社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (39)
2007－BIO－11 (39)
2007／12／21

大規模ゲノム比較システムの開発と応用

榊原 康文
慶應義塾大学理工学部生命情報学科
E-mail: yasu@bio.keio.ac.jp Webpage: http://dna.bio.keio.ac.jp/

概要：
近年の急速なゲノム配列の決定により，さまざまな生物種のゲノム配列データが利用可能になった．その結果，ゲノムを比較することにより，種の違いをゲノムに生じた遺伝子種類の違いや遺伝子構造の変化によって明らかにすることが可能になった．現在開発を行っている比較ゲノム解析システム Murasaki は，複数種のゲノム間での保存領域を高速に検出するシステムである．Murasaki が検出するのはアンカーと呼ばれる数十から数百塩基ほどの相同性の高い保存領域であり，エクソンや短い遺伝子くらいの単位である．Murasaki は徹底した高速化と高スケール化によって，(1) 3 種以上の複数種のゲノム配列の比較が可能，(2) 微生物ゲノムだけでなく，高等真核生物のゲノム比較も実用時間内で可能，(3) 配列パターンの出現頻度に基づくゲノム配列の統計言語的解析が可能，などの特徴を有している．また，Murasaki の出力を可視化するツール GMV を用いて，アンカーと既存のアノテーション情報や発現解析データなどを重ねて表示することができ，さまざまな比較ゲノム解析が簡単な操作で可能となる．本稿では，比較ゲノム解析システム Murasaki の紹介を行い，次に Murasaki を適用してオーソログ遺伝子の解析や偽遺伝子探索などを行った応用事例について紹介する．

Development of a large-scale comparative genome system and its application

Yasubumi Sakakibara
Department of Biosciences and Informatics, Keio University

Abstract :
As the number of whole genome sequences available continues to increase rapidly, the raw scale of the sequence data being used in analysis is the first hurdle for comparative genome analysis. When performing whole genome alignments, large-scale rearrangements make it necessary to first find out roughly which short well-conserved segments correspond to what other segments (termed anchors). Our novel algorithm, which we have implemented in a program called Murasaki (available at http://murasaki.dna.bio.keio.ac.jp), makes it possible to identify anchors of multiple large sequences on the scale of several hundred megabases (e.g. three mammal chromosomes) in a matter of minutes.

1. 複数種のゲノム配列の比較解析

生物の進化にともなって，ゲノム配列は突然変異により変化する．さらに，ゲノムレベルでは，再編成と呼ばれる逆位，転位，重複などのイベントがゲノム上で起こる（図１参照）．一方，タンパク質をコードする遺伝子領域などの機能的に重要な領域では，その機能を保存する必要があるため，変異や再編成などが起こり難い．そこで，現存する複数の生物種のゲノムを比較することにより，変化の多い部分と少ない部分を同定する．そして，「強く保存されている場所には遺伝子などの機能的に重要な領域が含まれている」と推測することにより，種をこえて進化においても保存されてきたオーソログ遺伝子などの発見と抽出が可能となる．一方，ゲノム全体を比較することにより，ある生物種では含まれている遺伝子が他種では存在しないなどのゲノムレベルにおける生物種間の差異を見つけることも可能となり，ゲノムを用いた生物種のプロファイリングという解析も行なうことができる．

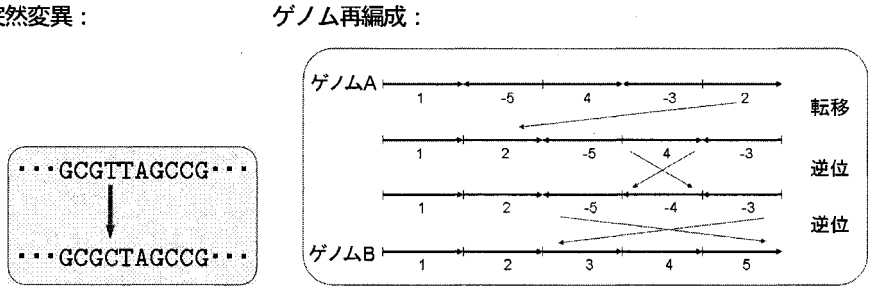

図１．（左）ゲノム配列上に起こる突然変異による塩基置換．（右）転位と逆位による２種ゲノム間のゲノム再編成．

2. 比較ゲノム解析の手順

比較ゲノム解析では，まずはじめにゲノム配列の比較解析を行なう．その典型的な計算手順は，

(1) 保存性の高い相同性領域を検索してアンカーを決定，
(2) アンカーからクラスターを計算してシンテニーブロックを同定する，
(3) シンテニーブロックを単位としてゲノム再編成の距離とシナリオを求める．

