向上、データベースアーカイブ容量を圧縮する方法を研究開発している。

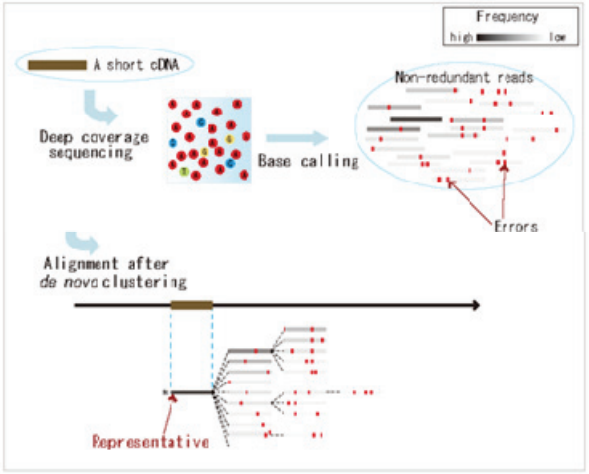

図３　FreClu: 短いタグをゲノムへアラインメントする前に頻度情報を利用してクラスタリングし、その後にアラインメントする

また、超高速シークエンサーの単位時間あたりの塩基解読量は 2006 年以降は年間約 4〜5 倍程度で推移している。このスピードは 1CPU の計算機の処理能力が 1.5 年間で約 2 倍になるという経験則(ムーアの法則)をはるかに凌駕しており、塩基解読量の増加に 1CPU あたりの性能向上は追いつかず、年々指数的に離されてゆく。そこで多数の CPU を使用してデータ処理を並列化して、指数関数的に増える塩基解読量に対応することが必要になる。そこで 2008-9 年度は東京大学情報基盤センターに導入された並列計算機を借用し、マルチモーダル統合バイオ DB の並列化を試みている。

3. まとめ

以上のように、革新的シークエンシング技術の普及に対応できるようにマルチモーダル統合バイオ DB の研究計画を前向きに軌道修正し、しかも短期間の間にツールキット UTGB (University of Tokyo, Genome Browser) も公開することができた（***Bioinformatics*** 誌に報告)[7]。マルチモーダル DB に関する最新の動向と今後の課題を議論するためのワークショップを、第 32 回日本分子生物学会年会（2009/12/11, 参加者約 300 名）にて開催した。研究成果はウェブサーバーとして公開されており、年間 142,594 (集計日数 360 日)、 1 日平均で約 396 の独立した訪問者(Visits) により利用されている。超高速 DNA シークエンサーが出力するデータを分析したウェブサーバーを、ヒト、カイコ、ショウジョウバエを対象に公開中である。主な論文発表成果は、***Science, Genome Research, Bioinformatics, Nucleic Acids Research*** に報告した。

4. 研究開発実施体制

代表研究者　森下　真一（東京大学大学院新領域創成科学研究科）

研究開発題目

(1)　マルチモーダル統合バイオＤＢの構築
　　グループリーダー　森下　真一（東京大学大学院新領域創成科学研究科）

(2)　出芽酵母を対象にしたマルチモーダル統合バイオＤＢの構築
　　グループリーダー　伊藤　隆司（東京大学大学院理学系研究科）

(3)　メダカを対象にしたマルチモーダル統合バイオＤＢの構築
　　グループリーダー　武田　洋幸（東京大学大学院理学系研究科）

5. 参考文献

[1] 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 SA***. 103(47):17846-51 (2006)

[2] Masahiro Kasahara(*), Kiyoshi Naruse(*), Shin Sasaki(*), Yoichiro Nakatani(*), Wei Qu, Budrul Ahsan, Tomoyuki Yamada, Yukinobu Nagayasu, Koichiro Doi, Yasuhiro Kasai, Tomoko Jindo, Daisuke Kobayashi, Atsuko Shimada, Atsushi Toyoda, Yoko Kuroki, Asao Fujiyama, Takashi Sasaki, Atsushi Shimizu, Shuichi Asakawa, Nobuyoshi Shimizu, Shin-ichi Hashimoto, Jun Yang, Yongjun Lee, Kouji Matsushima, Sumio Sugano, Mitsuru Sakaizumi, Takanori Narita, Kazuko

