系統解析用オーソログデータセット作成システムの開発

様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　５月１０日現在

機関番号：１３８０１
研究種目：若手研究（B）
研究期間：2010～2011
課題番号：22710184
研究課題名（和文）　系統解析用オーソログデータセット作成システムの開発

研究課題名（英文）　Development of new system for making ortholog dataset

研究代表者
堀池　徳祐（HORIIKE TOKUMASA）
静岡大学・若手グローバル研究リーダー育成拠点・特任助教（テニュア・トラック）
研究者番号：　20535306

研究成果の概要（和文）：全ゲノム配列が決定された生物を対象に全タンパク質のアミノ酸配列情報を用いてオーソログデータセットを作成するシステムを開発した。オーソログとは種分岐によって生じた相同遺伝子のことである。このシステムにより得られるデータセットは従来のオーソログデータセットと異なり、アウトパラログ（対象とする生物群が分岐する前に遺伝子重複によって生じた相同遺伝子）や遺伝子水平伝播によってもたらされた遺伝子を可能な限り取り除くため、大規模遺伝子情報を用いた系統解析に利用できる。

研究成果の概要（英文）：New system to construct ortholog dataset for organisms whose genome sequence data are available was developed. Ortholog is a homolog derived from speciation. The ortholog dataset is made with amino acid sequence data for all organisms by this system automatically. Since out-paralogs and horizontally transferred gene are removed from the dataset through the system, the dataset is suitable for phylogenetic analysis with massive amount of gene data.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	2,900,000	870,000	3,770,000
2011 年度	500,000	150,000	650,000
年度			
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：複合新領域
科研費の分科・細目：ゲノム科学・ゲノム生物学
キーワード：ゲノム進化学

１．研究開始当初の背景

近年、次世代シーケンサーの普及により、様々な種のゲノム配列が解明されてきた。これらの生物について、全ての遺伝子間で類似性を計算し、互いに最も類似性の高い遺伝子をオーソログのペアと見なし、そのペアを連結して作成したオーソログデータで種の系統を解析することが出来るようになった。しかし、原核生物の門の関係などの遠縁の生物間では研究報告によりそれぞれ系統関係が異なっていた。その原因として、使用されたオーソログデータセットに残存するアウトパラログの影響が示唆されている。OrthoMCL、DomClust、Gclust など、既に公開されているオーソログデータ作成プログラムは存在するが、元々未知タンパク質の機能推定を目的としているため、アウトパラログや水平伝播遺伝子が多少混在するデー

タセットが作成される。従って、確実にアウトパラログを除いたオーソログデータ作成法を新規に開発により、より信頼性の高い系統解析が可能になると考えられる。

２．研究の目的
これまで公開されてきたオーソログデータベースは機能予測への利用を主な目的としているため、パラログの混在は大きな問題ではなかった。しかし系統解析ではその基盤たるべきオーソログ配列データセットからパラログデータが完全に除去されることが望ましい。本研究はコンピュータプログラムでパラログをオーソログデータから除去する方法を新規開発し、配列データから簡単に系統解析用オーソログデータセットを作成するシステムを構築する事を目的とする。

３．研究の方法
(1)本システムのアルゴリズム
以下のプロセスを自動で実行するプログラムを作成した。

①前準備（データ収集、手作業で行う。）
通常系統樹を作成するときには分岐順序を知るために根を決定する必要がある。根を決定するには解析対象とする生物群の共通祖先と分岐したことが明らかな生物群をアウトグループとして加え、両者をつなぐ枝上に根があると仮定する。同様の理由で解析には予め互いにアウトグループとなりうる二群を用いる。完全長ゲノム配列が決定された二つの生物群の全遺伝子データ（FASTA 形式のアミノ酸配列データと GenBank 形式の遺伝子データ）を用意する。

