比較ゲノム解析のための情報変換モデルの構築と計算手法の開発

●榊原 康文[1)]　◆浅井 潔[2)]
1）慶應義塾大学理工学部　2）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＞

複数種のゲノム配列の比較解析を行うアルゴリズムを開発して，多種ゲノム間に保存されたシンテニー領域の同定と，ゲノム再編成などに見られる転移や逆位などのゲノム構造の変化を計算するシステムを構築する．また，開発されたシステムを用いて，領域内の他の研究班とも協力してさまざまな生物種に対して，遺伝子グループの再編成や進化などを解析することを目指す．応用の具体例として，ヒトゲノムにおけるサブテロメア解析を試みる．さらに，新世代シークエンサーを用いた比較ゲノムアセンブリに本開発手法を応用する試みも実施する．一方，比較ゲノム手法を用いて，同定が難しい機能性 RNA 領域をゲノムから網羅的に探索・発見するアルゴリズムの開発も行う．

＜ 2008 年度の研究の当初計画＞

2008 年，2009 年度の２年間の目標は，(1) 多種のゲノム配列から保存領域を高速に検出するシステム Murasaki のマルチプルアライメントを計算するための拡張，(2) Murasaki を検出精度，計算時間の両面から最適なものにしてシステムを完成させる，(3) アンカーを可視化するためのインタラクティブなツール GMV もさらにユーザの利便性を向上させてシステムを完成させる，(4) シンテニー計算に基づくゲノム再編成の解析，(5) 領域内の他の研究班と協力して霊長類や微生物ゲノムなどに対する比較ゲノム解析の計算機実験を行う．具体例としてヒトゲノムにおけるサブテロメア解析を実施する，(6) 比較ゲノムと機能性 RNA 同定アルゴリズムを組み合わせた新規機能性 RNA 領域のゲノム配列からの発見，を達成することである．さらに，比較ゲノムシステム Murasaki の新しい応用として，新世代シーケンサーのショートリードデータを用いた，リファレンスゲノムへのマッピングによる比較ゲノムアセンブリ手法の確立と微生物ゲノムの配列決定を試みる．

＜ 2008 年度の成果＞

比較ゲノムシステム Murasaki の並列化によってヒトを含む霊長類ゲノムの３種比較，さらに哺乳類ゲノム５種比較を全染色体丸ごと比較することが可能になった．24 本の染色体全部を丸ごと比較することができるようにするために，クラスターマシンと呼ばれる数十から数百，数千の CPU を並列に同時に動かすことのできる計算機を用いて，霊長類ゲノム比較などの非常に大きな仕事を多数の CPU に分散して手分けして処理する仕組みを開発した．並列化 Murasaki により，ヒト，チンパンジー，アカゲザル，マウス，ラットの５種を丸ごと比較解析した結果を図１に示す．さらに，Murasaki の最後の課題であったマルチプルアライメントを計算するアルゴリズムを開発してシステムに組み込むことが達成できた．アライメントを計算するアルゴリズムの概要は以下の通りである：(1) 従来の Murasaki の機能により多種間に保存されたアンカーを検出する，(2) BLAST スコア関数を用いてアンカー領域を拡張する，(3) オーバーラップする領域を結合する，(4) (2) と (3) をスコアが閾値より低くならない限り繰り返す，(5) 最終的に計算された最長領域をギャップ無し局所的マルチプルアライメントとする，(6) ギャップ無し局所的マルチプルアライメントから任意の２種間のギャップ付き局所的アライメントをアンカー情報と動的計画法を効率的に用いて計算する．本提案手法の最大の特徴は，マルチプルアライメントはギャップ無しのものだけを求めることとして，ギャップ付きアライメントは任意の２種間のものをギャップ無しマルチプルアライメントから計算する設定としたことである．実用的には，多種間の保存関係を見るためにはギャップ無しマルチプルアライメントで十分であり，一塩基レベルでの詳細な配列比較を行う場合には２種間のギャップ付きペアワイズアライメントが計算できれば十分である．実際，多くの既存手法が同様の方法を取っている．また，ギャップ付きマルチプルアライメントの計算は理論的には計算困難（NP 困難）な問題であり，それをゲノムレベルで求めることは実際不可能であり，意味がないことであると考えられる．上記のように拡張した Murasaki は，多種のゲノムから局所的マルチプルアライメントを計算する性能においても，既存手法より高い精度と計算速度を達成することができた．