次に，このようにして検出されたゲノム間の保存領域から，遺伝子領域の同定，オーソログやパラログなどのホモログ遺伝子の解析，機能性 RNA の発見やプロモータ領域の解析などの非コード領域の解析，を行う．ここで，「オーソログ」の定義は一般的に，２つの生物種に存在する配列相同性の高い遺伝子が，共通の祖先種では同一な遺伝子であり，現在も機能が保存されて同じ機能をもつ遺伝子群のことを指す，となっている．また，「アンカー」などの専門用語に関して，その詳細な説明は後述する．

3. 比較ゲノム解析システム：Murasaki

従来手法に対して，我々は，ゲノム再編成操作（逆位，転位，重複など）を考慮した複数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を開発している．比較ゲノムシステム Murasaki の重要な特徴は，(1) 2 種だけでなく 3 種以上の複数種のゲノム配列に適用できる，(2) ミスマッチを許すパターンを用いた高速で感度の高い保存領域の検出が可能，(3) 微生物ゲノムからヒトやマウスなどの

高等真核生物の億単位の長さのゲノムの多種比較まで適用できる，(4) 配列パターンの出現頻度情報に基づくゲノム配列の統計言語的解析が可能，などである．
具体的には，Murasaki が行なう仕事は，複数種のゲノム配列を入力するとアンカーと呼ばれる相同性領域を抽出する．ここで，アンカーとは，配列レベルで相同性の高い保存領域であり，数十から数百塩基ほどの相同領域で，エクソンや短い遺伝子くらいの単位である（図２を参照）．ゲノムを一つの長い文書と見なしたときに，単語に相当するものがアンカーであると考えてよい．

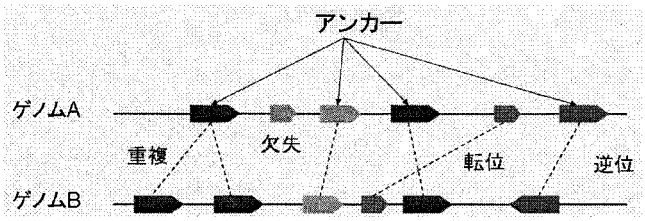

図２． ２種のゲノム間に保存されている複数のアンカーの様子を模式的に表した図.

Murasaki は，複数種のゲノム配列からアンカー領域を，適応的ハッシュ法とパターン（Pattern Hunter 的手法）を用いることにより計算する．その特徴として，(1) 検出する語（pattern と weight）の長さの制限は無い，(2) 多種（３種以上）のゲノムの比較が可能，(3) 100Mbp（ヒト染色体レベル）オーダーのゲノム配列の５種までは実用時間内で計算可能，(4) 配列パターンの出現頻度を用いた統計的解析が可能，がある．