②遺伝子水平伝播フィルタリング
中村らが 2004 年に発表した方法（Nakamura et al., *Nat.Genet.*, 2004）を用いて、解析対象の配列データセットから遺伝子水平伝播によってもたらされた遺伝子を推測する。推測された遺伝子のデータを以降の解析対象から外す。

③類似性スコアの計算(BLAST)
すべての配列間で BLAST 用いて、類似性スコアを計算する。

④アウトパラログフィルタリング
アウトパラログ（対象とする生物群が種分岐する前に生じた相同遺伝子）を簡便に取り除くために行う。BLAST 検索結果ファイルから、クエリー配列と同じ生物群にもかかわらず、もう一方の生物群よりも類似性の低い配列を BLAST の結果データから削除する。

⑤ベストヒットペアの連結
③で得られた BLAST の結果ファイルについて、最も類似性スコアの高い配列のペア（ベストヒットペア）をそれぞれの種間で検出し、オーソログペアとした。ここではそれらを単連結法で連結し、オーソログ候補データを作成する。

⑥系統樹作成とオーソログデータ抽出
各オーソログ候補データを用いて近隣結合法で系統樹を作成する。ここで得られた各系統樹には削除すべきパラログが混在する可能性があるため、樹形データを用いてパラログの検出、削除を行い、最終的にオーソログデータを抽出する。この時、二群間の単系統、多系統の違いを基準に系統樹の分別を行い、単系統のもののみをオーソログとする。多系統のものは枝を切断することで単系統になったものをオーソログとした。

⑦繰り返し作業
オーソログとならなかったデータについては BLAST の閾値を 10^{-10} から 10^{-10} 刻みで小さくし、⑤から⑥までを 10^{-200} まで繰り返す。この過程で相対的に低い類似性で繋がっていたアウトパラログが(5)の連結で繋がらなくなる。従って、繰り返し作業により単系統系統樹（オーソログ）が新たに得られる。

(2)シミュレーションテスト
人工的に作成したデータを用いてシミュレーションを行うことにより、本オーソログデータセット作成法の効果を検証した。

①アウトパラログのシミュレーション
祖先遺伝子で遺伝子重複が起こり、その後種分岐がおこり、ランダムに遺伝子欠失が起こる事を想定した系統樹データを作成した。この系統樹を元にタンパク質配列の進化をシミュレーションし（Indel-Seq-Gen を使用)、配列のデータセットを得た。系統樹切断のプロセスがない方法（従来型）であればアウトパラログが多すぎるため、オーソログデータセットを用いて正確な系統樹が作成できないと考えられる。そこで、本システムを用いて作成したオーソログデータセットと系統樹切断のプロセスがない方法（従来型）で作成した系統樹の再現率を計算し、比較した。

②遺伝子水平伝播のシミュレーション
アウトパラログと並び、系統推定を困難にする要素として遺伝子水平伝播が挙げられる。これまでに遺伝子水平伝播をランダムに起こすプログラムは公開されていなかったため、新規に作成した。具体的な手順は以下の通りである。各ステップ番号は図 1 に対応する。図 1、2 は論文(Horiike *et al.*, Bioinformation 2011)より引用した。

Step1：オリジナルの系統樹
進化速度が一定でないため、共通祖先から各末端（現存する配列）までの距離は等しくない。しかし、それぞれの現存する配列が共通祖先から分岐してから経過した時間は等しい。従って、共通祖先から各末端までの枝長を規格化することができる。

Step2：枝長の規格化
任意の結節 v からそれぞれの末端までの平均枝長を以下の式で再帰的に定義した（位置関係を図 2 に示す）。
$L(v) = (d(v_L)+L(v_L)+d(v_R)+L(v_R)) / 2$
この時、 v_L は v の左側の子結節、v_R は右側の子結節である。 v_p は v の親結節、d(v) は v と v_p の距離である。v が末端であるときには L(v)は 0 となる。