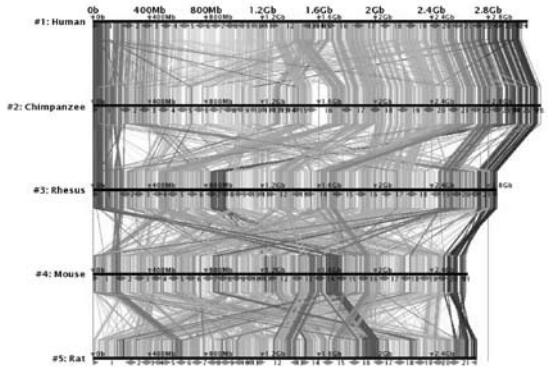

図１：哺乳類５種ゲノムの全染色体丸ごと比較

次に，比較ゲノムシステム Murasaki の応用の一つとして，Murasaki を機能性 RNA 配列解析のためのカーネル関数群と組み合わせることにより，任意の複数種のゲノム配列群から機能性 RNA 領域を高速に検出する手法を開発した．RNAz に代表される機能性 RNA 識別手法の多くは，入力として配列アラインメントを必要とする．そのため，機能性 RNA のゲノム探索では，複数種のゲノムを比較して保存領域を抽出する行程が重要となる．しかし，既存の比較ゲノム手法は塩基配列の相同性のみに注目しており，機能性 RNA の特徴である２次構造の保存を考慮していない．そこで，ゲノムの２次構造を高速に比較する新しい手法を提案した．２次構造の文字列表記法を利用し，RNALfold の２次構造予測の結果を文字列として直接比較する（図 2）．本手法の有効性を示すために，線虫 *C.elegans* のゲノムを同じ線形動物門に属する *C.briggsae* および *P.pacificus* と比較した．RNAz の主要な識別指標である MFE Z-score と SCI を用いた評価により，本手法で検出された領域は従来の比較ゲノム手法よりも熱力学的に安定で有意に保存された２次構造を形成しうることが示唆された（表 1）．

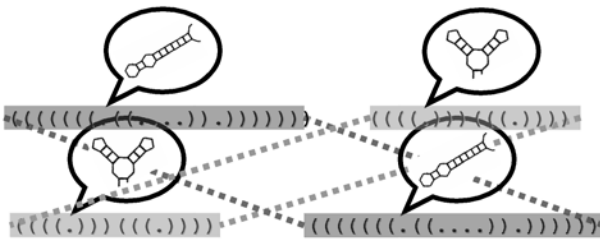

図２：２次構造の保存性を全ゲノム規模で検出

一方，機能性 RNA の解析のための新しいアルゴリズムの開発では，RNA 2 次構造の与えられたエネルギーモデルと，予測に適した評価尺度に対して，理論的に予測精度が最大となる予測手法を開発した．構造既知の RNA による計算機実験では，1 本の RNA 配列からの 2 次構造予測，RNA 配列群からの 2 次構造予測のどちらにおいても，既存のどの手法よりも精度が高いことが実証された．

次の Murasaki の応用として，最新のヒトゲノムのリファレンス配列を用いて，網羅的に全ゲノム規模でのサブテロメアのパッチワーク構造を解明した．全ゲノムにおけるパッチワーク構造の存在を観察し，サブテロメアがゲノム内で非常に柔軟性が高く，急速な進化を遂げる領域であることを確認した．また，本研究で同定された染色体間で重複した配列のブロックを調べることで，遺伝子ファミリーの多様化との関連性を明らかにした．その結果，嗅覚受容体 olfactory receptor（OR）遺伝子ファミリーの多様化に寄与していることや，FRG2（疾患 Facioscapulohumeral Dystrophy（FSHD）に関与する遺伝子の一つ）のような疾患関連遺伝子がサブテロメア領域内には多数存在することが確認された．

さらに，次世代シークエンサーを用いたショートリードからの微生物ゲノムの比較ゲノムアセンブリに Murasaki を適用する実験を行った．対象としたのは，まだゲノムが未解読である納豆菌ゲノムの配列決定で，枯草菌ゲノム配列をテンプレートとした．解析の手順は，Solexa を用いた納豆菌ゲノムのショートリードから，既存のアセンブリプログラム Velvet によりコンティグを生成して，Murasaki によりコンティグを枯草菌ゲノムに整列させた．さらに制限酵素の物理地図を用いた調整と残ったギャップ領域をサンガーシークエンサーを用いて埋めることにより，最終のドラフト配列を決定した．

