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，多種ゲノム間における信頼性の高い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，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．さらに，ゲノム比較から同定された保存領域の出現頻度情報に基づくゲノム配列の統計言語的解析を試みる．比較ゲノムシステムの開発と解析を行うときの基本的な考え方として，「情報変換モデル」という，１つの生物種のゲノムを１つの言語と見なして，比較ゲノムはゲノム言語からゲノム言語への変換（翻訳）であるという新しい考え方を提案する．一方，比較ゲノム手法を用いて，同定が難しい機能性 RNA 領域をゲノムから探索・発見するアルゴリズムの開発も行う．

＜ 2007 年度の研究の当初計画＞

平成 18 年度に開発した大規模なゲノム解析に用いることのできるゲノム比較システム Murasaki をさらに高速化，高スケール化する．目標はヒトゲノムなどの霊長類ゲノムを染色体ごとに分けて比較するのでなく，丸ごと比較できるシステムの開発である．同時に，Murasaki によって出力される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 の高スケール化と高機能化も行う．次に，(1) Murasaki のマルチプルアライメントを計算するための拡張，(2) シンテニー計算に基づくゲノム再編成の解析，(3) ヒトとマウス，ヒトとチンパンジー，また領域内の他の研究班とも協力して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，(4) 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目指す．

一方，比較ゲノムを用いた非コード RNA 領域の解析のためのシステム開発と，複数のアルゴリズムを統合するソフトウエアパッケージの開発を行う．

＜ 2007 年度の成果＞

平成 19 年度の最大の成果は，比較ゲノムシステム Murasaki の並列化とそれによってヒトを含む霊長類ゲノムの３種比較を全染色体丸ごと比較することが可能になったことである．比較ゲノムシステム Murasaki は従来のシステムに比べて優れた性能を発揮したが，１CPU を使った計算では残念ながらヒトの一本の染色体の大きさまでが比較の限界であった．そこで，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，ヒト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．多数の CPU を用いることにより高速化されると同時に，使用できるメモリの容量が増えることにより高スケール化できたことが最大のポイントである．霊長類ゲノムを多種で比較するときの問題点は膨大なデータ量を格納するメモリの容量不足であった．そこで仕事を多数の CPU に分散すると同時にデータの記憶も多数の CPU とそのメモリに手分けすることによりこの難題を解決した．その仕分けの方法に独自の巧妙な仕組みを開発した．この並列化された Murasaki を以後，並列化 Murasaki と呼ぶことにする．並列化 Murasaki により，ヒト，チンパンジー，アカゲザルの３種ゲノムとヒト，マウス，ラットの３種を丸ごと比較解析した結果を図１と２に示す．

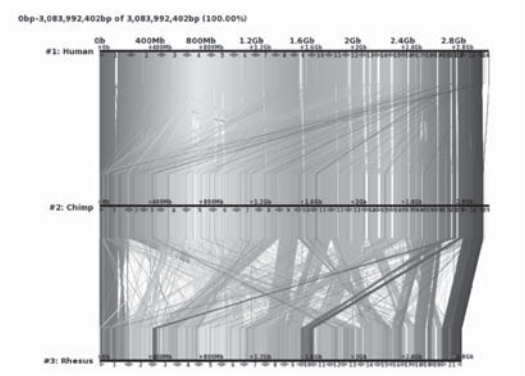

図１：ヒト，チンパンジー，アカゲザルの３種ゲノムの全染色体丸ごと比較

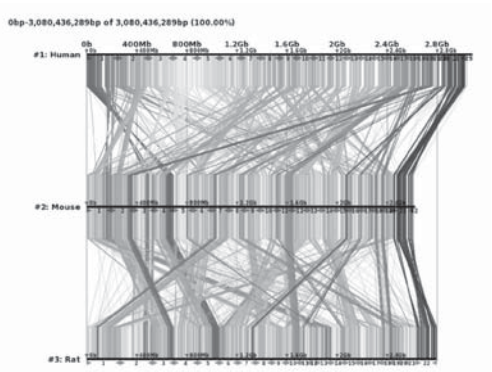

図２：ヒト，マウス，ラットの３種ゲノムの全染色体丸ごと比較

このように開発された Murasaki は，ゲノム特定の領域内や領域間でゲノム比較解析ツールとして使用されている．応用ゲノム K16 鈴木班との共同研究で，Murasaki と GMV を用いて，数種のマイコバクテリアゲノムと *M.leprae*（らい菌）ゲノムを比較してタイリングアレイデータを組み合わせることで，*M.leprae* の偽遺伝子の探索を行っている．さらに，ミヤコグサ根粒菌の共生アイランドの比較解析（佐伯（比較ゲノム B04 班田畑班)），霊長類ゲノムのテロメア解析（藤山（比較ゲノム B02 班)），マイコプラズマ 10 種の比較ゲノム解析（佐々木（応用ゲノム C04 見理班)），などの共同研究を実行中である．