Step3：相対進化距離の計算
遺伝子水平伝播がドナー、アクセプター間で起こる時間を決定するために、相対時間 T(v) をそれぞれの系統について計算した。v における相対時間である T(v)を以下の式で再帰的に定義した。
$T(v) = (d(v)+L(v)T(v_p)) / (d(v)+L(v))$
v が根である時には T(v) は 0 となる。一方、v が末端であるときには T(v) は 1 となる。上記の定義式は以下の等式を変形したものである。
$\{1-T(v_p)\}/\{1-T(v)\} = \{d(v)+L(v)\}/L(v)$

Step4：伝播位置の決定
遺伝子水平伝播を起こす時間は利用者が設定した時間の範囲内でランダムに選択される。時間の範囲は 0（根）から 1（末端）までの相対時間で設定できる。ドナーの位置は遺伝子水平伝播の起こる相対時間に存在するすべての系統から、ランダムに選ばれる。アクセプターの位置は決定された相対時間に存在するドナー以外の系統からランダムに決定される。

Step5：枝の移動
アクセプターサイトに新しい結節を作成し、ドナーサイトから先の枝を移動させる。

Step6：不要枝の削除
中途半端に残ったドナーサイトの上流の枝をドナーサイトに一番近い結節ごと削除する。

Step7：枝長情報の回復
系統樹の枝長情報を復活させる。

以上のプログラムを用いて、遺伝子水平伝播を起こしたデータを作成し、本システムの評価を行った。

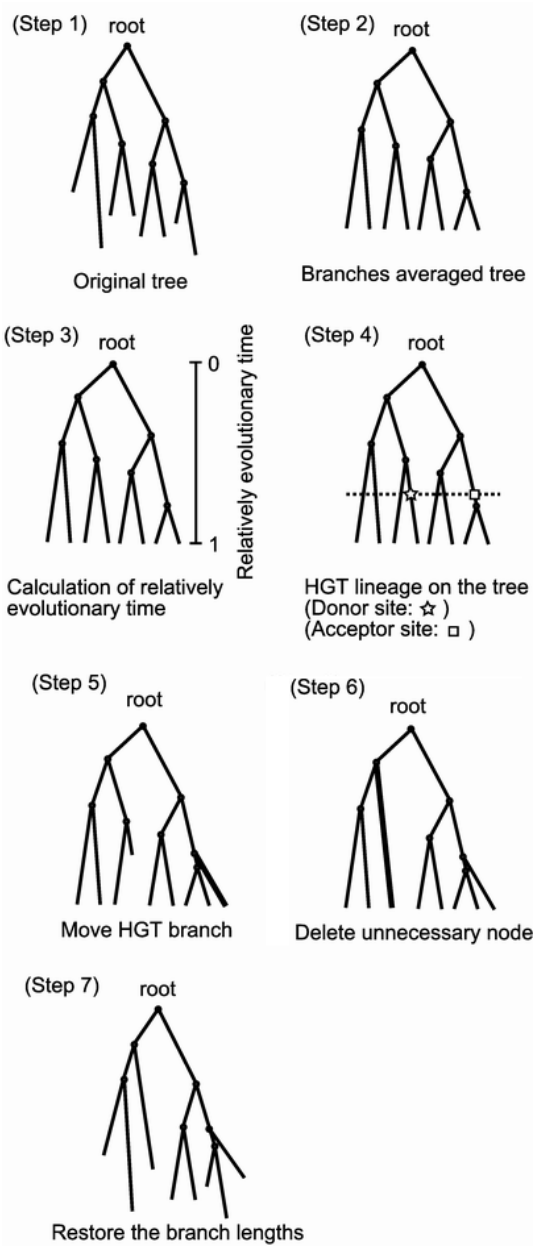

図 1　各ステップにおける系統樹

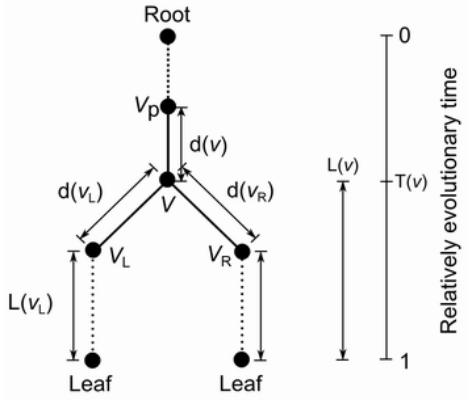

図 2　系統樹の枝長と相対時間を示す模式図

(3)オーソログ作成システムのプログラムパッケージ化
オーソログ作成システムはそれぞれのステップに応じた小さいプログラムの集合で成り立っている。ユーザーが使用する時に扱いやすいよう、これらをパッケージにまとめる。

４．研究成果
(1)オーソログデータセット作成システムについて
予定通り、システムは完成した。本システムはLinuxコンピュータで動作する。サンプルデータとして、アクチノバクテリア５種、ファーミキューテス８種についてオーソログデータを作成した所、354 のオーソログが得られた。計算時間は遺伝子水平伝播予測が最も長く、約７時間かかり、BLAST が約１時間、その他すべては約１時間かかった。あらかじめ多くの種について遺伝子水平伝播予測を実行しておけば、実際のオーソログデータを作成する時間を短縮できる。

(2) 遺伝子水平伝播シミュレーションプログラムについて
遺伝子水平伝播のシミュレーションを行うプログラムが存在しなかったため、新たに開発した。このプログラムに有根系統樹データを入力すると、設定した相対時間に遺伝子水平伝播を起こした系統樹データが出力される。このプログラムをウェブサイト（http://www.grl.shizuoka.ac.jp/~thoriike/HGT-Gen.html）に公開した。また、このプログラム開発に関する論文を発表した。