Ohishi, Shinobu Haga, Fumiko Ohta, Hisayo Nomoto, Keiko Nogata, Tomomi Morishita, Tomoko Endo, Tadasu Shin-I, Hiroyuki Takeda(#), **Shinichi Morishita(#)**, and Yuji Kohara(#). The medaka draft genome and insights into vertebrate genome evolution. ***Nature*** 447, 714-719 (2007)
* joint first authors. # joint corresponding authors.

[3] The International Silkworm Genome Consortium (including **Shinichi Morishita** as one of corresponding authors). The genome of a lepidopteran model insect, the silkworm Bombyx mori. ***Insect Biochemistry and Molecular Biology***, Vol. 38, Issue 12, 1036-1045 (2008)

[4] Taro L. Saito and **Shinichi Morishita.** Relational-Style XML Query. *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* ***(ACM SIGMOD)***, Vancouver, 303-314 (2008)

[5] Budrul Ahsan, (34 authors), Hiroyuki Takeda, Yuji Kohara, and **Shinichi Morishita**. UTGB/medaka: genomic resource database for medaka biology. ***Nucleic Acids Research***, Vol. 36, Database issue D747-D752, (2008)

[6] Shin Sasaki, Cecilia C. Mello, Atsuko Shimada, Yoichiro Nakatani, Shin-ichi Hashimoto, Masako Ogawa, Kouji Matsushima, Sam Guoping Gu, Masahiro Kasahara, Budrul Ahsan, Atsushi Sasaki, Taro Saito, Yutaka Suzuki, Sumio Sugano, Yuji Kohara, Hiroyuki Takeda, Andrew Fire(#), **Shinichi Morishita**(#) Chromatin-Associated Periodicity in Genetic Variation Downstream of Transcriptional Start Sites. ***Science***. 323(5912):401-4 (2009)
joint corresponding authors

[7] Taro L. Saito, Jun Yoshimura, Shin Sasaki, Budrul Ahsan, Atsushi Sasaki, Reginaldo Kuroshu and **Shinichi Morishita**. UTGB Toolkit for Personalized Genome Browsers, ***Bioinformatics*** 25(15):1856-1861 (2009)

[8] Wei Qu, Shin-ichi Hashimoto, **Shinichi Morishita**. Efficient frequency-based de novo short read clustering for error trimming in next-generation sequencing. ***Genome Research*** 19(7): 1309-1315 (2009)

[9] Budrul Ahsan, Taro L. Saito, Shin-ichi Hashimoto, Keigo Muramatsu, Manabu Tsuda, Atsushi Sasaki, Kouji Matsushima, Toshiro Aigaki, and **Shinichi Morishita**. MachiBase: a Drosophila melanogaster 5'-end mRNA transcription database. ***Nucleic Acids Research***, Vol. 37, Database issue D49-D53 (2009)

[10] Shin-ichi Hashimoto(*), Wei Qu(*), Budrul Ahsan, Katsumi Ogoshi, Atsushi Sasaki, Yoichiro Nakatani, Yongjun Lee, Masako Ogawa, Akio Ametani, Yutaka Suzuki, Sumio Sugano, Clarence C Lee, Robert C Nutter, **Shinichi Morishita**, Kouji Matsushima. High-resolution analysis of the 5'-end transcriptome using a next generation DNA sequencer. ***PLoS One*** 4(1):e4108. Epub (2009)

[11] Kuroshu RM, Watanabe J, Sugano S, **Morishita S,** Suzuki Y, Kasahara M. Cost-effective sequencing of full-length cDNA clones powered by a de novo-reference hybrid assembly. ***PLoS One***. 7;5(5):e10517 (2010)