次に，Murasaki により検出されたアンカーからシンテニー領域を計算するアルゴリズムを開発して，それに基づいて哺乳類ゲノムのゲノム再編成解析を行った．シンテニー領域は，短い保存領域が長いギャップ領域によって隔てられている領域である．そこで開発したアルゴリズムは，近隣に位置するアンカー領域を連結することによりシンテニー領域を検出する．シンテニー領域を計算するアルゴリズムの概略は，以下の３ステップから成り立っている：(1) アンカー領域の保存度とアンカー領域間のギャップの長さに基づいて，アンカー領域を連結するか否かを決定する．アンカー領域の保存度が高ければ高いほど，アンカー間のギャップ領域が短ければ短いほど，アンカー領域は連結基準を満たしやすくなる．(2) アンカー間に再編成が起きていない共線形性を満

たすアンカーを連結する．(3) アンカーの連結された領域を，シンテニー領域として出力する．このアルゴリズムによって同定されたシンテニー領域に基づいて，ヒト，チンパンジー，アカゲザル，イヌ，マウス，ラット，オポッサムの７種に対して再編成に基づく系統樹を計算した．その結果を図３に示す．再編成に基づく進化系統樹は，霊長目と食肉目のクレイドに対して，げっ歯目は遠縁であることを示した．

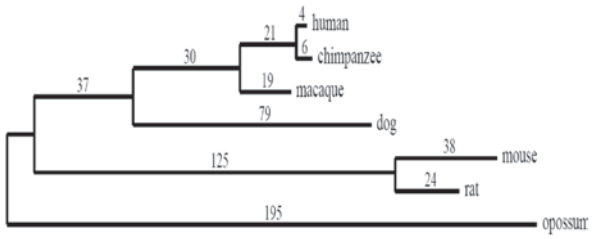

図３：再編成の回数に基づく胎盤性哺乳類の進化系統樹

一方，機能性 RNA の解析のための新たなアルゴリズムやシステムの開発も積極的に行った．その一つは，サポートベクターマシンを用いて機能性 RNA 配列の識別と探索を行うために，新しくカーネル関数を設計して，RNA ファミリーの識別に関する計算機実験を行い，さらにこの手法を適用して線虫ゲノム上の snoRNA を網羅的に予測した．次に，その予測候補に対して，既知 snoRNA には含まれないが確率値の高いものを新規 snoRNA 候補として，上位からプライマー設計を行い，50 個の候補に対して定量的 RT-PCR 実験を行った．その結果，Ct 値が 40 以下で発現しているという基準で評価した場合，プライマーを設計した 50 個の候補の内 33 個の発現を確認することが出来た．

＜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計算できる既存のシステムは存在しない．多種間ゲノム比較の代表的既存手法である Mauve は微生物ゲノムの大きさまでが適用の限界である．また，Pattern Hunter や BLASTZ などは２種間のゲノム比較のみ適用可能である．さらに今年度開発に成功した並列化 Murasaki 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 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＞

昨年度からの課題となっている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今年度も達成されなかった．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と考えています．したがって，直接マルチプルアライメントを計算するのではなく，ほぼ同様の結果を得ることができる全く異なるアプローチを取る必要があると考えている．

＜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を達成することができた．開発された並列化 Murasaki は世界に類を見ない高速高スケーラビリティの比較ゲノムシステムである．今後の課題は，このゲノム比較システムを積極的に利用した応用的課題，例えば，サブテロメア解析や遺伝子獲得・遺伝子損失などのゲノム reconcilation 問題など，に取り組んでいく予定である．

＜成果公表リスト＞

1) 論文／プロシーディング

1. 801221812
Sakakibara, Y., Ogawa, N., Popendorf, K., Asai, K., and K. Sato, Stem kernels for RNA sequence analyses, Journal of Bioinformatics and Computational Biology, 5(6), 1103-1122 (2007).
2. 801221818
Sato, K., Morita, K., and Sakakibara, Y., PSSMTS: position specific scoring matrices on tree structures, Journal of Mathematical Biology, 56, 201-214 (2008).
3. 801221823
Hachiya, T., Osana, Y., Popendorf, K., and Sakakibara, Y., OSfinder: an accurate orthology mapping program and its application to placental mammalian genomes, Proc. 2007 Annual Conference of Japanese Society for Bioinformatics (JSBi 2007), 2007.
4. 801221832
Nagamine, N. and Sakakibara, Y., Statistical prediction of protein-chemical interactions based on chemical structure and mass spectrometry data, Bioinformatics, 23(15), 2004-2012 (2007).
5. 801232142
Kiryu, H., Tabei, Kin, T., and Asai, K., Murlet: A practical multiple alignment tool for structural RNA sequences, Bioinformatics 23(13), 1588-1598 (2007).
6. 801232145
Terai, G., Komori, T., Asai, K., and Kin, T., miRRim: A novel system to find miRNAs with hight sensitivity and specificity, RNA 13:2081-2090 (2007).
7. 801272303
Kiryu, H., Kin, T., and Asai, K., Rfold: An exact algorithm for computing local base pairing probabilities, Bioinformatics Advanced Access published on December 4, 2007. doi:10.1093/bioinformatics/btm951
8. 801232136
Tabei, Y., Kiryu, H., Kin, T., and Asai, K., A fast structural multiple alignment method for long RNA sequences, BMC Bioinformatics 9:33 (2008).

2) データベース／ソフトウェア

1. 701121250
比較ゲノム解析システム Murasaki
http://murasaki.dna.bio.keio.ac.jp/
2. 801272312
ＲＮＡ配列多重整列プログラム MXSCARNA
http://mxscarna.ncrna.org/