＜国内外での成果の位置づけ＞

ヒトなどの高等生物のゲノム比較を多種間でゲノム丸ごと計算できる既存のシステムは存在しない．さらに並列化 Murasaki と今年度開発に成功した局所的アライメントは，霊長類ゲノムを全染色体丸ごとで多種間比較ができる唯一のシステムであり，並列化 Murasaki と局所的アライメント計算を開発した成果は非常に高いと評価できる．

＜達成できなかったこと、予想外の困難、その理由＞

Murasaki の最終課題となっていた検出されたアンカーからマルチプルアライメントを計算するアルゴリズムの開発は，困難を極めたが，多種間のギャップ無し局所的マルチプルアライメントと２種間のギャップ付きアライメントの計算というかたちで解決を得た．ギャップ付きマルチプルアライメントの計算問題は理論的には計算量困難な問題に属しており，またゲノム丸ごとのマルチプルアライメントを計算することはほぼ不可能である．したがって，愚直にマルチプルアライメントを計算することを求めるのではなく，ほぼ同様の結果を得ることができる今回のアプローチは実用的には十分であると考えている．

＜今後の課題＞

研究プロジェクト開始時からの最大の目標であった霊長類ゲノムの全染色体丸ごと多重比較の計算と局所的アライメントの計算を達成することができた．最後の１年間の課題は，このゲノム比較システムの完成とその応用的課題，霊長類の比較ゲノム解析や次世代シークエンサーを用いた微生物ゲノムやメタゲノム解析，に取り組んでいく．

＜成果公表リスト＞

1）論文／プロシーディング

1. 901111452（論文）
   K. Sato, T. Mituyama, K Asai, and Y. Sakakibara, Directed acyclic graph kernels for structural RNA analysis, BMC Bioinformatics, 9:318（2008）
2. 901111456（論文）
   M. Morita, Y. Saito, K. Sato, K. Oka, K. Hotta, and Y. Sakakibara, Genome-wide searching with base-pairing kernel functions for non-coding RNAs: computational and expression analysis of snoRNA families in Caenorhabditis elegans, Nu*CLE*ic Acids Research, doi: 10.1093/nar/gkn1054（2009）
3. 0805301625（論文）
   Okada, K., and Asai, K., Retention of genes involved in the adenohypophysis-mediated endocrine system in early vertebrates, Gene, 412（1-2）, 71-83（2008）
4. 0806231016（論文）
   Okada, K., and Asai, K., Expansion of signaling genes for adaptive immune system evolution in early vertebrates, BMC Genomics, 9:218（2008）
5. 0806230949（論文）
   Asai, K., Kiryu, H., Hamada, M., Tabei, Y., Sato, K., Matsui, H., Sakakibara, Y., Terai, G., and Mituyama, T., Software. ncrna.org: web servers for analyses of RNA sequences, Nu*CLE*ic Acids Research, 36: Web Server issue, W75-W78（2008）
6. 0901131759（論文）
   Hamada, M., Kiryu, H., Sato, K., Mituyama, T., and Asai, K., Prediction of RNA secondary structure using generalized centroid estimators, Bioinformatics Advanced Access published on December 18（2008）.
7. 901111501（プロシーディングス）
   Y. Saito, K. Sato, and Y. Sakakibara, Detection of secondary-structure conserved regions by genome comparison, The 2008 Annual Conference of Japanese Society for Bioinformatics（JSBi 2008）, T11（口頭発表）（2009）

2）データベース／ソフトウェア

1. 701121250
   Popendorf, K., Osana, Y., and Sakakibara, Y., 比較ゲノム解析システム Murasaki，ソフトウエア（バージョン 1.35），http://murasaki.dna.bio.keio.ac.jp/
2. 0806231106
   ＲＮＡ配列解析統合ウェブサーバ
   http://software.ncrna.org/

3）共同研究の状況

サブテロメアと個人ゲノム解析（藤山（比較ゲノム B02 班）），魚ゲノムマーカーの予測（成瀬（比較ゲノム B03 班）），粘菌ゲノムの比較解析（漆原（比較ゲノム B01 班））．