4. 比較ゲノム解析システム Murasaki を用いたマイコバクテリアゲノムの解析

Murasaki の最新バージョンの性能評価を行った結果，マイコバクテリア５種のゲノム比較において，複数種のゲノム比較が可能な唯一の既存システムである Mauve と同精度のアンカーを検出することができた．さらに，酵母と糸状菌のゲノム比較，ヒト，マウス，ラットの３種のすべての染色体のゲノム比較を行ない，有意な結果を得ることができた．各計算時間は，マイコバクテリアの比較において２種で 22 秒，３種で 42 秒，５種で３時間，酵母と糸状菌で 90 秒，ヒト，マウス，ラットの染色体の比較で 12 分，という結果が得られた．ヒトなどの高等生物の複数種のゲノム比較を計算できる既存のシステムは存在しないため（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．），Murasaki の性能評価結果は非常によいものと言える．
Murasaki によるマイコバクテリア５種，*M.tuberculosis CDC1551, H37Rv, M.avium, M.bovis, M.leprae,* のゲノム比較の結果を図４に示す．リンクプロット図においては，５種のゲノム間において，頻繁にゲノム再編成が起こっていることを観察できる．ドットプロットの図では，２種間の比較よりも多種間でゲノムを比較することにより，進化的によく保存された重要なシンテニー領域のみが精度よく検出されていることがわかる．すなわち，ノイズをフィルタリングする効果がある．一方，図５では，３種のゲノム上の遺伝子領域を比較することにより，らい菌において偽遺伝子となっている原因を解析することが可能となる．遺伝子 acyl-CoA dehydrogenase が，らい菌ゲノム上で偽遺伝子化しているが，その領域の中央部分の数箇所が欠落しており，またストップコドンも数箇所に挿入されており，完全なＯＲＦが構成されていないことが分かる．

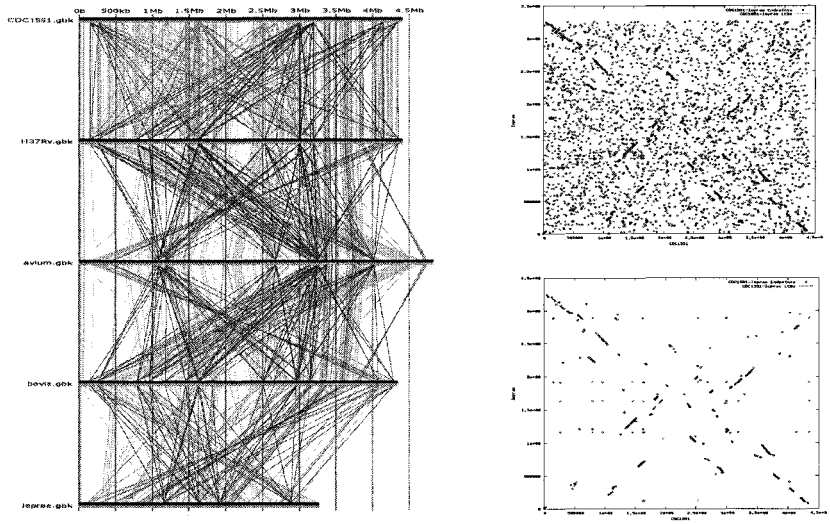

図4. マイコバクテリア5種, *M.tuberculosis CDC1551, H37Rv, M.avium, M.bovis, M.leprae,* のMurasakiによる比較解析の結果. (左) アンカーをリンクプロットした図, (右) 2種のゲノム比較によるドットプロット図 (上) と5種比較によるドットプロット図 (下).

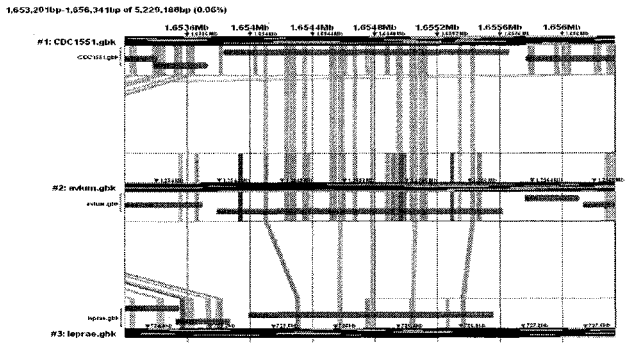

図5. ゲノム比較による偽遺伝子の見つけ方. マイコバクテリア 3 種, *M.tuberculosis CDC1551, M.avium, M.leprae* のゲノム比較において, 下段の *M.leprae* における偽遺伝子では中央部分の数箇所が欠落している.

謝辞：本研究は, 科学研究費補助金「特定領域研究」課題番号 17018029 の援助を受けている.

参考文献

[1] W. J. Kent. BLAT—The BLAST-Like Alignment Tool. Genome Res., 12(4): 656–664, 2002.

[2] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. Human-Mouse Alignments with BLASTZ. Genome Res., 13(1): 103–107, 2003.

[3] B. Ma, J. Tromp, and M. Li. Patternhunter: faster and more sensitive homology search. Bioinformatics, 18(3): 440–445, 2002.