(3)シミュレーションテストについて
アウトパラログが混在するデータを元にオーソログデータを作成し、系統樹の再現率を計算した所、系統樹切断のプロセスがない方法（従来型）では 42%だった再現率が 96%まで改善された。また、人工的に遺伝子水平伝播を起こした系統樹を元に作成した配列データから本システムを用いてオーソログデータセットを作成した所、水平伝播した遺伝子が50%含まれるデータにおいても93%の再現率が得られた。水平伝播した遺伝子を30%程度に減らせば 99%再現できることから、本システムに組み込んだ水平伝播遺伝子フィルタリングは精度を高めるために有効であるといえる。

(4)オーソログ作成システムのプログラムパッケージ化
本システムのプログラム群を簡単に実行できるよう、パッケージ化した。この際、汎用性と計算速度を向上させる為の改良も行った。また、多数存在するパラメータの中から変更する可能性が低いものを選び、固定値とすることにより、煩雑さを軽減した。論文掲載後にインターネットに公開する予定である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）
Tokumasa Horiike, Daisuke Miyata, Ryoichi Minai, Yoshio Tateno, HGT-Gen: a tool for generating a phylogenetic tree with horizontal gene transfer, Bioinformation, 査読有り、Vol.7(5), 2011, 211-213

〔学会発表〕（計７件）
堀池徳祐、遺伝子水平伝播シミュレーションプログラム(HGT-Gen)の開発、第 6 回日本ゲノム微生物学会年会、2012 年 3 月 11 日、東京

Tokumasa Horiike, Development of new method for making ortholog dataset, The 2010 Annual meeting of the society for molecular biology and evolution, July/27/2011, Kyoto

堀池徳祐、Development of ortholog dataset for phylogenetic analysis、BMB2010（第 33 回日本分子生物学会年会・第 83 回日本生化学会大会　合同大会)、2010 年 12 月 7 日、神戸

〔その他〕
ホームページ等
http://www.grl.shizuoka.ac.jp/~thoriike/research.html

６．研究組織
(1)研究代表者
堀池　徳祐（HORIIKE TOKUMASA）
静岡大学・若手グローバル研究リーダー育成拠点・特任助教(テニュア・トラック)
研究者番号：20535306